DISCOVERING DISEASE ASSOCIATED GENE-GENE INTERACTIONS: A TWO SNP INTERACTION ANALYSIS FRAMEWORK

Athos Antoniades

University of Cyprus

20 May 2011

DISCOVERING DISEASE ASSOCIATED GENE-GENE INTERACTIONS: A TWO SNP INTERACTION ANALYSIS FRAMEWORK

Athos Antoniades

A dissertation

Submitted in Partial Fulfilment of the

Requirements for the Degree of

Doctor of Philosophy

at the University of Cyprus

Recommended for Acceptance

By the Department of Computer Science

May, 2011

ACKNOWLEGMENTS

I gratefully acknowledge the contributions of all those who helped me complete this PhD thesis dissertation. First and foremost, I thank my dissertation supervisor; Dr.Constantinos Pattichis who guided me for the duration of my study helping me grow as a PhD student and a person with his insightful advice, critique and guidance.

I would also like to thank Dr. Lefkos Midletton who introduced me to the problem of two SNP interaction testing and who has helped me as a mentor advance my career as a computer scientist in the pharmaceutical industry in a way that enabled me to gain the necessary background information to achieve my goals.

I also thank my committee members, Dr. Christos Christodoulou, Dr. Pedro Trancoso, Dr. Vasilis Promponas and Dr. Dimitris Koutsouris who helped me ensure rigor in the qualitative part of my study and improve the technical writing aspects of my dissertation.

Moreover, I thank my colleagues, Dr. Paul Matthews and Dr. Nicholas Galwey, who have provided me with their knowledge and support in researching, developing and evaluating the proposed framework derived from my PhD work.

Finally, I thank my wife, Dr. Iolie Nicolaidou, my daughter Nefeli Antoniades and my parents Antonis and Aggeliki Antoniades for their continued support over the years.

CREDITS

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

CNV	Copy Number Variation
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DNA	Deoxyribo-Nucleic Acid
GWAS	Genome Wide Association Scans
HCC-HPC	Hybrid Cluster Cloud High Performance Computing
HLA	Human Leukocyte Antigen
HPC	High Performance Computing
LAN	Local Area Network
LDL	Low Density Lipoprotein
MDR	Mutlti-Dimensionality Reduction
MS	Multiple Sclerosis
SNP	Single Nucleotide Polymorphisms
WAN	Wide Area Network

Abstract

Most common diseases have a heritable component that is influenced by mutations on multiple loci, and by interactions between loci and with the environment. However, traditional genetic analysis techniques have focused on single locus effects. This is mostly due to the polynomial increase in computational capacity needed to attempt multi-loci interaction analyses, and the anticipated loss of power due to multiple testing. In this dissertation, a framework for performing a complete two single nucleotide polymorphism (SNP) interaction analysis of high dimensionality genome wide association scans (GWAS) is presented. The implementation of the framework utilizes diverse distributed computational resources to overcome the bottlenecks of each resource, harvesting enough capacity to analyze any of the GWAS in existence today within a reasonable time frame. Algorithmic approaches are proposed to improve the efficiency of the framework and improve its computational performance so that a brute force attack on the problem can be performed. The data is encoded in binary using a lossless algorithm that significantly reduces its size. Computationally efficient data mining measures for the Omnibus and Epistatic interaction effects are proposed and compared to traditional statistical techniques. An algorithm is proposed that optimizes the analyses of multiple response variables within the same GWAS. GenMSA, a multiple sclerosis (MS) dataset, is analyzed using the proposed framework with top results tested for replication using ANZgene, an independent MS, dataset. Some of the top results replicated, implicating SNPs in a region of known association to MS providing evidence to the validity of the proposed framework. Top results are further examined through a proposed approach that enables drilling into these results and studying correlation coefficient between each of the genotype combinations of the SNP and the signal level to each of the main and epistatic effects.

Chapter 1 Introduction

Most common diseases that have a heritable component such as diabetes, multiple sclerosis, schizophrenia and dyslipidemia are influenced by mutations on multiple loci, interactions between loci and interactions with the environment. In genetic studies, due to limitations in genotyping technology, traditionally only a small subset of the genome could be analyzed, so candidate gene studies were common. The introduction of affordable high throughput genotyping technologies (DNA chips) allows the assay of more than half a million single nucleotide polymorphisms (SNPs) per subject across the whole genome. Genetic association studies applying such technology allow investigation of the vast majority of common loci variants in the human genome; such studies are typically called genome wide association scans (GWAS). These GWAS provide an unprecedented opportunity to identify genetic variations associated with diseases.

However, traditional analyses techniques have focused on discovering single locus effects rather than multi-loci effects in GWAS studies. This approach has yielded rather limiting results in many studies because the model used is too simple and is being applied in a complex reality. Commonly used arguments for this practice are the increased computational requirements and the loss of power of detection following correction for multiple testing in multi-loci analyses. In the few cases where multiple loci interactions have been examined, only a subset of the search space has been analyzed. Typically loci with no or very small main effects (little or no single-marker association to the phenotype) were not included even though there is no reason to believe that they could not be involved in a strong multi-locus interaction associated with a complex disease [1,2,3].

Approaches that attempted to test for gene-gene interaction in GWAS data can be split into two categories, those that attempt to perform an exhaustive, or near exhaustive search, and those that attempt to use machine learning approaches derived from data mining to discover gene-gene interactions without going through the whole search space.

Another key issue that arises from current research is the need to develop new statistical measures to quantify the epistatic effect. There is a common acknowledgement in the field that end results need to be statistically interpretable, with p-values as the statistic of choice (the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true) [4,5,6]. However, the traditional analytical approaches to getting p-values for interactions are too slow to perform at the whole genome level. Therefore estimations have been proposed that attempt to provide close approximations to the actual statistic in a computationally efficient manner[2].

The evaluation of genetic results, and by extension the methods used to obtain them, are traditionally based on replication testing. Replication testing involves the use of two independent datasets to test if a hypothesis tested and found to be significant in one study replicated under the same parameters in a second independent study. A concise, replication methodology is necessary to be defined a-priory accompanied with a solid mathematical proof that the probability of reporting replicated results that are in reality false positive is extremely low.

1.1 Problem statement

With the complete sequencing of the human genome one might expect a plethora of drug targets for many diseases to be discovered [7]. This however isn't the case, since the analytical techniques applied to these data were focused on simple Mendelian diseases (diseases where a single genetic polymorphism was responsible for a phenotypical trait) while the majority of common diseases in humans are complex diseases (multiple genetic polymorphisms and environmental factors predispose a subject's disease status) [8].

The work presented in this thesis dissertation was focused around identifying the key problems surrounding the lack of a method performing a complete GWAS two SNP interaction test and addressing each one of them. Through this process several problems were identified, and each required an original contribution to be researched and developed in order for the proposed framework to get completed. In this section the key problems addressed in this dissertation are presented.

1.1.1 Data encoding

The need to reduce the size of the data is owed to the 100 fold increase in the genotyping capacity available today combined with the massive reduction in cost. Today's technology has the ability to analyze datasets up to 550,000 or even 1 million SNPs per subject with >99% accuracy, at a rate of >100 K genotypes per day and at a cost of around 20–30 cents per genotype [9,10,11].

Traditionally, the genetic data in these studies is stored in the QTDT (Quantitative and Discrete Traits) format introduced in the program QTDT and MERLIN [12]. Input files describe relationships between individuals in a dataset, store marker genotypes,

disease status and quantitative traits and provide information on marker locations and allele frequencies.

There was already a technique for compressing this data by utilizing a binary encoding PLINK [13]. However, that technique was focused at encoding SNPs as markers to be used in analyses that don't require the information of which DNA strand each allele belongs to. Today as more GWAS datasets become available researchers are developing innovative new methodologies to analyze them. Some of these methodologies are not relying so much on the SNPs as markers; rather they look at the sequence of genotyped alleles on each strand of DNA separately. The methodology used in PLINK [13] to encode the data looses the information of which strand holds each allele's genotype for heterozygote SNPs. This makes it impossible to run analyses that use strand information using the binary input format for GWAS.

1.1.2 Measure of epistasis

The established way of measuring epistasis with categorical response variables in genetic data is logistic regression [2,3,14]. However, logistic regression is not very fast and is therefore usually applied to test small subsets of the datasets. A faster computational approach is needed that will provide results comparable to logistic regression [2]. However the key aspects that make logistic regression the analysis of choice would need to be retained, such as testing for the null hypothesis generating p-values. Furthermore, a key issue in analyzing genetic data is the non-independence of markers, an issue that exists with logistic regression; any proposed methodology would need to not provide a positive bias towards loci that are non-independent [3]. The correlation between any proposed methodology and logistic regression is defined *a-priory* at 5% correlation between any proposed methodology and logistic regression

in order to accept it as producing sufficiently converging estimations to logistic regression.

1.1.3 Computing multidimensional contingency tables

In the majority of case-control studies with genetic data, more than one phenotype may be important even though the subject disease status (case - control) is typically the most interesting one. Many diseases tend to have different characteristics, or different distributions in sub-populations and typically, when performing genetic analyses, the goal is not only to identify genetic loci associated with the case-control status but also to identify the ones that are associated with subsets of subjects that exhibit other phenotypes (phenotypes are used as response variables) relative to the disease [4]. As an example, Multiple Sclerosis (MS) is more evident in females than males, it's not clear if this is due to environmental or genetic factors [15]. It is therefore logical to ask if a reported association is driven by males, females or both. Similarly, most diseases including MS, have subtypes, and it's typically interesting to test whether a specific effect is associated with one of the subtypes [11,16]. Such questions typically require a re-analysis of the dataset for each phenotype definition in order to perform data mining on each of the variables, and then collect all results together in a single table to interpret. However, traditional analyses also typically involve only univariate analysis that has a computational complexity of $O = n$, where n is the number of SNP's while in a complete two SNP interaction test as attempted in this work the computational complexity is $O(n^2)$. To put this in perspective, a univariate analyses on a typical personal computer today will take a few minutes. A complete two SNP interaction analysis will take several months if not years. Having to repeat the analyses for every phenotype of interest in the case of the univariate

analysis does not increase the requirements to an unacceptable level. In the case of two SNP interactions, however, it multiplies the complexity of an already challenging task by a factor equal to the number of extra response variables to be tested. Counting the values for a contingency table and the sums of all rows and columns is a known P hard problem [17]. Furthermore, the research in finding solutions to the problem is focused on getting approximate counts of the table using heuristics [18,19]. However this would result in loss of information in the case of GWAS analysis, since the number of subjects are limited compared to the statistical power needed, thus any approach which sacrifices statistical power in order to provide a linear increase in performance should be avoided.

1.1.4 Multiple testing problem

The multiple testing problem is a major issue in GWAS analyses that needs to be addressed [20]. Multiple testing, or multiple comparisons as it's sometimes referred to in statistics is a problem that occurs when a set of statistical inferences are considered simultaneously [20]. Errors in inference that fail to include their corresponding population parameters or hypothesis tests that incorrectly reject the null hypothesis are more likely to occur when one considers the set as a whole. The multiple testing problem is especially evident in data mining applications that test high dimensional datasets [21] such as the ones analyzed as part of this dissertation work. Traditionally, in genetic analyses heuristics such as the Bonferoni correction [22] were used to adjust for it [23]. Bonferoni correction is based on the idea that if an experiment is testing a dependent or independent hypothesis on a set of data, then one way of maintaining the set-wise error rate is to test each hypothesis at a statistical significance level of 1/n times of what it would be if only one hypothesis were tested, where *n* is

the number of tests performed. So if there is a need to compute the significance level of a set of tests n to be at most a, then the Bonferoni correction would be to test each of the individual tests at a significance level of (a/n) [24]. Statistical significance simply means that a given result is unlikely to have occurred by chance assuming the null hypothesis is actually correct (i.e. no effect) [25]. Bonferoni corrections tend to over-adjust, thus making identifying statistically significant p-values after correction for multiple testing a very difficult task [24,26]. Traditionally, the only test that was considered to be ideal for adjusting for multiple testing is replication testing [27].

The current norm in identifying a result as statistically significant or not is to a-priory set an arbitrary level of significance, up to which the results will be rejected as invalid [6]. In genetics, the typical level of statistical significance is either p<0.05 or p<0.01. This level of significance still involves a pretty high probability of error (5% or 1%) . Furthermore, the statistical measures used to estimate p, usually assume a normal distribution between the results and this only holds if the results are independent, but as it will be discussed in later sections in real life applications this assumption does not hold.

1.1.5 Computational complexity and high performance computing

The problem of testing for all two SNP interactions in GWAS has a growth rate of $O=n^2$ where n is the number of SNPs to be analyzed. Since the typical GWAS contain hundreds of thousands of SNPs and some even have over a million SNPs it's important to use a system capable of providing enough computational capacity both for current and for expected high dimensionality of GWAS. Therefore it's expected that in order to address the computational requirements of performing a complete two

SNP interaction test in a high dimensional GWAS, a high performance computing resource is required. This problem requires large computational capacity as well as large data storage and transfer capacity between the nodes of a distributed system [1,8,28].

1.2 Original contributions

The main contribution of this thesis is the proposed framework for a complete two SNP interaction testing in high dimensional GWAS. In order to overcome the problems associated with designing and developing a two SNP interaction framework, several original contributions in the fields of data mining, high performance computing and statistical genetics were made and are presented in the next sections of this dissertation.

1.2.1 Data encoding

A proposed methodology is presented that enables encoding of the GWAS data in a way that results in a lossless compression of the data compared to the traditional "MERLIN" format [12]. Furthermore, even though this method results in twice as large files compared to a similar approach presented in [13], it does not suffer from a loss of information. Although this issue does not affect univariate analyses, multivariate analyses may be affected. Furthermore the proposed methodology enables encoding of deleted SNPs, a type of polymorphism that is gaining momentum in being detected and combined in some analytical techniques with traditional SNP data.

1.2.2 Measure of epistasis

A new measure of epistasis was proposed in order to test for the epistatic effect, a term used in Biology to identify the interaction effect between two loci. The advantage of the proposed measure is that it's designed to be considerably faster than logistic regression, the established method of measuring epistasis, improving the efficiency of the proposed framework. Through the algorithmic steps to estimate the interaction effect, an established omnibus measure is also calculated that represents the total association between the categorical variables and a response variable.

To put this in data mining terms, the proposed methodology enables the evaluation of the probability that a multidimensional association rule is a false positive assuming the rule (null hypothesis) is true.

1.2.3 Computing multidimensional contingency tables

As a way to further improve the efficiency of the proposed framework, an algorithm that supports generating efficiently all contingency tables for each of the response variables to be tested is proposed. Assuming the number of response variables to be tested is φ, the proposed algorithm is nearly φ times faster than the traditional approach. The proposed algorithm succeeds this optimization by identifying the genotype combination between two SNPs of a specific subject only once rather than φ times.

In GWAS studies, response variables are usually phenotypes, and this approach enables the analysis of multiple phenotypes without significantly increasing the computational cost of the analysis compared to running just a single phenotype. Most common diseases have multiple subtypes; this is a key feature that will enable testing for epistatic effects associated with specific subtypes of diseases.

1.2.4 Evaluation of significance through replication

In this thesis dissertation we explore existing methodologies of testing for replication between independent genetic studies in order to identify statistically significant results [27,29]. However, we also propose some post-processing steps for replicated results that enable both the visualization of the distribution of each effect among the genotype combinations of the two SNPs as well as studying the correlation of the effects between the two studies.

When comparing p-values between two studies only the total association level of the test in each study is compared. The distribution of each effect among each of the genotype combinations of the two SNPs is not considered. If a replicated result is indeed a true positive, then it's expected that the signals for each effect will be similarly distributed between the two datasets. Failure to do so may be due to many factors and can't be used as a test for non-replication success. In the case of a high correlation of the distribution of effects (both main and interaction) to the genotype combinations of the two SNPs the confidence in the validity of the results will be significantly increased [3,5,27,29].

If an independent replication study is not available then the same approach of drilling down into results in order to visualize the direction and level of the signal of each effect on each of the genotype combinations is still helpful. Even though it will not help in increasing the confidence in the result, it will help identify the possible haplotypes that are more frequent in each phenotype class enabling researchers to generate new hypotheses that they may be able to test with follow up experiments.

1.2.5 Hybrid cluster cloud high performance computing (HCC-HPC) framework

The computing resources available for this work were composed of two different HPC systems. They each had distinct characteristics that disabled them from being used alone for this analysis [30]. A distributed analytical approach is proposed that enables the use of both HPC resources in parallel enabling improvements in efficiency of the proposed framework. The proposed approach is compared to using either HPC architecture alone, or both. This proposed distributed analytical approach is however not applicable to other problems since it's problem specific to the subject of complete two SNP interaction testing, thus it's labelled as a minor contribution to distinguish it from the other presented contributions.

The proposed computing framework is labelled as a hybrid cluster – cloud high performance computing framework (HCC-HPC). This is composed of an algorithm that is designed to work for two SNP interactions testing in such a way as to utilize both a cloud and a cluster grid to avoid the bottlenecks associated in each. This is achieved by breaking up the data into work nodes with each work node going through a serial pipeline from the analysts' personal computer to each of the two HPC systems and performing the computationally intensive part of the analysis on the computational cloud, leaving the parts that require both computational power and inter-process communication to the cluster. The performance of this Hybrid HPC is discussed in comparison to performing the analyses on either HPC alone or both.

Publications that were derived as part of this work are listed in Appendix A .

1.3 Structure of this dissertation

This dissertation is split into 7 chapters. In chapter 2 background knowledge will be presented to introduce the reader to key relevant concepts of this thesis. Chapter 3 presents a detailed literature review that was performed in the fields relating to the innovations introduced in this dissertation. The proposed framework is presented in detail in chapter 4. All results related to the actual analyses of the datasets as well as results related to the performance of the system are presented in chapter 5. A discussion of the results and the implications of this work are presented in chapter 6. Future work is presented on chapter 7.

Chapter 2 Background Knowledge

The goal of this chapter is to introduce some background information as to enable a clear understanding of the following chapters and a common terminology related to GWAS analyses. The first part of this chapter begins with some basic concepts in Molecular Biology, with a focus of explaining SNPs, the markers that the genetic data used in this dissertation is consisted of, and Linkage Disequilibrium, an effect that is of key importance in all analytical genetic approaches. It concludes with an introduction to Complex diseases and genetic and environmental factors that are both useful to define in order to understand the motivation for this work. The second part of this chapter focuses on statistical genetics, defining a key terminology and providing an explanation of key effects and analytical approaches relevant to this work.

2.1 Molecular biology and medical genetics

Molecular biology is the branch of biology which primarily deals with functions, characteristics and structures of three major macro-molecules DNA, RNA and proteins. In this section some basic biological mechanisms from molecular biology are presented to enable better understanding of the work in this dissertation. The general problem the proposed methodology tries to address falls within the field of human genetics. In human genetics the study is focused on inheritance in humans. Study of human genetics can answer questions about human nature, understand diseases and help in the development of effective disease treatments. However, to date these promises have not been met, and part of the problem as described in chapter one is the lack of a comprehensive framework to test for gene-gene interactions associated with complex diseases.

_____
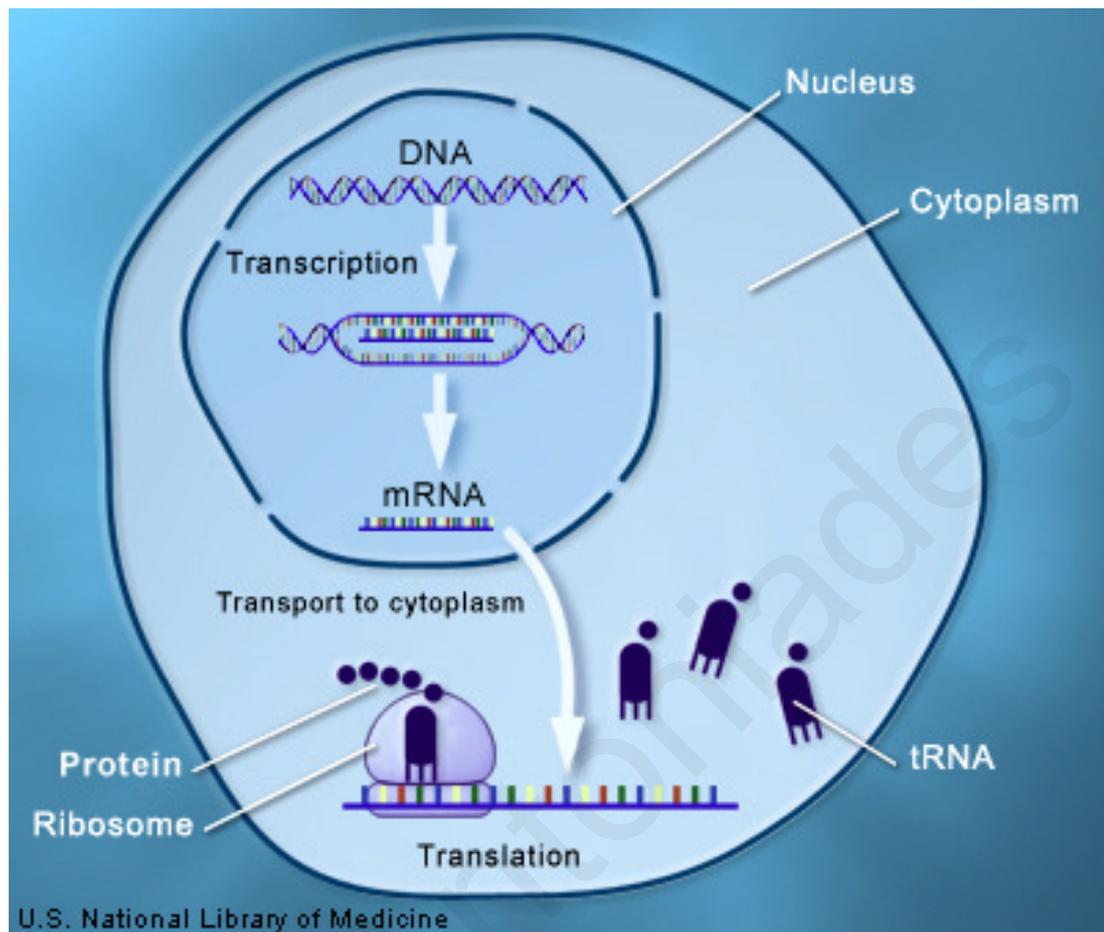
Figure 1 From DNA to proteins, the biological mechanism

Lister Hill National Center for Biomedical Communications, US National Library of Medicine, National Institutes of Health, Department of Health & Human Services, *Genetics Home Reference. Your Guides To Understanding Genetic Conditions*. United States: U.S. National Library of Medicine, 2011.

_____

2.1.1 Macromolecules DNA, mRNA and proteins

Deoxyribonucleic acid otherwise known as DNA is the building block of life. It contains the information the cell requires to synthesize proteins and to replicate itself. The central dogma of life; the coded genetic information hard-wired into DNA is transcribed (transcription is the process of creating a complementary mRNA molecule from a DNA segment) into mRNA molecules; each mRNA molecule contains the information for the synthesis of a particular protein. As Figure 1 demonstrated mRNA molecules can travel outside of the nucleus and into the cytoplasm, here they can be translated into proteins.

2.1.2 SNPs

A single-nucleotide polymorphism (SNP, pronounced snip) is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of a biological species or paired chromosomes in an individual [31].

All individuals have two strands, one inherited by each parent. Figure 2 depicts two strands. If we consider these to be from a single individual, then he has a heterozygote genotype for the SNP highlighted. Subjects who have the same allele on both strands are called homozygote. In the case shown in Figure 2 the two alleles of the SNP are C and T.

Although there are several types of polymorphisms that can occur on a disease causing gene, the majority of them can potentially be represented by one type of mutation, single nucleotide polymorphisms (SNPs). Therefore SNPs serve as biological markers for pinpointing a disease on the human genome map. This does not mean that the SNPs cause the disease even though some times that is the case. Simply put, due to the way that DNA is inherited, SNPs neighbouring a mutation (potentially

disease causing) of any type have a higher probability to be inherited together than normal distribution and linkage equilibrium would dictate. The reason for this effect is linkage disequilibrium, discussed in at length in a following section. Therefore the SNPs with a high probability of being inherited together with a disease causing polymorphism will carry that same association with the disease even though they are not the causative factor [31,32,11].

_____
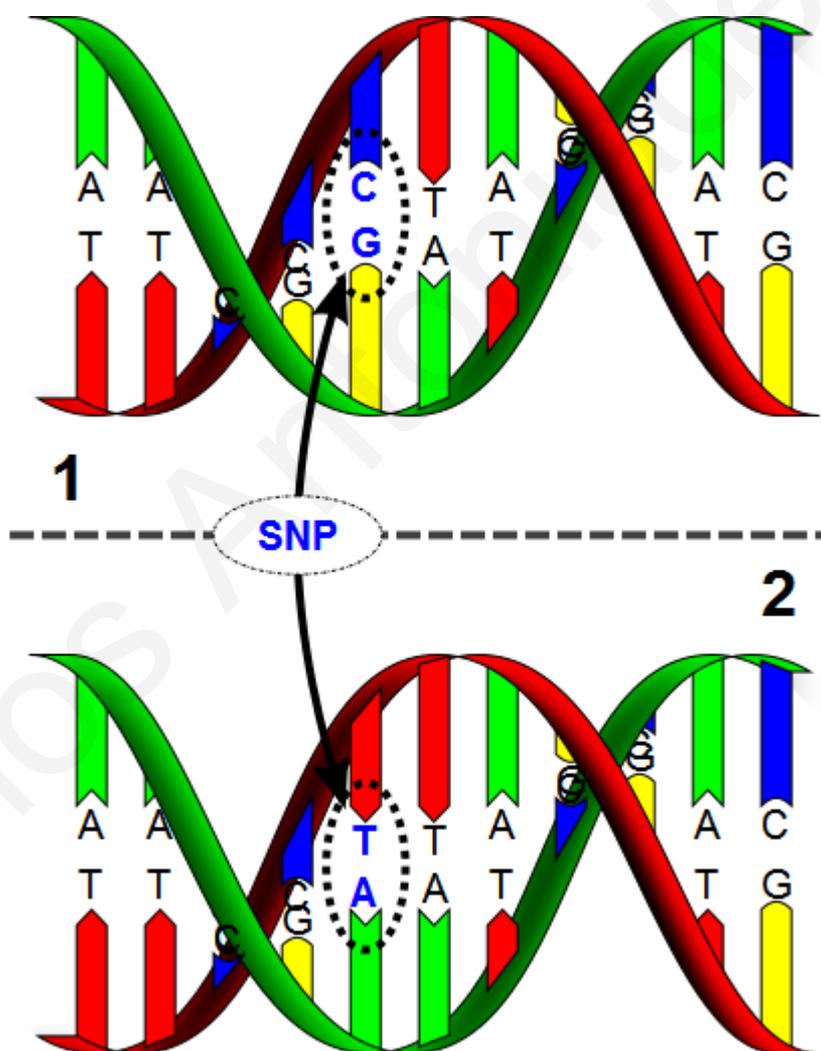
Figure 2 A SNP on DNA strands

_____

2.1.3 Genes

Genes are stretches of DNA that code for a type of protein or an RNA chain that has a function in the living organism [31]. Genes determine hereditary traits, such as the hair and eye colour or disease predisposition by providing instructions for how every activity in every cell of our body should be carried out [31,33]. Due to the multiplicity of polymorphism combinations inherited by each organism different forms of the same gene exist in subgroups of individuals. These forms are called alleles. When a gene is in the process of being transcribed into a functional protein the number of copies of mRNA sequences for that gene can be counted through "gene expression" experiments [34].

To better understand the functionality of genes, here's a classic example. A gene may enable a liver cell to remove excess cholesterol from our bloodstream. It does this by instructing the cell to make a particular protein. It is this protein that then carries out the actual work. In the case of excess blood cholesterol, it is the receptor proteins on the outside of a liver cell that bind to and remove cholesterol from the blood. The cholesterol molecules can then be transported into the cell, where they are further processed by other proteins [35].

Many diseases are caused by polymorphisms or changes in the DNA sequence of a gene. When the information coded for by a gene changes, the resulting protein may not function properly or may not even be produced at all. In either case, the cells containing that genetic change may no longer perform as expected. For example, we now know that mutations in genes code for the cholesterol receptor protein associated with a disease called familial hypercholesterolemia. The cells of an individual with this disease end up having reduced receptor function and cannot remove a sufficient amount of low density lipoprotein (LDL), or bad cholesterol, from their bloodstream.

A person may then develop dangerously high levels of cholesterol, putting them at increased risk for both heart attack and stroke [35].

2.1.4 Linkage disequilibrium

Linkage disequilibrium is the non-random association of alleles at two or more loci, not necessarily on the same chromosome [8,31]. In other words, linkage disequilibrium is the occurrence of some combinations of alleles or genetic markers in a population more often or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies. The amount of linkage disequilibrium is the difference between observed and expected (assuming random distributions) allelic frequencies [7,36].

Linkage disequilibrium can be visualized in a map that presents the LD between two SNPs by using an LD metric. LD metrics are tests of associations between two markers such as SNPs typically applied to a single group of subjects that is representative of the general population. HAPLOVIEW is the traditional software for performing and visualizing LD analyses [7,37]. A sample of HAPLOVIEW's output is presented in Figure 3.

The horizontal line at the top of the map represents a DNA strand. The vertical lines on it represent genotyped SNP locations. Below that all genotypes SNPs are put in series with a line connecting them to their position on the DNA strand. The LD map generates an equilateral triangle so that two of the vertices of the triangle are placed on the SNP location on the map and the third vertice is represented by a small rhombus. The colour of the rhombus represents the level of LD between the two SNPs. In this example, three categories are plotted, High LD, No LD, and Low LD as

red, blue and green respectively. All SNP pairs in this small region are calculated for LD.

Visualization of LD or any other genetic variation on genetic maps generated from published studies is possible through the HapMap project [36] that provides the functionality and tools to query any region of the human genome.
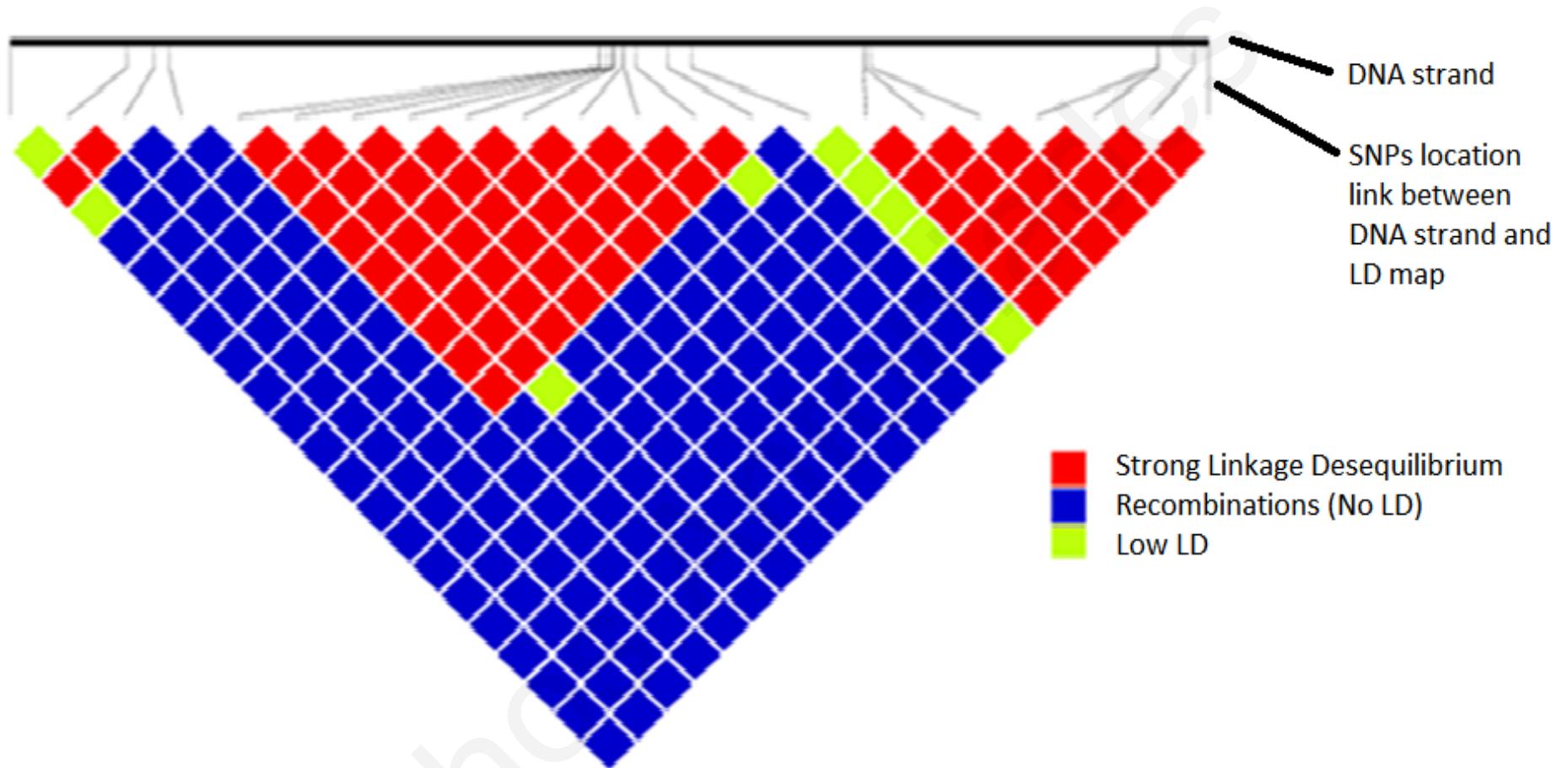
Figure 3 Linkage disequilibrium map example

Generated from HAPLOVIEW software on a sample dataset[37].

2.1.5 Complex versus Mendelian diseases

Mendelian diseases are diseases where one gene is responsible for one disease. These diseases are often rare (such as MCKD1 and MCKD2 researched by the Cyprus Institute of Neurology and Genetics) [38]. Linkage analyses were most often used, rather than association studies, to discover these Mendelian disease causing genes.

Complex diseases, otherwise known as multigenic diseases, are caused by more than one gene or SNP. Environmental or other factors may also play a role. Thus the problem of identifying causative factors in these diseases is far more complex since a combination of SNPs will predispose a person in a certain degree to a disease. Furthermore, it is believed that for some of the most common complex diseases (diabetes, Alzheimer, etc) genetic variations in many or all chromosomes are the causes [3,26,29].

In whole genome analyses, SNPs from the entire genome are involved. This type of research requires the analyses of up to hundreds of thousands or even millions of SNPs that may be associated to the disease by themselves, or through interaction with other genetic loci or environmental factors [3]. The increased number of SNPs needed for this type of research and the reduced number of patients' data that is usually available for analysis, due to the high cost of genotyping (determining the genotype of SNPs on their DNA), is what makes GWAS studies challenging to conduct. Nevertheless, multiple studies have being conducted in many complex diseases.

2.1.6 Genetic and environmental factors

For the most part, complex diseases are caused by a combination of genetic, environmental, and lifestyle factors, most of which have not yet been identified [15,34,39,40]. The vast majority of diseases fall into this category, including several

congenital defects and a number of adult-onset diseases. Some examples include alzheimer's disease, scleroderma, asthma, parkinson's disease, multiple sclerosis, osteoporosis, connective tissue diseases, kidney diseases, autoimmune diseases, and many more [41].

Scientists now know that complex diseases do not obey the standard Mendelian patterns of inheritance. Although we inherit genes associated with these diseases, genetic factors represent only part of the risk associated with complex disease phenotypes. A genetic predisposition means that an individual has a genetic susceptibility to developing a certain disease, but this does not mean that a person with a genetic tendency is destined to develop the disease. The actual development of the disease phenotype depends in a large part on a person's environment and lifestyle. While we cannot change our genes, we can alter our lifestyle and environment to prevent or delay the onset of such a disorder. Indeed, the interplay between genetic and environmental factors in complex disease continues to challenge the research community [39].

2.2 Statistical genetics

2.2.1 Response - explanatory variables terminology in genetics

The terms "dependent variable" and "independent variable" are used in similar but subtly different ways as part of the standard terminology in statistics. They are used to distinguish between two types of quantities being considered, separating them into those available at the start of a process and those being created by it, where the latter (dependent variables) are dependent on the former (independent variables).

However, in real life applications of statistical approaches in data mining, or statistical genetics, independent variables are rarely statistically independent, therefore the terms

response variable and explanatory variable are preferable instead of dependent and independent variable [42,43]. For the purposes of this dissertation the response variable (independent variable) will be composed of phenotypes, while the explanatory variables will be composed of SNPs. The terms response and explanatory variables will be used, but in the context of this dissertation they should be respectively equivalent in terms of statistics to dependent and independent respectively.

2.2.2 Regression analysis

In statistics, regression analysis includes any techniques for modelling and analyzing several variables, when the focus is on the relationship between a response variable and one or more explanatory variables. Regression analysis defines how the typical value of the response variable changes when any one of the explanatory variables is varied, while the other explanatory variables are held fixed [14]. Most commonly, regression analysis estimates the conditional expectation of the response variable given the explanatory variables — that is, the average value of the dependent variable when the independent variables are held fixed. Less commonly, the focus is on a quintile, or other location parameter of the conditional distribution of the response variable given the explanatory variables. In all cases, the estimation target is a function of the explanatory variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the response variable around the regression function, which can be described by a probability distribution [44].

In genetic studies, regression analysis can be applied to test for association or interaction (dependent on the model used) between multiple loci (explanatory

variables) and a phenotype (response variable) such as disease status (case-control). For categorical phenotypes logistic regression is typically applied to test for interaction between loci, while for continuous phenotypes linear regression can be used [45].

2.2.3 Testing the null hypothesis

Interpretation of statistical information can often involve the development of a null hypothesis (H_0) in that the assumption is that whatever is proposed as a cause has no effect on the variable being measured [45]. Hypothesis testing works by collecting data and measuring how probable the data are, assuming the null hypothesis is true. If the data are very improbable, usually defined a-priory as observed less than 5% or 1% of the time, then the experimenter concludes that the null hypothesis is false. If the data do not contradict the null hypothesis, then no conclusion is made. In this case, the null hypothesis could be true or false; the data give insufficient evidence to make any conclusion [14,46].

Traditionally, research conducted in the biological fields relies on testing the null hypothesis as a way to provide a comparable and repeatable measure of an effect. The traditional interpretation of testing the null hypothesis is p-values. P-values represent the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. A key feature of p-values is that the probability provides a combination of both the size of the effect (for example, confidence in data mining or odds ratio in statistics) and the size of the supporting evidence for it (the number of samples in a dataset, such as the support measure from data mining).

2.2.4 Statistical significance and multiple testing problem

The statistical significance of a result is the probability that the observed effect in a study occurred by pure chance, and that in the population from which the sample was drawn, no such relationship exist [47,48]. The higher the probability of error that is involved in accepting an observed result as a representative of the population, the less confidence there is that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population [14,42,46]. Specifically, a p-value of 0.05 (i.e.,1/20) indicates that there is a 5% probability that the relation between the variables found in our sample is a false positive. In other words, assuming that in the population there was no relation between those variables whatsoever, and we were repeating experiments, it's expected that approximately in every 20 replications of the experiment there would be one in which the relation between the variables in question would be equal or stronger than the 0.05 cut-off [6].

It follows that the more analyses you perform on a data set, the more results will meet by chance the conventional significance level. For example, if you perform 40 independent tests, then you should expect to find by chance that about two (i.e., one in every 20) tests are significant at the p-value < 0.05 level, even if the values of the variables were totally random and those variables do not correlate in the population. Some statistical methods that involve many comparisons and, thus, a good chance for such errors include some "correction" or adjustment for the total number of comparisons. However, many statistical methods (especially simple exploratory data analyses) do not offer any straightforward remedies to this problem. Therefore, it is up to the researcher to carefully evaluate the reliability of unexpected findings. In this

dissertation this will be referred to as the multiple testing problem, a common terminology used [22,23,24].

If there are very few observations (in the case of GWAS very few subjects), then there are also respectively few possible combinations of the values of the variables and, thus, the probability of obtaining by chance a combination of those values indicative of a strong relation is relatively high. This is the same problem as having to deal with low support but high confidence in data mining.

Consider this example from research on statistical reasoning [49]. There are two hospitals: in the first one, 120 babies are born every day; in the other, only 12. On average, the ratio of baby boys to baby girls born every day in each hospital is 50/50. However, one day, in one of those hospitals, twice as many baby girls were born as baby boys. In which hospital was it more likely to happen? The answer is obvious for a statistician, but as research [49] shows, not so obvious for a lay person: it is much more likely to happen in the small hospital. The reason for this is that the probability of a random deviation of a particular size (from the population mean), decreases with the increase in the sample size.

If a relationship between variables in question is small, then there is no way to identify such a relation in a study unless the research sample is correspondingly large. Even if our sample is in fact "perfectly representative", the effect will not be statistically significant if the sample is small. Analogously, if an association in question is very large, then it can be found to be highly significant even in a study based on a very small sample.

In the case of genetic analyses on high dimensional GWAS, the number of predictor variables is very high (on the dataset used to evaluate the performance of this system 550,000). However, the number of samples (in this case subjects) is comparably small

(roughly 2000). This is due, in one hand the difficulty of finding criteria matching cases of a disease under study and second on the large cost of associated with finding, evaluating, collecting data and consent forms from large sample populations. To make matters even worse, it is commonly believed that the current classification of most diseases is likely to have multiple different genetic factors contributing to them. In terms of the biological mechanisms involved there may be completely different genetic mechanisms involved that exhibit the same phenotypes and were thus classified together as a single disease. This reduces further the statistical power of detecting such effects in GWAS but it also provides a good argument for the analyses of disease subtypes [50].

2.2.5 Main effect

In the design of experiments and analysis of variance, a main effect is the effect of an independent variable on a dependent variable averaging across the levels of any other independent variables [51]. The term is frequently used in the context of factorial designs and regression models to distinguish main effects from interaction effects.

2.2.6 Epistasis

In genetics, epistasis is the phenomenon where the effects of one gene are modified by one or several other genes, which are sometimes called modifier genes. The gene whose phenotype is expressed is called epistatic, while the phenotype altered or suppressed is called hypostatic. Epistasis can be contrasted with dominance, which is an interaction between alleles at the same gene locus.

In general, the fitness increment of any one allele depends in a complicated way on many other alleles; but, because of the way that the science of population genetics was developed, evolutionary scientists tend to think of epistasis as the exception to the rule

[3]. In the first models of natural selection devised in the early 20th century, each gene was considered to make its own characteristic contribution to fitness, against an average background of other genes [1].

Epistasis and genetic interaction refer to different aspects of the same phenomenon. The term epistasis is widely used in population genetics and refers especially to the statistical properties of the phenomenon, and does not necessarily imply biochemical interaction between gene products. However, in general, epistasis is used to denote the departure from "independence" of the effects of different genetic loci. Confusion often arises due to the varied interpretation of "independence" between different branches of biology. For further discussion of the definitions of epistasis, and the history of these definitions, see [3]. For the purposes of this dissertation the term epistasis is used to refer to the statistical properties of the phenomenon.

The presence of epistasis can have important implications for the interpretation of statistical models. If two variables of interest interact, the relationship between each of the interacting variables and a third "dependent variable" depends on the value of the other interacting variable. In practice, this makes it more difficult to predict the consequences of changing the value of a variable, particularly if the variables it interacts with are hard to measure or difficult to control [5].

Chapter 3 Literature Review

A detailed review of the areas of research associated with the original contributions of this work is presented in this chapter. Existing proposed solutions to these problems are identified and described. Performance parameters are identified for every problem to enable a structure comparison methodology for the different approaches to addressing each problem. The first look is at different ways of encoding GWAS data. Then a look at the measures used to quantify epistatic effects. Ways of computing contingency tables are looked at followed with literature on significance testing through replication. Finally, attempts to address the problem of providing a 2 SNP interaction testing methodology are listed, split into two categories, those that focus on utilizing an HPC so that a near exhaustive search can be performed and those that rely on data mining techniques to identify significant effects while only spanning a small subset of the search space.

3.1 Data encoding

When the work for this framework started, the only widely available non proprietary format for storing GWAS data was the QTDT MERLIN format [12]. While however the work for research and development was taking place another format was developed by a different group referred to as PLINK binary format, after the software that it was introduced in [13]. Key limitations of the encoding in the PLINK binary format were discovered. Therefore through the work proposed in this dissertation an alternative format was developed and published. In this section the commonly used format introduced in the software MERLIN [12] and the newer binary encoding

introduced in PLINK [13] are discussed and in the following chapters a newer encoding that was proposed as part of this dissertation is introduced.

3.1.1 The MERLIN format

The file format described in MERLIN [12] is the one traditionally used for this type of data. It is split into three files: pedigree, map and data files. Pedigree files contain phenotypes for discrete and quantitative traits and marker genotypes for a specific number of subjects. They are usually white-space delimited files. The first (usually 6) columns contain information about the subject (Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype). The combination of the information of each subject must be unique. The next columns contain bi-allelic markers; typically SNPs. Marker genotypes are encoded as two consecutive integers, one for each allele, or using the letters "A", "C", "T" and "G". To denote missing alleles a sentinel value is used, typically "0". An example of a sample ped file is provided in Table 1. It's worth noting that the Merlin format is designed to support pedigree or lineage studies, were the subjects included in them are related. All studies relevant to this dissertation are case-control though and are designed to exclude related individuals, so each subject is from a single family family.

Map files contain information for each single nucleotide polymorphism. They are used to analyze genetic markers into the equivalent pedigree file. Each line per marker usually contains 3, 4 or 5 columns (chromosome, SNP identifier, morgans or centimorgans and base-pair position). Each column is separated by white space.

Dat files describe the pedigree file. They include one row per data item in the pedigree file, indicating the data type providing a one-word label for each item.

Table 1 A sample Merlin format pedigree file

Family	Person	Father	Mother	Gender	Disease	SNP1	SNP2	...	SNPn
1	1	0	0	m	1	AT	GA	...	CG
2	2	0	0	m	1	00	GA	...	CC
3	3	0	0	f	1	AA	AA	...	CG
4	4	0	0	m	1	TT	00	...	GG
5	5	0	0	f	2	AT	GA	...	00
6	6	0	0	f	2	AA	GA	...	CC

*0 is used to encode missing data.
* Disease is given as 1 for controls, 2 for cases

3.1.2 PLINK's Method for binary ped files

PLINK is an open source program offering a comprehensive range of basic large-scale whole genome association analysis methodologies. It has been widely adopted since high dimensionality GWAS have become available as it enables researchers to efficiently analyze these large datasets in a computationally efficient manner.

In PLINK there is an encoding format for transforming QTDT MERLIN data into binary formatted files. The approaches used in PLINK uses 2 bits for encoding bi-allelic markers with 4 possible states. PLINK uses the encoding for each genotype given in [13].

Testing on PLINK binary format showed that the exported binary file was 15 times smaller than the original file. The drawback however is that encoding of the heterozygote allele is the same regardless of what strand it's actually derived from. Therefore any analyses that rely on the sequence of the alleles on the strand will be missing this information.

One analysis technique that needs the lost information is imputation. Imputation analysis is the practice of "filling in" missing data with plausible values. It is a method for uncovering the genetic basis of human disease and it is used for inferring genotypes at observed or unobserved SNPs that can detect causal variants that have not been directly genotyped [9]. It is in essence an in-silico approach to discover the probability of the existence of a specific genotype for loci that haven't been directly genotyped but are known to be in LD with genotyped markers.

Table 2 PLINK binary format data encoding [13,52]

Allele On Strand +	Allele On Strand -	Marker Encoding
A	A	00
A	a	
a	A	01
A	a	11
Missing Data	Missing Data	10

3.2 Measures of epistasis

The gold standard for a statistic that measures the epistatic effect is regression analysis. Since in this dissertation the focus is on categorical response variables, the gold standard used will be logistic regression. There are two different measures that researchers are most interested in. These are testing for interaction between two factors, and test for association allowing for interactions [3,8,26,32]. In this research work, the proposed test for interaction is referred to as the epistasis test, while the omnibus test falls in the category of tests for association allowing for interaction. The difference is that epistasis only measures the interaction between the two SNPs, while the omnibus effect represents the combined effect of the main effects of the SNPs and the epistatic effect itself. Logistic regression can be used to derive both a test of interaction as well as a test allowing for interaction, thus it traditionally forms the standard by which to benchmark any new proposed methodologies [32].

However, logistic regression is computationally demanding. In order to develop any methodology that attempts to perform testing at the whole genome scan level even on a relatively small subset of SNPs faster measures are needed [32,53]. Thus some alternatives to logistic regression have been introduced. All of the proposed methodologies are compared to logistic regression. Similarly in this dissertation the proposed methodology will be compared with logistic regression as well.

Table 3 gives an overview of each of the proposed methods in the literature for performing a two SNP interaction test,. The table indicates for each test the following:

- Epistasis test: Does the method provide a statistic measure of the interaction between the two SNPs (epistatic effect).

- Allowing for epistasis test: Does the method provide a statistic measure of association between two SNPs allowing for the epistasis test, that is a test of the two main effects and the interaction between them combined as in the omnibus test proposed in this dissertation.

- Response variable: List of all possible types of response variables the method can analyze. Some methods such as regression can analyze more than one type of response variable (i.e. logistic regression tests for categorical response variables and linear regression for continuous response variables).

- Adjustment of covariates: Occasionally, a known bias may exist in a dataset, or it may be thought to exist. Adjusting by covariates removes this bias from the test performed.

- Marker type: In genetics there are two possible ways to consider SNP data. The first is the allelic approach, where each allele is scored independently, and the second is the genotypic approach where each genotype, that is the alleles on both strands for a specific SNP is considered.

- Tested on real data: Has the measure being tested on real data in a peer reviewed publication providing evidence to its performance.

- Replicated Result: Has an analysis applying the specific measure produced replicated statistically significant results in independent datasets.

- Bias: Are there any known biases in the proposed measure that may affect it's accuracy.

- Results interpretation: How are the results interpreted. The preferred interpretation is as the probability of getting the same result as a false positive (p-value).

- Computational requirements: The computational requirements of each method are ranked as high, medium or low based on published reviews of the methods.

- References: A list of references related to each method, either publications introducing or reviewing the methods.

Table 3 Comparison of existing measures of epistasis using statistical or data mining approaches

Measures	Epistasis test	Allow for epistasis tests	Response variable	Adjustment by covariates	Marker type	Tested on real data	Replicated results	Bias	Result interpretation	Scalability	References
Regression analyses	Yes	Yes	Categorical, linear	Yes	Genotypic Allelic	Yes	Yes	*LD, in some cases	p-values	Low	[13,54,55]
Odds ratio Multiplicative Interaction	Yes	No	Binary	No	Allelic	Yes	No	*LD, *MAF, size of dataset, Heterozygote effects	Approximated p-values	High	[3,32,56]
Case only analysis (χ^2	Yes	No	Categorical	No	Genotypic	Yes	No	Non linkage equilibrium	p-values,	High	[13,32]
Recursive Partitioning	No	Yes	Categorical	No	Genotypic Allelic	Yes	No	*LD, main effects	None, requires follow-up analyses	Medium, parameter dependant	[32,57,58]
Multi-dimensionality reduction (MDR)	No	Yes	Categorical	Yes	Genotypic Allelic	Yes	No	*LD, main effects	None, requires follow-up analyses	Low	[59,60]

*MAF= minor allele frequency

*LD= Linkage Disequilibrium

3.2.1 Logistic regression

The standard widely accepted measure of the effect of one or more terms in a regression model is provided by comparing the deviances obtained from fitting the model with and without the term(s) in question [55]. By using regression analysis, an allelic or genotypic test can be performed. Linear regression analyses can be used to perform interaction testing with continuous response variables rather than categorical. Implementations of epistasis test using logistic and linear regression models are available in the genetic analysis software PLINK [13] and are widely used [3,8,16,32]. This PhD dissertation is focused on categorical response variables and genotypic testing, therefore in order to provide a fair comparison between the proposed method and logistic regression the corresponding model of logistic regression for genotypic analysis of interaction is defined in the methods chapter, and implemented in R a statistical package [61].

3.2.2 Odds ratio multiplicative interaction measure

As a faster alternative to logistic regression, another measure of interaction for epistasis testing sometimes used is an odds ratio multiplicative interaction measure [13]. A recent publication [40] also proposed a new methodology for interaction testing that draws similarities to this measure. In that work, both a genetic and an allelic model of interaction testing based on odds ratio multiplicative interaction measure were presented and compared to a proposed pseudo-haplotype based measure [40]. They tested their hypothesis by analyzing GWAS in two independent datasets; however, they were not able to perform all possible two SNP interactions as in the proposed methodology in this dissertation. They cited the extreme computational complexity of the problem as the reason for that, although they do indicate it as their

future work goal to perform a complete test. They used a heuristic to limit the number of SNPs to those that were on genes reported to be involved in a total of 501 assembled pathways generated by a simple computational approach to reduce the number of interactions to test, as well as to increase the likelihood that any results found would be in genes involved in pathways of interest. Although they report discovering replicated results, these are only evident after they completely analyzed both datasets and compared the top results of both analyses. Since this involves the creation of two large lists of top results and the testing for overlap between them, it's not clear if the number they get as overlapping deviates significantly from what you would expect from a random association or not. An examination of the reported replicated interactions for evidence from the literature for their association to the disease is attempted in order to improve the confidence in the findings, however, since they limited the SNPs they tested by selecting pathways known to be associated with the diseases, it's expected that any false positive results would also be associated with the disease. The conclusion reported in this paper is that further research needs to be conducted, so that complete two SNP interaction testing is possible in a GWAS.

3.2.3 Case only analysis with chi-square

Another proposed way of performing epistasis testing is through a simple association test between two markers performed in just the cases, under the assumption of linkage equilibrium [32,53]. In theoretical terms, this analysis carries more statistical power, but it's this assumption of linkage equilibriums that doesn't hold in real data that provides that additional statistical power [32]. This approach can easily be extended to perform either genotypic analyses or allelic. In genotypic analyses a contingency table with size 3*3 will be created with the first SNP having its three genotypes on each of

the three columns, while the second SNP's genotypes will be represented in the rows. An allelic test follows a similar approach with a 2 by 2 contingency tables where the two alleles of each SNP are represented similarly in the corresponding rows and columns. A chi-square statistic or any other test for association applicable to a contingency table can reveal the level of association between the two SNPs.

This approach assumes non independence between LD, much like all other analyses presented on Table 3.

However, this method is affected to a considerably greater degree by non independence between markers. The test itself is a test of dependence, and the assumption here is that if there is dependence since there is an assumption of linkage equilibrium then that dependence must be caused by an epistatic effect associated with the disease the cases are subject to. This is not the case, even if two SNPs are on different chromosomes; they may still not be in linkage equilibrium. As an example, consider a case where two genetic locations do interact and polymorphisms combinations within these locations cause the subject's probability to reproduce to be significantly reduced compared to those with different genotypes at those locations. The two regions are not in equilibrium since less of the subjects would be expected with those specific rare polymorphisms. Thus in this analytical approach, those two polymorphisms will be reported as epistatic effects for the disease the cases were taken from, even though they may have nothing to do with the disease, and the frequencies of those two polymorphisms even though not in equilibrium would be very similar to a matching control population.

An implementation of this approach is presented in PLINK [13] called, case-only analysis. In an attempt to reduce the probability of having markers tested together for epistasis that are in LD, only tests on markers over a certain distance on the genome

are tested, however reviews of this method using real data revealed that the statistically significant results can be driven by non linkage equilibrium rather than true epistasis associated with the disease.

3.3 Computing multidimensional contingency tables

Often, data derived from natural sciences, fields that use a scientific method to study nature, come in the form of multidimensional tables of counts, referred to as contingency tables [62]. Generating contingency tables is a known P hard Problem [17]. The classic problem that scientists have worked on over the years has been how to compute the expected cell counts for the different statistical models used in analyzing contingency tables efficiently with an acceptable accuracy. In this work however, effort was placed in avoiding any methodology that diluted information derived from the already weak in terms of statistical power analysis of two SNP interaction testing in GWAS. Therefore, when computing multidimensional contingency tables, focus was placed on efficiently computing the tables with no statistical power loss. In order to do this we focused on the need to count multiple contingency tables for different response variables [1,16,41].

3.3.1 The classical data driven approach

Traditionally, when there is a need to count a multidimensional contingency table, the standard approach is to go through the data of each subject, and increment the count of the contingency table cell he belonged to. A complete pass through all the subjects will generate a contingency table for a response variable. However, as is often the case in analyzing data derived from natural science studies, there's more

than one response variable that is of interest. In current analytical frameworks of GWAS, hypothesis testing is grouped into response variables. The analysis of all hypothesis tests relating to a response variable is performed through a single pass of the dataset as presented in the form of pseudo code in Figure 4. However, when more than one response variable is involved in the hypothesis to be tested the process needs to be repeated, resulting in at least as many passes through the data as there are response variables.

Algorithm: Compute Contingency table
Input:

- **SNP1 genotypes for all subjects**
- **SNP2 genotypes for all subjects**
- **Array of response variables for all subjects**

Output:

- **3dTable: a four dimensional table with all contingency tables in it scored**

Description: This function passes through all the subjects in the dataset once and scores based on the 2 inputted SNPs genotypes each of the contingency tables in the 3d matrix.

1   FOR all Subjects
**   a.   Z=ResponseVeriableCategory(Subject)**
**   b.   X=SNP1genotype(Subject)**
**   c.   Y=SNP2genotype(Subject)**
**   d.   3dTable(Z,X,Y)++;**

Figure 4 Computing contingency tables, the traditional approach

3.3.2 The random walk approach, monte carlo sampling

Random walk approaches have been tested for this problem [17,19]. The basic concept behind all random walk approaches, is that a subset of the population can be counted that is sufficiently large to get an acceptable estimate of the actual count, had a complete count of the data taken place. The second type of random walk approaches, are not done to increase the computational efficiency, but to increase the statistical power. A recent example is the use of monte carlo sampling methods using Markov chains as presented in [18]. Neither of these methods however has gained popularity, mostly because they only provide a linear improvement in time based at a cost of not

testing the entire dataset. Since lack of statistical power is a key problem in this type of analysis, counting all subjects in the contingency tables takes priority over any method that only provides a linear increase in computational performance.

3.4 Evaluation of significance through replication

Replication testing is the cornerstone of evaluation of any finding in genetics [11,27,29,32,63]. Many reported strong genetic associations that made sense based on the knowledge of the relation of the underlying gene and the disease it was tested for were reported from a single study, but later failed to replicate in follow up analyses of independent datasets [29,27]. Even though failure to replicate doesn't necessary mean that the result is not a true positive as it could in theory be an effect with higher frequency in the population under study and therefore with more power of detection in that study. Generally, if a genetic marker is reported to have an association to phenotype in a single study if it's not replicated in an independent dataset it is not considered robust.

Replication testing methods can be split into two parts, the first being the verification of the effect size. This is easily performed by repeating the analyses with the same parameters in an independent dataset. However, even if there is statistical significance in both datasets, this may still be a false positive result if the distribution of the effect among the genetic markers categories is not the same between the two datasets. The next step is to try to acquire more confidence in the finding by examining if the effect is distributed similarly between all genotype combinations in both datasets. There are two ways to do this, merging the dataset and repeating the test, or examining the distribution of the effect across each of the genotype combinations.

3.4.1 Combine GWAS and repeat analyses

 Traditionally to get a more statistical power in an analysis, the independent datasets are combined and the analysis is repeated in the resulting dataset that is larger since it's an overset of the two dataset. The generated statistic will have considerably more statistical power since the sample size now is the combined of the two datasets, and if the effect carrying genotype combinations are matched in direction the resulting p-value should be stronger than either of the two independent datasets. If it is not, then the replication success itself is not considered proof of validity of the result. Even if the p-value of the combined dataset is statistically significant though, this does not necessarily mean that all genotype combinations carried the same signal in both datasets. It just indicates that the ones that did combined with the increased number of subjects and the resulting increase in power, provided stronger evidence for the effect.

3.4.2 Examine the distribution of effects across the genotype combinations

The second way of examining if the replication across two independent dataset was successful is to study the distribution of each effect among each of the genotypes in cases and controls in the two datasets. Assuming that the result is a true positive then the effect size should be similar in both dataset for all genotype + response variable combinations. In Univariate allelic analysis the odds ratios are used to study this for the main effect. However, simple odds ratios can't be used in the case of interaction testing since there are four degrees of freedom in Epistatic test, and 8 in the Omnibus test, and odds ratio's can be generated only when there are 2 degrees of freedom.

The information generated by examining the distribution of effects to the genotypes, is interesting to the interpretation of the results as well, as it will identify the key conditions that are associated with differences in disease predisposition.

3.5 Analytical frameworks for gene-gene interaction testing using GWAS data

Any solution to two SNP interaction problems in GWAS would need to address the problem of the high demand for computing requirements. Thus different researchers have attempted different approaches to this problem. In this section the most notable of these approaches will be presented so that they can in later sections be compared to and discussed in comparison to the proposed framework. Priority was given in approaches that attempted to perform exhaustive search, such as in the proposed framework, or near exhaustive search.

Table 5 provides an overview of the methods. These methods can be split into two broad categories, ones that rely on HPC computing resources to perform either exhaustive search, or near exhaustive search, and the second category is based on machine learning, or data mining techniques, that are attempting to identify effects of interest by going through only a small part of the search space. Finally, as part of this section, the two traditional techniques for filtering the data are explained, the candidate gene approach, and the main effect significance approach.

3.5.1 Two SNP interaction testing using HPC

3.5.1.1 Cluster HPC

In Marchini *et al* [54] highlighted the importance and feasibility of fitting interaction models using GWAS data. In that study a 10 node computing cluster was used to perform all pair-wise tests of association allowing for interaction on a simulated dataset of 300,000 loci in 1,000 cases and 1,000 controls. He quoted a time of 33 hours on that specific system. The PLINK web site [13], quotes 24 hours to test all pair-wise interactions at 1,000,000 loci with 500 subjects although it doesn't clarify what machine they used to verify this and their results are based on estimations using odds ratio interaction analyses that they recommend that they are validated trough follow up analyses with logistic regression.

It needs to be kept in mind when considering these numbers that the typical association study performed today can have 500,000 or even 1,000,00 markers using today's genotyping technology. And in the near future even higher dimensionality genotyping platforms are expected. In order to provide a level field for comparing the proposed methodology to the ones already conducted we estimate the runtime of each of the analyses if they had used the same dataset as the primary dataset used in this work that is composed of 550,000 SNPs with 1,000 case and 1,000 control subjects. The estimation is easy since the number of subjects (*s*) provides a linear change in time, and the number of SNPs (*n*) increases the number of pair wise combinations to test by:

$$\frac{n(n-1)}{2} * s$$

Based on this estimation, the Marchini *et al* approach on the same hardware 10 node clusters would take 4.62 days. While the PLINK approach would take 166.28 days.

Obviously, it would be best if all approaches could be tested on the same hardware using the same dataset with possibility of replication. But as described in 0 the proposed framework includes a custom hybrid cluster-cloud high performance computing platform.

3.5.1.2 GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies

GBOOST is an implementation of BOOST algorithm on a GPU (graphics processing chip) [56,64]. It's based on the Compute Unified Device Architecture runtime application programming interface (CUDA) [65]. GBOOST is reported as having a 40 fold performance improvement over BOOST's x86 implementation [64] (Table 4).

The statistical measure used as described is a derivative of logistic regression but with computational optimizations based on the Kirkwood superposition approximation [66] is one that attempts to perform a filtering based on the probability of a SNP pair to be significant. However, no definitive evidence is provided as to the efficiency of this filtering scheme understandably so since a true complete two SNP interaction would need to be performed to provide that information. Furthermore, the Kirkwood approximation that this approach is based on does not generally produce a valid probability distribution (the normalization condition is violated). Watanabe [67] claims that for this reason informational expressions of this type are not meaningful, and indeed there has been very little written about the properties of this measure.

The computational performance of BOOST was compared to PLINK in [64]. Results are shown in Table 4.

_____

Table 4 Reported time comparisons between PLINK BOOST and GBOOST

Number of SNPs	Number of Subjects	PLINK (3GHz CPU)	BOOST (3GHz CPU)	GBOOST NIVIDIA GTX285
1,000	5,000	106s	<2s	
5,000	5,000	2,703s	42s	1.04s
10,000	5,000	10,915s	170s	4.11s
351,542	5,003		60h	1.34h

These times were taken from [56].

PLINK is tested with the fast epistasis option (odds ratio based interaction test, not regression).

The reported times are based on a 3.0GHz CPU with 4Gbytes memory running Windows XP pro for BOOST and PLINK while for the GBOOST test an NIVIDIA GTX285 graphics card was used [64].

Xiang Wan et al., "BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies," *The American Journal of Human Genetics (AJHG)*, vol. 87, no. 3, pp. 325-340, September 2010.

Ling Sing Yung, Can Yang, Xiang Wan, and Weichuan Yu, "GBOOST: a GPU based tool for detecting gene-gene interactions in genome-wide case control studies.," *Bioinformatics*, vol. 27, no. 9, pp. 1309-1310, May 2011.

3.5.2 Data mining methods

Dealing with high dimensional data, and non-linear models using regression-based methods is often criticized when the data is expected to contain many interacting predictor variables resulting in sparse contingency tables (with many empty cells, or with cells with less than 5 counts) [68,69,70]. Data mining and machine learning methods have been applied as an alternative. These approaches do not attempt to fit a single pre-specified statistical model nor do they attempt an exhaustive search. They focus on stepping through the search space of possible models, not necessarily limiting to two SNP interactions but rather allowing for multi-way interactions. The goal is to identify the model that best fits the data in a computationally efficient manner. Based on this, in a review of methods for detecting gene-gene interactions in [32] Cordell argues, the distinction that is often made between data-mining and regression models is to some extend false. McKinney *et al* [69] provided a good overview of the most prominent machine-learning approaches for detecting gene-gene interaction. In the next few sections focus is given on approaches that have been applied to real data providing good performance indicators as to the validity of their results.

3.5.2.1 Recursive partitioning

Classification and regression trees are the basis of recursive partitioning approaches [58]. Trees are constructed based on rules that determine how well a split at a node (in this case representing SNP genotypes, or environmental factors) can differentiate observations with respect to the response variable. The traditional splitting rule is to use the variable that maximizes the reduction in a quantity known as the Gini impurity at each node [58]. Nodes are recursively split until either some preset stopping criteria

are met (such as the max number of SNPs involved in a rule) or if no further gain can be made (if all terminal nodes contain only cases or only controls). An example of the recursive partitioning approach steps is given in Figure 5.

Recursive Partitioning approaches do not test for interactions; rather they test for association allowing for interactions by computing paths through the tree that correspond to strong associations to the disease. Since however, recursive partitioning approaches at the very first stage, are relying on the main effect of SNPs to pick the first node to split, the likelihood of finding pure interactions in the absence of main effects can be missed [71].

In Figure 5, SNP 3 maximizes the reduction in the Gini impurity at the first node and is therefore chosen for splitting (according to the genotype at SNP 3) the original data set of 1,000 cases and 1,000 controls into two smaller data sets. Once a node is split, the same logic is applied to each child node (hence the recursive nature of the procedure). The splitting procedure stops when no further gain can be made (for example, when all terminal nodes contain only cases or only controls, or when all possible SNPs have been included in a branch) or when some preset stopping rules are met. At this stage, it is usual to prune the tree back (that is, to remove some of the later splits or branches) according to certain rules to avoid over fitting and to produce a final more parsimonious model [32].

_____
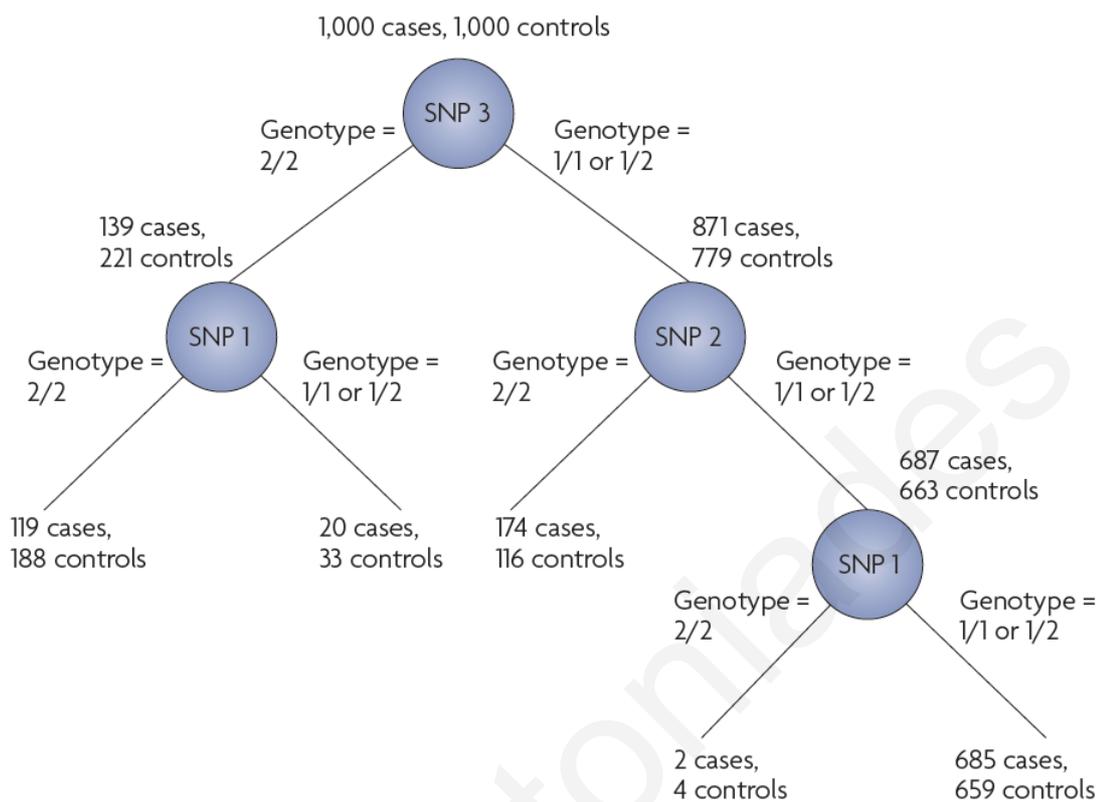
Figure 5 Recursive partitioning approach example

Heather J Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392-404, June 2009.

_____

3.5.2.2 Recursive partitioning with ensemble of trees (forest)

Substantial improvements in classification accuracy have been observed by growing an ensemble of trees rather than a single tree. A popular approach is the random forest approach which has been used in genetic studies [72,73]. In this type of analyses improvements in classification accuracy compared to the recursive partitioning approach of a single tree can be achieved by growing an ensemble of trees and letting them 'vote' for the most popular outcome class.

The most widely used ensemble tree approach is probably the random forests method [32,57]. A random forest is constructed by drawing with replacement several bootstrap samples of the same size (for example, the same number of cases and controls) from the original sample. An un-pruned classification tree is grown for each bootstrap sample, but with the restriction that at each node, rather than considering all possible predictor variables, only a random subset of the possible predictor variables is considered. This procedure results in a 'forest' of trees, each of which will have been trained on a particular bootstrap sample of observations. The observations that were not used for growing a particular tree can be used as 'out-of-bag' instances to estimate the prediction error. The out-of-bag observations can also be used to estimate variable importance in different ways including through use of a permutation procedure [32,69].

A key aspect of this approach is that there is no clear model identifying predictors either as associated or interacting. Therefore a follow up approach is required to actually test the important predictors output by the random forest approach. However, since the number of predictor variables (SNP) is small any method for testing for interaction or allowing for interaction can be applied without limitations in terms of computational complexity.

3.5.2.3 Multifactor dimensionality reduction

Multifactor dimensionality reduction (MDR) seeks to identify combinations of loci that influence a disease outcome in a way that would classify it as a test allowing for interaction rather than discovering pure interactions. MDR reduces the number of dimensions by converting a high dimensional multi locus model to a one-dimensional model, thus avoiding the issues of empty or small count contingency table cells.

The main problem with MDR, as with other exhaustive search techniques, is that it does not scale up to allow analysis of large numbers of SNPs. If an exhaustive search for the best n-locus combination (within each of ten cross-validation replicates) is performed, anything more than a two-locus screen on more than a few hundred variables will be computationally prohibitive [32]. For investigation of higher-order interactions, MDR is therefore perhaps best suited for use with small numbers of loci (up to a few hundred), which have perhaps been identified through some sort of informed filtering step.

MDR's output is a classification of genotypic classes as either high risk or low risk according to the ratio of cases and controls in each class. This approach is considered overly simplistic, and improvements that embed a more traditional regression-based approach allowing application of the method to continuous as well as binary traits and adjustment for covariates, have been proposed [74]. However these make the algorithm even slower increasing the problem of scalability.

MDR has been used to identify potential interacting loci in several GWAS, including breast cancer, type 2 diabetes, rheumatoid arthritis and coronary artery disease, although to date it is unclear whether any of these identified interactions have been replicated in larger samples [32].

3.5.3 Filtering approaches

Since in all of the methods mentioned above, computational complexity is a major issue, one of the classic ways of enabling testing for two SNP interactions in available computational resources was to limit the search space. Filtering of the number of SNPs is the most obvious choice, due to the exponential growth of the search space based on the number of SNPs. The first and most common perhaps approach to filtering the number of SNPs was actually derived from the early approach of identifying what SNPs to genotype. Before high dimensional GWAS studies were able to genotype hundreds of thousands of SNPs at a time at low cost the typical approach to identify key SNPs to genotype was through candidate gene studies [75,76]. The methodology of filtering to a small number of SNPs in candidate gene studies relies on identifying genes, or regions in the genome that are thought to be associated with a disease. These are typically identified through an extensive literature review of the functional biological and genetic knowledge of the disease. Then once all regions of interest are identified tagging SNPs within them would be computed. Tagging SNPs are SNPs that are estimated through their LD in the general population to represent the majority of polymorphisms within their LD region. A common framework for identifying tagging SNPs in multiple populations and regions was provided through HapMap [36]. This approach however, limited the search to regions of known interest; as a result, it was impossible to identify effects that involved regions outside the candidate gene regions. Also, another negative effect that emerged with these candidate gene approaches was that once the analysis was complete, no matter what the top result was, it was certain to have some association with the disease, as that was the inclusion criteria in the study. Therefore, the opportunity of looking for replicated

information discovery from other types of experiments relating top results with the disease was lost.

Another typical methodology for filtering of SNPs is to use the univariate analysis results to limit the number of SNPs to include in the analyses. In this approach, the main effect of each SNP in the GWAS is calculated, and then only SNPs that pass a preset threshold of main effect are included in the interaction tests. However, this approach will not test pairs of SNPs where one or both of the SNPs have no or very low detectable main effect in that specific study. There's no reason to believe that SNPs with low or main effects won't be involved in epistatic effects.

Table 5 Two SNP interaction testing frameworks

Multi-locus Interaction testing framework	Computing Platform	HPC Scalable	Measures	Largest number of tests recorded	Deterministic	Accessibility	Tested on Real data	Replicated statistical significance	References
Simple exhaustive search Marchini *et al*	10 node cluster	Yes	Simple association test allowing for interaction	300k SNPs 2000 subjects	Yes	Low	No	No	[54]
BOOST	Single core x86 cpu	No	Odds ratio based interaction measure	351k SNPs 5000 subjects	Yes	High	Yes	No	[56]
GBOOST	GPU	Limited Scalability	Binary odds ratio based interaction measure	351k k SNPs 5000 subjects	Yes	High	Yes	No	[64]
Recursive Partitioning (Tree)	Symmetric multiprocessing under development	No	Follow up analyses necessary to provide interpretable measure	Limited to only a few dozen SNPs	Depends on parameters.	High for small number of tests	Yes	No	[58]
Random Forest	Symmetric multiprocessing, in theory HPC compatible with many architectures	No	Follow up analyses necessary to provide interpretable measure	Limited to only a few dozen SNPs	No	High for small number of tests	Yes	No	[57]

Chapter 4 Two SNP Interaction Framework

This section describes the methodology of the proposed framework and its evaluation. The first section describes the functional requirements that were defined following the initial literature review for this work. The algorithmic approach for each original contribution is then presented beginning with data encoding that enables lossless data size reduction. Then, the proposed statistical measures for testing and allowing for epistasis are presented, followed by an algorithm that improves computational efficiency when calculating multidimensional contingency tables in a four-dimensional scoring matrix. An a-priory definition of the steps to verify the replicability of the results is also provided. Next an algorithm is proposed that enables the use of two independent HPC resources with different architectures in a way that significantly improves computational efficiency compared to using either HPC alone. Finally, the complete proposed framework for two SNP interaction testing is outlined in its entirety.

4.1 Functional requirements

Through the literature review it became evident that the following functional requirements are the most essential in any new methodology for multi-locus analyses to be widely adopted. Identifying and categorizing them helps in both providing the structure when presenting the methodology in this section as well as it helps with the evaluation of the proposed methodology in the chapters to follow.

1. Algorithmic

    1.1. Deterministic methodology: Needs to be a deterministic methodology that will always produce the same results in the same dataset.

1.2. Dimensionality scalability: It should be capable of running GWAS of at least 500,000 SNPs utilizing existing computational infrastructure and it should be capable of scaling up.

1.3. Avoidance of locus inclusion criteria bias: Quick heuristics that rely on the main effects of loci to determine if they will be included in the multi-locus analyses should be avoided as this is contradictory to biological knowledge.

2. Statistical Genetics

2.1. Capability to test for the epistatic as well as the omnibus effect.

2.2. Heterogeneity: Any effect of heterogeneity in the data such as linkage disequilibrium should be identified and accompanied with ways of detecting heterogeneity biased results.

2.3. Feature bias: The statistical measures used should not be biased towards any features of a locus such as the minor allele frequency.

3. Replication Testing

3.1. Availability: Repeating the analyses in a possible future replication database should be possible.

3.2. Exploration: The results produced by the method need to be presented in a user-friendly, familiar, and highly scalable environment that enables querying and visualization.

4.2 Data encoding

When designing the encoding of the data, focus was given to providing an encoding that doesn't reduce the amount of information in the data in any way. This would enable backing up the data in this encoded format and keeping it in the HCC-HPC after the analyses. This would also enable the performance improvement of future

analyses of any other type that uses the same encoding since if the analysis is performed on the same distributed high performance computing resources the data transfer of the data to nodes will be significantly reduced. Performance improvement of future analyses is only applicable when the data remains on the nodes and some of the same nodes are available for the analyses however.

Since some analyses such as imputation (a type of statistical analyses aimed at substituting missing values with ones estimated in-silico based on LD of the missing data with existing ones for every subject) requires knowledge of which strand the alleles of heterozygote SNPs exist on, we need to encode each allele on each strand separately for all cases. There are a minimum of three states each allele can be in, these include the two possible nucleotides (commonly denoted as A and a) as well as the possibility of missing data at that location. The smallest number of bits that can encode the 3 states of an allele is 2, however with 2 bits we can actually encode a fourth state. In many existing studies this may not be used, even though it will have no impact on the capacity requirements the databases will have for storage. However, we propose that the fourth state is set to denote markers that are in deleted regions. This utilizes the extra available coding to identify a deleted allele from a missing allele due to quality control concerns, or genotyping error when that information is available. This proposed encoding was peer-reviewed and published in [52].

An analytical technique that enables the detection of these deletions that has started being applied is copy number variants (CNV) analyses. It refers to the genetic trait of differences in the number of copies of a particular region (for example a gene) present in the genome of an individual [10, 11]. To perform CNV analyses most algorithms rely on raw data from the genotyping platform. CNV algorithms are able to detect

deletion regions as well as regions that are duplicated that may exist in some individual's strands.

However it should be clearly noted that CNV incorporates more information than just deleted regions. It can also detect regions that exist in more than one copy per strand. This information is not reflected in the proposed protocol; therefore methodologies that use this information would still rely on an external file for that information.

In the proposed format the data are structured as a two dimensional vector of alleles. The first dimension's size is equal to the number of strands the subject has. Typically in humans all chromosomes have two strands with the exception of X and Y chromosomes in males that each have 1 strand. In these cases a second strand can exist listing the alleles of the second strand on males as missing [52].

Each element in the vector of each strand will encode an allele on a single strand. The allele will be 2 bits long enabling encoding of a total of four states per allele missing data, nucleotide 1, nucleotide 2 or deleted. Table 6 presents the four different states that can be coded per allele. The term "Unknown" is used rather than the more typical "missing" to denote alleles that it's unclear what their genotypes are or if they are deleted so as not to confuse it with the deleted state that defines alleles that do not exist on that strand.

The two strands in each Subject's vector need to be perfectly aligned, that is, the i th element of each vector will point to the same marker's alleles one for each strand. To access the i'th marker's alleles the two bits at position i in each strand will carry a total of 4 bits, using the encoding column of Table 3. To access the i'th marker's alleles the two bits at position i in each strand will carry a total of 4 bits, using the encoding column of Table 3 as shown in Table 7.

Table 6 Proposed allelic Encoding

Allele	Encoding
Unknown	00
A	01
a	10
Deleted	11

Table 7 Proposed SNP genotype encoding

Allele 1	Allele 2	Encoding
Unknown	Unknown	0000
Unknown	A	0001
Unknown	a	0010
Unknown	Deleted	0011
A	Unknown	0100
A	A	0101
A	a	0110
A	Deleted	0111
a	Unknown	1000
a	A	1001
a	a	1010
a	Deleted	1011
Deleted	Unknown	1100
Deleted	A	1101
Deleted	a	1110
Deleted	Deleted	1111

4.3 Measure of epistasis

In order to provide a fair comparison between the proposed and traditional approaches to interaction testing, all test statistics need to be testing identical hypothesis as the proposed measures. Thus the statistical model used in the logistic regression analyses needs to be defined to match the test proposed. In PLINK, although a logistic regression measure is proposed, it is based on an allelic analyses, however this dissertation work has focused on a slightly more complex and perhaps closer to biological reality genotypic model, thus it needs to be defined.

The input to both the proposed and the test statistic are the contingency table for the tests. In the case of the main effect, a two dimensional contingency table made up of categories of the response variable as rows and the three possible genotypes of the SNP tested as columns. In Table 8 the subject type phenotype that defines if a subject is a case or a control is used as the response variable. The number of columns will always be three in this type of analyses (the number of genotypes of a bi-allelic SNP); however the number of rows depends on the number of categories in the response variable used. To identify the different counts of the cells the label U is given to represent main effect contingency tables.

Table 8 Contingency table for a single SNP

	aa	aA	AA
Case	U_{00}	U_{01}	U_{02}
Control	U_{10}	U_{11}	U_{12}

Table 9 is used directly to estimate the omnibus effect by performing a Pearson's chi square statistic analyses. It's in reality not a two but three dimensional as shown in Figure 6, with each SNP taking up its own dimension. When referring to cells on this table the letter *M* will be used.

Table 9 Contingency table for two SNPs (2d representation) and response variable disease status

	aa bb	aa Bb	aa BB	Aa bb	Aa bB	Aa BB	AA bb	AA bB	AA BB
Case	M_{000}	M_{001}	M_{002}	M_{010}	M_{011}	M_{012}	M_{020}	M_{021}	M_{022}
Control	M_{100}	M_{101}	M_{102}	M_{110}	M_{111}	M_{112}	M_{120}	M_{121}	M_{122}

_____
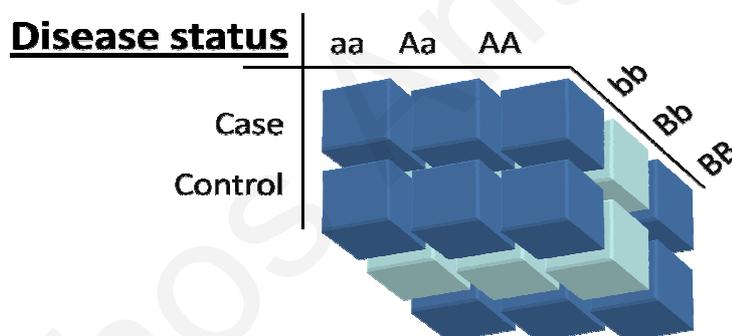
Figure 6 Contingency table for two SNPs (3d representation) and response variable disease status

_____

4.3.1 The proposed measure

The effect of a single marker's association to a phenotype (main effect) can be expressed by the following model [2][47]:

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \alpha_i + \varepsilon_i$$

where

p_i = observed proportion of subjects having the ith genotype at Locus A that display the phenotype

μ = overall log(probability) that a subject displays the phenotype.

α_i = main effect of the ith genotype at locus A.

ε_i = residual effect on the term on the left hand side of the equation (the response variable).

Most current research in genetics is concentrated on this type of effect, testing the following null hypothesis:

$H_0(1)$: $\alpha_i = 0$ for all i.

For the purpose of this dissertation study, this hypothesis is tested by the Pearson's chi-square test [77] for association between the row and column classifications in the contingency tables presented in Table 8 and Table 9.

There is a contingency table for each SNP that is used to estimate the Pearson's chi-square for the main effect (two degrees of freedom) of each SNP and a contingency table for each two SNP combination used to estimate the omnibus effect (eight degrees of freedom) with the proposed statistical measure as presented on Table 9.

However, it is possible to extend this model to represent the combined effect of two SNPs on the phenotype, interpreted in terms of the main effect of the first SNP, the

main effect of the second SNP, and the interaction between the two. The extended model proposed in this study can be represented as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

where

p_{ij} = observed proportion of subjects having the ith genotype at Locus A and the jth genotype at Locus B that display the phenotype

β_i = main effect of the jth genotype at Locus B

$(\alpha\beta)_{ij}$ = effect of interaction between the ith genotype at locus A and the jth genotype at locus B

ε_{ij} = residual effect on the response variable.

and the other terms defined as before.

The following additional null hypotheses may then be tested:

$H_0(2)$: $\beta_j = 0$ for all j

$H_0(3)$: $(\alpha\beta)_{ij} = 0$ for all combinations of i and j.

Chi-square statistics and p-values were obtained from the following significance tests:

    i.   an omnibus test of the main effects and interaction, i.e. a test of $H_0(1)$, $H_0(2)$ and $H_0(3)$ simultaneously, and

    ii.   a test of the interaction effect – also known to geneticists as the epistatic effect , i.e a test of $H_0(3)$ only.

The omnibus test combines both the additive and interaction effects between two SNPs and the phenotype [2]. It was performed using the chi square test for association between the row and column classifications in a contingency table of two SNPs such as the one in Table 9. Yates' correction for continuity [78] was used to prevent

overestimation of statistical significance when contingency tables have a less than 5 count.

Epistasis occurs when one genetic variant modifies the effect of another, i.e. when the two SNPs interact. The test of the epistatic effect is performed using the approximate additivity of chi-square statistics for orthogonal effects [46], to obtain the test statistic

$$\chi^2_{\text{epistasis, A.B}} = \chi^2_{\text{omnibus, A.B}} - \chi^2_{\text{main effect, A}} - \chi^2_{\text{main effect, B}}$$

and its degrees of freedom

$$\begin{aligned} \text{DF}_{\text{epistasis,A.B}} &= \text{DF}_{\text{omnibus,A.B}} - \text{DF}_{\text{main effect,A}} - \text{DF}_{\text{main effect,B}} \\ &= 8(x-1) - 2(x-1) - 2(x-1) \\ &= 4(x-1) \end{aligned}$$

where x is the number of categories of the phenotype [2].

The algorithmic steps of calculating the proposed omnibus and epistatic effects are focus around the Pearson's chi-square test. The pseudo code for the implementation of the chi-square test is shown in Figure 7 . The first step is to generate the totals of each row and column entry and of the entire table. The contingency table with all totals is shown in Table 10. The expected contingency table is generated that represents the expected contingency table under the null hypothesis. The expected contingency table along with all the formulas to generate it is shown in Table 11. The formula for estimating the chi-square statistic with Yates correction [78] is:

$$\sum_{0,0}^{c,r} \frac{\left(O_{c,r} - E_{c,r} - 0.5\right)^2}{E_{c,r}}$$

where O and E are the observed and expected tables respectively, c and r are the columns and rows, and the subtraction by 0.5 is done for the Yates correction.

The function that estimates the proposed omnibus and epistatic effect statistics is shown in Figure 8. This function first performs a direct call to the chi square function with the observed table of 2 SNP interaction (Table 9) as input. The result is the

proposed omnibus test. The next step is to estimate the main effects of the two SNPs. To do this there are two possible approaches. The first is to go through the steps as outlined in Figure 8, estimating the contingency table for each SNP by adding cells from the two SNPs contingency table and then performing a chi square on each resulting one SNP contingency table. The second is to estimate the main effect of each SNP at the beginning of the process and then just use those to estimate the proposed epistasis measure (Figure 8, Step 4) skipping two steps. Although this may seem like a nice optimization, in reality it's not, as in the first approach the contingency tables used to generate each SNP would reflect the missing data of both SNPs, while in the second case, missing data in one SNP would not be reflected in the contingency table of the second. The actual program developed implements both approaches, but preference is given to the first, as it provides more accurate results and a fairer, estimation of the runtime required.

_____

Algorithm: Chi-square
Input:
      Contingency table to calculate the Peasrson's chi-square (table noted as _O as in_ _observed table_)
Output:
      Chi square statistic
Description:
      This function will perform the chi square test on a contingency table of any size.

    // **First step is to estimate the contingency table totals.**
1   **FOR** c=0; c<**number of columns;** c++
      For r=0; r< **number of rows;** r++
            Σ_c+=$O(r,c)$
            Σ_r+=$O(r,c)$
    FOR all totals in rows t_r
      Σ_{all}+=t_c // **Table 10 shows the totals estimated**
2 // **Second step, generated expected table. E shown in Table 11**
    FOR c=0; c<**number of columns;** c++
      FOR r=0; r< **number of rows;** r++
            $E(c,r)$=$(\Sigma_c * \Sigma_r)$ / Σ_{all}
            Chi_Square= $[O(r,c)$- $E(c,r)$ -0.5 $]^2$/ $E(c,r)$ // **the -0.5 is for the Yates correction**

*All variables are considered to be initialized to 0.

Figure 7 Pseudo code for Pearson's chi-square test with Yates correction for

continuity

_____

Algorithm: OmnibusAndEpistasisTest
Input:
      Contingency table for 2 SNPs and a single response variable. (Labelled _M as in_ Table 9)
Output:
      Omnibus measure
      Epistasis measure
Description:
      This function will perform the chi square test on a contingency table of any size.

1.   // **First step is to estimate the Omnibus measure. This is a simple call to Chi-square.**
    Omnibus=**ChiSquare(**_O_**)** // **(Figure 7)**
2.   // **Estimate main effect observed tables (U) and chi square statistic of main effect for SNP1 adjusted for missing data of both SNP1 and 2** _Main1_
    For r=0; r< **number of rows;** r++
      U_{r0}= $O_{r,0,0}$+ $O_{r,0,1}$+ $O_{r,0,2}$
      U_{r1}= $O_{r,1,0}$+ $O_{r,1,1}$+ $O_{r,1,2}$
      U_{r2}= $O_{r,2,0}$+ $O_{r,2,1}$+ $O_{r,2,2}$
    Main1= **Chi-square(**_U_**)** //**main effect of first SNP**
3.   // **Estimate main effect observed tables (U) and chi square statistic of main effect for SNP2 adjusted for missing data of both SNP1 and 2** _Main2_
    For r=0; r< **number of rows;** r++
      U_{r0}= $O_{r,0,0}$+ $O_{r,1,0}$+ $O_{r,2,0}$
      U_{r1}= $O_{r,0,1}$+ $O_{r,1,1}$+ $O_{r,2,2}$
      U_{r2}= $O_{r,0,2}$+ $O_{r,1,2}$+ $O_{r,2,2}$
    Main2= **Chi-square(**_U_**)** //**main effect of second SNP**
4.   //**Estimate Epistasis**
5.   _Epistasis_= _Omnibus_- _Main1_- _Main2_

Figure 8 Pseudo code of omnibus and epistasis test function

_____

Table 10 Contingency table with totals estimated

	aa bb	aa Bb	aa BB	Aa bb	Aa bB	Aa BB	AA bb	AA bB	AA BB	Total
Case	M_{000}	M_{001}	M_{002}	M_{010}	M_{011}	M_{012}	M_{020}	M_{021}	M_{022}	Σ_{cases}
Control	M_{100}	M_{101}	M_{102}	M_{110}	M_{111}	M_{112}	M_{120}	M_{121}	M_{122}	$\Sigma_{controls}$
Total	Σ_{aabb}	Σ_{aaBb}	Σ_{aaBB}	Σ_{Aabb}	Σ_{AabB}	Σ_{AaBB}	Σ_{AAbb}	Σ_{AAbB}	Σ_{AABB}	Σ_{all}

Table 11 Table of expected values (E) with formulas for estimating it's cell values from the observed table (O)(Table 10)

	aa bb	aa Bb	aa BB	Aa bb	Aa bB	Aa BB	AA bb	AA bB	AA BB
Case	$\dfrac{\Sigma aabb \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma aaBb \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma aaBB \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma Aabb \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma AabB \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma AaBB \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma AAbb \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma AAbB \times \Sigma cases}{\Sigma all}$	$\dfrac{\Sigma AABB \times \Sigma cases}{\Sigma all}$
Control	$\dfrac{\Sigma aabb \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma aaBb \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma aaBB \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma Aabb \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma AabB \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma AaBB \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma AAbb \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma AAbB \times \Sigma controls}{\Sigma all}$	$\dfrac{\Sigma AABB \times \Sigma controls}{\Sigma all}$

4.3.2 The logistic regression model used as a test statistic

On the corresponding null hypothesis the change in deviance has a chi-square distribution with DF = number of terms dropped, and can be used as a test statistic. Thus $H_0(1)$ is tested by the change in deviance between the models:

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \alpha_i + \varepsilon_i$$

and

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \varepsilon_i$$

Similarly, the omnibus test of $H_0(1)$, $H_0(2)$ and $H_0(3)$ is given by the change in deviance between:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

and

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \varepsilon_i$$

$H_0(3)$ is tested by the change in deviance between:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

and

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

A known bias of Pearson's chi-square test is overestimation of the statistical significance of cells in the contingency tables with small values. Yates' correction for

continuity [78] is used when calculating the chi-square statistic to prevent this. This is a very simple, yet very robust method that only involves a single subtraction and it has been widely accepted as a valid correction.

The proposed methodology that relies on the additive property of Pearson's chi-square will be compared to this method referred to in this paper from now on as the logistic regression method.

4.4 Computation of multiple response variables

Genetic studies, apart from the genetic data of each subject in the study also contain a large amount of phenotypical data [4,16,79]. Some common examples are gender, race, blood type, height, age at the time the data was collected and also some family history data, such as diseases of the subject's parents, etc. Not all of these may be important for analysis but many are. Take gender as an example. It's well established that some diseases even though they occur in both genders, they are more frequent in one. This may imply that there are different genetic effects associated with each gender. The most common methodologies of analysis of genetic data provide ways to either analyze sub-sets of subjects that are expected to have common effects independently or combined. In traditional methodologies, this would require running the analyses once per response variable. But it's quite common that many response variables are of interest [16]. This has a linear but significant increase in the computational cost of the analysis.

An innovative methodology involving the use of a 4-dimensional scoring matrix to test all of the top results or if needed even of the entire dataset across multiple phenotypes with little impact on the computational cost is proposed in this work (Figure 9). The four dimensional matrix is composed of the following dimensions:

1. SNP1 genotype

2. SNP2 genotype

3. Response Variable Categories

4. Response Variable

The example in Figure 9 represents a test between two specific SNPs, the contingency tables for the response variables shown underlined and bold are estimated at the same time with a single pass through the dataset. Note that the number of categories The pseudo code of the implementation of the 4-dimensional scoring matrix is shown in Figure 10.

_____
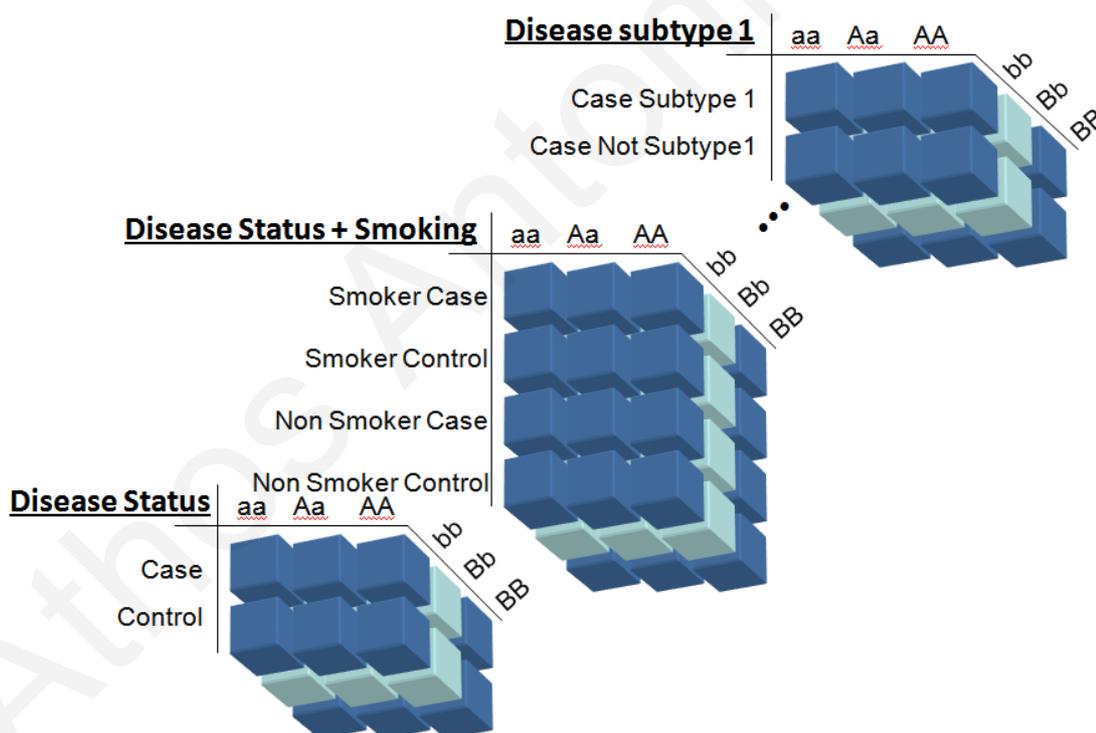
Figure 9 Contingency tables for two SNPs and multiple response variables(4d representation)

_____

_____

Algorithm: Score4dMatrix
Input:

- **SNP1 genotypes for all subjects**
- **SNP2 genotypes for all subjects**
- **Array of response variables for all subjects**

Output:

- **4dTable: a four dimensional table with all contingency tables in it scored**

Description: This function access all the subjects in the dataset once and scores based on the 2 inputted SNPs genotypes each of the contingency tables in the 4d matrix.

  FOR all Subjects
1. **//Determine genotypes of SNP1 and SNP2.**
       X=SNP1genotype(Subject)
       Y=SNP2genotype(Subject)
2. **//Loop through each response variable Identify the subject's category and score appropriate cell of contingency table**
       For all response variables K.
           Z=ValueofSubjectforReponseVariable(Subject, K);
           4dTable(K,Z,X,Y)++;

Figure 10 Pseudo code of the 4-dimensional algorithm for generating multiple

contingency tables on a single pass from the database.

_____

4.5 Evaluation of significance through replication

With the introduction of high throughput genotyping technologies and genome wide associations studies, a "blizzard of positive findings" claiming discoveries of associations between genetic variants and diseases was reported [76]. However, it soon became clear that the majority of these were probably false positives since they could not be replicated [27]. Ioannidis *et all* [29] pointed out that these findings from single association studies should constitute "tentative knowledge" and must thus be interpreted with exceptional caution.

In order to verify the reported results of this study and thus the validity of the analytical framework proposed, the top results of the analyses in the primary dataset will be tested for replication in an independent dataset. Since the number of tests performed in the second dataset will be very small, the multiple testing problem should not inhibit true positive findings from passing statistical significance.

Furthermore, replicated results will be examined more closely by looking at the distribution of each genotype combination in each dataset to each effect. Results where the strongest effect signals exist on the same genotype combinations in both studies analyzed will result in significantly more confidence in the replication success. To quantify this level of confidence a correlation coefficient will be estimated that will represent the level of correlation between each of the effects of the two SNPs. This approach will also yield information discovery that will relate each genotype combination to disease predisposition.

4.6 Hybrid Cluster Cloud High Performance Computing (HCC-HPC)

Two distributed HPC platforms were available for this work. Benchmarking each of the HPC platforms resulted in the identification of the bottlenecks associated with each HPC, as well as an estimation of the runtime for the given problem for each HPC system. The major differences of the two systems are the availability of computational processing power, ability to store large amounts of data, and the bandwidth available for network communication.

4.6.1 Dedicated LAN grid

The first HPC platform was composed of 200 dedicated backend machines in close geographical proximity linked together through a high speed Local Area Network (LAN), see Figure 11. All machines had access to the same shared storage facility through a 1,000 mbps LAN network, and had more than enough space available for the needs of this project. For the purposes of this proposal, these resources will be referred to as the dedicated LAN grid.

The dedicated LAN grid is composed of a blade server farm and a front end load balancing server. All machines on the dedicated LAN grid have access to the same data repository through high-speed connections. All other computers on the network can also access the data repository although not at the same high speed as the LAN grid. Data in the repository are stored, mirrored and backed-up so there's no need to move it after analysis is complete.

_____
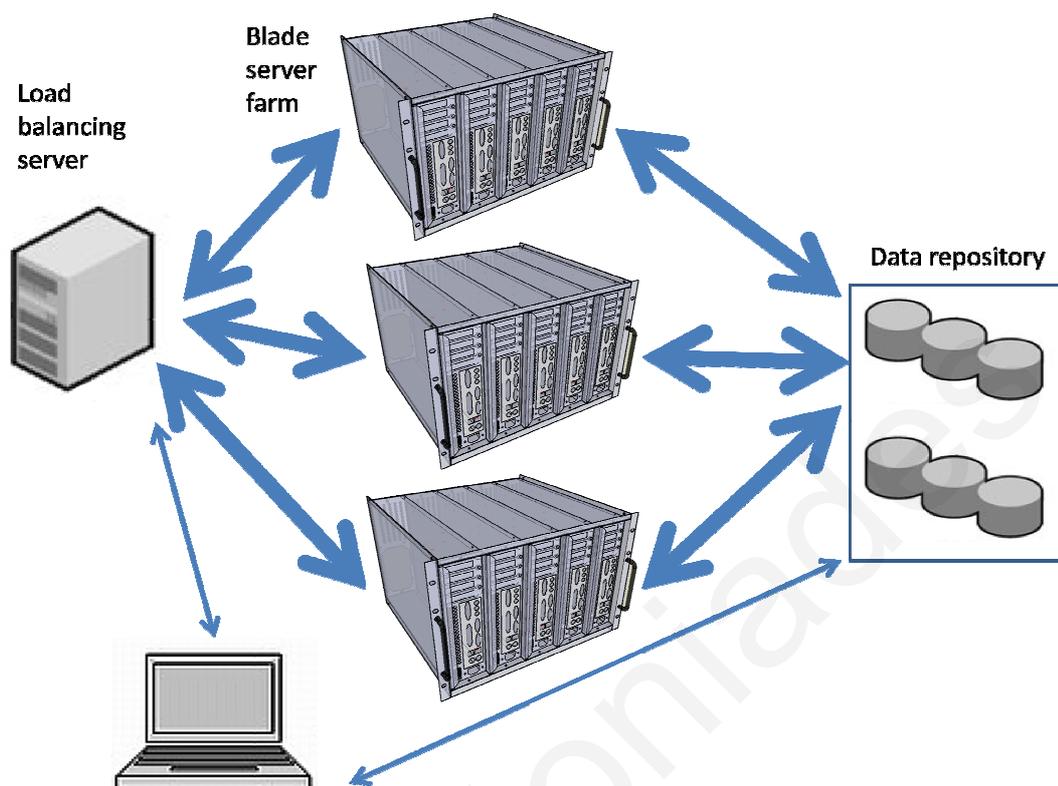
Figure 11 Architecture of the dedicated LAN grid

_____

4.6.2 WAN grid

The second computational platform is referred to as the residual Wide Area Network grid (WAN grid) and it relies on harnessing residual personal computer cycles to create a High Performance Computing (HPC) resource. This resources is implemented through the GridMP platform [85,94]. Figure 12 provides a simplified overview of the HPC's architecture. It is composed of a front end that monitors the available computational capacity on several thousand non-dedicated geographically dispersed nodes. These nodes are desktop computers concurrently being used by employees at the company Glaxo Smith Kline (GSK). The 1500 nodes with the most resources available were selected to execute, in the lowest possible priority, work submitted to

them by the front end of this grid. The nodes do not share a common data storage resource so all data are transferred to the nodes from the front end server and all results are transferred back to the front-end server. Since the network bandwidth utilized to transfer data between the nodes and the front-end is the same as the one used in the day-to-day operations of GSK it is necessary to reduce the load by as much as possible so as not to interfere with other potentially critical operations. Also, the amount of data storage available on the front-end is limited and shared between all applications that utilize the system. Since there is no centralized information of the number, schedule or the importance of other operations in GSK that utilize the same network strict rules were applied to limit the utilization of the WAN grid to applications that do not under any circumstances overload the system.

_____
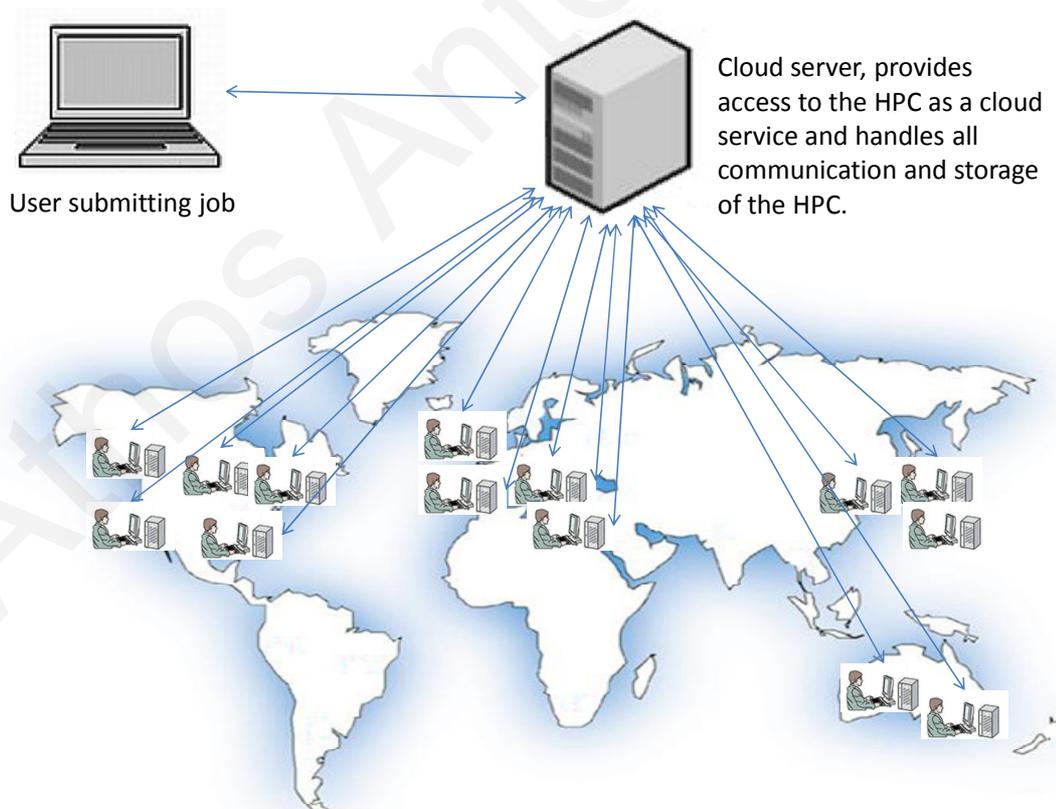
Figure 12 Architecture of the WAN grid
_____

The single WAN grid server monitors the available resources of thousands of machines around the globe that have an agent running on them responsible for reporting to the cloud server. When the WAN grid is requested to perform an analysis it selects in real time a subset of the computing nodes with the most available resources (processing cycles, main memory) and submits jobs to them. Communication is only allowed between each node and the server for security reasons.

4.6.3 Benchmarking of existing HPC resources

A random subset of the data was selected in order to run a test of the performance of the LAN and WAN grids for this specific application (1k, 5k, 10k, 20k SNPs for 1000 cases and 1000 controls). The algorithm used in this benchmark was the same as in the final system and since it's deterministic, the estimations on the execution time were consistent. However, since other users may at any point use either of the systems it was essential in order to provide a fair assessment to assume the optimal case where no other analyses are running on the systems during the tests. For the LAN grid, a high priority queue was used that does not allow any other analyses to take place until the benchmarking is complete. In the case of the WAN grid, it was impossible to monitor, limit or in any way control the user initiated applications. The cloud server did not submit any other work while the tests were running. Due to the high number of personal computers that were available in the WAN grid and the fact that the system selects only the 1500 (fixed number based on the licenses currently purchased) of them with the most available resources to use we expected that the average true capacity of the system would be relatively robust. In order to test this, the runtime of all analyses performed with the proposed framework where recorded, along with

detailed dates and times of the execution. Benchmarking using only the WAN grid wasn't allowed due to the restrictions in data transfer by the operator of the grid, therefore execution for that platform can't be presented. However, it is possible to estimate the execution times based on experiments using smaller samples.

4.6.4 HCC-HPC proposed framework

The residual WAN grid available for this project had significantly larger computational capacity available than the LAN grid but limited data transfer and storage capacity per. In order to utilize both HPC resources efficiently, a hybrid cloud HPC was designed and implemented as shown in Figure 13. This enabled the utilization of the residual WAN grid for the computational needs of the analysis while the LAN grid was used to post-process, annotate, merge and query the end results. The meta-scheduler breaks up the analysis that needs to be performed into work units. Each work unit is responsible for a small number of two SNP interactions. The data are split into files each containing the phenotype status and a subset of the SNPs for each of the subjects. The analysis is performed by composing each work unit to take as input two files, and perform all unordered two SNP combinations in them.

The maximum number of SNPs (m) in the files is estimated by

$$m = \sqrt{2w}$$

where w is the number of desired total work units to split the analyses. Having too many work units will increase the overhead of the WAN grid, whereas having too few will increase the runtime per work unit and therefore increase the frequency of data loss occurring from PCs on the WAN grid going offline. The range of acceptable

runtimes per work-unit for the system was obtained through trials on the actual system using worknodes of various sizes.

In order to identify how long a work node would run on average it was tested on a typical machine accessible by the WAN grid. Since the algorithm is deterministic, these estimations are quite accurate and deviation from them only has to do with different load availability on the computational resources.

_____
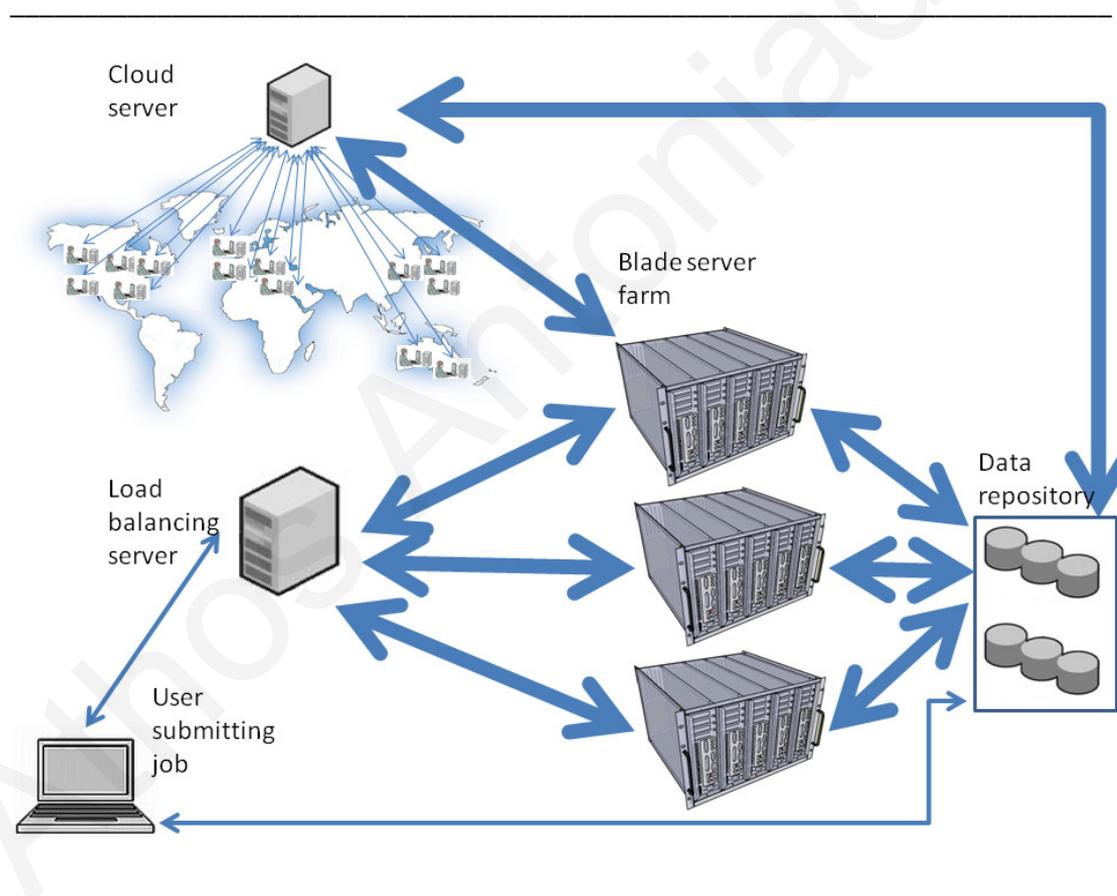
Figure 13 The structure of the purposed built HCC-HPC proposed system

_____

Thresholds for each measure and phenotype are used to identify interesting two SNP patterns. Each pattern that passes at least one threshold is reported once by the identifiers of the two SNPs. When a work unit has completed the analysis of a combination of two input files the results are compressed and transferred to a result repository. There, they are accessed directly by the LAN grid to generate the final result tables to be used for sorting and creating visualizations of the epistatic and omnibus effects based on a-priory defined thresholds.

_____

Algorithm: **High level view of Hybrid Cluster-Cloud algorithmic steps**
Input :
- **A GWAS study dataset with all the SNP data and unique subject identifiers**
- **The response variables data with subject identifiers to analyse in the entire genome.**
- **The response variables data with subject identifiers to analyse for annotation purposes on top results**
- **The thresholds to use for the first export of top results.**
- **A SNP annotation file**

Output :
- **A database with all 2 SNP interaction tests that passed the complete analyses threshold in at least one of the measures in any of the response variables analyzed in the entire genome.**
- **A single result file annotated with SNP information for the results the top results.**

Method

_____
                                           RUNING ON ANALYST'S PERSONAL COMPUTER

1. **//Define data pre-processing parameters**
      Define quality control parameters
      Binary Encoding if the data isn't already encoded
      Define number of segments to split the data into (w)*
      If data is not in permanent repository, encode and transfer there

_____
                                           RUNNING ON DEDICATED LAN GRID

2. **//Perform pre-processing and initiate analyses on HCC-HPC**

    Split data into $\left\lfloor \sqrt{2w} \right\rfloor$ segments*

    Initiate Analyses on HCC-HPC
    Wait to receive results from WAN grid

_____
                                           RUNING ON WAN GRID SERVER

3. **//Submit worknodes to computational nodes**
  For(all submitted worknodes)
      Pseudo code of selection algorithm for PC to use for running a worknode (Figure 15)

_____
                                           RUNING ON WAN GRID NODES

3 **Pseudo code of algorithm followed by the WAN grid agent on each node (Figure 16)**

_____
                                           RUNING ON WAN GRID SERVER

4 **Pseudo code of WAN grid handling of worknode state reporting (Figure 17)**

_____
                                         RUNNING ON DEDICATED LAN GRID

5 **Pseudo code of the LAN grid algorithmic steps (Figure 18)**

_____
                                       RUN ON ANALYST'S PERSONAL COMPUTER

6 **//Complete analyses transferring results and initiating visualization environment**
    Transfer data to analyst's selected machine
    Launch results query and visualization platform.

*w : number of work nodes to be created, system and dataset dependent
The orange box indicates the part of the algorithmic steps that take place in parallel.

Figure 14 A high level view of HCC-HPC algorithmic steps

_____

In Figure 14 the pseudo code of the algorithmic steps that take place on the HCC-HPC is presented detailing the steps in an abstract manner. Different parts of the algorithm are run on different machines, as indicated by the break lines. The parts of the algorithm that are run in parallel on all machines are highlighted in the orange box. The analysis begins at any personal computer that has access to the system. The data is pre-processed then transferred to the data repository where it will be permanently stored.

Once the above step is completed; all necessary data and parameters are stored in the permanent storage repository accessible by all machines except the nodes of the WAN grid. The web service controlling the Hybrid Cloud-Cluster HPC initiates the analysis on the dedicated LAN grid. At step 2the dedicated LAN grid will submit the data to the WAN grid. The WAN grid's architecture required that all data to be used in the analyses resides in a temporary area on the grid's main server thus this step is necessary. In parallel to this step as soon as all the data required for a worknode to be instantiated it's entered into the queue to be executed. This is important, since even though the data transfer between the data repository and the WAN grid's server is low, so long as it's on average faster to transfer the data than to analyze them on the WAN grid the total runtime isn't expected to be affected since the two steps are happening in parallel. The analyst from this point on has access to a web service that provides all information on the progress of the analyses as well as a detailed record of all messages (including errors) that are received.

The cloud server identifies the 1500 machines on the network with the most available resources. The WAN grid performs this step by continuously requesting updates from all machines on the network and keeping a list of the top 1500, the list can be considered near real time since the machines are only reporting to the server about once every minute. The algorithmic steps used to identify the machine to use for analyzing a specific worknode are shown in Figure 15. Each machine is looked up to see if it has been used in previous analytical iterations with the same data that contained either of the two parts needed for each work node. The machine in the list of the ones with the most available resources that already has the most data required to run the analyses will be selected and the worknode will be

submitted to it. Since data parts are cached (as per the functionality provided by the GridMP platform), there is an improvement in efficiency in terms of reducing data transfer as well as providing a redundant backup storage location. Once a work node is identified, and all relevant data is available on it the analyses is launched on that node in the lowest possible priority.

Algorithm: Selection of PC to use for running a worknode.
Input :
  Identifier of files needed for the analysis.
  Identifier of worknode
Output :
  Identifier of PC to be used to analyze the worknode

Identify *X* machines with most resources available

1 **If (among the *X* machines one has all data parts)**
   Submit worknode to machine with all data parts
2 **Else if (among the X machines one has 1 data parts)**
   Transfer remaining data part
   Submit worknode to the selected machine
3 **Else**   **//**
   Transfer both data parts to the machine with highest resources available
   Submit worknode to machine
4 **Else**
   Submit to worknode with Max Resources both data parts
   Launch worknode analyses on node.

Figure 15 Pseudo code of selection algorithm for PC to use for running a worknode

The WAN grid node that received all the relevant data to analyze a worknode will start the analysis as soon as all transfers are complete. The steps followed by the WAN grid's agent on the node are outlined in Figure 16. The agent will perform a parity check to verify that the data transferred without errors. Then it will decompress all necessary files, and instantiated the analyses wrapped within this agent. This is done for security issues and it's what causes the worknodes inability to communicate with any machine other than the WAN grid server. Once the analyses is finished, either due to an error or if it's successful, all outputted files, streams and exit code are packaged and send to the grid server.

Algorithm: Algorithm followed by the WAN grid agent on each node.
Input :
- **Location of local repository of all files needed for the analyses**
Identifier of worknode
Output :
- **All output of the program, including output streams, exit codes and files.**

1   A machine receives a worknode to analyze
Test if data received is not corrupt
Decompress data
Run analyses storing stdout and stderr to separate files
If process crashes, report error and all files generated.
If process returns, report return code to server along with all the result files produced, stdout and stderr files.

Figure 16 Pseudo code of algorithm followed by the WAN grid agent on each node

Possible outcomes fall in three categories successful, unsuccessful, error. The algorithmic steps followed on the WAN grid to deal with each category are shown in Figure 17. Successful worknodes are ones that returned a success completion message to the server and transferred all their results to the cloud server with no errors detected. Unsuccessful are work nodes that either returned an error code or resulted in an error for more than five times on different machines. In the case of returned error codes, each code represents a specific error testing criteria, such as errors in the input data; errors found in the analyses step and in general if conditions that were predicted to happen and were error tested will return a message to help identify the issue. All relevant debugging information is made available through the cloud server to the end user. Error worknodes are ones that stopped communicating with the server. This could be caused by the work node application crashing or getting killed on the node, or the machine might have gone off line. Errors always occur due to the nature of the cloud so each work node is run up to five times on different machines before it's reported as an unsuccessful indicating that perhaps there's something wrong with the program causing the crashes. As an example, if a worknode takes too long to run and affects the primary user the probability that the primary users on five different machines will either identify it and kill it or reboot their machines is high.

Algorithm: WAN grid handling of worknode state reporting.
Input :
- **Identifier of instantiation of a worknode on a specific PC.**

Output :
- **Update of all relevant databases including, overall worknode status reports, and output data.**

<u>**WORKNODE STATE REPORTING**</u>

1. **The grid server probes each machine.**
   <u>**Grid server probes nodes.**</u>
   a. **If a machine reports worknode completed**
      i. **Transfer data to cloud server database**
      ii. **Transfer data to cluster server database**
   b. **If machine reports worknode error**
      i. **Transfer all debugging data to cloud server**
      ii. **Send signal to halt operations on the cluster**
      iii. **Notify user of reported error.**
   c. **If machine stops responding.**
      i. **If worknode last running and incomplete on machine has not crashed more than 5 times.**
         1. **Remove machine from list of machines available.**
         2. **Mark worknode as not completed and resubmit it.**
         3. **Increment counter of worknode incomplete.**
      ii. **Else 5 machines running this worknode became unresponsive.**
         1. **Attempt to transfer any debugging data available to cloud server from all machines.**
         2. **Send signal to halt operations on the cluster.**
         3. **Notify user of reported error.**

Figure 17 Pseudo code of WAN grid handling of worknode state reporting

Whenever a worknode is reported to the server as successfully completed the results from it are transferred to the permanent storage area on the cluster grid, and the analytical steps on lines 3.a.ix and 3.a.x are submitted to run on the cluster. The process on the WAN grid finishes when all worknodes get a success or an unsuccessful status. The processes on the cluster grid are run in parallel to the grid (Figure 18), beginning as soon as successful results become available, but depending on the load of the cluster by other applications completion of the steps on this machine might be delayed. Therefore, there might be a short waiting period until all worknodes have successfully completed on the cluster as well.

Execution failures, or data loss are dealt by simply identifying them and reporting them to the user. There is an elemental error checking as well as testing return values in a similar way as the WAN grid; however there is no need to rerun the worknodes here since errors are very rare and usually have to do with something going wrong on the system level, or the worknode submitted.

Algorithm: LAN grid algorithmic steps.
Input :
- **Worknode identifier**
- **Link to result files on permanent data storage**

Output :
- **Annotated dataset of top results**
- **Report on overall execution of system**

   a. **For each new result to complete transfer to the permanent data repository**
        iii. **Decompress results**
        iv. **Create first annotated results based on predefined parameters and store in the appropriate annotated result database.**
   b. **Wait for all worknodes to complete on both grids.**
   c. **If unsuccessful worknodes recorded**
        v. **Report to user,**
        vi. **Break analyses.**
   d. **Else**
        vii. **Merge default datasets parts and return them in a user accessible location.**

Figure 18 Pseudo code of the LAN grid algorithmic steps

Once the entire process is completed the first results database is available, along with routines to create more result datasets specific to queries performed by the analyst (specific genetic regions, top results with specified thresholds, etc)

4.7 The proposed framework put together

In this section the entire framework is looked at as well as the steps to take in order to evaluate the results it produces and by association the accuracy of the proposed framework.

4.7.1 The analyses on the HCC-HPC

The proposed framework is provided as a cloud service to all analysts with appropriate access. In reality the framework is controlled via a web service. The researcher initiating the analysis submits the data to use, if these are not already in the permanent storage (if the dataset is new) then it's automatically transferred there, otherwise the existing copy is used. Quality control parameters are defined by the researcher and the data is pre-processed. Once this is completed the algorithm of the

hybrid cloud-cluster high performance computing system is initiated as described in figure 14. During this time, both HPC systems are utilized by the framework, the WAN grid running the computational part of each worknode, while the dedicated LAN grid as soon as results are delivered from the WAN grid, it decompresses them, re-analyzes just the potentially statistically significant ones both with response variables used in the primary analyses as well as ones defined as needed to be annotated on reported results only and generates the first database of the top results fully annotated.

4.7.2 Post processing results

The next step is query, visualize, and study the results. The first step is to adjust the top results for multiple testing using Bonferoni correction[5,46] and identify any that may pass significance after adjustment. Bonferoni correction as it has been stated previously is very stringent, meaning that it tends to over-adjust making finding statistically significant results very difficult. Also, the statistical power in these studies is relatively low due to the relative small number of subjects compared to the number of SNPs collected. Therefore it's quite likely that even results that don't pass statistical significance after Bonferoni adjustment for multiple testing may still be true positive effects.

4.7.3 Replication test

The first step when evaluating a newly proposed analytical framework is to quantifiably demonstrate how likely it is that the results it produces are true positive. In genetics, the classic approach for doing this as discussed in previous sections is through replication testing with an independent dataset [27,29]. A threshold is set to identify the top results with an independent replication dataset. This will both help in

discovering more evidence as to whether the top results are true positive or not but in extension it will also help validate the proposed framework in its entirety. Discovery of strong replicated results will indicate with a high degree of certainty that the proposed methodology is capable at detecting true positive effects [4,27,29].

For the replication test, since the goal of this dissertation is to provide evidence to the validity of the framework, only the top results in the dataset used for the primary analyses that passed a preset threshold are tested for replication in the independent replication dataset, rather than the entire replication dataset.

Chapter 5 Results

This chapter begins with a description of the datasets used to evaluate the proposed framework and continues with results generated from the analysis of these datasets. Results can be broken into two large categories. The first category refers to the Biological results, that relate to the findings of the system in relation to the actual analysis performed. The second category refers to the results related to the performance of the proposed analytical framework. The biological results are presented first, as they lay the foundation of the evidence that the method can produce innovative, statistically significant and replicable results. The second part will focus on each independent contribution of this dissertation in detail. Experiments were carried out to evaluate each of the contributions and the results are presented following the biological results in the same order as in the previous chapters.

5.1 Datasets used for the evaluation of the proposed framework

Two independent datasets were needed that would have the same subject inclusion criteria, and overlapping genetic markers genotyped. Two such GWAS were identified both designed to study genetic predisposition to Multiple Sclerosis (MS). These are presented in the next two subsections along with references to the first publications that provide their complete descriptions.

5.1.1 Primary dataset (GeneMSA)

The first dataset was first presented in [80]. The dataset after reported stringent quality control data filtering included 551,642 SNPs based on the Sentrix HumanHap550 BeadChip platform by Illumina (Illumina 550) in 978 MS cases and 883 matching controls. Susceptibility, age of onset, disease severity, as well as brain lesion load and normalized brain volume from magnetic resonance imaging exams were assessed for association with single genetic loci and are presented in [80]. An outline of the distribution of subjects and their relevant phenotypes as shown in [80] is presented in Table 13.

The analysis of this dataset revealed 242 statistically significant associations involving single SNPs including 65 within the MHC locus [80]. These results were tested for replication with another dataset, the IMSGC dataset and confirmed a role for GPC5 gene in disease risk.

5.1.2 Replication dataset (ANZgene)

ANZgene is a three year study that utilized the MS Research Australia (MSRA) Gene Bank and involved scanning the DNA of 1,618 people with MS and 3,413 people without MS (controls). The genotyping platform used in this GWAS was also made by Illumina, but it was the 300k model. ANZgene is a more recent GWAS with the first publication expected to be available in print in June 2011 [81]. This dataset will be used only to test for replication of the top GeneMSA case-control 2 SNP interaction results. The distribution of the phenotypes of interest are shown in Table 14, these are the only phenotypes that access was provided to for the ANZgene GWAS.

5.1.3 Addressing incomplete genetic marker overlap between primary and replication dataset

Since ANZgene was genotyped using a different genotyping platform than GeneMSA there is not complete overlap between the available SNPs. However the company that developed both platforms is the same (Illumina) and it seems that they chose to extend the platform used in the ANZgene analyses to create the one used in the GeneMSA analyses. Therefore the roughly 300k SNPs in ANZgene seem to exist in the 550k SNPs in GeneMSA but obviously there are about 250k SNPs in GeneMSA that are not in ANZgene. To test for replication of a two SNP interaction test, it is necessary that both SNPs be present in both datasets. This is expected to reduce the number of tests possible to test for replication to about 1/3.

There is an alternative method, sometimes used in genetics when there is a need to test for replication between two datasets that were genotyped on different platforms. This is called genotype imputation [63,82,83]. However to correctly analyze imputed data the probability that the imputed genotype is accurate needs to be taken into account [84], something that neither logistic regression nor the proposed methodology for the epistasis measure can do. Furthermore, imputation relies on LD between genotyped and imputed markers in order to predict each missing genotype with a high level of certainty. LD and its effect on the proposed results is one of the key elements of study in this dissertation. Relying on a method that is biased based on the level of LD between markers may impute markers and might have an undetectable effect on our final conclusions. Therefore for the purposes of evaluating the framework imputed genotypes will not be utilized.

5.2 Biological Results

The GenMSA analyses resulted in statistically significant results, even after Bonferoni correction for multiple testing was applied. The cut-off p-value of 10e-8 was used to test for replications in ANZgene. This cut-off was decided a posterior, since it would be impossible to correctly define how strong the expected top results would be due to the lack of knowledge of the level of dependence between markers and thus the inability to correctly adjust for multiple testing. Therefore the cut-off was put at 10e-8 so that it would include both results that were statistically significant in ANZgene as well as those that were high but perhaps lacked the statistical power in GeneMSA to reach statistical significance. Since ANZgene and GenMSA were genotyped using two different platforms (GeneMSA Illumina 550k, and ANZgene Illumina 300k) in order to test for replication of a result it was necessary that both SNPs involved in it would be available in the replication dataset.

In Figure 19, the x-axis represents the ordered location of the two SNPs involved in the hypothesis tested; the y-axis represents the epistasis p-value on a negative logarithmic scale. Two vertical lines represent the Bonferoni multiple testing correction for p-values 0.01 and 0.05. The results where both SNPs are available in both datasets are tested for replication and are indicated in Figure 19 as green, if they are found to be statistically significant in ANZgene or red if they are not statistically significant. Results where one or both of the SNPs were not available for analyses in ANZgene are indicated as yellow and replication testing was not performed on them. All of the results tested for replication that had both SNPs in or near to the Human Leukocyte Antigen system (HLA) region of chromosome 6 replicated successfully, while all results outside this region failed to replicate in the replication dataset.

The omnibus measure is composed of both the two main effects and the omnibus effect. In Figure 20 the omnibus effect (x-axis) is plotted against the Epistasis effect (y-axis) both in negative logarithmic order of the resulting p-value. The two vertical and horizontal lines represent the p-values of 0.01 and 0.05 following correction for multiple testing using Bonferoni correction for each respective measure. By plotting the omnibus measure vs the epistasis it is possible to see the relation between the epistasis and the two combined main effects since the omnibus measure is composed of both the main and epistatic effects.

Figure 21 has the same properties as Figure 20 but instead of GeneMSA the ANZgene replication results are plotted. Here all the tested replications are plotted. The failed replications clearly cluster near the origin with very small, if any, effects compared to the replication successes that are clearly statistically significant in both omnibus and epistasis measures.

One of the sub-phenotypes analyzed for the top epistasis results is the PRP (a rare subtype of MS). Patients with the PRP subtype were tested versus patients with Multiple Sclerosis but not PRP. The number of subjects that passed QC and belonged to the PRP group was very small, so there was very small statistical power to detect effects. In order to limit the search space in an informative way that would improve the statistical power of detection after adjusting for multiple testing, only tests that had a multiple testing unadjusted p-value of 0.01 in the MS case-control analyses were included in this analysis.

Females are associated with a higher predisposition to MS than males. It is not clear if this is due to genetic or environmental factors. In this project we spited the dataset in males and females and repeated the case-control analyses with subjects from each gender. Since the number of males was only about 1/3 of the total (numbers of males

and females) statistical power of detection was significantly higher in females than males.

The parameters of Figure 22 and Figure 23 are the same as in Figure 20. Figure 22 represents the analyses performed with only male subjects, while Figure 23 represents the analyses with only female subjects. Notice that none of the replicated results are strong when testing only males, while all of them seem to be driven by females.

All of the results that replicated involved SNPs that were on chromosome 6. A closer look revealed that all of these replicated tests were in or close to the HLA region. Figure 24 provides a visualization of the region and the detected epistasis effects. The horizontal axis represents the genetic location with base pair position on chr 6 indicated on the lower horizontal line. Above that line is a gene map of selected genes in the region. The horizontal line in the middle of the graph has a short vertical line indicating the location of every SNP that was genotyped in GeneMSA. All detected epistatic effects in this region that were tested for replication were included in this graph and were represented by diagonal lines forming an isosceles triangle connecting the two SNPs in each test. The height of the triangle represents the distance between the SNPs. The colour represents replicability with red being replication successful while blue being replication not possible due to not having both SNPs in ANZgene. None of the results in this region that came up in ANZgene and were tested for replication failed; therefore there is no colour to indicate replication failure.
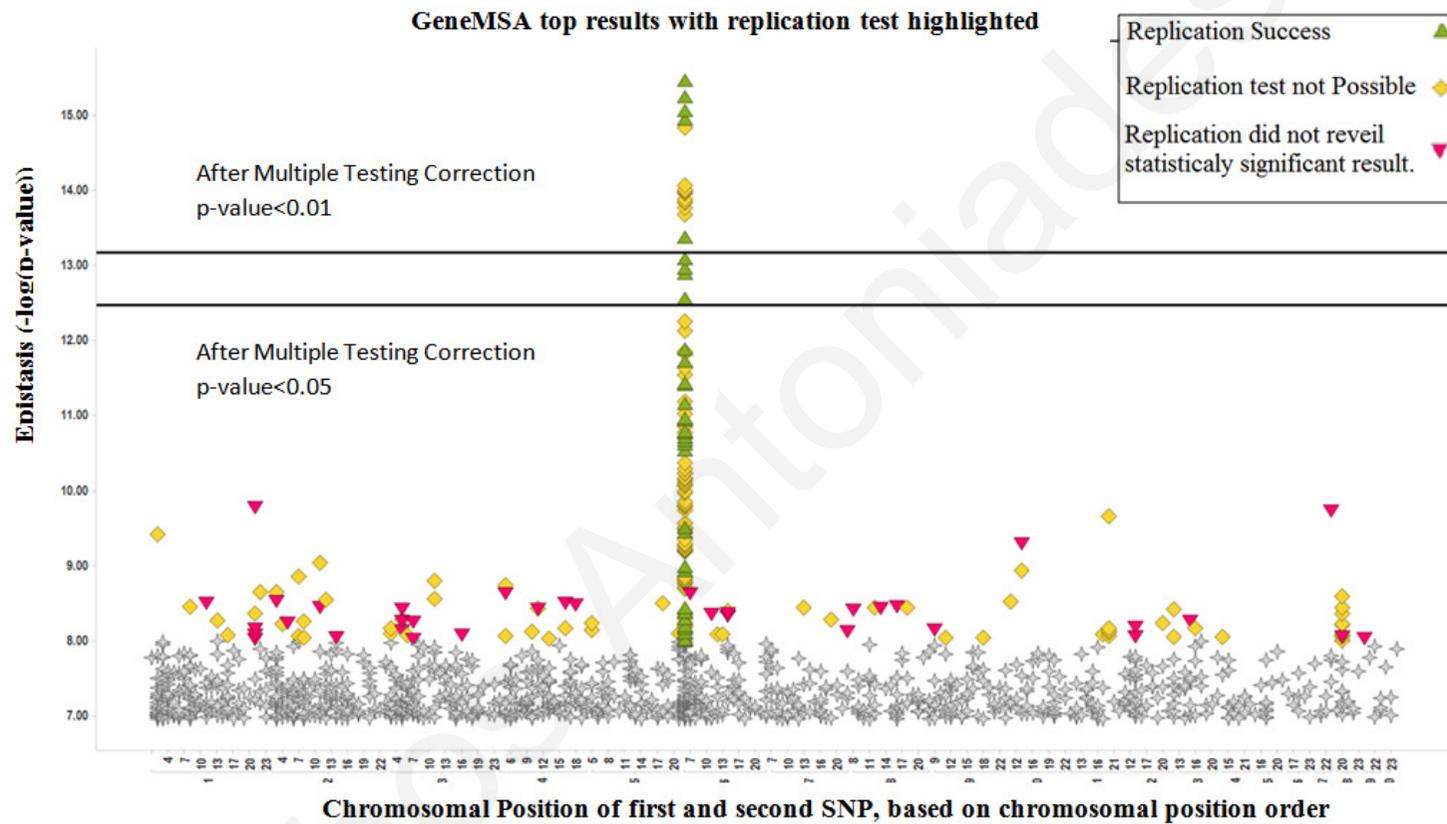
Figure 19 Top Epistasis results of complete GenMSA case-control analyses. A strong peak is visible were both markers are on chromosome 6 in the HLA region some passing statistical significance after multiple testing correction

GeneMSA top results replicated highlighted
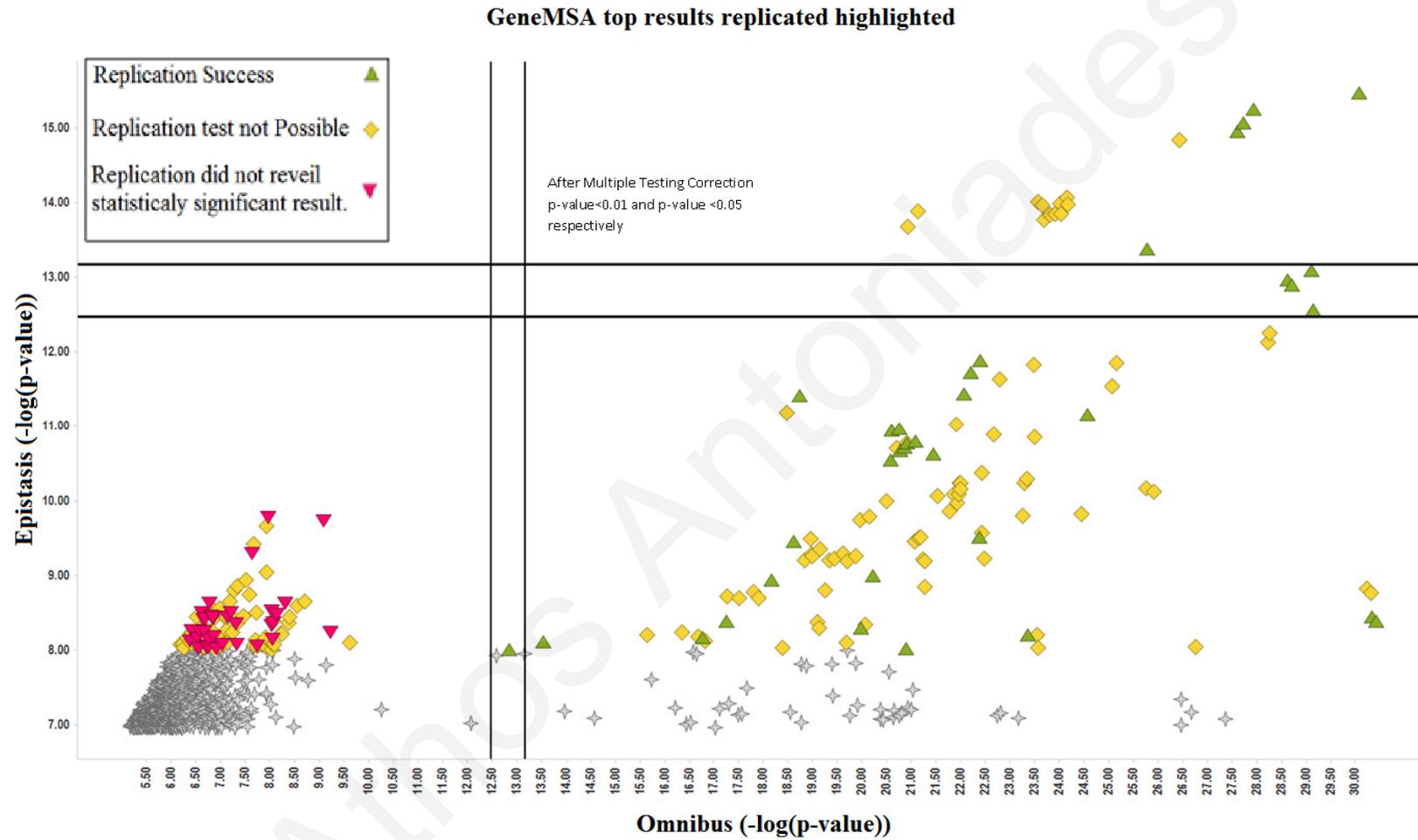
Figure 20 Epistasis vs Omnibus measures of the top Epistasis results in the complete GenMSA case-control analyses
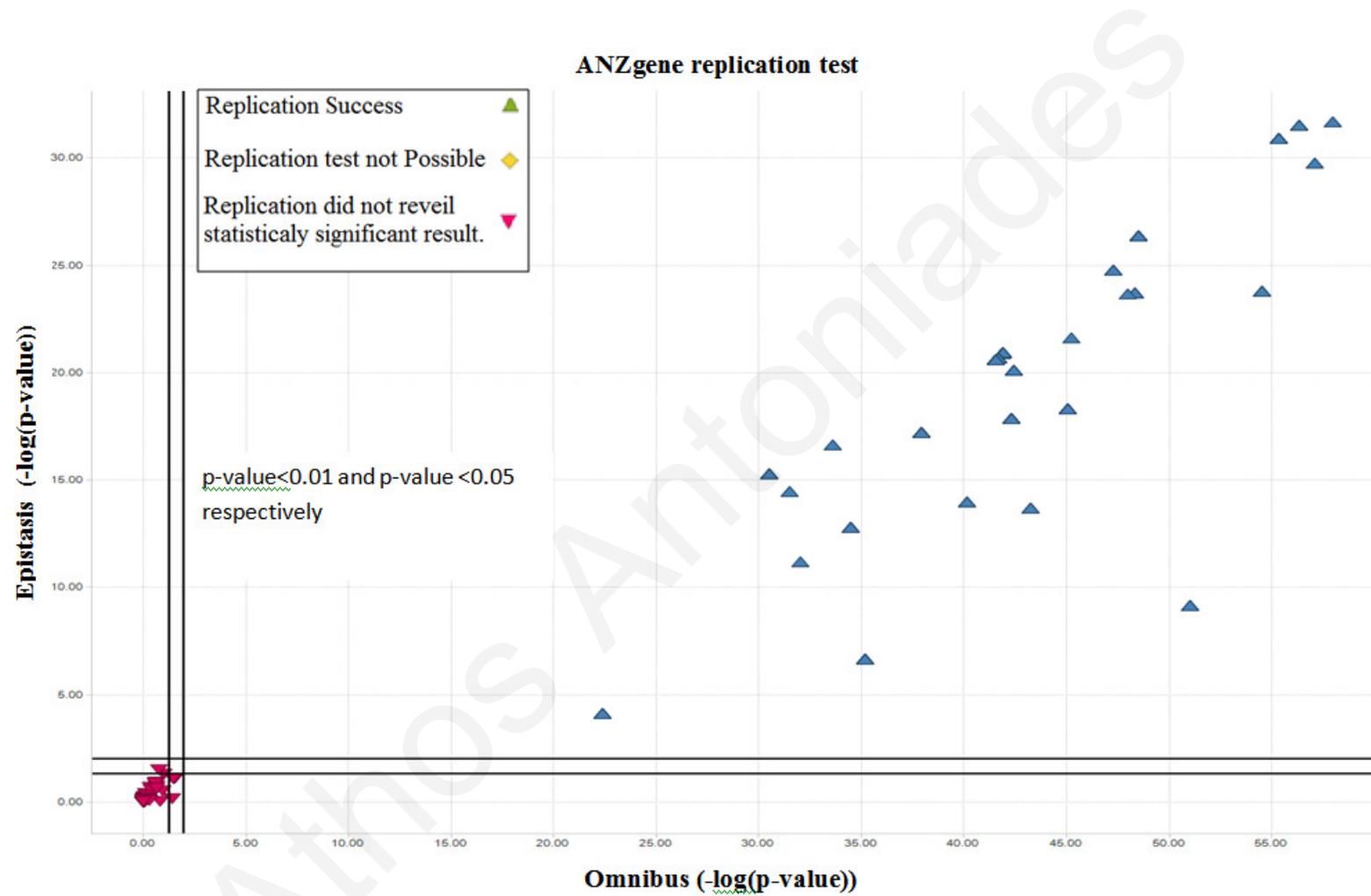
Figure 21 Epistasis vs Omnibus measures in ANZgene on the results tested for replication.
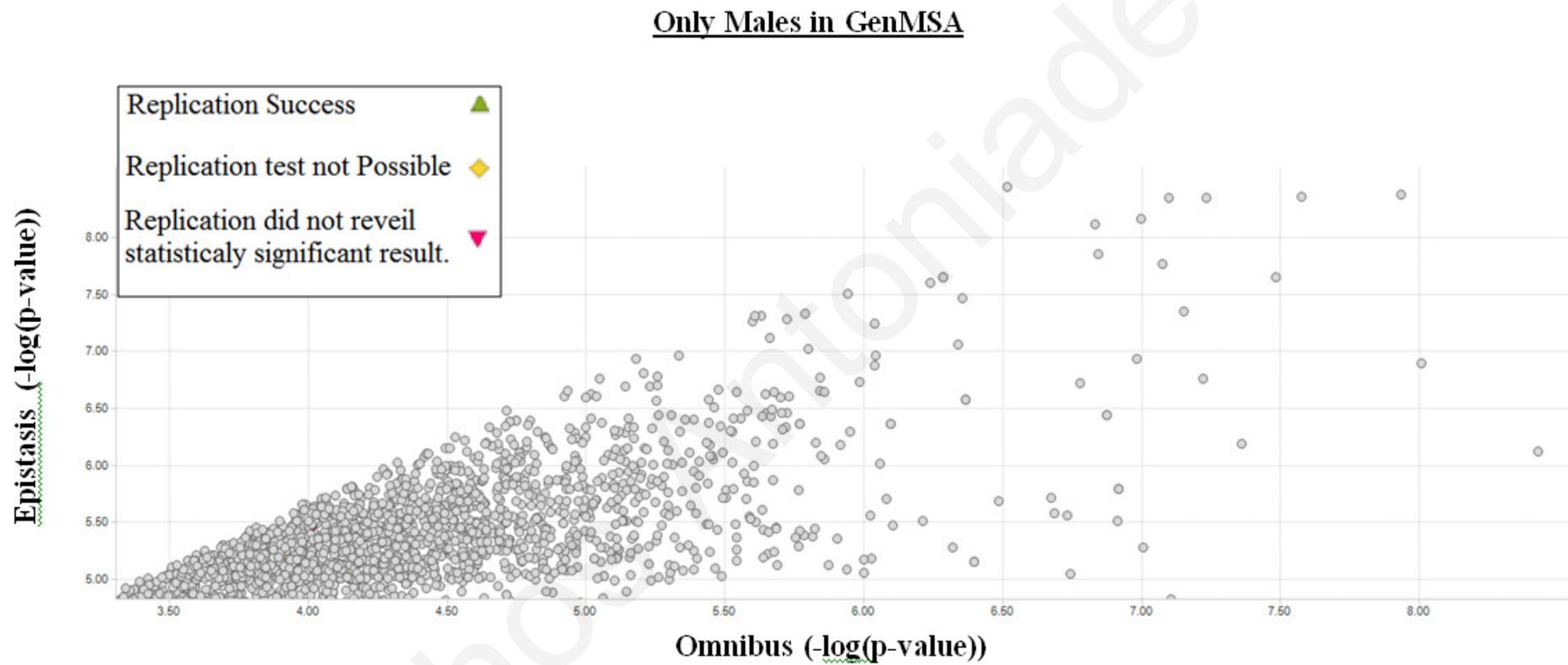
Figure 22 Epistasis vs Omnibus measures of the top Epistasis results in the analysis of only males in GenMSA
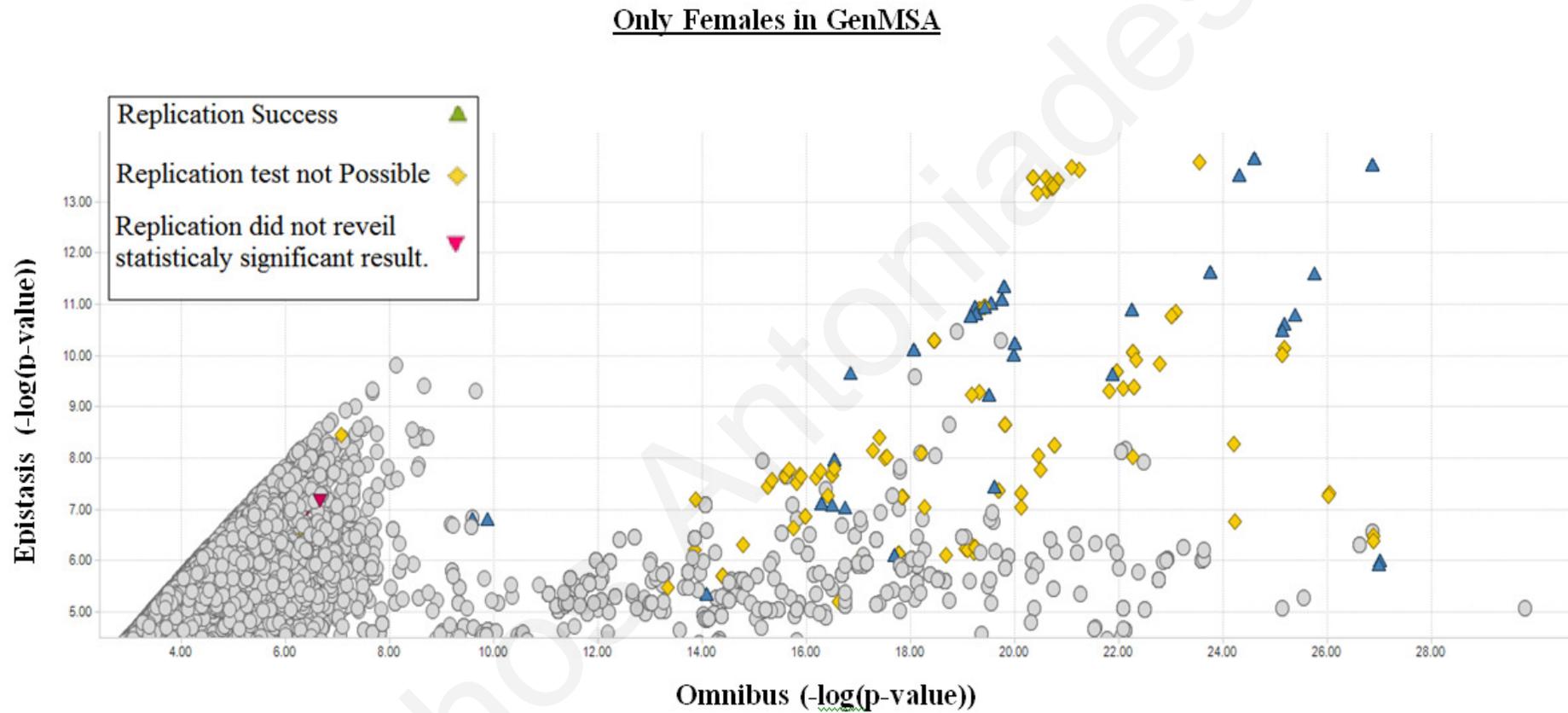
Figure 23 Epistasis vs Omnibus measures of the top Epistasis results in the analysis of only females in GenMSA

No replication possible
Replication Success
Replication Failure

Chromosome : base pair position

06:32,300,000   06:32,400,000   06:32,500,000   06:32,600,000   06:32,700,000   06:32,850,000
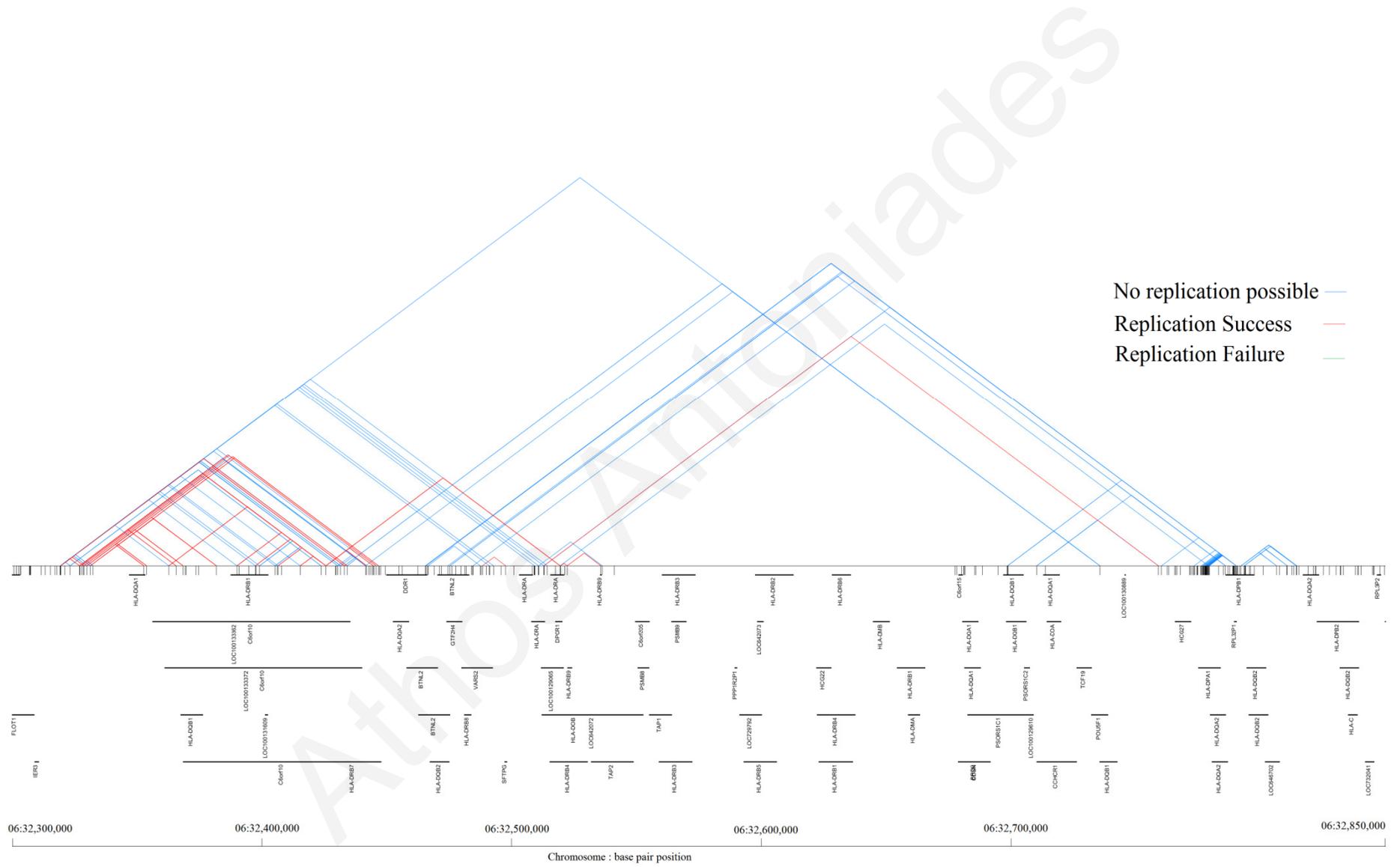
Table 13 Distribution of Subjects in GeneMSA

Variable	Total Patients	Controls	San Francisco Patients	Controls	Amsterdam Patients	Controls	Basel Patients	Controls
Female, n (%)	653 (67)	584 (66)	334 (69)	287 (66)	137 (60)	148 (62)	182 (69)	149 (71)
Male, n (%)	325 (33)	299 (34)	153 (31)	147 (34)	92 (40)	90 (38)	80 (31)	62 (29)
Female:Male ratio	2.0:1	2.0:1	2.2:1	2.0:1	1.5:1	1.6:1	2.3:1	2.4:1
Age at onset of disease, mean (range)	34 (11−61)		34 (11−61)		35 (16−60)		32 (11−59)	
Age at time of analysis, mean (range)	47 (21−69)	46 (21−74)	46 (21−69)	46 (23−69)	48 (25−67)	45 (24−68)	47 (21−69)	47 (21−74)
DRB1*1501+ individuals, n (%)	474 (48)	190 (22)	224 (46)	86 (20)	118 (52)	59 (25)	132 (50)	45 (21)
Disease duration (years) mean	11.5		10.7		11.4		12.9	
Median (range)	9.0 (0−47)		8.0 (0−46)		10.0 (1−34)		11.0 (0−47)	
Disease type, n (%)								
Relapsing remitting	659 (67.9)		343 (70.4)		141 (61.8)		175 (68.4)	
Secondary progressive	137 (14.1)		46 (9.5)		46 (20.2)		45 (17.5)	
Primary progressive	72 (7.4)		18 (3.7)		29 (12.7)		25 (9.8)	
Clinical isolated syndrome	100 (10.3)		79 (16.2)		10 (4.4)		11 (4.3)	
Unknown	3 (0.3)		1 (0.2)		2 (0.9)		0 (0)	
DMT, n (%)[a]	432 (54)		245 (63)		60 (32)		127 (58)	
EDSS, n (%)[b]								
<3	440 (55.4)		279 (71.9)		58 (31.0)		103 (47.0)	
3 to <6	271 (34.1)		86 (22.2)		95 (50.8)		90 (41.1)	
6−6.5	68 (8.6)		22 (5.7)		24 (12.8)		22 (10.1)	
≥7	15 (1.9)		1 (0.3)		10 (5.4)		4 (1.8)	
Mean (SD)	2.74 (1.81)		2.06 (1.64)		3.74 (1.72)		3.07 (1.67)	
MSSS, n[c]	794		388		187		219	
Median	2.97		2.23		4.82		3.17	
Mean (SD)	3.45 (2.37)		2.70 (2.21)		4.81 (2.23)		3.61 (2.21)	
T2 lesion load/mm^3, n[d]	791		387		185		219	
Median	2580		2164		3001		3276	
Range	0−81317.2		0−64258.3		0−81317.2		0−34984.5	
nBPV/cm^3, n[e]	753		371		176		206	
Mean (SD)	1530.8 (86.11)		1511.6 (81.91)		1554.5 (83.17)		1544.9 (88.55)	
Range	1225−1767		1225−1704		1360−1750		1309−1767	

[a]Disease modifying treatment (DMT). Test of association with site: $\chi^2 = 50.64$; $P < 0.0001$.
[b]Expanded disability status scale (EDSS). Test of association with site: $\chi^2 = 102.17$; $P < 0.0001$.
[c]Multiple sclerosis severity scale (MSSS). Test of variation among sites (Kruskal−Wallis test): $\chi^2 = 107.82$; $P < 0.0001$.
[d]T2 lesion load/mm^3. Test of variation among sites (Kruskal−Wallis test): $\chi^2 = 18.84$; $P < 0.0001$.
[e]Normalized brain parenchymal volume (nBPV). ANOVA of variation among sites: $F = 19.77$; $P < 0.0001$.
a−e analyses include only subjects with relapsing remitting and secondary progressive disease.

Table 14 Distribution of Subjects in ANZgene

	Entire dataset	Males	Females	*Primary Progressive MS	*Not Primary Progressive MS
Cases	1618	445	1173	407	1211
Controls	1988	757	1231	NA	NA
Total	3606	1202	2404	NA	NA

*Primary progressive is an MS disease subtype.

NA: Not applicable.

Table 15 Results of MS cases vs controls in primary and replication datasets.

SNP identifiers		Annotations SNP1		Annotations SNP2		Primary Dataset Case Control		Replication dataset Case Control	
SNP1	SNP2	Gene SNP1	Chrmosomal Position SNP1	Gene SNP2	Chrmosomal Position SNP2	Omnibus	Epistasis	Omnibus	Epistasis
RS3129890	RS9268832	HLA-DRA	06:032522251		06:032535767	29,1	13,1	58	31,69
RS3130320	RS6457536		06:032331236	C6ORF10	06:032381743	27,6	15,0	56,3	31,55
RS3130320	RS6935269		06:032331236	C6ORF10	06:032368328	27,7	15,1	55,3	30,94
RS3115553	RS3130320		06:032353805		06:032331236	27,9	15,3	55,3	30,94
RS3130320	RS3130340		06:032331236		06:032352605	27,9	15,3	55,3	30,94
RS3129890	RS7192	HLA-DRA	06:032522251	HLA-DRA	06:032519624	29,1	12,6	57,1	29,76
RS3096700	RS6935269		06:032329760	C6ORF10	06:032368328	22,2	11,7	48,5	26,38
RS3096700	RS3115553		06:032329760		06:032353805	22,4	11,9	48,5	26,38
RS3096700	RS3130340		06:032329760		06:032352605	22,4	11,9	48,5	26,38
RS3115553	RS3096700		06:032353805		06:032329760	22,4	11,9	48,5	26,38
RS3130340	RS3096700		06:032352605		06:032329760	22,4	11,9	48,5	26,38
RS3096700	RS6457536		06:032329760	C6ORF10	06:032381743	22,1	11,4	47,3	24,8
RS4959089	RS926070		06:032327703		06:032365544	28,6	13,0	54,5	23,8
RS926070	RS9296015		06:032365544		06:032326967	28,7	12,9	54,5	23,8
RS3129943	RS3130320	C6ORF10	06:032446673		06:032331236	24,6	11,2	48,3	23,71
RS2395174	RS9275141		06:032512856	LOC650557	06:032759095	18,6	9,5	48	23,69
RS3135363	RS6932542	LOC642071,	06:032497626		06:032488240	20,9	8,0	45,3	21,62
RS2076537	RS547261	C6ORF10	06:032425613	C6ORF10	06:032390011	20,7	11,0	41,9	20,93
RS2076537	RS547077	C6ORF10	06:032425613	C6ORF10	06:032397296	20,6	10,9	41,8	20,9
RS2076537	RS9268132	C6ORF10	06:032425613		06:032362632	21,1	10,8	41,7	20,7
RS2076537	RS9268368	C6ORF10	06:032425613	C6ORF10	06:032441933	20,6	10,5	41,5	20,62
RS2076537	RS9268384	C6ORF10	06:032425613	C6ORF10	06:032444564	20,8	10,7	41,5	20,62

SNP identifiers		Annotations SNP1		Annotations SNP2		Primary Dataset Case Control		Replication dataset Case Control	
SNP1	SNP2	Gene SNP1	Chrmosomal Position SNP1	Gene SNP2	Chrmosomal Position SNP2	Omnibus	Epistasis	Omnibus	Epistasis
RS1033500	RS2076537	C6ORF10	06:032415360	C6ORF10	06:032425613	20,9	10,7	41,5	20,62
RS2076537	RS9405090	C6ORF10	06:032425613	C6ORF10	06:032406350	20,9	10,8	41,5	20,62
RS9405090	RS2076537	C6ORF10	06:032406350	C6ORF10	06:032425613	20,9	10,8	41,5	20,62
RS3096700	RS3129943		06:032329760	C6ORF10	06:032446673	20,2	9,0	42,4	20,13
RS2076537	RS3130320	C6ORF10	06:032425613		06:032331236	30,1	15,5	45,1	18,31
RS2076537	RS3096700	C6ORF10	06:032425613		06:032329760	25,8	13,4	42,3	17,87
RS2076537	RS3115573	C6ORF10	06:032425613		06:032326821	18,7	11,4	38	17,22
RS2076537	RS3130315	C6ORF10	06:032425613		06:032328663	18,7	11,4	38	17,22
RS3115573	RS2076537		06:032326821	C6ORF10	06:032425613	18,7	11,4	38	17,22
RS3130315	RS2076537		06:032328663	C6ORF10	06:032425613	18,7	11,4	38	17,22
RS1077393	RS2242660	BAT3	06:031718508	BAT2	06:031705732	12,9	8,0	33,6	16,65
RS2076537	RS2239804	C6ORF10	06:032425613	HLA-DRA	06:032519501	17,3	8,4	30,5	15,3
RS1046089	RS1077393	BAT2	06:031710946	BAT3	06:031718508	13,5	8,1	31,5	14,48
RS2395150	RS6907322	C6ORF10	06:032434023	C6ORF10	06:032432923	22,4	9,5	40,1	13,99
RS2395150	RS411326	C6ORF10	06:032434023		06:032319295	23,4	8,2	43,3	13,71
RS3115573	RS411326		06:032326821		06:032319295	21,4	10,6	34,5	12,82
RS3130315	RS411326		06:032328663		06:032319295	21,4	10,6	34,5	12,82
RS411326	RS3115573		06:032319295		06:032326821	21,4	10,6	34,5	12,82
RS411326	RS3130315		06:032319295		06:032328663	21,4	10,6	34,5	12,82
RS1053924	RS2269425	PRRT1,	06:032228693	LOC653870	06:032231617	18,2	8,9	32	11,2
RS3129941	RS4959089	C6ORF10	06:032445664		06:032327703	30,3	8,4	51	9,156
RS3129941	RS9296015	C6ORF10	06:032445664		06:032326967	30,4	8,4	51	9,156
RS2076537	RS2395150	C6ORF10	06:032425613	C6ORF10	06:032434023	20,0	8,3	35,2	6,65
RS3115573	RS6907322		06:032326821	C6ORF10	06:032432923	16,8	8,2	22,4	4,116
RS3130315	RS6907322		06:032328663	C6ORF10	06:032432923	16,8	8,2	22,4	4,116

SNP identifiers		Annotations SNP1		Annotations SNP2		Primary Dataset Case Control		Replication dataset Case Control	
SNP1	SNP2	Gene SNP1	Chrmosomal Position SNP1	Gene SNP2	Chrmosomal Position SNP2	Omnibus	Epistasis	Omnibus	Epistasis
RS6907322	RS3115573	C6ORF10	06:032432923		06:032326821	16,8	8,2	22,4	4,116
RS6907322	RS3130315	C6ORF10	06:032432923		06:032328663	16,8	8,2	22,4	4,116
RS3746115	RS761226	MATK	19:003734070	PAK7	20:009769007	6,7	8,0	0,74	1,512
RS2170239	RS17085		08:014919295		13:047312488	6,8	8,4	1,02	1,319
RS12705099	RS866482		07:100356394	CADPS	03:062568930	6,5	8,0	1,48	1,119
RS12705099	RS833632		07:100356394	CADPS	03:062569490	6,7	8,2	1,53	1,094
RS206789	RS11706107	DEPDC1B	05:059939728		03:074070374	6,7	8,4	0,66	0,954
RS4326096	RS11706107		05:059926035		03:074070374	6,6	8,3	0,61	0,918
RS918060	RS1914854	LOC400958	02:065009962		03:111227300	8,0	8,5	0,63	0,773
RS7527828	RS11909106	WDR64	01:239915670		21:014852044	7,3	8,1	0,34	0,706
RS555017	RS13235422	MBOAT1	06:020293030		07:017020565	8,3	8,6	0,66	0,646
RS2268474	RS2059421	TSHR	14:080594160	PLEKHB2	02:131592500	6,8	8,0	0,42	0,568
RS1202201	RS2165662	MBOAT1	06:020270857	CNTN5	11:099647982	7,3	8,4	0,91	0,512
RS876594	RS550391	PRMT8	12:003531820		13:068988875	6,9	8,2	0,14	0,402
RS201247	RS2874146	OFCC1	06:010122547	TRA@	14:021444239	8,0	8,3	0,25	0,383
RS1206988	RS2874146	OFCC1	06:010132793	TRA@	14:021444239	8,1	8,4	0,21	0,354
RS16945681	RS6139898		18:003926384	CRLS1	20:005962954	7,7	8,0	0,15	0,314
RS6429288	RS2822767		01:239897499		21:014844910	6,9	8,0	0,05	0,248
RS6845037	RS2644262		04:182355503	FHOD3	18:032477564	8,1	8,5	0,31	0,193
RS10909918	RS663003	PRDM16	01:003198235		11:118328358	6,6	8,5	0,13	0,188
RS6429288	RS11909106		01:239897499		21:014852044	8,0	9,8	0,06	0,181
RS9894630	RS7889974	SPAG9	17:046526413		X::140992709	9,1	9,7	1,41	0,176
RS17661532	RS6472762	ANK1	08:041710431		08:074300848	6,6	8,4	0,03	0,17
RS3750736	RS4238282	OIT3	10:074341617	LOC644725	13:110436314	7,6	9,3	0,02	0,131
RS9316160	RS368416		13:045210726		15:072301706	6,4	8,3	0,23	0,119

SNP identifiers		Annotations SNP1		Annotations SNP2		Primary Dataset Case Control		Replication dataset Case Control	
SNP1	SNP2	Gene SNP1	Chrmosomal Position SNP1	Gene SNP2	Chrmosomal Position SNP2	Omnibus	Epistasis	Omnibus	Epistasis
RS4238641	RS2029347		16:017629616	KIAA1458	04:048108676	7,2	8,5	0,03	0,096
RS6877916	RS10490018		05:135269537	TRPM8	02:234582471	9,2	8,2	0,27	0,069
RS10428541	RS10461060		05:052760915	ATP8A1	04:042242207	6,8	8,6	0,06	0,063
RS6429288	RS2178933		01:239897499	SAMSN1	21:014831652	6,5	8,2	0,01	0,058
RS6477190	RS7024902		09:007371520		09:032261795	8,0	8,1	0	0,038
RS10500737	RS6761029		11:011002835		02:170345142	7,1	8,4	0	0,035
RS1971156	RS1513536	CNTN5	11:099263024		04:060111325	6,8	8,4	0,82	0,031
RS1466971	RS9830450	CNTNAP4	16:075056976	ADAMTS9	03:064607296	7,0	8,1	0,01	0,008
RS2410936	RS4334611		05:106501648	FOXP1	03:071289593	6,8	8,1	0	0,001
RS10773806	RS9548097		12:129693670		13:037384449	6,7	8,1	0,07	3E-04
RS1548577	RS5910109		07:019015687	ODZ1	X::123734843	6,4	8,1	0,01	9E-05
RS4524788	RS4786850		08:084793776	A2BP1	16:006285925	6,8	8,5	0,02	0

5.3 Data Encoding

To evaluate the proposed binary encoding of the data and the PLINK binary encoding format the size of the resulting files was considered in relation to the original QTDT MERLIN format. Also the amount of information lost through the encoding of the original QTDT MERLIN format and the ability of each format to retain different types of information available today was computed. Table 16 provides a summary of the comparison.

Table 16 Results of Data Compression

QTDR MERLIN	Proposed Protocol	PLINK 's protocol
Information Loss for bi-allelic markers		
Used as Reference	None	strand location of heterozygote alleles Aa,aA Missing one of the two alleles A0, 0A, a0,0a
Added Information capability		
tri or quad allelic markers	Alleles deleted from a specific strand	None
Size*		
Pedigree File 3.6 GB		.bed file : 229.6 MB
		.fam file : 30 KB
Map File 12.6 MB	One binary file 496 MB	
		.bim file : 14.1 MB
Total: 3.612 GB		**Total** : 243.7 MB

*Using a dataset containing **1804** subjects with **532578** SNPs per subject.

5.4 Measure of epistasis

In order to compare the proposed method for testing for epistasis and testing for association allowing for epistasis with the respective logistic regression models, a random subset of 218 SNPs were selected from a study that contained half a million SNPs and 1868 subjects described in [63]. All pairs of the randomly selected SNPs were tested on both methods with omnibus and epistasis as well as the 218 main effects. The Pearson's correlation was then estimated for the main, omnibus and epistasis test performed with each approach. The cut-off for the Pearson's correlation is set a-priori for accepting two tests as having a reasonable agreement to r > 0,95 [47,61].

An effort was made to keeping all variables that affect runtime for each test constant. The same computer was used for all analysis, the analysis were set up to run sequentially, while no other program was running on the computer. To make sure the implementation of each method was fair, the statistical package R was used to implement both methods [61]. R provides all required functions for these tests pre-implemented and open sourced; all scripts used in the benchmarking of the two solutions are also available.

In order to evaluate the performance of the proposed measures of interaction testing a random sample of genetic data was analyzed and the results of the proposed method were compared to those of logistic regression. Although both the logistic regression and the Pearson's chi-square test are well established methods for obtaining estimations of the main effect of a single marker it is important to compare the two in this instance since the chi-square test for the main effect forms the basis for estimating the proposed epistasis effect measure between two markers. In Figure 25 a

comparison of the main effect estimations is presented. The Pearson's correlation between the two main effect measures is 0.98755.

The Pearson's chi-square test performed on a contingency table for 2 SNP genotypes is compared in Figure 26 with the omnibus test performed using the logistic regression model fitting approach. The Pearson's correlation for the two omnibus measures is 0.99287.

The comparison between the tests of the interaction between the two SNPs for association with the categorical phenotype, estimated with the two methods, is presented in Figure 27. The Pearson's correlation for the two epistasis measures is 0.9616.

The complete set of all 2 SNP tests (23.6k tests) took 332.5 seconds to run with the proposed methodology based on the Pearson's chi-square test, while the Logistic Regression took 6628.37 seconds to run. This constitutes an increase in speed of 1993% or roughly twenty times faster than the currently commonly used approach of logistic regression.

The two approaches' results have a reasonable agreement with a correlation coefficient greater than 0.95 for each of the main effect, omnibus and epistasis effect tests. Non-independence between SNPs can occur due to Linkage Disequilibrium, and that will result in wrong results for the epistasis but not the omnibus test since this affects the additive property of Pearson's chi-square. However, well established methods exist that can test for LD between markers and therefore detection of false positive results because of this issue are minimized.
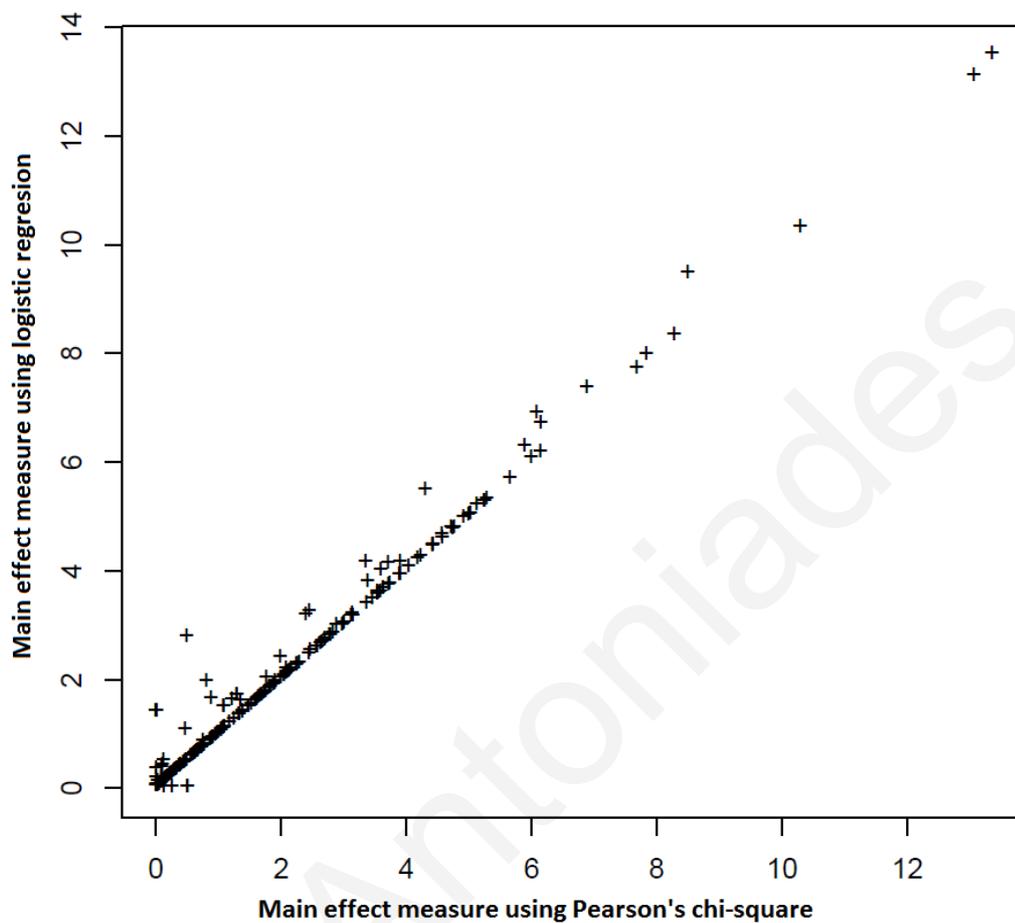
Figure 25 Main effect comparison between logistic regression (y-axis) and Pearson's chi-square test with Yates correction for continuity (x-axis).
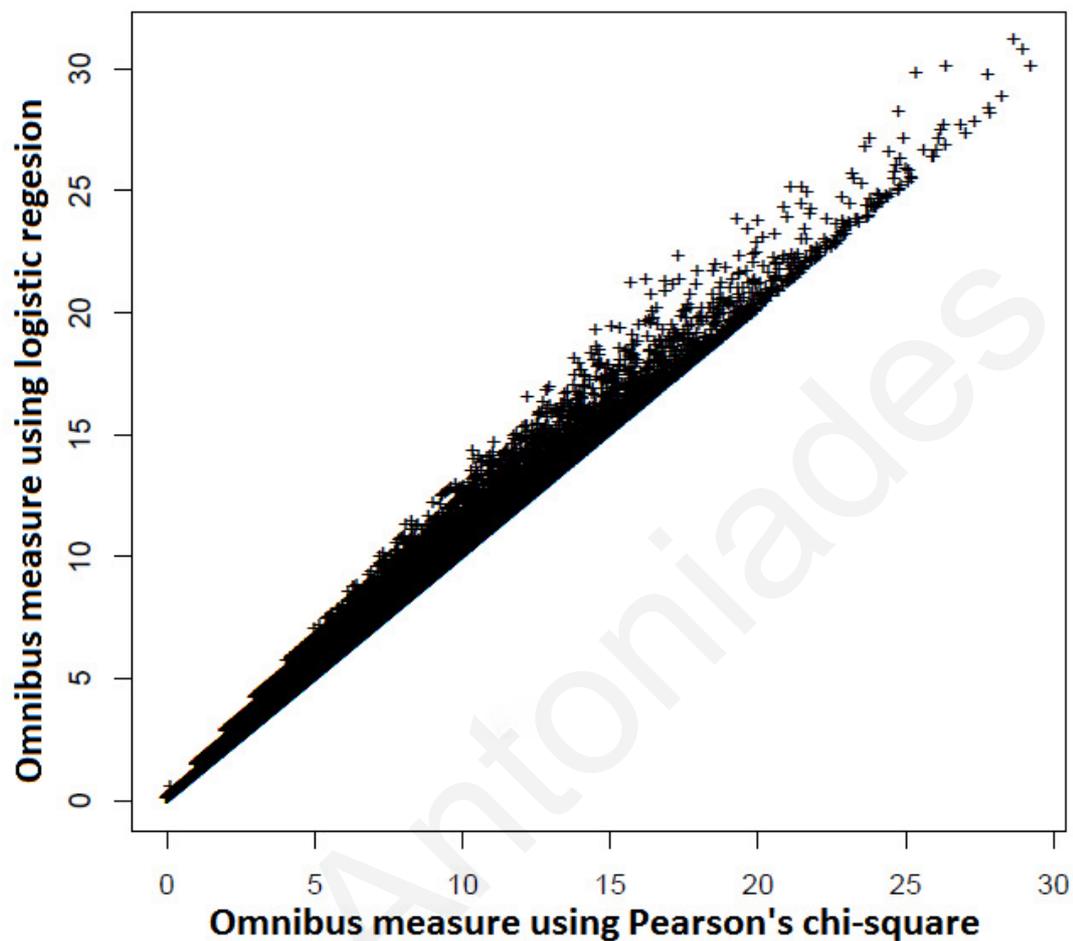
Figure 26 Comparison between the omnibus measures produced using the Pearson's chi-square test with Yates correction and the logistic regression model fitting test.
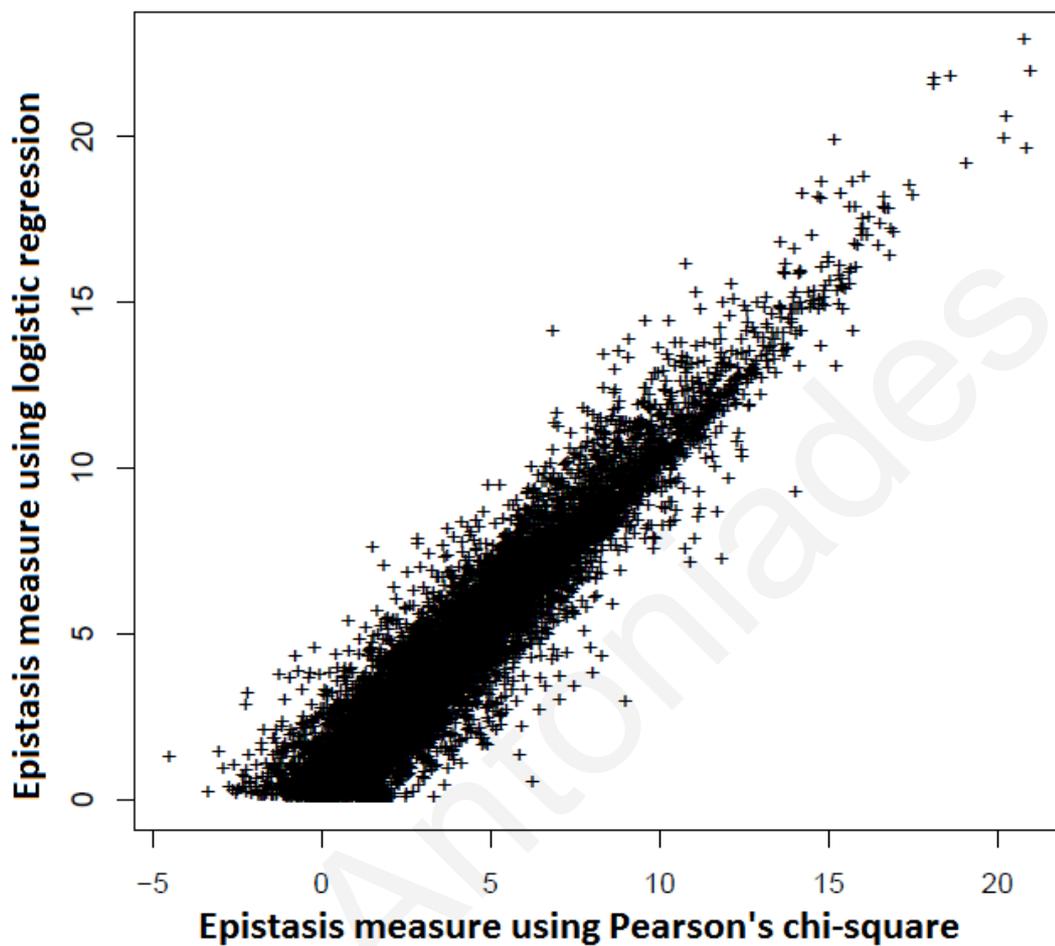
Figure 27 Comparison between the Epistasis measures produced using the Pearson's chi-square test with Yates correction and the logistic regression model fitting test.

5.6 Computation of multiple response variables

Testing the efficiency of the proposed algorithm for computing the multidimensional contingency tables involved the analysis of 1, 5 and 10 response variable within the same dataset (the primary dataset) was carried out. The variables were all randomly generated to match the distribution of the case-control data in the real data. The analysis of each group of different phenotypes were run in sequence once using the proposed algorithm (Figure 10) and then the sequence was repeated for a total of 3 times to get the average runtime s. For the traditional approach (Figure 4) [4,50] since the analyses is identical when analyzing 5 phenotypes or the first 5 of a total of 10 phenotypes, in order to reduce the consumption of computational capacity on the HCC-HPC the phenotypes were run 3 times generating average runtimes for the first 1, the first 5 and all 10. The results are shown in Table 17.

Table 17 Performance of proposed contingency multidimensional contingency table computing algorithm

Response variables	Proposed algorithm	Traditional approach
1	315min	311min
5	311min	1584min
10	320min	3320min

- All times given in minutes

5.7 Evaluation of significance through replication

The replication test provides two ways to test if the results replicate between two independent datasets. The first and more obvious approach is the replication of high significance results in both datasets. The second and often overlooked approach is the replication of the effects between the different genotype combinations. When testing a pair of SNPs for interaction the total number of genotype combinations is nine (columns in Figure 9). Each of the effects (omnibus representing the two main effects and epistasis, and also the epistatic effect alone) should be distributed in the same manner between the two datasets. Figure 28 shows an example of such a test for correlation. The graphs represent the odds ratio for each genotype for a specific response variable and effect. The X and Y axis represent the main and replicating dataset respectively. For a valid replication we expect the majority of the genotype combinations, especially the more frequent ones, to be on the *x=y* diagonal.

Another way to visualize if a test replicated between two datasets across all genotypes with strong associations to disease is shown in Figure 29. In this figure the algorithmic steps are followed to estimate the chi square for the omnibus measure between the two SNPs as described in section 4.3.1(page 77). In this figure though, in order to get a visualization of the distribution of the measure across the genotype combinations for each cell

$$\sum_{0}^{r} \frac{(O_r - E_r - 0.5)^2}{E_r}$$

The result is a chi square measure of the effect of each column (genotype combination). In order to add some more information in the visualization a sign is

added to indicate if a genotype combination is more frequent in cases or controls. Positive indicates that the genotype combination's effect is associated with predisposition towards having the disease (being a case) while negative with being healthy (control).
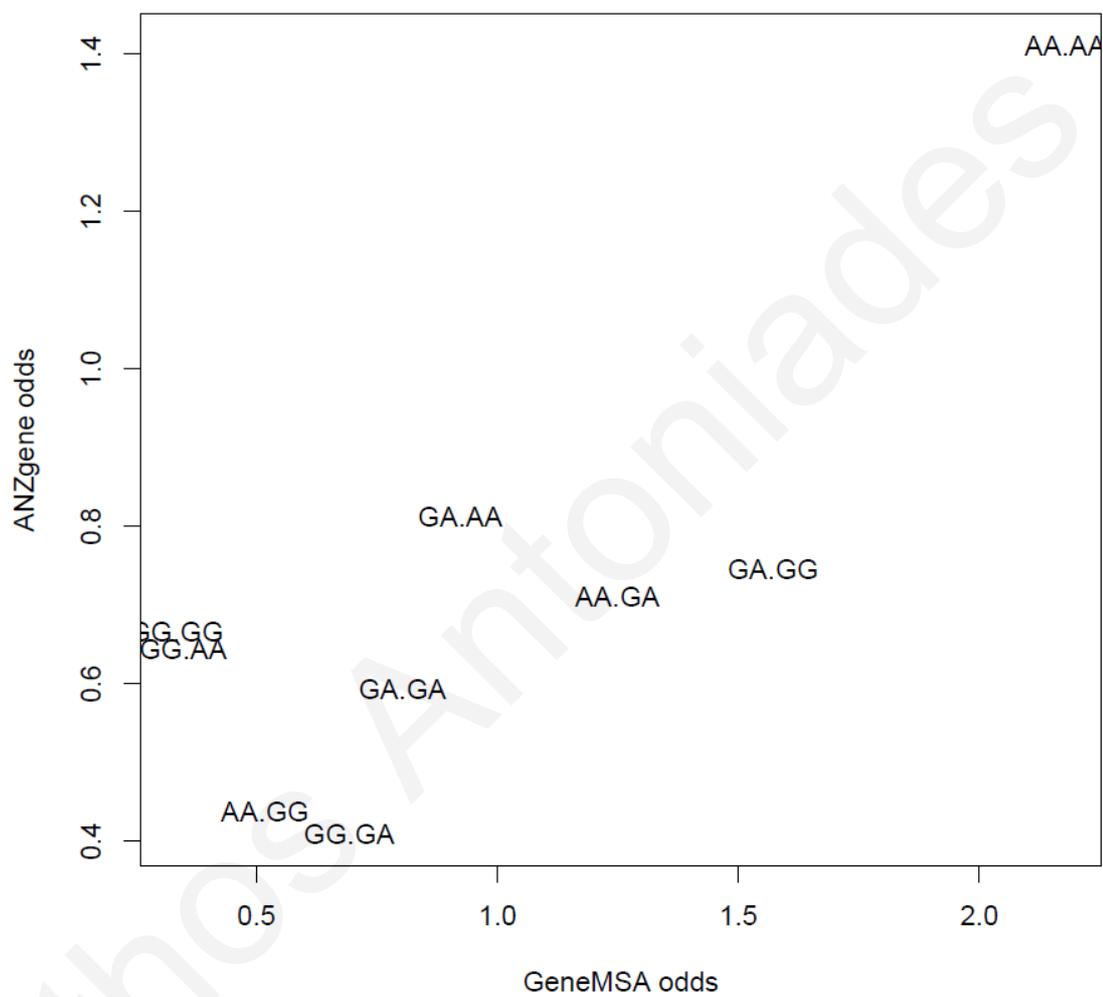
**Figure 28 Replication test for correlation of distribution of effects on genotype combinations between the two diseases**

Chi square measure (arbitrary units) per genotype combination sign based on effect direction (cases negative, control positive)
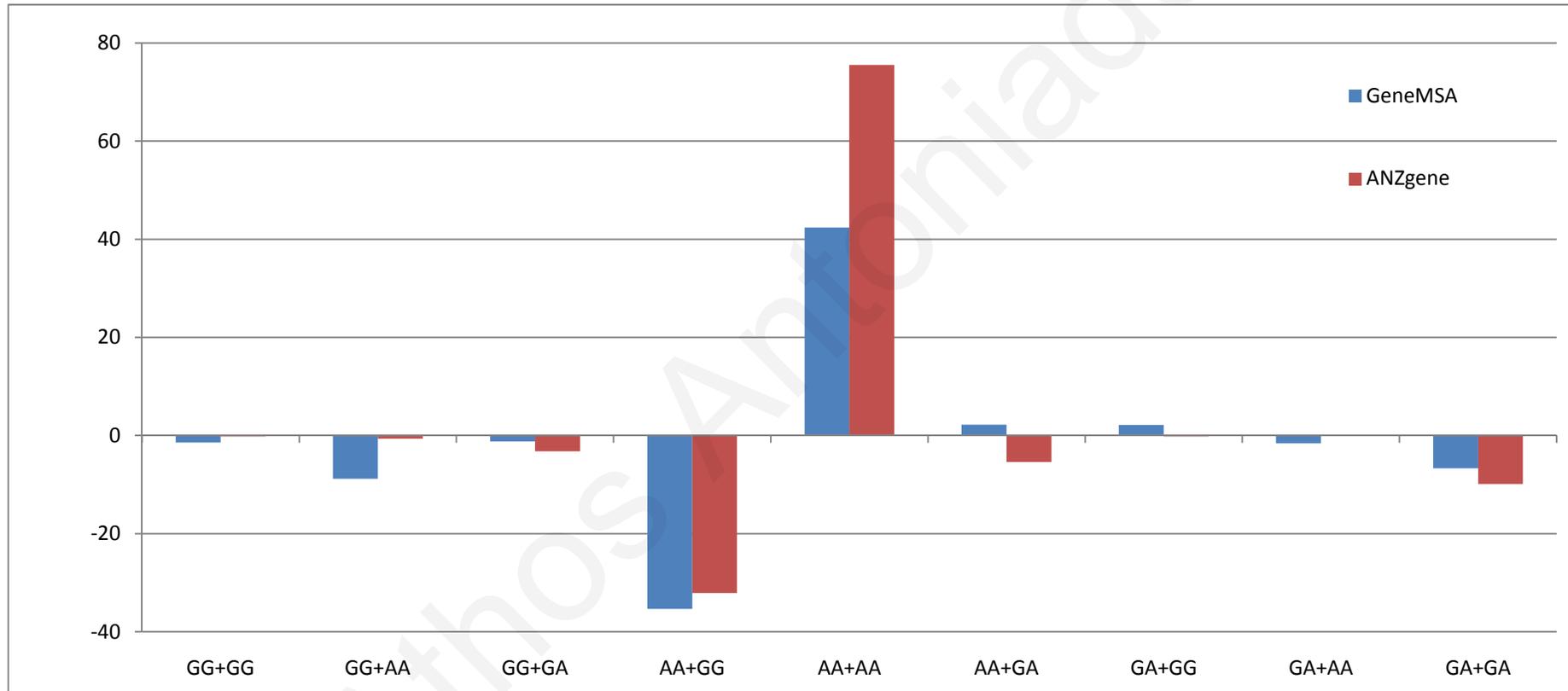
Figure 29 Bar chart of the chi-square metric per genotype in the two datasets signed for phenotype predisposition direction at each

column.

5.8 Hybrid cluster cloud high performance computing (HCC-HPC)

Each HPC system was found to have significant disadvantages [30] that disabled it from being used alone for this analysis. Specifically the WAN grid had very limited data bandwidth between the nodes and the grid's server and did not allow for inter-node communication, while the LAN grid did not have these disadvantages but has significantly less computational resources available.

However, it was found that their advantages and disadvantages were complementary, suggesting that utilizing both resources in parallel utilizing a new purpose built load balancing algorithm with access to both systems could provide the computational capacity necessary to carry out the complete 2 SNP interaction testing of a GWAS more efficiently. Existing models of Hybrid Multi-Cloud architectures [85] were examined; these did not provide enough control to the developer as to what grid each part of the analysis would execute on.

A benchmarking application was developed that replicates the algorithmic approach of this analysis enabling the estimation of the total runtime on each HPC system and identifying the bottlenecks that would be encountered if the analysis were to be performed on it.

A feasibility study focused on the computational complexity and statistical power of detecting significant gene-gene interactions associated with disease was performed in the beginning of this work. The results indicated the need to develop fast algorithmic approaches to perform complete analyses of whole genome data producing robust statistically accurate measures of the probability that each hypothesis tested is a false

positive. The load capacity of the available HPC systems (one relying on load balancing clusters, and the other relying on a cloud utilizing desktops simultaneously used by their primary users [85]) was tested in order to identify significant bottlenecks associated with each [86].

The results were performed with a 2000 subject dataset (1000 case and 1000 controls) and with a datasets with 1k, 5k, 10k and 20k SNPs. The WAN grid server required that all results are transferred to the server until the analysis is complete therefore to avoid filling up its storage space (only 200GB were available for all projects running on it and for both input and output files) all results were deleted after they were transferred to the server. This may be a technical limitation, but it's an important one that is worth mentioning since the required analyses required several orders of magnitude larger storage capacity that would not be possible under the current architecture even if it was to undergo a conservative upgrade. Thus the runtimes include the data transfer to the server, but the limitation of the disk size was bypassed. The results are shown in Table 18.

The results in the Table 18 were further expanded to include estimations based on the 20k SNPs dataset of how long each HPC would take to run in 300k SNPs 550k SNPs and 1000k SNPs respectively with the same number of subjects (2000). These estimations were computed by taking the 20k SNP datasets runtimes and estimating the runtime per work node. Since the computational requirements per worknode remain constant (two 1000 SNPs subsets and all 2000 subjects) the only variable that changes when the number of markers in the GWAS is increased is the number of worknodes. Estimating the number of work nodes each theoretical GWAS dataset would generate is done through the steps provided in section 4.6.4. However since each HPC system has a different total number of nodes, the estimations reflect the best

case scenario for each HPC system assuming all nodes are utilized for performing the proposed analysis.

The results indicated that neither HPC could perform the analyses by itself. The WAN grid required a large amount of data to transfer to and from the nodes, as well as to store the data on the grid's main server. Since the machines were spread across a large geographical region and were connected through a dedicated private network, the impact on performing two SNP interaction testing on the cloud server alone was significant to the available networking resources of the network. The majority of connections between sites were T1 connections (1.5Mbit upload and 1.5Mbit download), with some having T3 (44Mbit) connections. Within each site the majority of computers were connected on a 100Mbit LAN network with a few exceptions with 1000Mbit LAN speed. Since all of these and other machines share these connections, and since high priority communications took place constantly throughout the network, initiating an analysis that would have had consumed too much of this resource would have had an impact on the global operations of the company.

Thus a preliminary assessment on the amount of data needed was performed. This was done by analyzing a randomized sample dataset of 5000 SNPs on the cloud alone. The analyses was identical to the one proposed in the HCC-HPC section of chapter 4, with the exception that now all results that were returned to the server would be fully calculated and annotated. In order to reduce the transfer of data and keep a level field, results transferred were limited to those with a p-value for epistasis less than 0.01. The result was a file with a size of 1,9 Gbytes. Since the method is deterministic and the dataset used had a normal distribution, we can estimate the minimum size needed for the 550,000 SNP dataset. That's 2,30e4 Gbytes, i.e. several orders of magnitude higher than the limit of 200GB on the server's hard drive. It's important to note that

this number only represents the 1% distribution of random top results, in reality true positive results will require extra space (although they may not be that many for the epistatic effect alone).

With runtimes for the proposed framework on the 300k, and 550k SNP GWAS available it was also possible to estimate the expected system runtime of the proposed framework with the current hardware for larger datasets that may become available in the future (Figure 30).
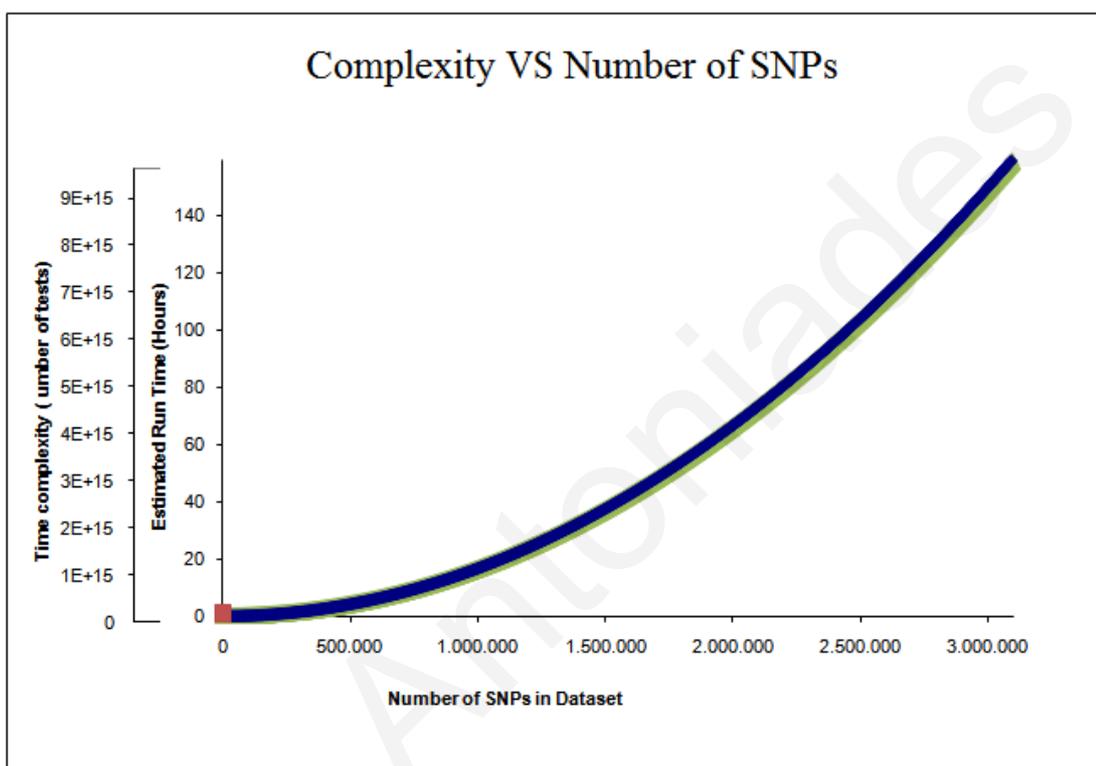
Figure 30 Estimated runtime based on the number of tests for GWAS up to 3.1 million SNPs.

Table 18 Experimental and estimated runtime on each HPC and proposed framework.

Number of SNPs	Number of tests	Number of Work units	LAN grid (200 nodes)		WAN grid (1500 nodes)		Proposed HCC-HPC	
			CPU time*	System time*	CPU time*	System time*	CPU time*	System time*
Times derived experimentally with 1000 SNP size work units on a dataset with 1000 case and 100 controls								
1,000	5.00e05	1	3.01	3.01	5.4	5.40	3.5	3.68
5,000	1.25e07	12	34.9	2.9	65.16	5.43	44.16	3.7
10,000	5.00e07	50	154	3.00	272.5	5.45	185	3.79
20,000	2.00e08	200	618	3.09	1092	5.46	727	3.64
Times Estimated by assuming linear increase with respect to number of tests performed with 20k SNPs test								
300,000	4.50e10	45000	139090	695	179999	164	163.5	109
550,000	1.51e11	151250	467499	2337	604999	550	549000	366
1,000,000	5.00e11	500000	1545453	7727	2727270	1818	1818000	1212

* All times are given in minutes.

Chapter 6 Discussion

This chapter begins with a discussion on the biological results generated through the use of the proposed framework. Focus is given on the impact of the biological result's replication success with statistically significant results to the confidence of the proposed methodology in performing complete 2 SNP interaction testing in GWAS datasets.

The rest of this chapter goes through each of the innovations proposed as part of this framework and discusses the results of specific experiments aimed at evaluating the performance of each.

6.1 Biological Results

The proposed methodology's results in the analysis of the primary dataset yielded statistically significant results even after adjusting for multiple testing. This is a great achievement since none of the previously conducted two SNP interaction frameworks succeeded in discovering statistically significant results after adjusting for the multiple testing problem with the stringent Bonferoni correction.

Replication testing was conducted on all results that passed the threshold of –log p-value > 8 in the primary dataset and had both SNPs represented in both datasets. This resulted in a total of 84 tests that were repeated in the replication dataset, an independent MS study. The a-priory expected number of replications under the null hypothesis was that less than 1 (0.84) tests would replicate by chance with the

threshold for statistical significance set at a p-value of 0.01. However, a total of 49 tests were replicated having statistically significant values in the second dataset as well. Moreover, nearly all of these tests revealed very strong epistatic effects with many orders of magnitude greater than the a-priory defined threshold for statistical significance. An even closer look into the distribution of effects on genotypes between all of the 49 replicated tests, revealed that the distribution of effects between the genotype combinations of each replicated two SNP pair were nearly identical, providing further evidence that the replicated results are probably true positive.

Furthermore, a closer look at the replicated results indicates that they all come from within a very specific region commonly referred to as the Human Leukocyte Antigen (HLA) region. The HLA region has been considered as a primary suspect for immunodeficiency diseases such as MS, with many reported associations across many studies[15,80,81] (see Figure 31).

To summarize the following evidence were found supporting the success of the proposed framework:

(1) Analysis on the primary dataset yielded statistically significant results even after Bonferoni correction for multiple testing.

(2) The top 84 results that were made up of SNPs also genotyped in the replication dataset were tested for replication with 49 replicated statistically significant results.

(3) An examination of the distribution of the effect size on each of the genotype combinations of the two SNPs revealed matching distributions between the two datasets for the replicated, but not for the non-replicated results.

(4) All the replicated results implicate SNPs in and around a region called

HLA, known to be associated with MS (see Figure 31).

The key to note here is that all the above evidence is independent. Thus the probability of the combination of the above evidence existing by pure coincidence is highly unlikely. This clearly meets all criteria for validation testing of the results [29,5,1,54,87]. Comparing to all other methods in the literature that attempted to perform two SNP interaction testing, this is the first method based on the literature review conducted as part of this work that has provided statistically significant results and replication success in an independent dataset [32,54].
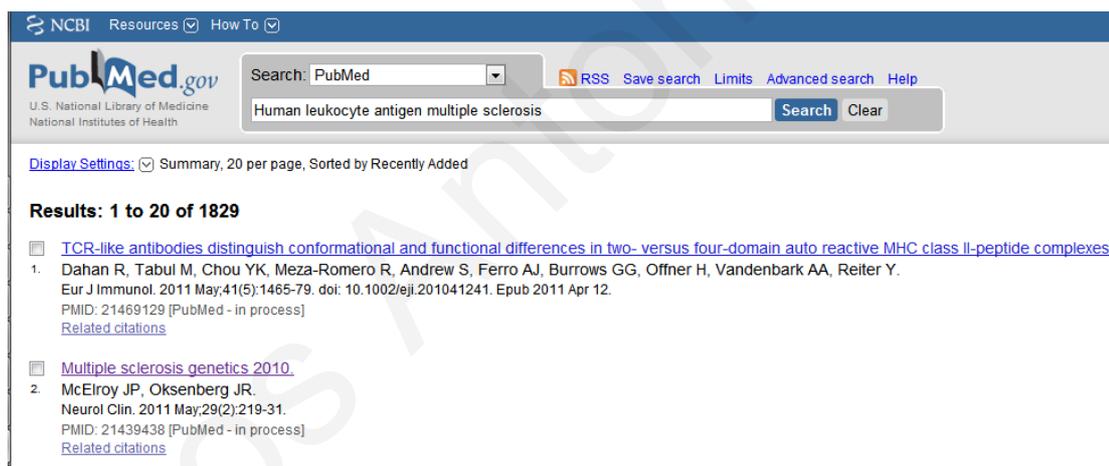
Figure 31 PubMed search linking Human Leukocyte Antigen to multiple sclerosis where 1829 publications were found

6.2 Data Encoding

Compressing data requires a balance between readability, loss of information and required reduction in data size. In order to make a decision on how to balance these parameters it is important to keep in perspective how the data will be used.

6.2.1 Information Loss

On one hand loosing information due to encoding is undesired; however, if the encoding is used to simply speed up analyses then it's not an issue so long as the original un-encoded file is kept for future analyses. However if an algorithm that needs the information of which strand each heterozygote marker's alleles are on is developed[83], or if there are deletion regions overlapping the markers in either strand, then utilizing the encoding methodology proposed in this dissertation will in a single table or file encode all of that information efficiently. Another issue is the actual storage of the data for long term use or for transferring over the internet. Utilizing a non-lossy approach to compressing the data that incorporates all genetic information into a single file can reduce the resources required[13,52].

6.2.2 Added Information Capability

In this dissertation the focus was on bi-allelic markers as they are the ones that current high throughput genotyping technologies are able to genotype. However, it should be noted that the format of QTDT Markers enables encoding of tri or quad allelic markers as well since each allele is encoded as an ASCII character. Neither PLINK's neither binary ped file nor the format proposed could handle tri or quad allelic markers. Large datasets with tri-allelic markers do not exist today (to the knowledge

of the author) while information on deleted markers is available through CNV analyses and other methodologies available to the various genotyping platforms. Tri-allelic and quad allelic markers were ignored in this encoding since they are very rare and unlikely to be included in GWAS analyses in the future[31,36].

6.2.3 Size of encoded data

The compression rate can easily be estimated since both encodings are deterministic, however an actual test performed using an average dataset size of 1806 subjects and 532,579 markers is also provide as an example. The PLINK binary ped file was able to compress the file to $1/15^{th}$ of its original size while the proposed method compressed the file to exactly double the size of that achieving just $1/7.5^{th}$ of the original size[13,52]. However, both encodings produce enough of a compression to overcome the issue of the large data since the data can now fit on the average computer's physical memory (RAM) as today's typical computer has at least 1 GB available[52].

6.3 Measure of epistasis

The proposed measures of epistasis and the omnibus measure (testing for association allowing for epistasis) were compared to the equivalent logistic regression test. Both the correlation between the results, as well as the computing of runtime of each approach was evaluated. First the issue of correlation with logistic regression, the traditional analytical approach for this type of analyses is addressed.

The a-priori defined threshold for the Pearson's correlation was passed in the omnibus and epistasis measures indicating that there is a reasonably good agreement between the two approaches to measuring main, omnibus and epistasis effects.

An expected problem of the proposed methodology might be caused by non-independence between markers. This dependence would cause epistasis test results estimated relying on the Pearson's chi-square additive property as described in this work to be incorrect. However, well established techniques exist for testing for dependence between markers, commonly caused by Linkage Disequilibrium [37]. It's also possible to test for Linkage Disequilibrium in various populations based on other datasets that have been fully analyzed and made the results available through the HapMap project [36]. The omnibus measure does not utilize the additive property of the Pearson's chi-square test, therefore its validity is not affected by LD [47].

Runtimes of the two methods were compared, while keeping all other parameters that affect runtime of the analyses constant. The proposed Pearson's chi-square approach to estimating epistasis was twenty times faster. This could be translated in savings in terms of computational resources needed to perform specific analyses or in the case where more tests are needed to be performed than the resources can provide, a significant boost by a factor of 20 for the number of possible tests to perform on a given computational resource.

Furthermore, the proposed methodology provides both a measure of interaction (epistasis) as well as a measure of association allowing for interaction (omnibus), through the analyses of a single contingency table.

Measures	Epistasis test	Allow for epistasis tests	Response variable	Adjustment by covariates	Marker type	Tested on real data	Replicated results	Bias	Result interpretation	Scalability	References
Proposed Method	Yes	Yes	Categorical	No	Genotypic	Yes	Yes	Negative on LD for epistasis	p-values	High	[2]
Regression analyses	Yes	Yes	Categorical, linear	Yes	Genotypic Allelic	Yes	Yes	*LD, in some cases	p-values	Low	[13,54,55]
Odds ratio Multiplicative Interaction	Yes	No	Binary	No	Allelic	Yes	No	*LD, *MAF, size of dataset, Heterozygote effects	Approximated p-values	High	[3,32,56]
Case only (χ^2)	Yes	No	Categorical	No	Genotypic	Yes	No	Non linkage equilibrium	p-values,	High	[13,32]
Recursive Partitioning	No	Yes	Categorical	No	Genotypic Allelic	Yes	No	*LD, main effects	None, requires follow-up analyses	Medium, parameter dependant	[32,57,58]
Multi-dimensionality reduction (MDR)	No	Yes	Categorical	Yes	Genotypic Allelic	Yes	No	*LD, main effects	None, requires follow-up analyses	Low	[59,60]

*MAF= minor allele frequency
*LD= Linkage Disequilibrium

Table 19 Comparison proposed versus existing measures of epistasis using statistical or data mining approaches

6.4 Computation of multiple response variables

In tests performed with the analysis of a single response variable or multiple response variables, the expected reduction in runtime was successfully recorded as expected (Table 17). The minor deviations between the computational resources available to the analyses during each test seem to provide a stronger effect to the runtime than adding more than one response variables to test the proposed methodology. However, using the traditional approach the computational time of the analysis was consistently linearly increased with the addition of each response variable. This result was consistent both in terms of the primary analyses of the complete whole genome scan, as well as in terms of the annotation stage where response variables were used to simply annotate top results from other analyses. Since the algorithm does not attempt to estimate but rather attempt to completely compute all contingency tables there's no loss of power associated with this increase in efficiency.

This algorithm only provides a performance boost when more than one contingency table based on the same explanatory variables and different response variables is needed. In GWAS associations studies this is the typical case, thus the performance increase is valuable since even though linear, it has a pragmatic impact on the runtime as it can enable researchers to include a number of the response variables in a single analysis run.

6.5 Evaluation of significance through replication

Replication testing is considered the holy grail of verification of reported genetic associations to disease [8,27,29,32]. In order to perform a replication test, an independent GWAS dataset is needed that has near identical inclusion criteria (same disease definition) and the same genetic markers (genotypes). For the purposes of this dissertation, the validity of results needs to be determined in order to provide by extension the validity of the proposed framework that was used to derive them.

The replicating dataset had matching subject inclusion criteria, had more subjects, but unfortunately it was genotyped on a platform that only included 300k SNPs compared to the 550k SNPs in the primary dataset. However, both genotyping platforms were generated by the same company (Illumina) and it seems that the 550k platform is in fact an extension of the 300k platform. That is the Illumina 550k platform contains all 300k SNPs from the Illumina 300k platform plus 250k extra SNPs.

This limited the replication tests to only those that were composed of SNPs genotyped in both platforms. However, since the goal of the replication test was primarily to provide evidence as to the performance of the proposed framework that is not a problem as long as the remaining replicable tests produce successful replications. The results with markers not in the replication dataset will be published in order to provide the opportunity to researchers conducting new studies to attempt to replicate them in future studies of other independent datasets. The threshold of -log p-value >8 was set as the threshold for performing replication tests.

In order to gain further evidence of the replication success between the two datasets, two types of figures were generated for all replicated results (Figure 28 and Figure 29).

Figure 28 presents the odds ratio between the two datasets on a specific result plotted with each dataset on an axis. We expect that if a result is indeed a true positive then the odds ratios of all genotype combinations will be presented across the y=x diagonal, as is the case in the example of Figure 28.Figure 29 presents the chi-square metric across each genotype combination in each dataset with a sign added to indicate if the effect indicated predisposition towards being either a case or a control. It's clear from Figure 29 that the genotype combinations with the strongest effect sizes match in both effect size and in predisposition sign between both datasets even though the measures used to identify top results and test for replication do not consider the distribution of effects across the genotypes. It's also useful to note that in this graph the predisposing genotypes are identified, a key question that needs to be answered as a post processing step in all genetic analyses results[8,25] as it can be used to generate further hypothesis on the biological mechanisms that form the causality of the detected epistatic effect.

6.6 Hybrid cluster cloud - high performance computing (HCC-HPC)

6.6.1 Performance of WAN grid alone

Utilizing the WAN grid alone for this analysis was prohibitive for multiple reasons. First, the server that was controlling the entire system only had a 200GB storage space and a requirement that all input, output data and debugging data be stored on it for all analysis taking place on the WAN grid. That is, not only analyses using this proposed framework. The size of the resulting data are several orders of magnitude larger. Secondly, the data would need to be transferred to and from the nodes of HPC that are scattered around the world and connected through the private network of one of the biggest companies in the world, generating that amount of traffic on the network is expected to have an impact on the day to day operations of the company and was thus prohibited to even try by the IT (Information Technology) administration. Thirdly, and perhaps more significant, the results showed, that for large datasets using the proposed HCC-HPC platform the runtime was reduced. This is because considerably less time was spent waiting for data transfers to complete when using only the WAN grid alone.

6.6.2 Performance of dedicated LAN

The dedicated LAN performed faster than either the WAN grid or the proposed HCC-HPC platform but only for very small datasets where the number of work nodes created was less than 200. The reason is that it only has 200 processing nodes, while the WAN grid had 1500, and by extension the HCC-HPC can utilize both the 1500

WAN and the 200 LAN grid simultaneously. When the number of work nodes created to analyze a dataset exceeds 200 the analyses on the LAN grid significantly slows down since it can only run 200 work nodes in parallel while the WAN grid can analyze 1500.

Considering that the benchmarking was conducted under ideal load conditions for the LAN grid given that it was the only application running (tests utilized a special high priority queue that suspended all other work on the HPC until all jobs in the priority queue finished). The average usage of the system is at around 80%, and genetic analyses were considered highly risky exploratory work. So the slots allocated for analysis were actually of very low priority, that only ran when no other queues request resources.

6.6.3 Performance of the HCC-HPC proposed platform

Overall the HCC-HPC performed better than using either HPC alone providing the necessary computational resource to perform the complete 2 SNP interaction testing framework in a reasonable time.

The intermediary results coming from the WAN grid that are highly compressed and include all results that pass a threshold of p-value < 0.01 in either omnibus or epistatic measures in any of the analyzed phenotypes are stored in a permanent storage to enable quick future queries testing hypothesis that involve a subset of the search space. These are accessible to all machines in the private network equipped with a web service that enables query and visualization of all results generated in real time. This way, even though the primary 2 SNP interaction analyses was only interested in the top results in this case to provide evidence to the effectiveness of the method, in the future it' quite possible that researchers will want to data mine the results buried

within the top results to test individual hypothesis they may have. Multiple testing only applies to the number of tests a researcher performs, thus if someone were to generate a list of tests that he had a-priory defined a hypothesis for, then the multiple testing will only apply to the number of tests he will perform, and not to the entire set of possible tests performed to generate the dataset. This is assuming of course that there was no bias towards the generation of the hypothesis tested by the actual results of this analysis.

The proposed HCC-HPC system has provided the computational requirements needed to perform the analyses in a reasonable time frame and considerably faster than using either of the two HPC platforms it utilizes individually. However, it is applicable only to the problem of testing all pair wise SNPs in a GWAS study for an effect, and even though it is an innovative contribution, it's only a minor one.

6.7 The proposed framework versus other gene-gene interaction testing methodologies

Table 19 provides an overview of the proposed and existing gene-gene interaction methodologies. It's clear that from all tests conducted so far with any methodologies the proposed method utilizes the most computational resources. It's also the only one that provides a complete framework for storing all results with a p-value of 0.01 or less on any measure and response variable analyzed giving the opportunity to researchers to test their future hypothesis without needing to re-run the analyses. Furthermore the WAN grid used that provided the majority of the computational resources is actually rarely utilised and relies on redundant computing resources on personal computers. Thus although it's a large resource, it's impact to the cost, and

power consumption are minimized since it's designed to identify machines that are already switched on and running (otherwise their resources would had been wasted). In doing so, the proposed framework can within a reasonable time span, analyze even the largest reported GWAS datasets in existence today (1 million SNPs).

The proposed method is scalable and provides both a measure of interaction (epistasis) and also a measure of the association allowing for interaction (omnibus). It's the only method to have discovered statistically significant results, after Bonferoni correction for multiple testing. Also, it's the only method that has replicated its top findings in an independent replication dataset yielding statistically significant results in the second replication dataset as well.

Marchini et al. [54], proposed an exhaustive search for association allowing for interaction only on simulated data. The proposed framework presented in this thesis dissertation allowed testing for epistasis (epistasis measure)as well as testing for association allowing for epistasis (omnibus measure).

The BOOST [56] and the complementary GBOOST [64] algorithms compute epistasis through the use of odds ratios. Furthermore, even though these algorithms were tested on real data, and also tested for replication they did not succeed in generating statistically significant results after Bonferoni correction. The replication test although it did produce some top results in both datasets, the number of these was not greater than what was expected assuming no true positive effects were included in them [56,64].

Recursive Partitioning Tree, and also Random forests techniques [57,58], offer some key features, specifically, they enable testing for more than 2 SNP interactions and they could also be used to test for interactions between multiple genetic loci, and environmental factors. Their key disadvantage though is that they can only be applied

to small subsets of data, and not to an exhaustive search of a GWAS. To date, statistically significant or replicated results have not been discovered with either method [57,58].

6.8 Potential pharmaco-genetic impact of the proposed framework

Drugs function by binding and selectively interacting with specific pharmaceutical targets. Each drug has its own target(s) that is a protein that it can bind to or interact with to modulate its function in a way that inhibits, stimulates or modifies its action[88]. With the complete sequencing of the human genome one might expect a plethora of drug targets for many diseases to be discovered[7]. This however isn't the case, since the analytical techniques applied to this data were focused on simple Mendelian diseases (diseases where a single genetic polymorphism was responsible for a phenotypical trait) while the majority of common diseases in humans are complex diseases (multiple genetic polymorphisms and environmental factors predispose a subject's disease status)[8].

The proposed analytical framework attempts to discover evidence of interactions between genetic polymorphisms. Thus, looking for epistatic effects, that are closer to what is expected to be the underlying causative effect of complex diseases.

Once statistically significant and replicated results are identified (as is the case in the work presented in this dissertation), these can be used to generate new hypothesis that can lead to:

New drug targets and by extension to new drugs that attempt to inhibit, stimulate or modify the action of one or both of the proteins generated by the epistatic effect detected[3,32,88].

Personalized medicine, by identifying interactions associated with adverse events or drug susceptibility, and thus creating the possibility to provide efficient widely available genetic tests that will help doctors decide what drug to give to a patient so as to increase the likelihood of positive outcome [3,32,79].

Table 20 Two SNP interaction testing frameworks and the proposed method

Multi-locus Interaction testing framework	Computing Platform	HPC Scalable	Measures	Largest number of tests recorded	Deterministic	Accessibility	Tested on Real data	Replicated statistical significance	References
Proposed method	Computing cloud of 1500 PC nodes, cluster with 200 nodes	Yes	Interaction, allowing for interaction	550k SNPs, 2000 subjcets	Yes	Low	Yes	Yes	
Exhaustive search Marchini *et al*	10 node cluster	Yes	Association test allowing for interaction	300k SNPs 2000 subjects	Yes	Low	No	No	[54]
BOOST	Single core x86 cpu	No	Odds ratio based interaction measure	351k SNPs 5000 subjects	Yes	High	Yes	No	[56]
GBOOST	GPU	Limited Scalability	Binary odds ratio based interaction measure	351k k SNPs 5000 subjects	Yes	High	Yes	No	[64]
Recursive Partitioning (Tree)	Symmetric multiprocessing under development	No	Follow up analyses necessary to provide interpretable measure	Limited to only a few dozen SNPs	Depends on parameters.	High for small number of tests	Yes	No	[58]
Random Forest	Symmetric multiprocessing, in theory HPC compatible with many architectures	No	Follow up analyses necessary to provide interpretable measure	Limited to only a few dozen SNPs	No	High for small number of tests	Yes	No	[57]

Chapter 7 Conclusions and Future Work

7.1 Conclusions

The proposed framework presented in this thesis dissertation looks at all possible pair-wise interactions between the loci studied, performing an exhaustive search of a GWAS dataset. The proposed framework for performing complete two SNP interaction test on a GWAS dataset is broken up in the key original contributions of this work that had to be developed to address key issues relating to the problem and also provide computational efficiencies. Each original contribution performance on addressing it's corresponding problem was evaluated through experiments designed at exposing the difference between the proposed approaches and the current approaches used in the field as described in the literature.

The framework, involves first encoding the data in a lossless, binary format that significantly reduces its size. The proposed format encodes genotypic data into a binary format in order to compress it and at the same time preserve all the information relating to the bi-allelic markers and the strand location of each allele. However it does double in size the resulting datasets from existing encoding methodologies that are lossy, but looses information only necessary to certain type of analytical approaches. The analytical approaches that would benefit from the proposed format of encoding are primarily the ones that take into account the strand on which heterozygote alleles are based on. That is, the existence of the marker on a specific DNA strand, identifying if the second wasn't available due to genotyping errors or a deletion over the marker on that strand. Due to the need to use two bits per allele for

encoding while only needing 3 states for each allele we were left with an available fourth state. That fourth state we proposed that is used to denote markers that were deleted as that information is becoming commonly available from genotyping platforms available already; however, future researchers may choose to use the fourth state to code a different state an allele can be in. Also in cases where compression of the data in a non-lossy way for storage, backup or data transfer is needed, the methodology proposed would be preferable.

The proposed frameworks addresses the problem of identifying a measure of epistasis and a measure of association allowing for epistasis by proposing a new measure and algorithmic approach to estimating it based on Pearson's chi-square association test and it's additive property. The proposed method, which utilizes the Pearson's chi-square to estimate main effects and omnibus tests and uses its additive property to estimate the epistasis effect test is twenty times faster than the currently commonly used approach of logistic regression. The two approaches' results have a reasonable agreement with a correlation coefficient greater than 0.95 for each of the main effect, omnibus and epistasis effect tests. Non-independence between SNPs can occur due to LD, and that will result in wrong results for the epistasis test only since this affects the additive property of the Pearson's chi-square test. However this error results in a negative bias towards SNP pairs with LD between them. Even though this is not ideal, considering all other currently published methods for performing 2 SNP interaction most of them exclude SNP pairs that are in LD. Since this problem is not one that other methodologies address, the fact that this method only has a negative bias is an advantage since it can be looked at as the minimum score of the interaction between the two SNPs.

In GWAS studies, one key common observation is that more than a single response variable is interesting and needs to be analysed. In this dissertation, an algorithmic approach was introduced that enabled the analyses of multiple response variables on the same explanatory variables with minimal increase in computation time compared to analyzing a single variable at a time. The algorithm is based on counting the contingency tables, a p-hard problem, simultaneously with a single pass through the dataset for explanatory variables rather than a pass for each test.

The framework presented in this dissertation study addresses the need for a high performance computing system capable of performing a test on all possible pair-wise interactions between two SNPs through the use of a proposed hybrid cluster-cloud high performance computing framework (HCC-HPC). The combined effect of two SNPs on a phenotype can be interpreted in terms of the effect of a single marker's association to the phenotype, commonly referred to as the main effect of the first SNP, the main effect of the second SNP, and the interaction between the two, commonly referred to as the epistatic effect. The total of these effects is considered as the omnibus measure, a term used in statistics to identify tests that are composed by multiple independent effects as in this case. The interaction between the two SNPs is also considered through a proposed new epistasis measure using logistic regression to compare and evaluate the performance of the proposed measure of epistasis.

The proposed framework is used to perform analyses on GeneMSA, a GWAS with multiple sclerosis matching cases and controls that is used as the primary dataset. Stringent statistical significance thresholds were defined a-priory adjusted for the multiple testing problem. The validity of reported statistically significant results, and also of the top results with a p-value less than 10e-8 were tested for replication in an independent GWAS called ANZgene. Both the analyses of the main dataset and the

replication test of the top results of the main dataset revealed statistically significant results after Bonferoni correction for multiple testing. This provides evidence for the validity of the replicated statistically significant results, and by extension it also provides the evidence that the proposed framework, can discover validated through replication statistically significant results.

7.2 Future Work

7.2.1 Measure of epistasis for n SNPs

In the proposed framework SNP pairs were tested for epistatic effects with success. A follow up question could be are there any three SNP epistatic effects associated with the disease. The search space grows exponentially when increasing the number of SNPs in every test, however, filtering, dimensionality reduction and other data mining techniques or heuristics could be applied to limit the number of SNPs to test to a small subset. However, even though the technology to reduce the search space is available, a computationally efficient statistical test to apply in testing for n SNP interaction does not. The epistasis measure proposed and accompanied algorithm that also produces the omnibus measure can be expanded to test for n SNP interactions. This work is ongoing and preliminary results indicate that the performance increase compared to applying a logistic regression model is increased even more than the 20 times increase in the 2 SNP approach.

7.2.2 Distributed Neural Network (NN) approach for subject cluster discovery

Through the application of the two MS dataset to the two SNP analytical framework proposed a list of replicated statistically significant 2 SNP interactions were derived. These replicated interactions all lie on or very close to a region on chromosome six known to be associated with immunodeficiency disorders. This region is in high LD and also has many genes in it making it difficult to interpret the results from two SNP interactions to two gene interaction effects. The problem then arises on how to further analyze the results in the region in order to be able to identify exactly how many independent disease factors exists.

As part of the future work of this dissertation, an analytical work is being designed that involves the creation of an efficient distributed Self Organizing Feature Maps (SOFM) framework. This framework attempts to cluster the subjects based on the genotypes in the region of interest. SOFMs were selected due to their unsupervised learning behaviour; this provides a way to test the produced results. If it successfully identifies clusters with mostly cases than controls, then we know that the cluster is driven by genetic markers associated with the disease.

This method is already under development, some preliminary results are shown in Figure 32. The analysis was run on only cases, only controls and on all subjects, 10 times. On all runs the largest cluster in both cases and in all subjects seemed to be the same involving 42 cases and 9 controls.

7.2.3 Gene-gene interaction testing in a box

One of the key problems of the proposed framework is that it relies on two specific HPC resources that belong to a corporation. In doing so, the availability of the system is reduced only to that cooperation. In an attempt to provide an alternative that can be accessible by any group of researchers new iterations of the proposed framework need to be created for different types of computing resources available. A currently considered approach is the use of a single server machine with smp x86 processors as well as multiple GPU processors. GPU's have already being applied in performing two SNP interaction testing by others (GBOOST program) using alternative analytical frameworks [64].

The goal is to develop a system prototype based on the compute unified device architecture CUDA[65] that utilizes the GPU cores for the analytically intensive part and the x86 CPU's for dealing with annotating, sorting and storing the results. Once the system is up and running, estimation on its performance will reveal if it's a direction worth pursuing. The system once completed will be used to provide an open analytical platform for analyzing GWAS.

50x50 network controls

Neuron Coordinates		Subjects in Cluster
X	Y	
18	0	13
5	4	10
14	7	10
12	0	9
24	24	9
33	34	9
10	10	7
12	6	7
24	17	7

50x50 network cases

| Neuron Coordinates		Subjects in Cluster
X	Y	
18	2	42
35	49	27
5	5	23
39	42	20
2	1	17
30	6	10
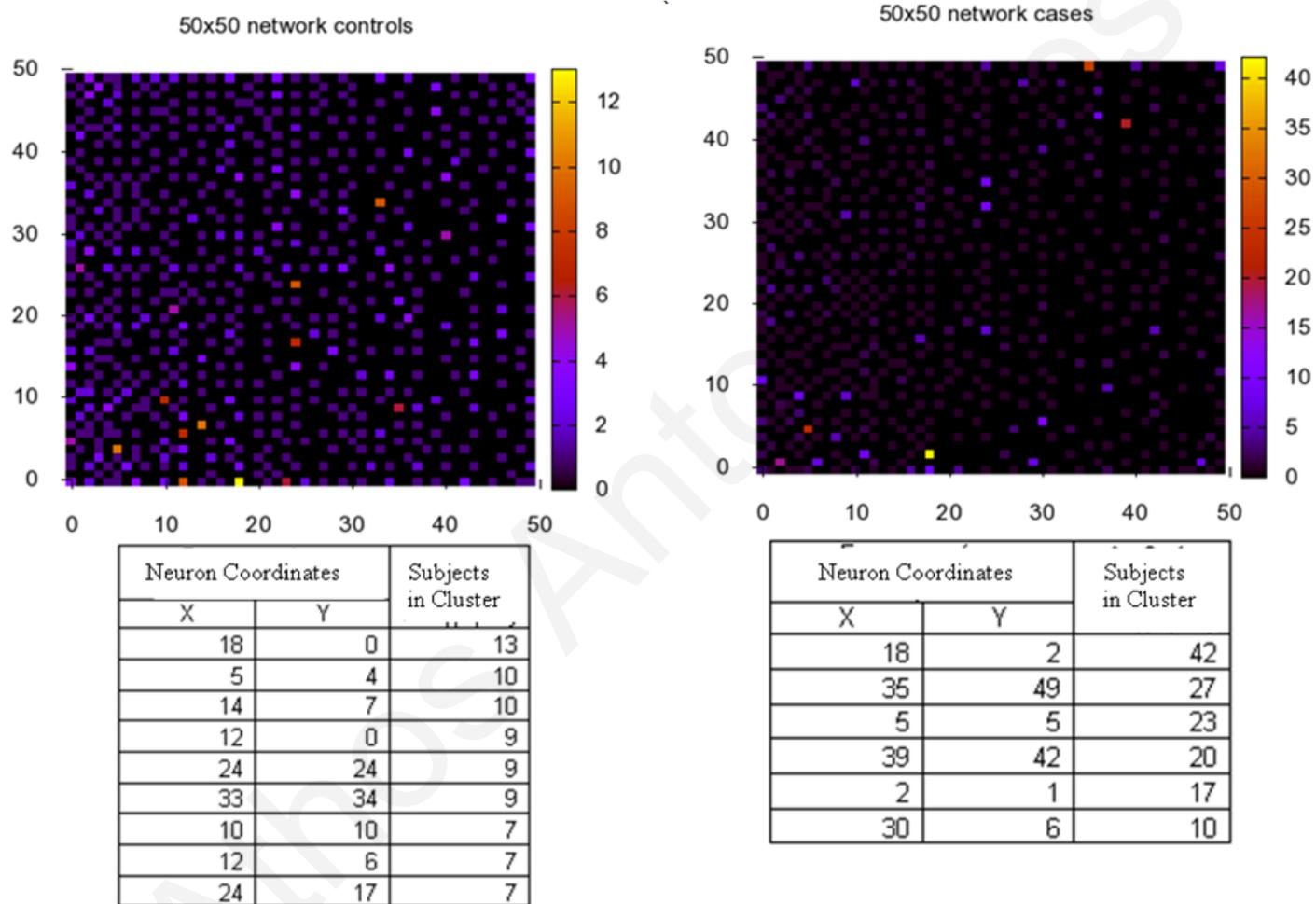
Figure 32 Sample run of Neural Network for clustering subjects based on their genetic data at specific loci.

Bibliography

[1]   A. G. Heidema et al., "The chalenge for genetic epidimiologists: how to analyze large number of SNPs in relation to complex diseases", *BMC Genetics*, no. 7, pp. 23, 2006.

[2]   A. Antoniades, P. Matthews, C. Pattichis, and N. Galwey, "A computationally fast measure of epistasis for 2 SNPs and a categorical phenotype", in *IEEE Engineering in Medicine and Biology Society*, Buenos Aires, 2010, pp. 6194-6197.

[3]   H. J. Cordell, "Epistasis: what it means, what it doesn't mean , and statistical methods to detect it in humans", *Human Molecular Genetics*, no. 11, pp. 2463-2468, 2002.

[4]   I. R. Dohoo et al., "An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies", *Preventive Veterinary Medicine*, no. 29, pp. 221-239, 1997.

[5]   D. R. Cox and N. M. Reid, *The Theory of Design of Experiments*: Chapman & Hall/CRC, 2000.

[6]   S. Stigler, "Fisher and the 5% level" *Chance*, vol. 21, no. 4, pp. doi:10.1007/s00144-008-0033-3, Nov. 2008.

[7]   International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome", *Nature*, no. 409, pp. 860-921, Feb. 2001.

[8]   A. T. Thorton-Wells, J. H. Moore, and J. L. Hainess, "Genetics, statistics and human disease: analytical retooling for complexity", *Trends in Genetics*, vol. 12, no. 20, pp. 640-647, Dec. 2004.

[9]   S. Jenkins and N. Gibson, "High-throughput SNP genotyping", *Comparative and Functional Genomics*, vol. 3, no. 1, pp. 57-66, 2002.

[10]   P. Y. Kwok, "High-throughput genotyping assay approaches", *Pharmacogenomics*, vol. 1, pp. 95-100, Feb. 2000.

[11]   J. N. Hirschhorn and M. J. Daly, "Genome wide association studies for common diseases and complex traits", *Nature Review Genetics*, vol. 6, pp. 95-108, 2005.

[12]   J. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon, "Merlin-rapid analyses of dense genetic maps using sparse gene flow trees", *Nature Genetics*, vol. 30, pp. 97-101, 2002.

[13]   S. Purcell et al., "PLINK: a toolset for whole genome association and population based linkage analysis", *American Journal of Human Genetics*, vol. 81, pp. 559-575, Sep. 2007.

[14]   D. A. Freedman, *Statistical Models: Theory and Practice.* Cambridge, United Kingdom: Cambridge University Press, 2005.

[15]   M. J. Greer and A. P. McCombe, "Role of gender in multiple sclerosis: Clinical effects and potential molecular mechanisms", *Journal of Neuroimunology*, doi:10.1016/j.jneuroim.2011.03.003, pp. Epub PMID: 21474189, Apr. 2011.

[16]   M. Firmann et al., "The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome", *BMC Cardiovascular Disorders*, vol. 8, no. 6, Nov. 2008.

[17]   M. Dyer and J. Mont, "Sampling contingency tables", *Random Structure and Algorithms*, vol. 10, no. 4, pp. 487-506, 1997.

[18]   I. Bezakova, B. Bhatnagar, and D. Randall, "Sampling and counting contingency tables using Markov Chains", in *15th Annual International Conference on Computing and Combinatorics (COCOON)*, NY, USA, 2009, pp. 5609.

[19]   M. Cryan and M. Dyer, "A polynomial-time algorithm to approximately count contingency tables when the number of rows is constant", *Journal of Computer and System Sciences*, vol. 67, no. 2, pp. 291-310, Sep. 2003.

[20]   R. G. Miller, *Simultaneous Statistical Inference*, 2nd ed., Verlag, Ed. New York, USA: Springer, 1981.

[21]   X. Zhu, *Knowledge Discovery and Data Mining: Challenges and Realities*, ISBN: 978-759904252-7, Ed. New York, USA: Hershey, 2007.

[22]   T. V. Perneger, "Concept: multiple comparisons", *British Medical Journal (BMJ)*, no. 316, pp. 1236-1238, Apr. 1998.

[23]   C. W. Dunnet, "A multiple comparisons procedure for comparing several treatments with a control", *Journal of the American Statistical Association*, no. 50, pp. 1096-1121, 1955.

[24]   H. Abdi, "Bonferroni and Šidák corrections for multiple comparisons", in *Encyclopedia of Measurement and Statistics*. Thousand Oaks, California: Salkind, 2007.

[25]   J. D. Storey and R. Tibshirani, "Statistical significance for genome-wide studies", *Proceedings of the National Academy of Sciences*, vol. 16, no. 100, pp. 9440-9445, 2003.

[26]   J. Gayán, "A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis", *BMC Genomics*, no. 9, pp. 360, July 2008.

[27]   J. P. Ioannidis et al., "Replication validity of genetic association studies", *Nature Genetics*, vol. 29, pp. 306-309, 2001.

[28]   A. J. G. Hey, "High performance computing-past, present and future", *Computing & Control Engineering Journal*, vol. 8, no. 1, pp. 33-42, Aug. 2002.

[29]   J. P.A. Ioannidis, "Commentary: grading the credibility of molecular evidence for complex diseases", *International Journal of Epidemiology*, vol. 3, no. 35, pp. 572-578, 2006.

[30]   I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid. Enabling scalable virtual organizations", *International Journal of Supercomputing Applications*, vol. 3, no. 15, pp. 2-13, 2001.

[31]   T. Strachan and A. Read, *Human Molecular Genetics*, 4th ed.: Garland Science, 2010.

[32]   H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases", *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392-404, June 2009.

[33]   F. Crick, "Central dogma of molecular biology", *Nature*, vol. 227, no. 5258, pp. 561-563, Aug. 1970.

[34]   A. Antoniades et al., "Discovering genetic polymorphism associated with gene expression levels across the whole genome", in *Engineering in Medicine and Biology Society IEEE EMBC*, Mineapolis, MN, USA, 2009, pp. 5466-5469.

[35]   D. F. Wyszynski et al., "Relation between atherogenic dyslipidemia and the Adult Treatment Program-III definition of metabolic syndrome (Genetic Epidemiology of Metabolic Syndrome Project)", *American Journal of Cardiology*, vol. 95, pp. 194-202, Jan. 2005.

[36]   International HapMap Consortium., "The International HapMap Project", *Nature*, vol. 426, no. 6968, pp. 739, Dec. 2003.

[37]   J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "HAPLOVIEW: analysis and visualization of LD and haplotype maps", *Bioinformatics*, Jan. 2005.

[38]   C. Stavrou, C. C. Deltas, T. C. Christophides, and A. Pierides, "Outcome of kidney transplantation in autosomal dominant medullary cystic kidney disease type 1", *Nephrolysis, Dialisies Transplantation*, vol. 18, no. 10, pp. 2165-2169, Oct. 2003.

[39]   J. D. Badano and N. Katsania, "Beyond Mendel: An evolving view of the human genetic disease transmission", *Nature Reviews Genetics*, vol. 3, pp. 779-789, 2002.

[40]   X. Wu et al., "A novel statistic for genome-wide interaction analysis", *Public Library of Science Genetics (PLOS Genetics)*, vol. 6, no. 9, pp. e1001131, Sep. 2010.

[41]   J. D. Hunter, "Gene-enviroment interactions in human diseases", *Nature Reviews Genetics*, vol. 6, pp. 287-298, 2005.

[42]   S. B. Everit, *Cambridge Dictionary of Statistics*. Cambridge: Cambridge University Press, 2002.

[43]   Y. Dodge, *The Oxford Dicrionary of Statistical Terms*. Oxford, England: Oxford University Press, 2003.

[44]   R. G. Mogull, *Seconds-Semester Applied Statistics*, Kendall/Hunt Publishing Company, Ed., 2004.

[45]   A. R. Fisher, *The design of Experiments*, 8th ed. Edinburgh: Hafner, 1966.

[46]   D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. London, United Kingdom: Chapman and Hall, 1974.

[47]   B. S. Everitt, *The Analysis of Contingency Tables, Second Edition*. London, United Kingdom: Chapman and Hall, 1992, pp. 164.

[48]   R. Hubbard and M. J. Bavarri, "Confusion over measures of evidence versus errors in classical statistical testing", *The American Statistician*, vol. 57, no. 3, pp. 171-178, Aug. 2003.

[49]   R. E. Nisbett, G. T. Fong, D. R. Lehman, and P. W. Cheng, "Teaching reasoning", *Science*, vol. 238, no. 4827, pp. 625-631, Oct. 1987.

[50]   K. Miller, "Bringing the fruits of computation to bear on human health: It's a tough job but the NIH has to do it", *Biomedical Computational Review*, vol. 5, no. 2, pp. 18-28, 2009.

[51]   M. D. MCBurney and L. T. White, *Research Methods.* California: Wadsworth Learning, 2004.

[52]   A. Antoniades, L. Loizou, A. Aristodimou, and C. Pattichis, "A binary format for genetic data designed for large whole genome", in *IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2008,* Athens, 2008, pp. 4-10.

[53] Q. Yang, M. J. Khoury, F. Sun, and W. J. Flanders, "Case-only design to measure gene-gene interaction", *Epidemiology* , vol. 10, no. 2, pp. 167-210, Mar. 1999.

[54] J. Marchini, P. Donnely, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases", *Nature Genetics*, vol. 37, pp. 413-417, 2005.

[55] R. D. Cook and S. Weisberg, "Criticism and influence analysis in regression", *Sociological Methodology*, vol. 13, pp. 313-361, 1982.

[56] X. Wan et al., "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies", *The American Journal of Human Genetics (AJHG)*, vol. 87, no. 3, pp. 325-340, Sep. 2010.

[57] L. Breiman, "Random forests", *Machine Learning*, vol. 5, no. 32, 2001.

[58] L. Breiman, H. J. Freidman, A. R. Olshen, and J. C. Stone, *Classification and Regression Trees*, CRC, Ed. New York, USA: Chapman and Hall, 1984.

[59] L. Baston, M. Reilly, J. D. Rader, and S. A. Foulkes, "MDR and PRP: a comparison of methods for highorder genotype–phenotype associations", *Human Heredity*, vol. 58, pp. 82-92, 2004.

[60] H. J. Moore, "Computational analyses of gene-gene interactions using multifactor dimensionality reduction", *Expert review mollecular diagnostics*, vol. 4, pp. 795-803, 2004.

[61] R Foundation for Statistical, *R: A Language and Environment for Statistical Computing*, 3900051070th ed. Vienna, Austria: R Development Core Team, 2005.

[62] S. E. Fienberg, "The analysis of multidimensional contingency tables", *Ecology*, vol. 51, no. 3, pp. 413-433, May 1970.

[63] L. J. Scott et al., "Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry", *Proceedings of the National Academy of Sciences USA*, vol. 18, no. 105, pp. 7501-7507, Apr. 2009.

[64] L. S. Yung, C. Yang, X. Wan, and W. Yu, "GBOOST: a GPU based tool for detecting gene-gene interactions in genome-wide case control studies", *Bioinformatics*, vol. 27, no. 9, pp. 1309-1310, May 2011.

[65] NVIDIA, "NVIDIA compute unified device architecture programming guide version 2.1", NVIDIA, Technical Report 2008.

[66] H. Matsuda, "Physical nature of higher order mutual information: Intrinsic correlations and frustration", *Physical Reviews*, vol. 62, no. 3, pp. 3096-3102, Sep. 2000.

[67] S. Watanabe, "Information theoretical analysis of multivariate correlation", *IBM Journal of Research and Development*, vol. 4, pp. 66-82, 1960.

[68] D. M. Ritchie et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer", *American Journal of Human Genetics*, vol. 69, no. 1, pp. 138-147, July 2001.

[69] A. B. McKinney, M. D. Reif, D. M. Ritchie, and J. H. Moore, "Machine learning of detecting gene-gene interactions: a review", *Applied Bioinformatics*, vol. 5, no. 2, pp. 77-88, 2006.

[70] H. J. Moore and M. S. Whilliams, "New strategies for identifying gene-gene interactions in hypertension", *Annals of Medicine*, vol. 34, no. 2, pp. 88-95, 2002.

[71] A. B. McKinney, E. J. Growe, J. Guo, and D. Tian, "Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis", Public Library of Science *Genetics (PLOS Genetics)*, vol. 5, no. 3, pp. e1000432, Mar. 2009.

[72] L. K. Lunetta, B. L. Hayward, J. Segal, and P. Van Eerdewegh, "Screening large-scale association study data: exploiting interactions using random forests", *BMC Genetics*, vol. 5, no. 32, 2004.

[73] A. Beureau et al., "Identifying SNPs predictive of phenotype using random forests", *Genetic Epidemiology*, vol. 28, no. 2, pp. 171-182, Feb. 2005.

[74] S. Lee, Y. Chung, R. Elston, Y. Kim, and T. Park, "Log-linear model based multifactor-dimensionality reduction method to detect gene–gene interactions", *Bioinformatics*, vol. 23, no. 1, pp. 2589-2595, 2007.

[75] J. M. Kwon and A. M. Goate, "The Candidate Gene Approach", *Alcohol Research and Health*, vol. 3, no. 24, pp. 164-172, Fall 2000.

[76] R. N. Hoover, "The evolution of epidemiology research: from cottage industry to "big" science", *Epidemiology*, vol. 1, no. 18, pp. 13-17, 2007.

[77] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Philosophical Magazine*, vol. 5, no. 50, pp. 157-175,

1900.

[78]   F. Yates, "Contingency table involving small numbers and the χ2 test", *Journal of the Royal Statistical Society*, no. 1, pp. 217-235, 1934.

[79]   H. K. Kroemer and H. E. M. Schwabedissen, "A piece in the puzzle of personalized medicine", *Clinical Pharmacology and Therapeutics*, vol. 1, no. 87, pp. 19-20, Jan. 2010.

[80]   S. E. Baranzini et al., "Genome-wide asociation analysis of susceptibility and clinical phenotype in multiple sclerosis," *Human Molecular Genetics*, vol. 18, no. 4, pp. 767-778, Nov. 2009.

[81]   C. O'Gorman et al., "Familial recurrence risks for multiple sclerosis in Australia.," *Journal of neurology, neurosergery and psychiatry*, vol. PMID: 21551470, p. Epub, May 2011.

[82]   T. Marwala, *Computational Intelligence for Missing Data Imputation, Esitmation, and Management Knowledge Optimixation Techniques*: Information Science Reference, 2009.

[83]   J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnely, "A new multipoint method for genome-wide association studies via imputation of genotypes", *Nature Genetics*, vol. 39, pp. 906-913, 2007.

[84]   B. N. Howie, P. Donnely, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies", Public Library of Science *Genetics, (PLOS Genetics)* , vol. 5, no. 6, pp. e1000529, 2009.

[85]   Univa & Intel, "Reference architecture: How to build a hybrid multi-cloud", Univa Intel, White Paper AST-0001110, 2010.

[86]   B. Jacob, M. Brown, K. Fujui, and N. Trivedi, "Benefits of grid computing", in *Introduction to Grid Computing*.: IBM Redbooks, 2005, ch. 2, pp. 7-18.

[87]   D. V. Zaykin and L. A. Zhivotovski, "Ranks of genuine associations in whole-genome scans", *Genetics*, vol. 171, pp. 813-823, Oct. 2005.

[88]   C. R. Prüll, "Caught between the old and the new--Walther Straub (1874-1944), the question of drug receptors, and the rise of modern pharmacology", *Bulletin of the History of Medicine*, vol. 3, no. 80, pp. 465-554, Fall 2006.

[89]   J. H. Moore and S. M. Williams, "Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis", *Bioessays*, no. 27, pp. 637-646, 2005.

[90]   C. Kessler and J. Keller, "Models for parallel computing: review and perspectives", *PARS⬚Mitteilungen*, vol. 24, no. 0177-0454, pp. 13-29, Dec. 2007.

[91]   M. Huerta, F. Haseltine, and Y. Liu. "National Institute of Health Working Definition of Bioinformatics and Computational Biology". 2000 [Online last access May 2011]. http://www.bisti.nih.gov/docs/compubiodef.pdf

[92]   I. Foster. Designing and Building Parallel Programs. ISBN 9780201575941, Dec. 1995.

[93]   B. Delvin and N. Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping", *Genomics*, vol. 29, no. 2, pp. 311-322, 1995.

[94]   A. Ashok, N Wilner, "Grid in action: Harvesting and reusing idle compute cycles: How United Devices Grid MP helps this happen at the UT Grid project" UNIVA. GridMP datasheet. 8 june 2005

[95]   A. R. Bailey, *Design of Comparative Experiments*. Cambridge: Cambridge University Press, 2008.

[96]   D. R. Cox, *Planning of Experiments*, New York: Wiley, 1958.

[97]   Lister Hill National Center for Biomedical Communications, US National Library of Medicine, National Institutes of Health, Department of Health & Human Services, *Genetics Home Reference. Your Guides To Understanding Genetic Conditions*. United States: U.S. National Library of Medicine, 2011.

[98]   P. Hogeweg, "Simulating the growth of cellular forms", *Simulation*, vol. 31, no. 3, pp. 90-96, 1978.

[99]   P. Hogeweg and D. B. Searls, "The roots of bioinformatics in theoretical biology", Public Library of Science *Computational Biology (PLOS Computational Biology)*, vol. 7, no. 3, pp. e1002021, 2011.

[100]   W. Li and J. Reich, "A complete enumeration and classification of two-locus disease models", *Human Heredity*, vol. 50, pp. 334-349, 2000.

Appendix A Publications

Journals:

F. C. Calboli, F. Tozzi, N. W. Galwey, A <u>Antoniades</u>, V Mooser, M Preisig, P. Vollenweider, D Waterworth, G. Waeber, M. R. Johnson, P. Muglia and D. J. Balding. "A genome-wide association study of neuroticism in a population-based sample". *Public Library of Science PLoS One.* 2010 July 9;5(7):e11504.

L. J. ScottJ, P Muglia, X. Q. Kong, W. Guan, M Flickinger, R. Upmanyu, F. Tozzi,J. Z. Li, M. Burmeister, D. Absher, R. C. Thompson, C. Francks, F. Meng, <u>A. Antoniades</u>, A. M. Southwick, A. F. Schatzberg, W. E Bunney, J. D. Barchas, E.G. Jones, R. Day, K. Matthews, P. McGuffin, J. S. Strauss, J. L. Kennedy, L. Middleton, A. D. Roses, S. J. Watson, J. B. Vincent, R. M. Myers, A. E. Farmer, H. Akil, D. K. Burns and M. Boehnke "Genome Wide Association and meta-analysis of bipolar disorder in individuals of European ancestry". *Proceedings of the National Academy of Sciences of the United States of America (PNAS),* 2009 May 5;106(18):7501-6. Epub 2009 Apr 28.

Currently working on:

<u>A. Antoniades</u>, P. M. Mathews, J. Rubio, J. Stankovich, L. Middleton, N. Galwey and C. Pattichis. "A distributed complete 2 SNP interaction analyses framework applied to Multiple Sclerosis". Draft under Review, by co-authoring institutions and stake holders. To be submitted IEEE Trans Computational Biology

<u>A. Antoniades</u>, P. M. Mathews, J. Rubio, J. Stankovich, L. Middleton, N. Galwey and C. Pattichis. "The complete 2 SNP interaction analysis of two independent datasets reveals replicated statistically significant novel interactions in Multiple Sclerosis". Under writing, by co-authoring institutions and stake holders. To be submitted PNAS Genetics.

Conferences:

<u>A. Antoniades</u>, I. Kalvari, C. Pattichis, N. Jones, P. Matthews, E. Domenici and P. Muglia. *"*Discovering genetic polymorphisms associated with gene expression levels across the whole genome*". IEEE Engineer in Medicine and Biology Conference, 2009*, pp. 5466-5469 September 2~6 2009.

<u>A. Antoniades</u>, L. Loizou, A. Aristodimou, C. Pattichis: "A Binary Format for Genetic Data Designed for Large Whole Genome Studies that Enables both Marker and Strand Based Analyses*". IEEE International Conference on Bioinformatics and Bioengineering*, doi:10.1109/BIBE.2008.4696674, Athens, Greece, Oct 8th 2008

A. Antoniades, N. Galwey, R. Gibson, D. Waterworth, F. Tozzi, P. Muglia and C. Pattichis: "Cracking gene-gene interactions associated with common diseases in high dimensional whole-genome data sets through the use of distributed computing resources". *Glaxo Smith Kline Scinovations forum*, Italy, Verona, Sept 2008

O. Ray, A. Antoniades, I. Demetriades and A. Kakas:"Abductive Logic Programming in the Clinical Management of HIV/AIDS". *European Conference on Artificial Intelligence*, Italy,pp. 437–441, Aug 28th - Sept 1st 2006

B. A. Julstrom and A. Antoniades: "Two hybrid evolutionary algorithms for the rectilinear Steiner arborescence problem". *ACM Symposium on Applied Computing*, SAC2004 pp. 980-984, Nicosia, Cyprus, March 14-17 2004

B. A. Julstrom and A. Antoniades: "Three Evolutionary Codings of Rectilinear Steiner Arborescence". *Genetic and Evolutionary Computing Conference*, GECCO2004, pp. 1282-1291,Seattle, Washington, June 26 – 30 2004