Increasing the Discovery Power and Confidence Levels of Disease Association Studies: A Survey

Technical Report 2014-620

Layan Nahlawi

Supervisors:
Dr. Parvin Mousavi, Dr. Hagit Shatkay
Supervisory Committee:
Dr. Dorothea Blostein, Dr. James Stewart

School of Computing, Queen's University
Kingston, Ontario, Canada
July 10, 2013

Table of Contents

1 Introduction **1**

2 Disease Association Studies **4**
2.1 Disease Association Studies Data 4
2.2 History of Disease Association Studies 5
2.3 Categorization of Disease Association Studies 6
2.4 Standard Statistical Approaches for Disease Association Identification 7

3 Challenges in Disease Association Studies **11**
3.1 Missing Heritability 11
3.2 Shortcomings of Traditional Statistical Testing 12
3.3 Inter-Population Applicability 12
3.4 Replication Issues 12

4 Modifications to Current techniques in Disease Association Studies **14**
4.1 GWAS Data 14
4.1.1 Quality Control and Design Recommendations for GWAS 15
4.1.2 Consolidation of GWAS Data 16
4.2 Preprocessing of GWAS Data 16
4.3 Significance Evaluation 17
4.4 Association Discovery Techniques 19
4.4.1 Statistical Modeling 19
4.4.2 Computational Approaches 21

5 Integrative Approaches for Disease Association Studies **24**
5.1 Meta-Analyses 25
5.2 Pathway Analysis 28
5.3 Interaction and Function Information 31
5.4 Gene Expression Analyses 32

6 Conclusion and Future Work **35**

List of Figures

2.1 An example of a genotype sequence for three SNPs along with their possible haplotypes. 5

2.2 Association studies time line. The time line shows the beginning of association studies with Fisher's model , which led to genetic screening. Genetic screening then paved the way for candidate gene association studies and later to GWAS 6

2.3 The categorization of association studies. Disease association studies comprise two categories: genetic and non-genetic (such as environmental). Genetic associations can be performed either on a candidate gene or on a genome-wide scope, whereas non-genetic association is performed by analyzing epigenetics data. 7

3.1 Summary of challenges in association studies (precisely GWAS). Missing heritability, statistical testing problems, inter-population applicability and failure of replication are the challenges discussed in this report. 13

4.1 An example of a decision tree where each node represents a SNP and branching is based on the allele value at each node.The leaves denote the classes: Either *case* or *control*. 22

5.1 Meta-analysis of GWAS can be performed in different ways. The first consists of combining previous GWAS and using them to discover novel association. The second is the cumulative analysis where the results of previous GWAS are used as prior for another GWAS. The third is the replicative meta-analysis where results from one GWAS are replicated in another data sample. 26

5.2 A diagram obtained from Baranzini *et al.* [3]. It summarizes the proposed framework for the integration of network-based pathway analysis with GWAS. 30

5.3 A diagram obtained from a publication by Cusanovich *et al.*[20]. It summarizes the flow of gene expression analysis (in 1a)) , GWAS (in 1b)) and the integration of both results (in 2)) to identify novel associations (in 3)) that would have been missed if GWAS alone was performed. 34

List of Tables

1.1 Different categories in addressing shortcomings in disease association studies 2

2.1 An example of an genotypic contingency table for a recessive penetrance disease (e.g. Daltonism) with major allele 'A' and minor allele 'T' in a dataset of 250 cases and 225 controls. Major homozygous or heterozygous genotypes do not manifest the disease whereas a minor homozygous genotype manifests it. 8

2.2 Comparison of common statistical tests for association studies 10

Chapter 1

Introduction

The majority of common terminal diseases are influenced by multiple genetic and environmental factors [62, 15, 78]. Cancer, Multiple Sclerosis, Alzheimers disease and Diabetes are a few examples of such diseases [25, 7, 58, 71, 86]. These diseases are known as polygenic due to the additive and interactive contribution of multiple genes in the disease risk as opposed to monogenic diseases affected by a single gene [15, 99]. Even though uncovering the main causes of disease is deemed difficult due to the complexity of gene-gene and gene-environment interactions, major research efforts aim at identifying disease risk factors, especially genetic ones [29, 69, 21, 92].

Over the past decade, disease association studies have been used to uncover the susceptibility, aetiology and mechanisms of action pertaining to common diseases [42, 57, 56]. In disease association studies, genetic data is analyzed in order to reveal the relationship between different types of mutations on the DNA strands, known as variants, and a disease of interest [92, 57]. The ultimate goal of association studies is to facilitate susceptibility testing for disease prediction, early diagnosis and enhanced prognosis [62, 69]. Susceptibility testing and disease prediction are particularly important for diseases that can be prevented by diet, drugs or change in lifestyle [62]. The discovered associations assist in understanding the molecular mechanisms influenced by the reported variants, and in identifying important risk factors [56]. Moreover, genetic risk factors are used to reveal the correlations of common variants with functional genes and their coded proteins. These correlations are in turn utilized to uncover possible target loci for better therapeutic treatments [69].

Different types of disease association studies analyze different types of data, genetic and non-genetic. However, the most common form of association studies utilizes genetic common variant data, specifically single nucleotide polymorphism (SNP) data [48]. A SNP is a mutation in a single nucleotide that appears in at least 1% of the population [25, 9]. The hypothesis behind using SNP data in association studies is that common variants influence common disorders [30]. Additionally, common variants are easier to sequence and analyze due the higher possibility of being encountered in the assayed population. As a consequence, common variants are involved in the majority of reported disease associations.

Current association studies suffer from several shortcomings. First, the reported associations of common variants do not explain all the variability in the genetics of complex disease. Therefore, common variant's contribution in the variability of a disease, known as its effect size on disease risk, is considered to be modest [56, 68]. The unexplained portion

Category	Sub-Category
Modifications to current association studies techniques	Quality control of the analyzed data
	Significance evaluation
	Association testing approach
Integrative approaches for association studies	Integrating several association studies for meta-analysis
	Integrating interaction and functional information with genetic data
	Integrating gene expression data with genetic data
	Complementing association studies with pathway information

Table 1.1: Different categories in addressing shortcomings in disease association studies

of the variability in disease genetics, known as the "missing heritability", is believed to exist in associations of rare variants that are hard to capture with the current association studies data [56, 34]. Second, a pressing issue in association studies is the reproducibility and replication of results in both the studied population (intra-population) and other populations (inter-population) [57, 56]. Populations of European ancestry have been extensively analyzed in disease association studies and a significant number of informative associations was reported [57]. In order to generalize such findings and use them in further analysis, reproducing and replicating the results on both intra-population and inter-population basis is needed [57]. Third, statistical hypothesis testing is used to make inferences about the association of variants with the disease being studied. The statistical significance of these inferred associations is reported using *p-values*, which reflect the rate of false positive associations [56, 48]. Using such methodology of testing and significance evaluation does not convey the statistical power of the study as well as the false discovery rate (the number of reported associations that are actually unassociated) [56, 83, 65, 6]. The statistical power of a study is its ability to detect associations that are actually real and it is affected by several factors such as the sample size, the sampling distribution mean and standard deviation.

Recently, considerable research efforts have been directed toward providing strategies to overcome the challenges of association studies mentioned above. Some of these efforts focus on machine learning techniques to detect disease associations such as using decision trees and random forests [60]. Others aim to enhance existing testing and evaluation methods for disease association studies, for instance by substituting the use of *p-values* by q-values to account for false discovery rate [83]. Yet other research efforts propose techniques for post-processing the findings of association studies or complementing these findings with additional information [18, 31, 3, 11]. Such techniques suggest the integration of functional information, gene expression or pathway data with association studies [18, 31, 3, 11]. These latter techniques attempt to answer questions about disease mechanisms [33]. Finally, research about alternative methods for association studies seeks to find non-genetic factors in disease susceptibility such as the processes that modify the function of genes without altering the nucleotide sequence itself [73, 44].

This report surveys the literature that addresses the shortcomings of current methods the identify genetic disease associations. In addition, it reviews the suggested solutions that either enhance some aspect of the methodologies, or complement them. The surveyed topics cover two categories of techniques, as summarized in table 1.1. The first category, *Modification to current association studies techniques*, comprises techniques that aim to pre-process association data, and the methods used for testing/evaluation of disease associations. The second category, *integrative approaches for association studies*, consists of techniques that aim to integrate additional sources of information into association studies such as functional information, gene expression and pathway data.

The rest of this report is organized as follows. Chapter 2 introduces the history of disease association studies and presents their different types. The standard statistical approaches used in association studies are also explained in this chapter. Chapter 3 discusses four of the most prominent challenges faced in identifying genetic disease associations. Chapter 4 describes several techniques that aim to improve different aspects of disease association methods. Chapter 5 discusses various techniques for integrating additional information into association studies. Finally, Chapter 6 summarizes the survey and provides concluding remarks about possible research directions.

Chapter 2

Disease Association Studies

The idea of disease association studies originated from the work of the eugenicist statistician Ronald Fisher on a mathematical model of genetic contributions to complex phenotypes [92]. Based on Mendelian rules, Fisher demonstrated that continuous traits (such as weight and height) and susceptibility to disease stem from the joint contribution of different genes [24]. This model has led to extensive research in genetic screening in order to identify genetic risk factors in predisposition to different types of diseases. Complex diseases are known to be affected by interacting genetic and environmental factors [99, 57]. Therefore, the discovery of genetic causality and mechanisms of actions through which disease develops requires a thorough analysis of genetic data and educated interpretation of disease association studies.

The interest in association studies is justified by the importance of their findings. As mentioned in Chapter 1, the ultimate goals of association studies are prevention of disease, early diagnosis, improved prognosis and the discovery of new therapeutic targets [38].

2.1 Disease Association Studies Data

Genetic markers data, i.e. specific parts of DNA such as genes with known properties and locations [44], is the type of data analyzed in genetic association studies. The analysis of such genetic marker data in disease association studies leads to the detection of associations between common DNA variants and the disease or trait being studied [56]. Different types of genetic markers such as Copy Number Variations (CNV) and SNP are used in disease association studies. As defined in Chapter 1, SNPs are single nucleotide mutations. Single nucleotides are the building blocks of the DNA sequences, where a nucleotide is represented as one of four values {A, T, C, G}. CNVs are modifications in the number of nucleotides at certain loci on the DNA, they can be either duplications or deletions of regions of nucleotides [23]. Both kinds of mutations (i.e. SNP and CNVs) are caused by several heritable and environmental factors. In the following sections, the term SNPs is used interchangeably with "markers" or "variants" since SNPs are the most common type of variants used in association studies.

The majority of association studies are performed using bi-allelic variants where each variant has two alleles, major and minor. A major allele appears more frequently in a population whereas a minor allele appears less frequently in the same population. The genetic

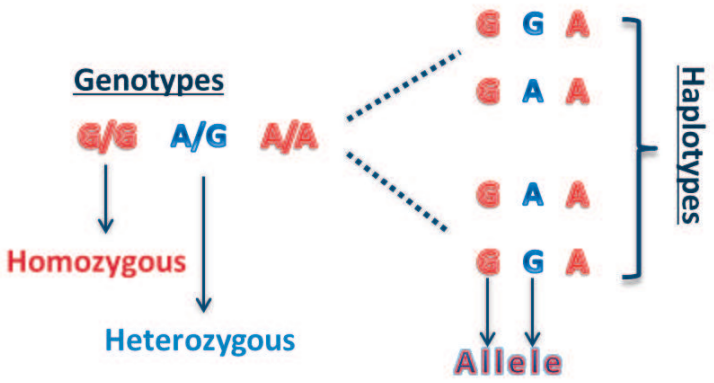

Figure 2.1: An example of a genotype sequence for three SNPs along with their possible haplotypes.

data can be represented as genotypes or haplotypes. Genotypes are the succession of pairs of alleles for each location (locus) on the DNA [44]. A genotype is called homozygous when its pair of alleles has identical values (e.g. AA or TT). However when the pair of alleles have different values (e.g. AC or GA), the genotype is called heterozygous. Haplotypes are sequences of statistically related single allele values at successive locations on the DNA [9, 104]. Figure 2.1 presents an example of a sequence of genotypes and its derived haplotypes. It also highlights the difference between homozygous and heterozygous genotypes and demonstrates alleles on haplotypes.

2.2 History of Disease Association Studies

Following Fisher's model of disease susceptibility, genetic screening started to gain popularity in 1950s where the genetic contribution in predisposition to disease was sought [92]. The tobacco industry funded one of the earliest research projects related to the predisposition to lung cancer in 1954 [92]. Since 1990s, disease susceptibility research evolved into candidate gene association studies where genetic variants on candidate genes were analyzed [74]. In candidate gene association studies, researchers identify possible target genes of a disease using knowledge about disease pathways and sequence these genes for analysis in association studies. The cost and scarcity of sequencing resources in the past were the reasons for bounding the search to candidate genes. Due to the incomplete understanding of disease mechanisms,the process of acquiring pathway information and then identifying target genes is difficult. Yet with the advent of recent biotechnologies, sequencing the whole genome became a reality in 2001 [90] and the first genome-wide association study (GWAS) was performed in 2005 to investigate the genetic risk factors of age-related macular degeneration [39]. Since then, GWAS is considered the de facto standard in association studies. Searching for disease associations on a genome-wide scale requires more computational resources than searching on a candidate gene scale . At the same time, the size and coverage of GWAS data increase the probability of encountering genetic variants in association with disease. The evolution of association studies is depicted in Figure 2.2, starting from the Fisher's model in

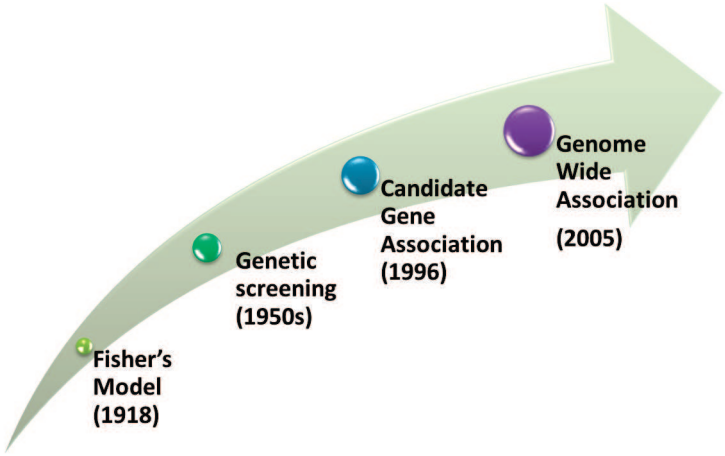

Figure 2.2: Association studies time line. The time line shows the beginning of association studies with Fisher's model , which led to genetic screening. Genetic screening then paved the way for candidate gene association studies and later to GWAS

1918 until the first genome-wide association study in 2005 [24, 39].

2.3 Categorization of Disease Association Studies

Disease association studies can be categorized based on two criteria: *risk factor type*; *scope of the studies*. Risk factors are either genetic or non-genetic. For genetic association studies, the scope can be either limited to the analysis of variants on candidate genes (candidate gene scope) or broadened to cover all variants on the genome (genome-wide scope). The scope of non-genetic association studies is limited to epigenetic association studies [73]. Epigenetics refers to the processes that modify gene functions without causing a mutation in the DNA sequence [44]. Review of epigenetics studies is beyond the scope of this report. Figure 2.3 shows a graphical representation of the categories of disease association studies. The discussion in the following sections is limited to genetic association studies on a genome-wide level, known as GWAS.

The data analyzed in association studies vary. Some studies are based on the comparison of SNP frequencies between a predefined number of individuals manifesting the disease (referred to as cases) or the trait being studied and a number of individuals who do not manifest the trait (referred to as controls). This type of studies is known as case/control studies. When the case/control data is collected from related individuals in a family, the association study is referred to as a family-based association study [14]. When case/control data is collected from unrelated individuals, the association study is known as population-based association study [14].

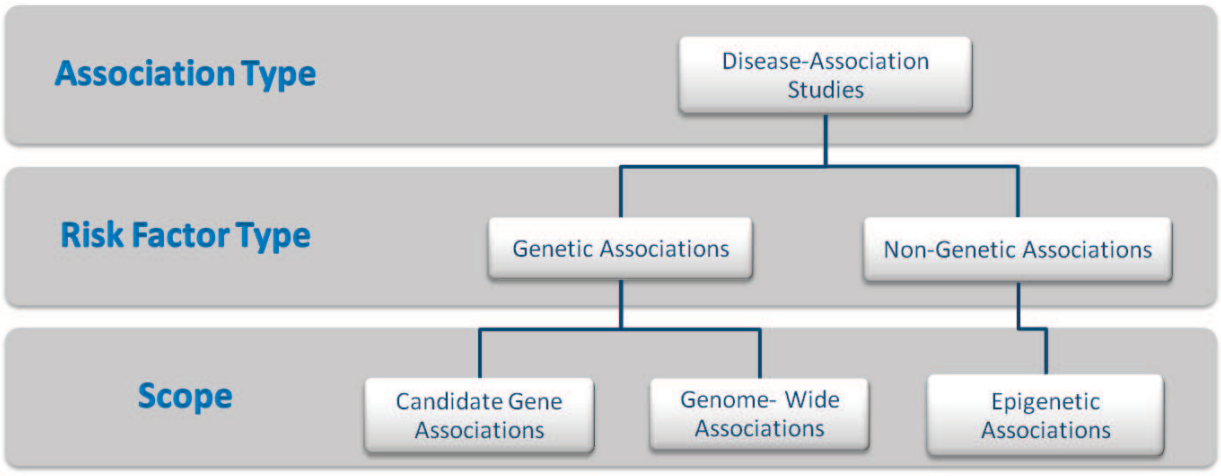

Figure 2.3: The categorization of association studies. Disease association studies comprise two categories: genetic and non-genetic (such as environmental). Genetic associations can be performed either on a candidate gene or on a genome-wide scope, whereas non-genetic association is performed by analyzing epigenetics data.

2.4 Standard Statistical Approaches for Disease Association Identification

Case/control association studies explore the difference in allelic (haplotype data) or genotypic (genotype data) frequencies of a SNP between cases and controls. For instance, if a certain allele has a higher frequency in cases than in controls, the difference in frequency suggests an association of the disease with this specific allele [48].

The standard approach, known as single-marker-based association testing, consists of inspecting the association of a certain disease or trait with a single SNP at a time [94]. Another approach, multi-marker-based association testing, is also used to seek patterns in allelic or genotypic frequencies by analyzing groups of SNPs together. This latter way of testing increases the ability of association studies to discover disease risk factors due to Linkage disequilibrium (LD) patterns that provide a mean to find associations with non-assayed variants [48, 2]

A variety of statistical tests is used to infer association between a disease and a SNP. The allelic or genotypic frequencies of a SNP are compared in order to reject or accept the null hypothesis, that is, the absence of association between the studied disease and the SNP being inspected. The frequencies are summarized in several ways using contingency tables. For allelic frequency comparison, the contingency table contains the counts of major and minor alleles in cases and controls [14]. For genotypic frequency comparison, the counts of major/minor homozygous and heterozygous genotypes of cases and controls are summarized

Genotype counts:			
	AA + AT	TT	Total
Cases	**190**	**60**	**250**
Controls	**200**	**25**	**225**
Total	**390**	**85**	**475**

Table 2.1: An example of an genotypic contingency table for a recessive penetrance disease (e.g. Daltonism) with major allele 'A' and minor allele 'T' in a dataset of 250 cases and 225 controls. Major homozygous or heterozygous genotypes do not manifest the disease whereas a minor homozygous genotype manifests it.

in the contingency table [14]. Disease penetrance is the probability of a disease to be manifested in individuals who carry certain alleles [89]. Additive, dominant and recessive models are examples of disease penetrance models used in association studies [66]. The presentation of allelic/genotypic information in a contingency table depends on the model of disease penetrance adopted in the study. For example, color blindness or Daltonism is a disease with a recessive penetrance model, which means that a person has to carry two minor alleles in order to manifest the disease. An example of a genotypic contingency table for Daltonism is presented in Table 2.1 where 'A' is the major allele and 'T' is the minor allele. Due to the recessive penetrance model of Daltonism, major homozygous and heterozygous genotypes do not manifest the disease unlike minor homozygous genotypes.

Using contingency tables, the independence of alleles/genotypes or their trend is tested using one of several statistical tests. If as a result of statistical test an allele or genotype is associated with cases or controls, the null hypothesis is rejected and an association between this allele and the disease being studied is inferred. Consequently, an individual is believed to have a risk factor for a disease if this individual carries the identified allele associated with disease manifestation.

Several statistical tests are used to infer disease risk factors. These tests are categorized into three main classes: *goodness-of-fit, likelihood-based* and *regression-based*. One of the most frequently used tests is the Chi-squared (χ^2), a *goodness-of-fit* test; it infers the presence of an association or its absence by assessing the dependence or independence of rows and columns in a contingency table [14, 88]. In order to use the χ^2 test, cases and controls are presented as categorical data but do not need to be ordered according to major/minor homozygous or heterozygous. This test cannot be used when the disease penetrance model follows a trend in the variations of allele/genotype frequencies (i.e. when the disease risk increases with the number of a certain allele in a given genotype such as 'T'). For example, in a certain genotype (GG or GT or TT), if the disease risk increases from none in the 'GG' genotype, since there is no 'T' allele, to high risk in the 'TT' since there is two 'T' alleles. Thus, the variation of allele frequencies is believed to have a certain trend that is associated with the disease. Hence, the disease penetrance model has an impact on the results of the χ^2 test [14].

Another commonly used test is the Cochran-Armitage (CA) trend test. It tests whether the variation in the number of a certain allele, between different ordered categories in the contingency table, follows an upward or downward trend in cases versus controls [1, 14]. Unlike the χ^2 test, the CA is applied to genotype data that is ordered according to one

of the disease penetrance models [14]. For instance, the above example of Daltonism has a recessive penetrance model, therefore the order of genotypes in a CA contingency table should reflect the number of recessive alleles in each genotype (i.e. the order of genotypes should be $GG \rightarrow GT \rightarrow TT$). CA is more robust than χ^2 when the dependencies between the analyzed genotypes is due to the characteristics of the sampling population [48]. In general, the χ^2 and the CA tests are used to test for associations when adopting a single-marker-based testing.

Associations can also be inferred using *likelihood ratio* tests. This type of testing was initially designed for association testing in family-based data [48], but it was extended to support case/control data [22]. The advantage of likelihood ratio tests is their applicability in multi-marker-based testing. In these tests, an association is based on the likelihood of having specific haplotypes (i.e. a group of SNPs with specific allele values) in cases versus controls [14, 61]. The likelihood of different haplotypes in cases and controls is usually estimated using maximum likelihood estimation.

Logistic regression, a *regression-based* method, is another approach for association testing yet a more computationally expensive one [102, 32, 98]. In this approach, SNPs (predictors) are used to predict the disease status (case or control) using several disease penetrance models [32]. Due to its complexity, this method is predominantly used with single-marker-based testing. Applying logistic regression as a multi-marker-based testing is made more feasible when the relationship between SNPs and the disease is summarized in a model and used in the application of the regression [102].

All the above mentioned statistical methods for association studies are summarized and compared in table 2.2. The comparison is based on the type of the statistical test used, the applicability of a method to single or to multiple markers, the type of data needed for the method and other special characteristics of each method.

Method	**Statistical Base**	**Marker Base**	**Data type**	**Characteristics**
χ^2	Goodness-of-fit	Single-marker	Allellic/ Genotypic	Widely used method
			Unordered/ Categorical	Not applicable if penetrance model has a trend
CA	Linear Regression	Single-marker	Allellic/ Genotypic	find trend in penetrance model
			Ordered/ Categorical	robust when allelic/genotypic frequencies changes over generations
Likelihood Ratio Test	Likelihood	Single/multi-marker	Allellic /Genotypic	uses maximum likelihood estimation
Logistic Regression Test	Logistic Regression	Single/multi-marker	Allellic /Genotypic	high complexity
				explores different penetrance models

Table 2.2: Comparison of common statistical tests for association studies

Chapter 3

Challenges in Disease Association Studies

Despite the success in discovering numerous markers associated with a wide range of diseases, associations studies face several challenges. Knowing and addressing these challenges pave the way for more accurate and novel discoveries of disease risk factors in the future. The challenges reviewed in this report are problems in current methods of discovering disease associations that are frequently discussed in the literature. They span a variety of topics, from the quality of the data to the details of obtaining and replicating the results.

3.1 Missing Heritability

The genetic variants identified by analyzing GWAS data can not currently explain a significant amount of the genetic variance present in complex diseases [80]. The total portion of disease phenotypic variations attributed to additive genetic factors is known as the heritability of this disease [47]. The "missing heritability"is the term used to refer to the unexplained portion of the genetic variance in diseases [57, 54]. In literature this is attributed to:

i- the lack of statistical power to detect associations of genetic markers explaining a smaller portion of disease risk than the currently reported ones [18].

ii- rare variants that are not being captured by sequencing technologies. Current biotechnologies enable the sequencing of millions of common variants on the DNA.

iii- the influence of gene-environment interactions, epistasis (i.e. the effect of several genes on another single gene) and epigenitcs [18, 54].

iv- other forms of genetic variants, such as copy number variations, and the lack of their analysis [54].

In general, there is no consensus on the reasons for the failure to capture a significant portion of the genetic variation responsible for individual differences in complex disease susceptibility [54].

3.2 Shortcomings of Traditional Statistical Testing

The use of some traditional parametric statistical approaches hinders the process of modeling the complexity of genetic architecture of diseases [60]. Such methods overlook the effect of different types of interactions such as gene-gene interactions on disease, due to their limited ability to uncover high-order non-linear interactions [60, 12]. The traditional approach of comparing the distributions of genetic markers in case/control studies uses hypothesis testing and *p-values* to make statistical inferences [56]. However, this approach does not express the statistical power of the study, the false discovery rate (the percentage of reported significant associations that are truly unassociated with the disease), or the number of likely true positives [56, 83, 5]. This inference method sets a single threshold on *p-values* in order to report an association as significant. The more stringent the threshold the more significant is the association. Given the large number of assayed variants, multiple use of statistical testing, that is, multiple comparisons in association studies leads to a large number of false positives in the reported results [5]. In addition, stringent *p-value* thresholds prevent discovering many causal variants with modest effect sizes [26]. Hence, translating current discoveries in association studies to clinical knowledge is considered a problematic task due to the number of false positives.

3.3 Inter-Population Applicability

In the past decade, the majority of association studies were performed on populations of European ancestry [57, 76]. The ubiquity of technological resources and research institutions and their concentration in some parts of the world as compared to others is one of the reasons for the focus on European population. Additionally, some population specific characteristic such as allele-frequency variations has also favored the focus on association studies in European populations [76]. The findings of studies in one population may not be applicable to other populations [57]. Thus, an association found in one population with a specific effect size does not necessarily translate into the same association in another population. The absence of association across populations can be attributed to environmental and epigentic factors. It can also be due to the difference in effect sizes and external influences effecting different populations. The lack of applicability across populations is yet to be explored and investigated. The importance of tackling this problem stems from the need to understand the effect of reported associations on diverse populations in order to provide improved universal health care [76].

3.4 Replication Issues

An essential step after performing an association study is the replication of results in order to avoid spurious associations [56, 38, 60]. Replication is considered a validation of the discovered associations and an evaluation in additional independent groups of cases and controls [56]. As mentioned in Section 3.2, multiple statistical tests performed in association studies lead to a significant number of false-positives, which are highly manifested with the massive amount of data in GWAS [69]. The importance of replication arises from the need

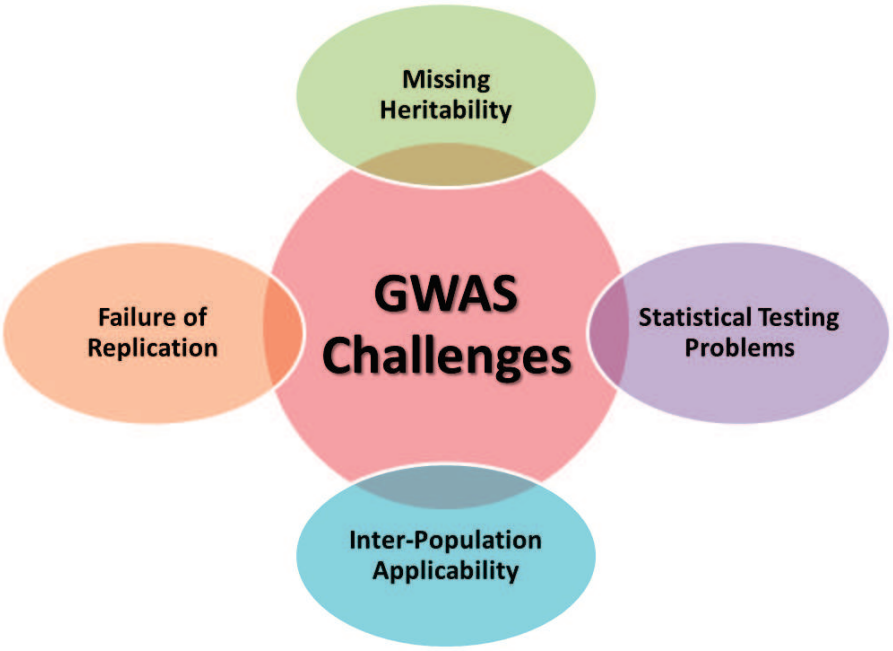

Figure 3.1: Summary of challenges in association studies (precisely GWAS). Missing heritability, statistical testing problems, inter-population applicability and failure of replication are the challenges discussed in this report.

to have reproducible associations [56]. However, replication efforts face some difficulties in reaching the desired outcome. Failure to replicate association results can be attributed to several factors. Poor design of an association study, such as the selection of non-population-based controls or errors in genotyping, may lead to the discovery of spurious associations. Heterogeneity of the studied population, i.e. difference, between the original and replication studies is another cause of replication failure [56, 38]. Heterogeneity in the studied populations can be caused by sampling biases or difference in allelic frequency within/among populations [38]. Heterogeneity may also stem from epidemiological reasons, where some genetic variations are affected by environmental factors [38]. Heterogeneity may affect the results of replication in two different ways. First, it may lead to spurious associations, which are reported associations but are not truly associated. Second, it may affect the estimated magnitude or extent of association as reported in the discovery and replication studies [38].

The aforementioned challenges are summarized in Figure 3.1. The following chapters lists emerging techniques that aim to overcome several aspects of the discussed shortcomings, in particular those of the analysis of GWAS, and to explain additional factors involved in disease. The discussion in the rest of the report focuses on GWA studies rather than on the general concept of association studies.

Chapter 4

Modifications to Current techniques in Disease Association Studies

In this chapter, we survey the literature pertaining to modifications to the techniques currently used in GWAS. Such modifications include either modifying some steps in the process of association discovery or increasing the genetic coverage and discarding of unwanted structure/ relatedness in the GWAS data.

The discussed techniques change the standard way of performing GWAS and are categorized into three groups:

- Methods increasing the coverage of GWAS data by either using imputation techniques (which will defined later in Section 4.1) or aggregating several datasets. In addition, this category includes guidelines and recommendations about the GWA study design that aim to minimize the bias that usually arise in GWA studies.
- Preprocessing methods for SNP selection designed to assist with choosing the most informative group of SNPs to be included in the GWAS.
- Methods that focus on significance evaluation of statistical testing for associations. In this category, we summarize the modifications to the standard parametric statistical approaches.
- Statistical and computational methods that aim to improve the association discovery technique itself.

4.1 GWAS Data

Due to the advancements in sequencing technologies and imputation techniques, massive amount of data are available for GWA studies [92, 50]. Imputation techniques are statistical or computational methods used to predict the values of non-assayed variants depending on the phenomenon of LD, which is the existence of dependencies between SNPs residing in close proximity. Consequently imputation techniques increase the genetic coverage by increasing the number of SNPs included in the association study [56]. At the same time, increasing the amount of available data results in higher statistical power of GWA studies. Selecting

the group of cases and controls to be included in a study is a critical process. This selection is influenced by many factors such as the population used and the way of ascertaining that a case really manifests the disease or that a control is really free of the disease. In some circumstances when the case/control selection introduces a bias towards cases or controls, increasing the amount of data provides a way to improve the accuracy of an association study [56]. Increasing both the number of individuals involved in the GWA study and the genetic coverage (number of assayed SNPs) are two important factors that increase the power of an association study [53]. By ensuring a wider genetic coverage, GWA studies are more likely to discover various causal variants with different effect sizes. However, an increased genetic coverage does not lead to the discovery of variants with smaller effects sizes since their discovery depends on the method used to find the disease associations.

Even though in GWA studies, the number of assayed SNPs may reach 100,000 to 1,000,000, this coverage is still incomplete since the number of genetic variants exceeds 10 million [50]. Marchini *et al.* [55] proposed to impute non-assayed data using a population genetic model and tested the influence of imputed data on the power of the association study. Li *et al.* [50] emphasized the importance of such imputation techniques since they enable the investigation of diseases associations with non-assayed genetic variants. Yet, the authors commented about imputation techniques by anticipating that in the era of genome-wide sequencing, imputation techniques will be mainly used for combining genetic data from different resources for analysis in the framework of meta-analysis of GWA studies [50].

In the following subsections, we discuss different GWA data related topics: quality control for GWAS Data, design recommendations and aggregation of GWA data from multiple sources.

4.1.1 Quality Control and Design Recommendations for GWAS

The design of a GWA study is a critical step in the process of discovering disease risk factors and avoiding spurious disease associations. It mainly affects the quality of the data in terms of robustness in the face of different sources of bias, for example population stratification or cases misclassification [69, 56]. Population stratification is the difference in allele frequencies between cases and controls caused by systematic ancestry differences and not as a result of a disease [72]. The discovery of spurious associations may be the result of population stratification. Therefore, special attention must be paid to the selection and classification of participants in a study (i.e. cases vs. controls). Controls should be drawn from the same population as the cases and the classification of the population into cases vs. controls should avoid selecting individuals whose disease status is not clearly defined in order to avoid spurious associations [69]. Moreover, choosing to sequence variants that belong to genomic areas with lower information redundancy leads to a larger genetic coverage, which in turn increases the power of the study [56].

Price *et al.* [72] proposed the use of principal component analysis (PCA) to correct for population stratification [72]. Adjusting the genotype data by removing possible ancestry correlation is required to avoid spurious associations caused by population stratification. The authors extracted the principal components of a matrix representation of genotype data in order to reveal the directions of variation in the genetic data, which are ancestry-based variations [72]. Then, all genotype data are adjusted by using the Eigen values of the

principal components, which represent the directions of variation in the data [72].

The quality of the assayed genetic variants used in GWAS also influences the power of the study [56, 50]. Sequencing and genotype calling techniques differ in the quality of their resulting sequences. The assayed genetic variants with poor quality genotypes (reflected by the genotype calling rates) should be filtered out of GWA data [8]. Additionally, variants demonstrating a significant deviation from the Hardy-Weinberg equilibrium, a property of variants that reflects how genotype frequencies defer from generation to another in a population, should also be excluded from the GWA data [56, 50]. The above procedures ensure a better quality of the assayed variants, which helps avoid spurious associations [56, 50].

4.1.2 Consolidation of GWAS Data

With the purpose of increasing the amount of genetic data available for GWA studies, a consortium of several institutions and foundations such as the National Human Genome Research Institute (NHGRI) called for consolidating the research efforts in GWA studies [53]. They provide the means to combine data from multiple resources and make the integrated dataset available, thus supporting more accurate analysis and more powerful GWAS [53]. The consortium also introduces some improvement to the ascertainment of cases/controls by adopting strict ascertainment schemes for each disease being studied [10]. In addition, it emphasizes the importance of using a single control group for the analysis of different diseases as shown in the GWAS of seven common diseases [10], unlike previous GWAS where different control groups were used for each disease [69, 53]. Another project, 1000 Genomes project, aims to sequence the whole genome of thousands of individuals and make this data available for researchers [91]. This project also facilitates the identification of rare genetic variants using improved imputation techniques to estimate the values of missing variants. Thus, the power of association studies is expected to increase with the availability of such high resolution genetic data [91].

4.2 Preprocessing of GWAS Data for Dimension Reduction

SNP selection is the process of choosing the most informative variants to be included in a GWA study. The informative variants, known as tag SNPs, are a group of variants that summarize the information conveyed by the whole set of SNPs. The selection process is performed by informatively decreasing the dimension of the SNP dataset to facilitate association testing. This selection does not defeat the purpose of increasing the coverage of the GWAS dataset since the group of selected SNPs convey the information hidden in the original set of SNPs but without redundancy. Selecting a subset of SNPs to be assayed and analyzed is considered a way of saving sequencing resources and sometimes providing a higher quality data [50, 19].

In the literature, different categories of approaches are proposed for SNP selection. Dimension reduction methods borrowed from the data mining literature as well as machine learning techniques are used for SNP selection [60]. SNP selection can be performed depending on the intrinsic properties of the variants or their prediction accuracy. Keating *et*

al. used the correlations among different groups of SNPs due to LD, an example of SNP properties, as a criteria to select a representative variant for each group of correlated SNPs [37]. Miclaus *et al.* used the deviation from the Hardy-Weinberg disequilibrium, another example of SNP properties, for the selection of the most infromative set of variants [59]. As for prediction accuracy, it can be used for SNP selection in two ways:

- Select the set of variants that can accurately predict the value of other SNPs in the data. The discussion in this section is limited to this type of SNP selection techniques.
- Select the set of variants that can best classify the data in the desired disease outcome.

SNP selection techniques that optimize prediction accuracy are used to decrease the number of variants to smaller subset with minimal redundancy in the information. Prediction techniques provide a mean for selecting the most informative SNPs according to how accurate a select group of SNPs can predict the values of non-selected group [19]. Chuang *et al.* [13] adopted a Genetic Algorithm (GA) with the predication accuracy of k-nearest neighbor (KNN) as a fitness function in order to select the most informative SNPs. Davidovich *et al.* [19] demonstrated that selecting SNPs for an association study using the prediction accuracy also increases the accuracy of the study. The main idea behind Davidovich *et al.*'s work is to compare two methods of SNP selection: the first uses SNP correlation as a selection criteria and the second uses prediction accuracy, in order to evaluate how the power of association is affected by either selection criteria [19].

4.3 Significance Evaluation

Reaching an accepted significance level in a genome-wide context is at the core of GWA studies [33]. The standard approach to ensure genome-wide significance is setting a *p-value* threshold of less than or equal to 10^{-7} [33]. Lower *p-value* threshold such as the stricter 10^{-10} [33] may result in more significant associations, yet it introduces a lot of stringency and may lead to missing associations with smaller effect sizes and contribute to the missing heritability [56]. Genome wide association studies require a large number of statistical testing, thus a correction of *p-value* thresholds is needed to avoid an increased number of false positives. The most widely used correction is the Bonferroni correction where the *p-value* threshold is divided by the number of performed tests [69, 5]. However, Pe'er *et al.* [70] raised the point that the use of the simple Bonferroni correction overlooks the fact that variants associated with the disease being studied are not all independent.

In addition to the fact that many reported associations do not meet strict *p-value* thresholds, using only *p-values* to report association significance does not provide all the information needed about a study [65, 6]. The statistical power of the association study as well as the confidence in the reported associations greatly depend on the amount of the data, which is not conveyed in the *p-value*. For instance, a *p-value* of an association between a disease and a certain variant may have the same value in two different studies where the sample sizes, the effect sizes, and the segregation into cases and controls are greatly different [65]. However, the evidence of association in both of these studies may not be equal [65]. Hence, when using *p-values* alone, the quantification of confidence in the reported associations becomes difficult [82]. The number of true positives and the false discovery rate both reflect

the significance of reported associations and are also not accounted for by the use of *p-values* [56, 83].

Evaluating the significance of disease associations using other measures than *p-value*s is the basis for several studies [83, 82]. The following subsections explains the use of the false discovery rate and a bayesian approach as an alternative to the use of *p-values*.

False Discovery Rate and q-Values

In the context of GWAS, False Discovery Rate (FDR) is the proportion of statistically significant discovered associations that are not truly associated with the studied trait [83, 69]. Reaching the smallest number of false positives (F) and the largest number of true positives (T) is important; FDR is calculated as the ratio of F to the total number of statistically significant variants ($F+T$) as in Equation 4.1 [83].

$$FDR = \frac{\#\ of\ False\ Positive\ Variants}{\#\ of\ Statistically\ Significant\ Variants} = \frac{F}{F+T} \tag{4.1}$$

The false positive rate is the probability of reporting the variants that are not associated with the trait being studied, as significantly associated with these traits [69]. As mentioned above, the use of false positives rate as the only significance measure overlooks several aspects of the assessment of significant associations. Tenesa *et al.* [85] used the false discovery rate as an assessment of their reported risk loci in association with colorectal cancer. Storey *et al.* [83] proposed the use of *q-values* instead of *p-values* as a measure of significance. The *q-value* is an FDR-based measure that provides more information than a *p-value* about significant associations and facilitates further analysis [83]. The *q-value* of a certain variant is defined as the proportion of false positives one gets when calling this variant statistically significant[83]. Whereas the *p-value* is the probability of a non-associated variant to be as or more extreme than the observed probability, the *q-value* is the expected proportion of false positives among all variants to be as or more extreme than the observed proportion [83].

Bayesian Approach

Bayesian approaches were developed for association studies in order to alleviate the problems of *p-values* using Bayes' rule[82]. In Bayesian approaches, the ratio between prior and posterior information about two compared data models is used to measure the strength of evidence favoring one of these models [56]. Bayesian measure of evidence can be directly compare associations within or across studies [82]. Such approaches allow and justify the incorporation biological information in the GWAS analysis using a quantitative way [82]. While using the Bayesian approach in GWAS, the evidence of association and its significance is measured using a quotient known as *Bayes Factor* (BF), as shown in Equation 4.2 [56, 82]. The larger the BF the greater the evidence of association. The BF is the ratio between the probability of the variants to be associated with the disease and the probability of the variants to be observed without the presence of the disease being studied [82]. The

BF is calculated according to the following equation:

$$BF = \frac{P(data \mid H_1)}{P(data \mid H_0)} \ . \tag{4.2}$$

The distribution of data under H_0 and H_1, in Equation 4.2, strongly depends on the selected genetic model (additive, dominant, recessive, etc ...). The use of the BF avoids the stringency of a single *p-value* threshold by assigning probabilities to each of the alternate hypotheses and provides a simpler way of comparing evidence of association across studies [56]. Craddock *et al.* [17] used the BF to assess the evidence of association of copy-number variations with eight different diseases. Additionally, Yasuno *et al.* [103] also used BF to evaluate the significance of the loci they reported to be associated with intracranial aneurysm.

Despite advantages of BF, a common and valid concern arises when deciding to use it as the significance measure. In order to be able to calculate the BF, the parameters of the prior-distribution, the plausible values of our belief about the possibility of a SNP to be associated with a disease, and the disease genetic models are required to be pre-defined. One way of determining these parameters and values is to use previous GWAS results, if there exist such studies [82].

4.4 Association Discovery Techniques

Single-marker-base association testing is the traditional statistical approach in GWA studies. The data is tested, one SNP at a time, and a *p-value* is calculated to decide between accepting the null hypothesis (absence of association) or the alternative hypothesis (presence of association). Despite the success in discovering thousands of disease-associated SNPs using the traditional approach, improved association approaches are required to overcome the challenges faced in GWAS such as the missing heritability. The majority of association testing approaches in the literature are designed to increase the confidence in the reported associations. Additionally, they aim to eliminate the effect of possible structures (such as individual relatedness) in the data causing inflation in association results or leading to spurious associations. Some of the approaches are based on single-marker methods and provide a way to increase the power of discovery of associations studies, for example, by decreasing the complexity and speeding up the testing process. Others are multi-marker based, where data from a group of SNPs is tested for association discovery. We categorize the different approaches into two main types: statistical and computational.

4.4.1 Statistical Modeling

The statistical approaches surveyed here use linear modeling and its variations such as the mixed linear model and generalized linear model to model the various factors affecting the disease status (also known as the disease outcome). Le *et al.* [45] use a generalized linear model to represent the relationship between the disease status of individuals (represented as a vector Y) and the genetic data (reperesented as a matrix X). The model assigns a coefficient for each SNP represented as column in X to designate its contribution in the disease status. To test the null hypothesis of no association, the authors use the a statistic

test, which tests the influence of two or more samples, here SNPs, on a certain outcome. A separate statistic is calculated for each SNP in the data and the overall statistic is the average of all SNP statistic values. The authors propose a multi-marker method where the effect of groups of SNPs is tested as opposed to testing the effect of single SNPs only. Accordingly, a *p-value* is calculated for each SNP and an overall *p-value* for each group of SNPs is then computed. A Bonferroni correction was applied to correct for multiple testing.

In order to account for different factors in disease risk, Kang *et al.* [35] proposed modeling the variance of the disease status as a linear mixed model (LMM). The contribution of risk factors such as the environment and epistasis are included in the variance model. In addition, they justified their proposal by emphasizing the importance of such models in estimating the unwanted relatedness between individuals. The variance of the disease phenotype $Var(Y)$ is modeled as a linear combination of the variance of the additive genetic effect and the environmental variance weighted by the probability of two variants at certain locus to be identical by decent. The implementation of such model is computationally expensive due to the multiple calculation of variances. Yet, the authors made some assumptions to speed-up the computation of variances in order to facilitate its use with very large genetic data and called their method Efficient Mixed Model Association Expedited (EMMAX) [35]. With these assumptions come approximations, which may cause a decrease in the statistical power when compared to exact calculations.

Zhou *et al.* [108] proposed a new efficient implementation of linear mixed models as a single-marker-based approach. Their implementation is faster than previous ones and performs exact calculations instead of approximations. In the literature, several approximation methods were adopted to make LMM applicable on a genome-wide scale. The method proposed by Zhou *et al.* [108] is called Genome-wide Efficient Mixed Model Association (GEMMA). Their method modifies an existing method known as Efficient Mixed Model association (EMMA) proposed by Kang *et al.* [36]. To speed up the testing process, GEMMA avoids the iterative eigen decomposition used in EMMA by substituting it by one eigen decomposition and simple vector-matrix multiplication. The vector of disease status is modeled as a linear combination of covariates, the marker genotypes weighted by their effect size and the relatedness information in the data. The association testing method requires the calculation of the maximum-likelihood estimates of the model parameters. Since calculating the exact estimates is computationally expensive, especially because of the size of the data, the method instead proceeds by optimizing the log-likelihood functions for each of the analyzed SNPs [108].

A multi-marker association test was proposed by Kwee *et al.* [43]. Despite the multivariate aspect of their proposed statistical test, its advantage relies in its abilities to produce test statistics with a lower amount of calculation than traditional multivariate approaches without sacrificing the power of the study. The genotype data is modelled using a non-parametric function known as a kernel and its relationship with the disease outcome is represented as a combination of the modelled genotype data and their estimated covariates. The parameters in the proposed semi-parametric model are estimated using a maximum-likelihood approach.

The above surveyed literature lacks a discussion about the scalability and the required computational time of the used methods. Although in some of the papers, the authors mention the applicability of their method on a genome-wide scale, the information about exact amount of data that the method can analyze and a complexity analysis is missing.

With the advancement in biotechnology and the increasing amount of available data, even the term genome-wide scale is supposed to be clearly defined in order to be able to assess the scalability and applicability of the proposed method. For instance, it is required to know how a proposed testing approach will perform with an increase in the amount of SNPs from 500k to 3 million (where both number are usually claimed to have a genome-wide coverage). Moreover, in some of the surveyed papers, the authors claim that their proposed method is faster/more efficient than a previously devised method. Yet, they overlook the need to clearly describe the extent of improvement their method provides in terms of computational resources (time and space). For instance, GEMMA [108] is claimed to be faster than EMMA [36] but the discussion is missing the information about how much faster it is.

4.4.2 Computational Approaches

Using parametric statistical approaches for association testing entails modeling the relationship between a phenotype and genetic markers under assumptions that are unrealistic most of the time such as assuming independence between the effects of genetic markers despite the important role of epistasis. In addition, parametric statistical approaches do not model high-order non-linear interactions that play an important role in disease risk factors [60]. Thus, computational approaches such as machine learning and data mining methods are proposed instead [60].

Moore *et al.* [60] discussed the use of decision trees and random forests for testing associations between a number of SNPs and a disease phenotype. In the data mining literature, these two classifiers are used for distinguishing between classes based on information carried by the attributes (i.e. features and characteristics; here SNPs). A decision tree is a predictive model used to classify an object (here an individual) into a predefined set of classes depending on the values of the object's attributes (SNPs) [75]. The genetic model underlying the disease risk is represented as a decision tree or a random forest where the values of associated SNPs are used to predict the disease phenotype. An example decision tree is shown in Figure 4.1. Each node in the tree, known as a predictor, represents a SNP for example SNP5 in Figure 4.1. The nodes at the very last level of the tree, known as leaves, represent the classification outcome, for example in Figure 4.1 the leaves are either *case* or *control*. While building the tree, going from one level to another, a process known as branching, is performed based on the possible values the node in the previous level can have. For example to go from level 1 to level 2 in Figure 4.1, the value of SNP1 can be either "a" or "b" and these values lead to node SNP10 or SNP5 respectively in the next level . The proposed method is a multi-marker-based association testing as the decision tree's classification outcome depends on a set of SNPs rather than on a single one.

The trees are built using the following steps:

- Select a portion of the data to be used as a training set for building the tree classifier;
- Randomly choose a subset of SNPs from the training set to be used as predictor of the disease class at each node of the tree;
- Decide on selecting the nodes to be added at a certain level of the tree and their branching out by choosing the highest information gain caused by each of the selected

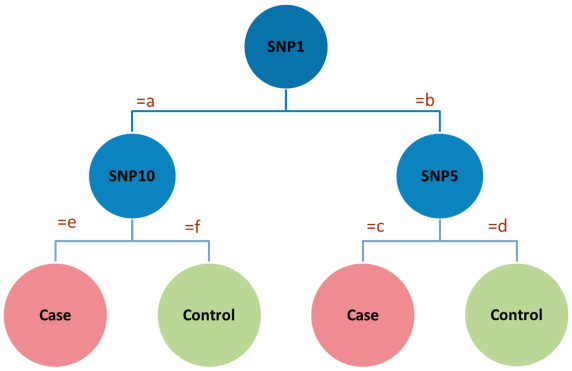

Figure 4.1: An example of a decision tree where each node represents a SNP and branching is based on the allele value at each node.The leaves denote the classes: Either *case* or *control*.

SNPs [41];

- Recursively repeat the process of SNP subset and best branching selection until the stopping criteria is satisfied then branch out the nodes in the previous level into leaf nodes that represent the disease phenotype (classification outcome).

One of the main advantages of decision trees is the fact that they are self-explanatory [75]. However decision tree classification strongly depends on the existence of features with high effect on the classification outcome. When the classification outcome is affected by many complex interactions, decision trees classification tends to get less accurate [75].

Random Forests are an ensemble of decision trees constructed by randomizing the branching at each node [75]. The generation of such an ensemble aims to increase the accuracy of classification by providing a voting scheme where the best classification accuracy is reported as the random forest accuracy [75]. Moore *et al.* generated the random forest after performing a bootstrap sampling of the SNP data [60]. This method has been applied to genetic data taken from patients suffering from asthma, rheumatoid arthritis, glioblastoma and age-related macular degeneration [60]. Yet, the authors of this paper overlooked discussing the complexity of the proposed methods and did not mention the amount of data used in the application of the methods.

Li *et al.* [49] proposed a framework for multi-marker-based association testing using Principal component Analysis (PCA). They modeled the effect of each SNP on the phenotype of interest using a linear combination of allele frequencies. The effects of a predefined number of SNPs are aggregated as a score for the gene containing the SNPs. Genotype scores are assigned weights proportional to the linkage disequilibrium (LD) between the corresponding pairs of SNPs. The information about LD is collected from the International HapMap project [49]. In order to further decrease the redundancy in the information, carried by the genotype scores and caused by LD, PCA is applied to the matrix representation of genotype scores with dimensions $n \times M$ where n is the number of individuals in data and M is the number of SNPs in the reference dataset. Principal components are independent linear combinations of the gene scores. Each principal component (PC) demonstrates a direction of high variation

in gene scores and is calculated according to Equation 4.3:

$$PC_l = \sum_{k=1}^{M} e_{l,k} S_k, \tag{4.3}$$

where e_l is the Eigen vector, $l = 1, ..., m$, m is the number of PCs and k is the index of the group of SNPs such that $1 \leq k \leq M$. Finally using a linear regression analysis the principal components are tested for association with the phenotype of interest (disease) [49]. The authors applied their proposed approach on a data of 550K SNPs and 2527 individuals, however, they did not discuss the scalability of their proposed approach.

Wang and Elston [97] also proposed a multilocus association test but instead of PCA it is based on weighted Fourrier transformation coefficients. They transform the numerical representation of sequence X of m SNPs for individual i by means of a discrete Fourier transformation (FT) , which maps the data into its frequency components [97]. The authors claim that the lowest-frequency FT component compresses the information and variation relayed by the whole SNP data, thus this FT component can be used to test for associations. To emphasize the importance of the lower frequency FT components, FT components are assigned weights inversely proportional to the order of the component in the FT decomposition. Consequently, the weight scores are used to test the association of the group of SNPs with the disease of interest. In comparison with standard testing methods, the FT based test shows an increased power when associations exist between the disease and the low-frequency components. The proposed method has been applied to simulated datasets that contain small amounts of SNPs (4, 10 and 22 SNPs), the fact that casts doubts on the scalability of the method on a genome-wide scale knowing the computational demands of FT.

The matrix decompositions needed for PCA and FT are computationally expensive and require a significant amount of time to be performed, especially when the dimension of the data is very large. The reviewed literature for computational methods does not discuss the scalability of the proposed approaches. The increasing amount of available data and the need to increase the coverage of the genetic information make scalability a pressing requirement. Hence, the category of computational association testing approaches is a potent resource of improvement for GWAS where the applicability of algorithms capable of analyzing ultra-large scale data in GWAS is yet to researched.

Chapter 5

Integrative Approaches for Disease Association Studies

Researchers commonly adopt GWAS as the first step in identifying culprit genes and revealing disease risk factors. As discussed so far in this report, current GWAS results are often not sufficient to identify the details of genetic and molecular mechanisms pertaining to disease due to the missing heritability. Therefore, several approaches were developed to provide further analysis of the reported results. Some approaches aim to combine the evidence of association with genetic functional and pathway data. Other approaches are designed to uncover hidden low-effect associations caused by epistasis [12], whereas another group of approaches tackle the discovery of disease risk factors by analyzing non-genetic factors at the molecular level [73]. All these approaches have the potential to increase confidence in reported disease associations and to provide additional information to support the discovery process.

The approaches discussed in this chapter do not change association testing methodologies. Instead, they take advantage of available biological and medical knowledge to provide additional sources of information for GWAS in order to assist in identifying disease risk factors. Some of these approaches are post-processing of GWAS results used to explain how the reported associations influence the disease and its symptoms and help the clinical interpretation of GWAS [12].

In this chapter, we classify the aforementioned approaches into four categories. The first aggregates GWAS results in form of meta-analysis. The second encompasses the approaches that couple GWAS with pathway analysis. The third is the integration of functional information about genes with GWAS. The last category comprises methods that combine gene expression data with GWAS. It is worth mentioning that some recent research efforts are directed toward integrating more than one source of information with GWAS in a single framework [96, 28, 79, 77]. For instance, Schadt discussed the use of molecular networks to bridge the gap between GWAS findings and clinical medicine [77]. These networks are build by integrating different types of data: DNA-variation, gene expression, protein interaction proteomic and metabolic data [77]. He described the importance of molecular networks in elucidating complex interactions between genes that affect disease etiology and susceptibility. Schadt presented a probabilistic causal networks as an alternative to interaction networks that are built using only gene expression [77]. The former type of networks is capable of

integrating different sources of data and providing means to infer the effect of one gene on another and its effect on diseases. Bayesian network construction approaches are used to build probabilistic causal networks. The advantage of the Bayesian approach is the ability to incorporate causal relationships using Bayes rule [77]. Following a similar integrative approach of multiple types of data, a framework, called Integrated Complex Traits Networks (iCTNet), was recently proposed by Wang *et al.* [96]. iCNet integrates information from different sources of data about: disease-associations, protein-protein interactions , and relationships between disease-tissue, tissue-gene and drug-gene [96].

5.1 Meta-Analyses

Meta-analysis of GWAS aims to aggregate the results of different GWAS, examining the same research hypothesis for a common disease while varying either the number of assayed genetic markers/individuals or the study designs. This type of analysis is designed to boost the accuracy of GWAS and to explore the variability in effect sizes and genetic models across studies [7, 42, 82, 100]. In other words, meta-analysis studies serve to pool evidence in order to increase the chance of finding more disease associations and to decrease the number of false positives [12]. The purpose of performing meta-analysis differs from one study to another. Some studies assess previously reported associations and investigate the reasons of failure to replicate results [52]. Others provide a way to increase the number of discovered causal genetic markers with a moderate to small effect on the disease of interest by combining different sources of genetic data previously analyzed in GWAS [42]. The combination of genetic data assists in discovering rare variants with large effect on the disease being studied and helps in finding the missing heritability [47]. To perform a GWAS meta-analysis, the information obtained from the different studies included in the analysis is summary association statistics about variants in each of the GWAS such as odds ratios or effect sizes [105]. Hence, in the rest of this section, the term studies is used to refer to the information obtained from the studies and included in the meta-analysis.

Cantor *et al.* [12] discussed meta-analysis studies as an approach to asses associations across GWAS without the need to re-analyze the original data. Their meta-analysis was designed to prioritize previously reported associations. They differentiated between different types of meta-analysis (as summarized in Fig 5.1):

i) Cumulative meta-analysis where previous GWAS are used as priors for a new GWAS.

ii) Combinatorial meta-analysis where previous GWAS are combined together to discover novel associations.

iii) Replicative meta-analyis where previous GWAS results are replicated using another data sample.

Combinatorial Meta-Analyses

Combinatorial meta-analysis of GWAS can be performed using different methods. The traditional approach is combining *p-values*, however this approach has several limitations

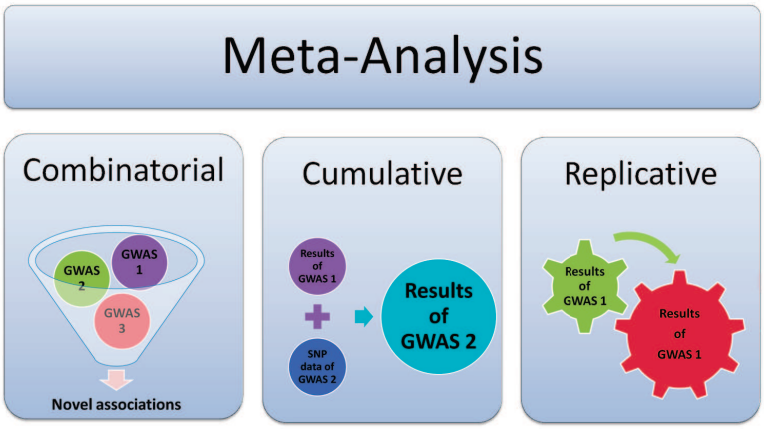

Figure 5.1: Meta-analysis of GWAS can be performed in different ways. The first consists of combining previous GWAS and using them to discover novel association. The second is the cumulative analysis where the results of previous GWAS are used as prior for another GWAS. The third is the replicative meta-analysis where results from one GWAS are replicated in another data sample.

including the inability to provide effect sizes, as discussed by Zeggini *et al.* [105]. The combined studies are weighted using either the sample size or the inverse of estimated variances since studies with low precision (high variance) should have a lower influence on the meta-analysis than the studies with higher precision (low variance) [105]. One way of combining and weighing summary results or datasets of GWAS is using a Bayesian meta-analysis approach [12, 82]. Bayesian hierarchical models are an example of Bayesian meta-analysis where external LD information is used as the priors for studies. These hierarchical models allows for the inclusion of various sources of information, such as results of previous GWAS and SNP functionality as priors [12, 82]. After performing the meta-analysis, SNPs are prioritized either according to meta-analysis *p-values* or Bayes factor (described in Section 4.3).

Cumulative Meta-Analyses

In the literature, there are various methods used for cumulative meta-analysis of GWAS. Using a gene-based approach, Liu *et al.* present a cumulative meta-analysis method based on a Monte-Carlo simulation [51]. The *p-values* of n SNPs on a gene are obtained from previous GWAS and are used to assess the association of the gene with the disease of interest. The *p-values* of SNPs from different studies are gathered and the genes in which the analyzed SNPs occurred are identified. Each gene is ranked according to its SNP association *p-values* converted to chi-square statistic. The proposed Monte-Carlo approach is used to account for the different linkage disequilibrium structures between SNPs on the analyzed genes and it avoids performing computationally expensive permutations to estimate the distribution of these structures under the null hypothesis (absence of association). The authors claim that their proposed method is faster than previous meta-analyses and can be performed without having the individual genotype information from the different studies.

Willer *et al.* proposed two different approaches to perform cumulative meta-analysis and

implemented them in a tool, called METAL [100]. The two approaches are used to combine evidence of association across different studies. The first approach converts the previously calculated *p-values* into z-scores (a score that reflects how many standard deviation a result is away from the mean of a normal distribution), where low values in z-scores reflect low risk associations and high values designate high risk associations [100]. The second approach uses the calculated standard error for each included GWAS to weigh the effect-size estimates of all reported associations in the study and these weighed values are used as z-scores. Associations are then compared according to their weighed effect sizes, and overall *p-values* are calculated using the z-scores [100]. The second approach can only be used when the effect sizes and/or standard errors are given in consistent units across studies whereas the first approach can be used even if the estimates are in different units (for a detailed explanation of when to use either of these approaches see reference [100]). The authors claim that, asymptotically, both approaches are equivalent when analyzing a trait with the same distribution across the data.

Replicative Meta-Analyses

Replicative meta-analysis of GWAS can been carried out using different approaches. Lohmueller *et al.* [52] performed a replicative meta-analysis on 301 published studies of common disease, which cover 25 reported associations in order to discover inconsistency when attempting to reproduce association studies. The analysis led the authors to believe that common variants with modest or low effect play an important role in common disease susceptibility due to their discovered associations. As in all meta-analysis studies, the associations to be included go through a thorough selection process, which depends on their significance and reproducibility in the original study. Association studies with *p-values* lower than a predefined threshold are compared to *p-values* of similar studies generated by simulation. A χ^2 goodness-of-fit test is used to test each association's homogeneity of results across all studies. In case of persistently heterogeneous associations between the analyzed studies, heterogeneity is attributed to either flaws in the methodology or population stratification.

Bertram *et al.* [7] performed a replicative/cumulative meta analysis of Alzheimer's disease GWAS. In their study, they re-assessed the significance of one of the most reported associations with Alzheimer's and discovered an additional 20 associations with modest effect on Alzheimer's. The meta-analysis was performed using the estimates of the odds ratios and 95% confidence intervals obtained from the Alzheimer's disease published studies.

The work surveyed above about meta-analysis focuses on combining the results of several GWAS in different ways in order to find novel associations or replicate previously reported ones. Yet, it overlooks the benefit of integrating other sources of information with the combined GWAS results such as prior knowledge about pathways, epistasis, protein interactions in order to prioritize the previously reported disease-associations. For example, the confidence in a disease association with a variant occurring on a gene that has been reported to be highly expressed in tissues affected by the disease of interest is greater than the confidence in an association that its only evidence is a *p-value* lower than a certain threshold. Therefore, such a prioritization approach provides the means to accumulate evidence about associations beyond the mere fact of being statistically significant. Associations supported by prior-knowledge will have a higher impact on the meta-analysis, and associations without this support will have less of an impact.

5.2 Pathway Analysis

Disease risk is commonly believed to be influenced by a plethora of interacting genes, cell pathways and complex molecular networks [94]. Biological cell pathways is a network of genes and molecules that interact in order to accomplish the activities required for a specific cell process such as apoptosis (i.e. programmed cell death). Therefore integrating pathways information with GWAS data has much potential for discovering novel disease risk factors and increasing the confidence in previously discovered ones as well as facilitating the clinical interpretation of disease etiology and mechanisms of action. Pathway collections are publicly available in several databases such Kyoto Encyclopedia of Genes and Genome (KEGG), Biocarta and Gene Ontology (GO) [94]. Recently, pathway based approaches are used for the analysis of GWAS in either of two ways:

i) Instead of focusing on the disease associations that are manifested by strong evidence (very low *p-values*), pathway-based approaches looks for disease risk factors among groups of SNPs (mapped to their closest genes) with small evidence of association. These approaches are designed to search for the associations between a disease of interest and a group of genes where variants (SNPs) belonging to these genes are in modest association with the disease. These groups of genes usually belong to the same pathway.

ii) Association results are post-processed by taking advantage of the publicly available pathway information. When the analysis reveals that a gene, where a variant associated with a disease occurs, is involved in a certain pathway, the clinical interpretation of the symptoms and mechanisms of action of this disease become more comprehensible. In addition the confidence in association results is increased when these results are supported by pathways information since the evidence of association is beyond just statistical significance.

Wang *et al.* proposed the use of a modified Gene Set Enrichment Analysis (GSEA) with GWA data [93]. GSEA is a method originally devised for the analysis of microarray gene expressions, it identifies groups of genes belonging to the same pathway that are differentially expressed instead of focusing on single genes with strongest evidence of differential expression ([84, 93]. Using pathway-based approach with GWA results facilitates the interpretation of these results for identifying complex-disease risk. The proposed approach calculate the statistic of association and *p-values* for all SNPs in a GWAS data and assigns them to their closest gene while disregarding SNPs 500kb away from any gene. The highest *p-value* of SNPs residing on a gene is assigned to this gene as its statistic value. The SNPs are assigned to gene sets extracted from the following pathway databases: BioCarta, KEGG and GO, where a gene set is a group of SNPs that share a common biochemical pathway or are coexpressed in previously published experiments [84]. All genes in each gene set are ranked according to their *p-value* obtained from their associated SNPs. Afterwards, an Enrichment Score (ES) is calculated to reflect the overrepresentation of a gene or a gene set at the top of the list of ranked *p-values* [93, 84]. Pathways related to the highly enriched gene sets are identified at the end of the analysis. These highly enriched gene sets constitutes the discovered disease risk factors that can not be discovered by only looking at SNPs with *p-values* below certain

threshold. This proposed approach was applied to Crohn Disease (CD) data by Wang *et al.* [95] where they identified and replicated the association of genes in different pathways with CD. The identified pathways and their corresponding cellular processes help provide interpretations to certain treatment problems CD patients are facing [95].

Zhong *et al.* [107] proposed a modification to the pathway-based approach presented by Wang *et al.* [93]. They use well-defined SNPs associated with gene expression levels in different type of tissues. These SNPs result from studies about the genetics of gene expressions (GGE) instead of studies about disease associations [107]. The SNPs are tested for associations with Type 2 Diabetes (T2D) in a GWAS. The associated SNPs are then mapped to genes from pathways available in KEGG database. Finally ES scores are calculated and highly enriched gene set are linked to their corresponding pathways [107].

Aiming to identify overlapping and over-represented pathways from different studies about T2D etiology, Elbers *et al.* proposed the use of a pathway-based approach [20]. They combine the information of SNPs associated with T2D discovered in different GWA studies into one dataset in order to find common associations missed by the original GWA studies. These SNPs are then mapped to LD blocks, which contain several genes. The authors defined the LD blocks as the groups of SNPs that have a correlation coefficient greater than a certain threshold [20]. SNPs that are successfully mapped to LD blocks with genes are included in the pathway-based study whereas SNPs mapped to blocks with no genes are excluded [20]. The authors then use a gene network tool called *Prioritizer* in order to build functional gene networks using a bayesian approach and based on prior knowledge obtained from publicly available databases. The built gene networks consist of genes functionally related to other genes that contain SNPs in association with T2D. Consequently, these networks offer the opportunity to assess the susceptibility to T2D affected by genes that would have been missed without searching for associations using these gene networks [20]. Finally, they compared five pathway-classification tools (KEGG, GATHER, DAVID, PANTHER, BioCarta) to check the over-representation of some of the susceptibility genes in the identified pathways.

Similarly, to asses and identify the polygenic basis of seven common diseases (bipolar disorder, coronary artery disease, Crohns disease , hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes), Torkamani *et al.* adopted a pathway-based approach to further analyze the Wellcome Trust Case Control Consortium (WTCCC) GWAS results [87]. The SNPs associated with the seven diseases are assigned to the their nearest gene. The highest SNP *p-value* on each gene is considered as the gene's weight. Then using the *MetaCore* software, the identified genes are tested for enrichment in Maps, Diseases, GO processes and GeneGO to determine the biological pathways and networks shared by the seven diseases [87].

Lebrec *et al.* presented a Bayesian linear model to integrate GWAS results with pathway and gene annotation data from GO [46]. The purpose of their approach is to use external sources of biological information in prioritizing candidate genes instead of using only GWAS results for prioritization. The authors use the same approach proposed by Wang *et al.* [93] to map associated SNPs to genes. They then build a matrix X of size $n \times p$ where n is the number of genes, p is the number of pathways and X_{ij} is set to one if gene i belongs to the pathway p, and to zero otherwise. Afterwards, they model the effect of gene i noted as μ_i on the disease of interest using their proposed approximate model [46].

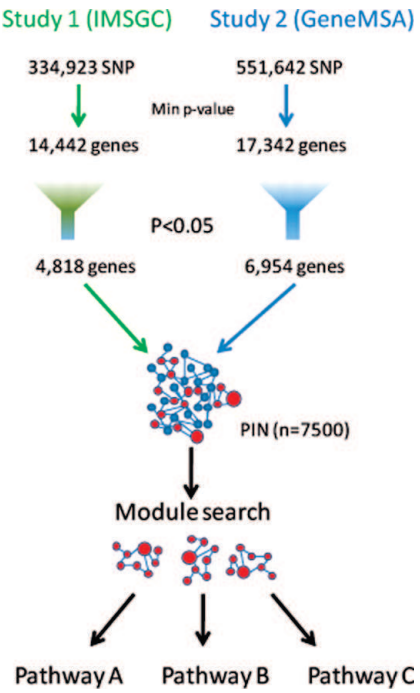

Figure 5.2: A diagram obtained from Baranzini *et al.* [3]. It summarizes the proposed framework for the integration of network-based pathway analysis with GWAS.

Baranzini et. al introduced a network-based pathway analysis for Multiple Sclerosis (MS) [3]. Their work is based on the hypothesis that genes containing SNPs with modest effect on disease can be identified in case they all belong to the same biological pathway or mechanism [3]. Their analysis adopts a protein interaction network (PIN) analysis with GWAS. Two GWA studies are performed on two different SNP data published by Baranzini et al. and the International Multiple Sclerosis Genetics Consortium [4, 16], where *p-value*s are calculated for SNPs demonstrating association with MS. Then SNPs are filtered according to a *p-value* threshold of less than 0.05, and the ones with *p-values* less than the threshold are included in further analysis. Disease-associated SNPs are mapped to genes, and their *p-value*s are used by the well-known open-source software called *Cytoscape* plugin *jActive modules* to extract the protein interaction network [3]. Afterwards, sub-networks of interacting gene products that are associated with MS are identified. Finally, significant gene modules are included in an enrichment test using KEGG pathways. The proposed framework is summarized in Figure 5.2, obtained from the publication of Baranzini *et al.* [3].

The methods presented above use a variety of pathway databases. Authors usually choose one database over another according to their familiarity with the database and sometimes opt to combine pathway information from different databases. However, pathway information differs from one database to another, for example Notch signaling pathway in KEGG is different from the one in Reactome as the entities involved in this pathway are not identical in both pathway databases. Therefore, the results of a study using pathway information from a certain database will not necessarily be validated if another database is used. Hence, there is a need to provide either a mapping scheme between different pathway databases or a standardized presentation of pathways in a centralized database in order to ensure

consistency in GWAS using pathway analysis.

5.3 Interaction and Function Information

In this section, the surveyed approaches focus on the integration of genes/markers interaction and functional information with GWAS. The purpose of this integration is to increase the power of GWAS by decreasing the burden of multiple association testing when providing means to group variants in sets according to previously reported knowledge regarding their interactions and functionality. Incorporating additional resources of information enables performing a more comprehensive analysis in GWAS that leads to increased evidence in reported associations beyond the sheer evidence of *p-values*. The literature reviewed in this section varies in the type of the integrated information, which is either: gene-gene interactions (epistasis) or functional relatedness of genes or SNPs functionality [67].

Epistasis is an effect in which some genes mask other genes' contribution to certain phenotypes. This gene-gene interaction has an important role in disease risk and poses a great challenge to GWAS. Analyzing the combinatorial effect of pairs of marker/genes on disease, to account for epistasis, requires a tremendous amount of computational power [11]. However, the literature contains a significant amount of information regarding epistasis. Therefore, integrating previously acquired knowledge about gene interactions offers an opportunity to group SNPs related through gene epistasis and increase the evidence about associations with modest/rare effects. Bush *et al.* [11] proposed a system of biological knowledge integration, called Biofilter, where multi-SNPs (a group SNPs) are modeled and weighted before being used in GWAS. It integrates various sources of data that offer information about gene interactions, including pathways, protein-protein interactions and gene expression. All the studied SNPs are mapped to their corresponding genes. Then data about interactions of gene pairs is collected from publicly available databases such as GO, KEGG and Reactome. Relationships between multiple SNPs are inferred from their genes interaction information using pair-wise combinations of SNP-related-genes and these relationships are called "multi-SNP models". Each multi-SNP model is weighed according to the proposed "implication index", which quantifies the knowledge base support [11]. The implication index is calculated by summing up the number of databases that provide evidence of gene×gene interactions. Finally, weighted multi-SNP models are analyzed in a GWAS to uncover the associations between the disease of interest and these models.

Due to gene-gene interactions and the fact that complex diseases are influenced by multiple genes, the analysis of risk factors necessitates using multilocus analysis approaches [67]. With the purpose of reducing the cost of multiple testing, Pan proposes a gene network-based weighing model for detecting disease association with multiple loci. Pan's approach assumes that interactions between gene products (protein protein interaction (PPI)) is a sign of interactions between genes themselves and if one of the interacting genes is associated with a disease, the other should also be associated. In Pan's study, genes included in the analysis are ranked according to the probability of being a disease-causing gene as reported by recent studies. Using PPIs information, networks of the most probable disease-causing genes are built. Following the aforementioned assumption and using a network of 23 interacting genes, a network of disease-causing genes (DCG) is formed. According to the assumption

that disease-causing genes tend to be closely interconnected in gene-networks, the network of DCG is extended by adding direct neighbors (first-order neighbors, 1ON) of these genes according to their PPIs. The idea of first order neighbors is then generalized to K-order neighbors (kON) where higher order neighbors are included to form larger groups of DCG. kON will be used to perform a multi-marker based GWAS, which will reduce the number of association tests.

Wu *et al.* proposed an approach based on a distance measure to group SNPs into SNP-sets [101]. The proximity of SNPs to genomic features, such as genes, provides a measure for forming the SNP-set, which is then used in GWA multilocus testing. Integrating gene proximity information with GWAS, as a criterion to form multi-SNP groups, helps in decreasing the effect of multiple testing and enables the detection of the influence of epistasis interactions on disease risk [101].

As a post-processing step, Sohns *et al.* suggest ranking genes of associated genetic markers using a combination of Gene Set Enrichement Analysis (GSEA) and Hierarchical Bayes Prioritization (HBP) [81]. The two combined approaches aim to increase the power of selecting genes associated with the disease of interest. GSEA is a method originally developed for gene expression analysis and its application on GWAS data has been explained in Section 5.2. In Sohns *et al.*'s work, HBP uses prior covariates of each variant in the dataset to rank these variants. Two types of regression models are applied to estimate the relationship between the prior covariates and the observed association statistic of variants, then the estimated regression coefficients are used to rank the variants. The first is a logistic model for variant-prior probabilities , which are calculated from variant distance to genes and gene set information. The second is a linear model applied to estimate association strength [81].

Despite their ability to increase the power of GWAS, the reviewed approaches are computationally expensive and require a significant amount of data preprocessing. In addition, there is a lack of a discussion about the scalability of these approaches especially when attempting to apply them on datasets with larger genetic coverage.

5.4 Gene Expression Analyses

Microarrays experiments are used to measure gene-expression levels. The starting point in gene expression analysis is finding the genes that are differentially expressed [27]. Next, the analysis usually proceeds to find patterns and clusters among differentially expressed group of genes. Gene-expression analysis has been an active research area for the past two decades utilizing myriads of statistical and machine learning methods [27]. With the goal of identifying novel genes pertaining to disease risk, gene-expression analysis is combined with GWA studies. In this section, we discuss the use of gene expression data as complementary information to SNP data or as a pre/post processing of GWA data .

Naukkarinen *et al.* explained the importance of integrating genome-wide expression data with SNPs data in order to find the missing heritability in disease [64]. The authors proposed a framework of three stages to integrate gene expression analysis with an association study on obesity and Body Mass Index (BMI) [64]. The first stage consists of performing a gene expression analysis to identify the genes differentially expressed in tissues related to BMI. In the second stage, genes correlated with obesity are identified using a stringent Pearson's

correlation threshold. From the first two stages, two groups of genes are formed: genes that are both differentially expressed and correlated with obesity are referred to as *reactive genes* and genes that are only correlated with obesity are considered *putative causative* genes. Both groups of genes are analyzed in the last stage where a GWA study is performed on the SNPs located within these genes. Then the GWAS results are replicated using the data of the GenMets study (a previously published GWAS about the genetics of metabolic syndrome) to increase the confidence in the reported associations. The integration approach proposed by Naukkarinen *et al.* forms here a preprocessing step of GWAS data that limits the analysis to SNPs within differentially expressed genes.

Gene-expression analysis can also be integrated with GWAS as a post-processing approach to increase the evidence in the reported associations or highlight novel ones. As a post-processing step, Hsu *et al.* proposed using gene-expression analysis to further analyze the results obtained by a GWA study for Osteoporosis [31]. They first performed a GWAS analysis using a regression model, then replicated the top associations in two independent cohorts. The replication phase generated two sets of results, the first is the group of significantly associated loci and the second is the group of suggestive association loci that were selected according to two *p-value*s thresholds. Reported suggestive association loci were prioritized according to their differential expression in Osteoporosis related tissues. The top prioritized genes in this set of suggestive association loci and the set of significantly associated loci are used in a GSEA to find if the selected genes are over-represented in certain biological/functional pathway then the gene-expression abundance of these genes is measured [31].

Murphy *et al.* performed an study to uncover the relationship between genome-wide SNP data and gene expression levels using linear models [63]. As a consequence, SNPs that have strong association with the abundance of expression of gene, which are referred to as expression SNPs (eSNPs), are mapped to their corresponding genes, and a gene set is formed [107]. The gene-expression levels were measured for blood lymphocytes. Using the associated gene set, an enrichment test was performed to compare the associated genes with previously curated catalog of GWAS of common diseases [63]. Similarly, Zeller *et al.* also undertook an association study between genome-wide SNP data and the Transcriptome of monocytes using ANOVA [106]. Afterwards, the associations between genes that are differentially expressed and cardiovascular disease were tested using linear regression models. Finally, SNPs associated with differentially expressed genes were mapped to their corresponding genes and their functional classification was performed using ontology analysis [106].

Aiming to identify novel genetic associations with small effects, Cusanovich *et al.* proposed a complimentary approach to combine GWAS results with gene expression analysis in Lymphoblastoid cell lines (LCLs) for individuals suffering from Asthma [18]. The authors identified differentially expressed (DE) genes for individuals that have high and low lymphocyte count (a physiological Quantitative Trait (QT) associated with asthma). In parallel, they explored the associations between genome-wide SNP data and lymphocyte count and mapped the associated SNPs to their corresponding genes. Consequently, two sets of genes resulting from the two previous analyses were integrated to identify the genes that are related to asthma and at the same time were missed by GWAS alone. Finally the intersection of these two sets of genes was labeled as a disease risk after performing an enrichment test and comparing it to previous asthma GWA studies [18]. The flow of analyses proposed by

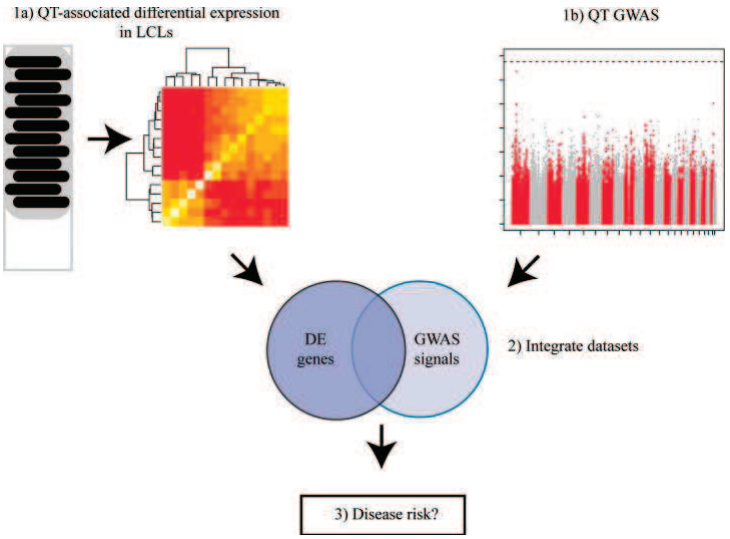

Figure 5.3: A diagram obtained from a publication by Cusanovich *et al.*[20]. It summarizes the flow of gene expression analysis (in 1a)) , GWAS (in 1b)) and the integration of both results (in 2)) to identify novel associations (in 3)) that would have been missed if GWAS alone was performed.

Cusanovich *et al.* [18] is summarized in Figure 5.3.

Gene expression data used in the work described above pertains to genes/tissues that have been previously reported to be related to the disease being studied. Nowadays myriads of sources of information about various complex diseases are available. Research oriented towards using gene expression levels measured within gene/tissue of diseases that share similar symptoms and/or pathways, provides an opportunity to reveal novel disease-associations that were difficult to find when examining gene expression data related only to one disease.

Chapter 6

Conclusion and Future Work

In this report, we discussed the importance of association studies as a tool to reveal disease risk factors. The discussion was focused on genome-wide association studies (GWAS) since the state of the art in genotyping technologies enables the acquisition of genome-wide SNP data. We presented the history of genetic screening and association studies, which started with Fisher's model of disease susceptibility and has been transformed into GWAS. In addition, we differentiated between the various types of association studies and their different design methods and data. The standard statistical approaches used in GWAS were also described and compared. Moreover, the challenges currently faced in GWAS were discussed along with the ways to address them.

The major goal of this report is to present the different approaches designed and implemented to improve GWA analysis and to increase its ability to discover disease risk factors. We discussed the research oriented towards modifying and improving the techniques used for the discovery of disease associations. Moreover, we explained several alternative approaches used for the significance evaluation of disease associations. We also presented different recommendations about the design of GWA studies and discussed several approaches that aim to increase the genetic coverage and discard unwanted structure pertaining to population stratification or unreported individual relatedness in GWAS data. Furthermore, we reviewed the literature that discusses integration of external information sources with GWAS. Incorporating additional information can be done either as preprocessing of GWA data or as post-processing. We discussed the four different integration approaches: meta-analyses, pathway analysis, gene expression analysis and interaction and function information.

Despite the wide range of proposed approaches to improve GWAS approaches, the discovery of disease risk factor of common complex disease is still far from complete. The acquired knowledge about current association loci explains a small portion of phenotypical and heritable disease variation. In addition, the translation of discovered association into clinical knowledge that answers questions about the etiology and chain of mechanisms of complex disease remains unclear. Therefore, the improvement of current GWAS approaches requires additional effort in taking advantage of the plethora of biological, physiological and medical information.

The continuing advancement in biotechnology point out the necessity to adopt scalable improvement approaches. Current GWA data spans hundreds of thousands to million of genetic markers. Yet, efforts to increase the genetic coverage will lead to even larger

GWA datasets, which also contain rarer genetic variants. Therefore, GWAS improvement approaches should be robust enough to detect rare disease associations and the search for associations is supposed to be able to scale well with the increasing size of genetic datasets.

Research in the biomedical field provides a wide spectrum of information sources. The number of publicly available databases, which contain sequences, gene expression, tissue information, protein-protein interactions and pathways, is increasing rapidly. Novel GWAS approaches should be designed to make use of the variety of available information in order to increase the ability of GWAS to discover novel associations and to boost the confidence in the discovered associations. The integration of different information sources into the analysis of GWAS is a key factor towards bridging the gap between GWAS discoveries and their application in the clinical world.

The majority of surveyed approaches were based on statistical analysis except for a few association testing approaches, which were based on machine learning and data mining techniques. Most of these statistical techniques propose the use of complex models to be able to reflect all different aspects of disease risk. However, the model complexity jeopardizes the scalability of the approach and increases the computational power needed for applying the model. New approaches should be designed to improve GWA studies while taking in consideration the scalability, coverage and integrability issues. Literature in computer science and machine learning contains extensive amount of research addressing pattern recognition, data mining and signal processing. Using this research provides a great opportunity to improve GWAS approaches. Future research in GWAS is supposed to take advantage of the state of the art of computational methods that are significantly scalable, efficient and effective.

A future direction in GWAS is the consolidation of more than one of the approaches discussed in this report. For instance, it is important to address the fact that the presented meta-analysis do not use pathway, functional and gene expression information to prioritize the previously reported GWAS results. Currently, we have access to a GWAS data of 500K SNPs for 1857 patients of Multiple Sclerosis (MS). In the catalog of published genome-wide association studies, there are numerous reported MS-SNP associations that can be harvested and prioritized in order to serve as prior knowledge for a new MS GWAS.

Our future work plan is implement a two-stage MS GWAS inspired by recent work of Kostem and Eskin [40]. They proposed to accelerate the process of association discovery using a two stage testing framework. First, they select a group of SNPs, referred to as proxy SNPs, by searching each LD region for the group of proxies that minimizes the number of required tests for GWAS. In each LD region, they calculate the association statistics for all proxies. If these statistics are lower than a certain threshold they test the remainder SNPs in this region, otherwise the region is discarded. Our proposed work-flow incorporates an integrative meta-analysis and a selective association testing approach. The integrative meta-analysis will collect the previously reported MS associations and prioritize them according to the relationship between their genomic locations and well known MS related pathways. These prioritized associations provide information about important genomic locations that are likely to hold more associations with MS. As a consequence, the 500k MS SNP data will be partitioned into blocks of potential areas of association according to the genomic locus of the previously reported prioritized associations. As for the selective association testing approach, the SNPs belonging to the blocks containing the highly prioritized previous associations will

be included in the set of SNPs to be tested for association with MS.

Our proposed two-stage study tackles the lack of integration of pathway analysis into meta-analysis data. In addition, it provides a testing approach that minimizes the computational and time complexity of GWAS by limiting the testing process to only potential areas of associations. Instead of exhaustively testing all SNPs in the 500K dataset, our proposed approach is to test a SNP only if this testing step is supported by previously reported association results and pathway information. Thus, our future work plan builds on previously published studies and takes advantage of publicly available knowledge about MS in order to increase the chance of discovering disease-causal variants/genes. As a consequence, our approach contributes to the solution of the missing heritability problem in MS and facilitates a clearer understanding of this disease's etiology and mechanism of actions.

Bibliography

[1] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[2] P. I. W. De Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nature genetics*, 37(11):1217–1223, 2005.

[3] S. E. Baranzini, N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B. M. J. Uitdehaag, L. Kappos, and C. H. Polman. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human molecular genetics*, 18(11):2078, 2009.

[4] Sergio E. Baranzini, Joanne Wang, Rachel A. Gibson, Nicholas Galwey, Yvonne Naegelin, Frederik Barkhof, Ernst-Wilhelm Radue, Raija LP Lindberg, Bernard MG Uitdehaag, and Michael R. Johnson. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human molecular genetics*, 18(4):767–778, 2009.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.Series B (Methodological)*, pages 289–300, 1995.

[6] James O. Berger and Thomas Sellke. Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–122, 1987.

[7] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi. Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. *Nature genetics*, 39(1):17–24, 2007.

[8] Hector Corrada Bravo and Rafael A. Irizarry. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, 66(3):665–674, 2010.

[9] Anthony J Brookes. *Single Nucleotide Polymorphism (SNP)*. John Wiley & Sons, Ltd, 2001.

[10] Paul R. Burton, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, and Nilesh J. Samani. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[11] W. S. Bush, S. M. Dudek, and M. D. Ritchie. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 368. NIH Public Access, 2009.

[12] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing gwas results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

[13] L. Y. Chuang, Y. J. Hou Jr, and C. H. Yang. A novel prediction method for tag snp selection using genetic algorithm based on knn. *World Academy of Science, Engineering and Technology*, 53, 2009.

[14] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[15] David E Comings. Polygenic disorders. In *eLS*. John Wiley & Sons, Ltd, 2001.

[16] International Multiple Sclerosis Genetics Consortium, D. A. Hafler, A. Compston, S. Sawcer, E. S. Lander, M. J. Daly, P. L. De Jager, P. I. de Bakker, S. B. Gabriel, D. B. Mirel, A. J. Ivinson, M. A. Pericak-Vance, S. G. Gregory, J. D. Rioux, J. L. McCauley, J. L. Haines, L. F. Barcellos, B. Cree, J. R. Oksenberg, and S. L. Hauser. Risk alleles for multiple sclerosis identified by a genomewide study. *The New England journal of medicine*, 357(9):851–862, Aug 30 2007.

[17] N. Craddock, M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, S. Robson, D. Vukcevic, C. Barnes, D. F. Conrad, and E. Giannoulatou. Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720, 2010.

[18] D. A. Cusanovich, C. Billstrand, X. Zhou, C. Chavarria, S. De Leon, K. Michelini, A. A. Pai, C. Ober, and Y. Gilad. The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Human molecular genetics*, 2012.

[19] O. Davidovich, G. Kimmel, E. Halperin, and R. Shamir. Increasing the power of association studies by imputation-based sparse tag snp selection. *Communications in Information & Systems*, 9(3):269–282, 2009.

[20] C. C. Elbers, K. R. van Eijk, L. Franke, F. Mulder, Y. T. van der Schouw, C. Wijmenga, and N. C. Onland-Moret. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genetic epidemiology*, 33(5):419–431, 2009.

[21] Carol Emslie and Kate Hunt. *Genetic Susceptibility*. John Wiley & Sons, Ltd, 2001.

[22] D. Fallin, A. Cohen, L. Essioux, I. Chumakov, M. Blumenfeld, D. Cohen, and N. J. Schork. Genetic analysis of case/control data using estimated haplotype frequencies: application to apoe locus variation and alzheimer's disease. *Genome research*, 11(1):143–151, 2001.

[23] L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

[24] Ronald Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Londonranscations of the Royal Society of Edinburgh*, 52:399–433, 1918.

[25] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L. Y. Ch'ang, W. Huang, B. Liu, and Y. Shen. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[26] G. Gibson. Hints of hidden heritability in gwas. *Nat Genet*, 42(7):558–560, 2010.

[27] J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[28] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[29] K. Hemminki, J. L. Bermejo, and A. Forsti. The balance between heritable and environmental aetiology of human disease. *Nature Reviews Genetics*, 7(12):958–965, 2006.

[30] K. Hemminki, A. Forsti, and J. L. Bermejo. The common disease-common variant hypothesis and familial risks. *PLoS one*, 3(6):e2504, 2008.

[31] Y. H. Hsu, M. C. Zillikens, S. G. Wilson, C. R. Farber, S. Demissie, N. Soranzo, E. N. Bianchi, E. Grundberg, L. Liang, and J. B. Richards. An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits. *PLoS genetics*, 6(6):e1000977, 2010.

[32] D. J. Hunter, P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, and A. Hutchinson. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, 39(7):870–874, 2007.

[33] J. P. A. Ioannidis, G. Thomas, and M. J. Daly. Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, 10(5):318–329, 2009.

[34] W. Ji, J. N. Foo, B. J. O'Roak, H. Zhao, M. G. Larson, D. B. Simon, and C. Newton-Cheh. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature genetics*, 40(5):592–599, 2008.

[35] H. M. Kang, J. H. Sul, N. A. Zaitlen, S. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[36] Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[37] Brendan J. Keating, Sam Tischfield, Sarah S. Murray, Tushar Bhangale, Thomas S. Price, Joseph T. Glessner, Luana Galver, Jeffrey C. Barrett, Struan FA Grant, and Deborah N. Farlow. Concept, design and implementation of a cardiovascular gene-centric 50 k snp array for large-scale genomic association studies. *PloS one*, 3(10):e3583, 2008.

[38] M. J. Khoury, L. Bertram, P. Boffetta, A. S. Butterworth, S. J. Chanock, S. M. Dolan, I. Fortier, M. Garcia-Closas, M. Gwinn, and J. Higgins. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. *American Journal of Epidemiology*, 170(3):269, 2009.

[39] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, and S. T. Mayne. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[40] Emrah Kostem and Eleazar Eskin. Efficiently identifying significant associations in genome-wide association studies. In *Research in Computational Molecular Biology*, pages 118–131. Springer, 2013.

[41] SB Kotsiantis, ID Zaharakis, and PE Pintelas. Supervised machine learning: A review of classification techniques. *Frontiers in Artificial Intelligence and Applications*, 160:3, 2007.

[42] CS Ku, EY Loy, Y. Pawitan, and KS Chia. The pursuit of genome-wide association studies: where are we now? *Journal of human genetics*, 55(4):195, 2010.

[43] L. C. Kwee, D. Liu, X. Lin, D. Ghosh, and M. P. Epstein. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, 2008.

[44] E. Lawrence. *Henderson's dictionary of biology/[edited by] Eleanor Lawrence*. Prentice Hall, 2005.

[45] B. Nguyen le, S. J. Diskin, M. Capasso, K. Wang, M. A. Diamond, J. Glessner, C. Kim, E. F. Attiyeh, Y. P. Mosse, K. Cole, A. Iolascon, M. Devoto, H. Hakonarson, H. K. Li, and J. M. Maris. Phenotype restricted genome-wide association study using a gene-centric approach identifies three low-risk neuroblastoma susceptibility loci. *PLoS genetics*, 7(3):e1002026, Mar 2011.

[46] J. Lebrec, T. Huizinga, R. Toes, J. Houwing-Duistermaat, and H. van Houwelingen. Integration of gene ontology pathways with north american rheumatoid arthritis consortium genome-wide association data via linear modeling. In *BMC proceedings*, volume 3, page S94. BioMed Central Ltd, 2009.

[47] S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 2011.

[48] C. M. Lewis and J. Knight. Introduction to genetic association studies. *Cold Spring Harbor Protocols*, 2012(3):pdb. top068163, 2012.

[49] M. Li, K. Wang, S. F. A. Grant, H. Hakonarson, and C. Li. Atom: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*, 25(4):497–503, 2009.

[50] Y. Li, C. Willer, S. Sanna, and G. Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387, 2009.

[51] J. Z. Liu, A. F. Mcrae, D. R. Nyholt, S. E. Medland, N. R. Wray, K. M. Brown, N. K. Hayward, G. W. Montgomery, P. M. Visscher, and N. G. Martin. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145, 2010.

[52] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature genetics*, 33(2):177–182, 2003.

[53] T. A. Manolio. Collaborative genome-wide association studies of diverse diseases: programs of the nhgri's office of population genomics. *Pharmacogenomics*, 10(2):235–241, 2009.

[54] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, and A. Chakravarti. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[55] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[56] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, JP Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[57] M. I. McCarthy and J. N. Hirschhorn. Genome-wide association studies: potential next steps on a genetic journey. *Human molecular genetics*, 17(R2):R156, 2008.

[58] JP McElroy and JR Oksenberg. Multiple sclerosis genetics. *Advances in multiple Sclerosis and Experimental Demyelinating Diseases*, pages 45–72, 2008.

[59] Kelci Miclaus, Russ Wolfinger, and Wendy Czika. Snp selection and multidimensional scaling to quantify population structure. *Genetic epidemiology*, 33(6):488–496, 2009.

[60] J. H. Moore, F. W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

[61] AP Morris, JC Whittaker, and RN Curnow. A likelihood ratio test for detecting patterns of disease-marker association. *Annals of Human Genetics*, 61(4):335–350, 1997.

[62] A. G. Motulsky. Genetics of complex diseases. *Journal of Zhejiang University-Science B*, 7(2):167–168, 2006.

[63] A. Murphy, J. H. Chu, M. Xu, V. J. Carey, R. Lazarus, A. Liu, S. J. Szefler, R. Strunk, K. DeMuth, and M. Castro. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood cd4 lymphocytes. *Human molecular genetics*, 19(23):4745, 2010.

[64] J. Naukkarinen, I. Surakka, K. H. Pietilainen, A. Rissanen, V. Salomaa, S. Ripatti, H. Yki-Jarvinen, C. M. Van Duijn, H. E. Wichmann, and J. Kaprio. Use of genome-wide expression data to mine the "gray zone" of gwa studies leads to novel candidate obesity genes. *PLoS genetics*, 6(6):e1000976, 2010.

[65] Tuan V. Nguyen. Interpretation of randomized controlled trials of fracture prevention. *IBMS BoneKEy*, 6(8):279–294, 2009.

[66] J. Ohashi, S. Yamamoto, N. Tsuchiya, Y. Hatta, T. Komata, M. Matsushita, and K. Tokunaga. Comparison of statistical power between 2x2 allele frequency and allele positivity tables in case/control studies of complex disease genes. *Annals of Human Genetics*, 65(2):197–206, 2003.

[67] W. Pan. Network-based model weighting to detect multiple loci influencing complex diseases. *Human genetics*, 124(3):225–234, 2008.

[68] J. H. Park, M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, S. J. Chanock, J. F. Fraumeni Jr, and N. Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108(44):18026–18031, 2011.

[69] T. A. Pearson and T. A. Manolio. How to interpret a genome-wide association study. *JAMA: the journal of the American Medical Association*, 299(11):1335, 2008.

[70] I. Pe'er, R. Yelensky, D. Altshuler, and M. J. Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology*, 32(4):381–385, 2008.

[71] B. A. J. Ponder, G. I. Evan, K. H. Vousden, J. Taipale, P. A. Beachy, P. Blume-jensen, T. Hunter, J. Hoeijmakers, L. A. Liotta, and E. C. Kohn. Cancer genetics. *Cancer*, 411(6835), 2001.

[72] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[73] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541, 2011.

[74] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science-AAAS-Weekly Paper Edition*, 273(5281):1516–1517, 1996.

[75] Lior Rokach. *Data mining with decision trees: theory and applications*, volume 69. World scientific, 2007.

[76] N. A. Rosenberg, L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic, and M. Boehnke. Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, 11(5):356–366, 2010.

[77] E. E. Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, 2009.

[78] N. J. Schork. Genetics of complex disease. *American journal of respiratory and critical care medicine*, 156(4):S103–S109, 1997.

[79] Solveig K. Sieberts and Eric E. Schadt. Moving toward a system genetics view of disease. *Mammalian Genome*, 18(6-7):389–401, 2007.

[80] H. C. So, A. H. S. Gui, S. S. Cherny, and P. C. Sham. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genetic epidemiology*, 2011.

[81] M. Sohns, A. Rosenberger, and H. Bickeboller. Integration of a priori gene set information into genome-wide association studies. In *BMC proceedings*, volume 3, page S95. BioMed Central Ltd, 2009.

[82] M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

[83] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440, 2003.

[84] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, and Eric S. Lander. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[85] A. Tenesa, S. M. Farrington, J. G. D. Prendergast, M. E. Porteous, M. Walker, N. Haq, R. A. Barnetson, E. Theodoratou, R. Cetnarskyj, and N. Cartwright. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature genetics*, 40(5):631–637, 2008.

[86] G. Thomas, K. B. Jacobs, P. Kraft, M. Yeager, S. Wacholder, D. G. Cox, S. E. Hankinson, A. Hutchinson, Z. Wang, and K. Yu. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11. 2 and 14q24. 1 (rad51l1). *Nature genetics*, 41(5):579–584, 2009.

[87] A. Torkamani, E. J. Topol, and N. J. Schork. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–272, 2008.

[88] N. Turner. Chi-squared test. *J Clin Nurs*, 9:93, 2000.

[89] D. Urbach, M. Lupien, M. R. Karagas, and J. H. Moore. Cancer heterogeneity: origins and implications for genetic association studies. *Trends in Genetics*, 2012.

[90] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, and R. A. Holt. The sequence of the human genome. *Science Signalling*, 291(5507):1304, 2001.

[91] M. Via, C. Gignoux, and E. G. Burchard. The 1000 genomes project: new opportunities for research and social challenges. *Genome Med*, 2(3), 2010.

[92] H. M. Wallace. *Genetic screening for susceptibility to disease*. Wiley Online Library, 2009.

[93] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283, 2007.

[94] K. Wang, M. Li, and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, 2010.

[95] K. Wang, H. Zhang, S. Kugathasan, V. Annese, J. P. Bradfield, R. K. Russell, P. Sleiman, M. Imielinski, J. Glessner, and C. Hou. Diverse genome-wide association studies associate the il12/il23 pathway with crohn disease. *The American Journal of Human Genetics*, 84(3):399–405, 2009.

[96] Lili Wang, Pouya Khankhanian, Sergio Baranzini, and Parvin Mousavi. ictnet: A cytoscape plugin to produce and analyze integrative complex traits networks. *BMC bioinformatics*, 12(1):380, 2011.

[97] T. Wang and R. C. Elston. Improved power by use of a weighted score test for linkage disequilibrium mapping. *The American Journal of Human Genetics*, 80(2):353–360, 2007.

[98] S. Wellek and A. Ziegler. Cochran-armitage test versus logistic regression in the analysis of genetic association studies. *Human heredity*, 73(1):14–17, 2012.

[99] Andrew OM Wilkie. *Polygenic Inheritance and Genetic Susceptibility Screening*. John Wiley & Sons, Ltd, 2001.

[100] C. J. Willer, Y. Li, and G. R. Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[101] M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

[102] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[103] K. Yasuno, K. Bilguvar, P. Bijlenga, S. K. Low, B. Krischek, G. Auburger, M. Simon, D. Krex, Z. Arlier, and N. Nayak. Genome-wide association study of intracranial aneurysm identifies three new risk loci. *Nature genetics*, 42(5):420–425, 2010.

[104] N. A. Zaitlen, H. M. Kang, M. L. Feolo, S. T. Sherry, E. Halperin, and E. Eskin. Inference and analysis of haplotypes from combined genotyping studies deposited in dbsnp. *Genome research*, 15(11):1594–1600, 2005.

[105] Eleftheria Zeggini and John PA Ioannidis. Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10(2):191–201, 2009.

[106] T. Zeller, P. Wild, S. Szymczak, M. Rotival, A. Schillert, R. Castagne, S. Maouche, M. Germain, K. Lackner, and H. Rossmann. Genetics and beyond: the transcriptome of human monocytes and disease susceptibility. *PLoS One*, 5(5):e10693, 2010.

[107] H. Zhong, X. Yang, L. M. Kaplan, C. Molony, and E. E. Schadt. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *The American Journal of Human Genetics*, 86(4):581–591, 2010.

[108] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.