# ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

## ΣΧΟΛΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

### ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ


ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

# ΜΟΝΤΕΛΑ ΑΝΑΠΤΥΞΗΣ ΓΕΝΕΤΙΚΩΝ ΒΙΟΔΕΙΚΤΩΝ ΜΕ ΧΡΗΣΗ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ


**Κωνσταντίνα Αποστόλου**


Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

**Πειραιάς, Σεπτέμβριος 2022**

# ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

## ΣΧΟΛΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

### ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ


ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

# ΜΟΝΤΕΛΑ ΑΝΑΠΤΥΞΗΣ ΓΕΝΕΤΙΚΩΝ ΒΙΟΔΕΙΚΤΩΝ ΜΕ ΧΡΗΣΗ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ


**Κωνσταντίνα Αποστόλου**


Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

**Πειραιάς, Σεπτέμβριος 2022**

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη Συνέλευση του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή της σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Μπερσίμης Σωτήριος, Αναπληρωτής Καθηγητής (Επιβλέπων)

- Μπάγκος Παντελής, Καθηγητής

- Πολίτης Κωνσταντίνος, Αναπληρωτής Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

# UNIVERSITY OF PIRAEUS

## SCHOOL OF FINANCE AND STATISTICS

## DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE



## MASTER PROGRAM IN APPLIED STATISTICS


# GENETIC BIO-INDICES DEVELOPMENT USING DATA ANALYTICS


**By**

**Konstantina Apostolou**



Master Thesis submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Applied Statistics

**Piraeus, Greece, September 2022**

**Έχω διαβάσει και κατανοήσει τους κανόνες του ΠΜΣ που περιέχονται στον Οδηγό Συγγραφής ΔΕ και ιδιαίτερα όσα συνιστούν λογοκλοπή. Δηλώνω ότι η παρούσα διπλωματική εργασία αποτελεί προϊόν αποκλειστικά δικής μου προσπάθειας, υπό την καθοδήγηση του επιβλέποντος καθηγητή, ενώ για όλες τις πηγές που χρησιμοποιήθηκαν περιλαμβάνονται οι αντίστοιχες αναφορές.**

# Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλλαν στην εκπόνησή της.

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή μου, τον αναπληρωτή καθηγητή Σωτήριο Μπερσίμη, για την εμπιστοσύνη που μου έδειξε εξ' αρχής αναθέτοντάς μου το συγκεκριμένο θέμα και με την επιστημονική του καθοδήγηση, τις υποδείξεις του, την επιμονή του, τη συνεχή υποστήριξή του και το αμείωτο ενδιαφέρον, που έδειξε από την αρχή έως το τέλος, με βοήθησε να το φέρω εις πέρας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την στήριξη και τη βοήθεια που μου έχει προσφέρει καθ' όλη τη διάρκεια των σπουδών μου.

# Μοντέλα Ανάπτυξης Γενετικών Βιοδεικτών με χρήση Αναλυτικής των Δεδομένων

**Σημαντικοί όροι:** Γονιδιωματική , Γονιδιωματική Ιατρική, Εξατομικευμένη Ιατρική, Μελέτες συσχέτισης ευρέος γονιδιώματος, Πολυγονιδιακή Βαθμολογία Κινδύνου

## Περίληψη

Από την πρώτη αλληλούχιση του ανθρώπινου γονιδιώματος το 2003, η ανθρώπινη γενετική έχει υποστεί μια πραγματική επανάσταση λόγω της μεγάλης διαθεσιμότητας συνόλων δεδομένων με γενετικές πληροφορίες και της εισαγωγή της βιοπληροφορικής σε αυτόν τον τομέα. Οι νέες γενετικές ανακαλύψεις επιτρέπουν την κατανόηση του τρόπο με τον οποίο τα γονίδια αλληλεπιδρούν με διάφορους παράγοντες του τρόπου ζωής ή του περιβάλλοντος, σε μια πορεία προς ένα πιο αποτελεσματικό κλινικό έλεγχο. Οι μελέτες συσχέτισης σε όλο το γονιδίωμα (GWAS) αποσκοπούν στον εντοπισμό συσχετίσεων γονότυπων με φαινότυπους, ελέγχοντας για διαφορές στη συχνότητα αλληλομόρφων γενετικών παραλλαγών μεταξύ ατόμων. Ο τυπικός στόχος των GWAS είναι ο εντοπισμός τόπων που φιλοξενούν αιτιολογικές παραλλαγές και η χρήση γενετικών παραγόντων κινδύνου για να γίνουν προβλέψεις σχετικά με το ποιος διατρέχει κίνδυνο και να αναπτυχθούν νέες στρατηγικές πρόληψης και θεραπείας. Μια απλή και διαισθητική προσέγγιση για τη μετατροπή των γενετικών δεδομένων σε ένα προγνωστικό μέτρο επιδεκτικότητας σε ασθένειες είναι η συγκέντρωση των επιδράσεων αυτών των τόπων σε ένα ενιαίο μέτρο, το πολυγενετικό σκορ κινδύνου (PRS). Παρουσιάζονται επίσης κάποιες έρευνες που έχουν πραγματοποιηθεί για την εύρεση βιοδεικτών.

# Genetic Bio-Indices Development using Data Analytics

**Keywords:** Genomics, Genomic Medicine, Precision Medicine, Genome-Wide Association Studies, Polygenic Risk Score

## Abstract

Since the first sequencing of the human genome in 2003, human genetics has undergone a real revolution due to the wide availability of datasets with genetic information and the introduction of bioinformatics in this field. New genetic discoveries are enabling an understanding of how genes interact with various lifestyle or environmental factors, in a move towards more effective clinical control. Genome-wide association studies (GWAS) aim to identify genotype-phenotype associations by testing for differences in the frequency of allelic genetic variants between individuals. The typical goal of GWAS is to identify loci harboring causal variants and use genetic risk factors to make predictions about who is at risk and to develop new prevention and treatment strategies. A simple and intuitive approach to converting genetic data into a predictive measure of disease susceptibility is to aggregate the effects of these loci into a single measure, the polygenic risk score (PRS). Some research that has been carried out to find biomarkers is also presented.

# Contents

# CHAPTER 1

# HUMAN GENOMICS

## 1.1 Introduction

Since the human genome was first sequenced in 2003, human genetics has undergone a veritable revolution. The growth in computing power, the explosion in the availability of datasets with genetic information, and the infusion of bioinformatics into this field have changed our views about how we think about disease and behavior. The relevance of genetics has also penetrated disciplines far beyond its original homes in biology, epidemiology, and the medical sciences to gain relevance in new areas across the biomedical, social, and psychological sciences. As of 2019, around 4,000 genetic discoveries have been published, linking the genetic basis of thousands of traits ranging from height, type 2 diabetes, and body mass index (BMI) to coffee consumption, depression, neuroticism, and even the age when you have your first child. Researchers in the biomedical sciences can now estimate the genetic component of many major diseases such as type 2 diabetes, breast cancer, or cardiovascular lar disease. More importantly, new genetic discoveries allow them to understand how genes interact with different lifestyle or environmental factors in a move toward more effective clinical screening and interventions. The goal of statistical genetics is to explain population variation or, in other words, to ask why humans differ in their health out-comes, behavior, or appearance. [58]

## 1.2 Definitions

DNA (deoxyribonucleic acid) is the molecule that makes up the genetic material con-tained within our bodies' cells. As Figure 1.1 illustrates, two long DNA chains, composed of simpler molecular units (called nucleotides), coil around each other to form a double helix. DNA contains the genetic instructions that tell each cell which proteins to make. A genome is the complete set of genetic material of an organism or, in other words, the entire set of DNA contained within the nuclei of somatic cells in the human body. The size of each organisms' genome is the total number of bases in one representative copy of its nuclear DNA. As the figure shows, a gene is a section of DNA found on a chromosome that consists of a particular sequence of nucleotides at a given position on a given chromosome that in turn codes for a

specific protein (or an RNA molecule). A gene is a segment of DNA that tells the cell how to make a certain protein. Humans are estimated to have 20,000 to 25,000 genes. [58]

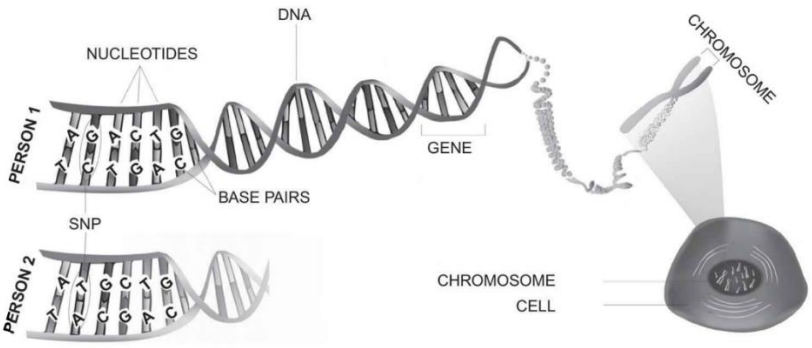


**Figure 1.1 Organization of DNA in the cell nucleus [58]**

DNA strands are polymers, which are made up of many repeating units called nucleotides. Nucleotides form the structure of DNA and consist of one of four nitrogenous bases-cytosine (C), thymine (T), adenine (A), and guanine (G)-plus a molecule of sugar (deoxyribose) and a phosphate molecule. The sugar and phosphate molecules on the nucleotides alternate but also form the backbone of the DNA strands. One of the four dif ferent nitrogenous bases-A, T, C, or G-joins to each sugar. Recall from Figure 1.1 that DNA is in the form of a double helix. Each base links to a base on the opposite end of the strand in the double helix. Humans are thus composed of diploid cells or in other words, pairs of chromosomes with one set of chromosomes inherited from each parent. Since we are diploids, DNA's two strands are complementary to each other or in other words they follow complementary base pairing rules. Complementary base pairing means that A always pairs with T and C always pairs with G, forming base pairs. The two strands are complementary to each other and therefore contain the same information. As figure 1.1 also illustrates, it is the order of these bases along a single strand that comprises the genetic code.[58]

A chromosome is a single molecule of DNA that comprises part of the genome. It consists of nucleic acids and protein and is found in the nucleus of somatic cells and carries genetic information in the form of genes. As Figure 1.1 demonstrates, chromosomes are central to our understanding of genetics. Humans have 23 chromosome pairs (i.e., 46 chromosomes) in total consisting of 22 autosomal chromosomes and one pair of sex chromosomes, two Xs for females (XX) and an X and a Y for males (XY). Autosomal chromosomes are the numbered

chromosomes that are not related to sex determination or, in other words, chromosome 1 through 22. [58]

The term genotype is defined as the observed genetic sequence information and can be thought of as a categorical variable. Humans carry two homologous chromosomes, which are defined as segments of deoxyribonucleic acid (DNA), one inherited from each parent, that code for the same trait but may carry different genetic information. Thus, in its rawest form in humans, the genotype is the pair of DNA bases adenine (A), thymine (T), guanine (G) and/or cytosine (C) observed at a location on the organism's genome. This pair includes one base inherited from each of the two parental genomes and should not be confused with the pairing that occurs to form the DNA double helix. Genotype data can take different forms across the array of genetic association studies and depend both on the specific organism under investigation and the scientific questions being considered. For example, in humans, most SNPs are biallelic, indicating there are two possible bases at the corresponding site within a gene (e.g., A and a). Furthermore, since humans are diploid, each individual will carry two bases, corresponding to each of two homologous chromosomes. As a result, the possible genotype values in the population are AA, Aa and aa. In the context of genotype data, it might be possible to test the null hypothesis that cholesterol levels are the same for individuals with genotype AA and genotype aa. In the expression setting, the null hypothesis may instead be framed as the gene expression level being the same for individuals with cardiovascular disease and those without cardiovascular disease. [50] A measure of disease status or disease progress, referred to as the trait or phenotype, is also collected for analysis. Notably, in population association studies, it generally treats the genotype as the predictor variable and the trait as the dependent variable.[50]

Genes are defined simply as regions of DNA that are eventually made into proteins or are involved in the regulation of transcription; that is, regions that regulate the production of proteins from other segments of DNA.[50]

A mutation is a permanent change in the sequence that makes up a gene. Mutations can affect a single base pair or multiple genes across a large segment of a chromosome. There are two types of gene mutations. The first is the hereditary mutation, which is inherited from a parent, is present for an individual's entire life, and is in almost all cells in the body. It is also often referred to as germ line mutation, which is a mutation that will be inherited by the offspring of

the organism. The second type is somatic or acquired mutations that occur during an individual's lifetime and exist only in certain cells. These mutations are generally related to environmental factors. This could include, for example, smoking or exposure to ultraviolet radiation from the sun. They can also occur if there is an error during DNA replication before or during cell division. These acquired mutations in somatic cells are not passed on to the next generation. [58]

A genetic polymorphism refers to the variation in the DNA sequence between individuals. The possible variants of a polymorphism are referred to as alleles. A variation must be present in at least 1% of a population to be classified as a polymorphism. Such a variable site is commonly referred to as a single-nucleotide polymorphism (SNP).[50] In population-based association studies, the fundamental unit of analysis is the single-nucleotide polymorphism (SNP). A SNP simply describes a single base pair change that is variable across the general population at a frequency of at least 1%.[50] To date, they have not found polymorphisms at every site in the genome. This is due to the fact that only a selection of people have been genotyped but also that variation at some sites cannot be tolerated.[58]

The most common form of human genome variation is SNPs, and they can be used to associate genotypic variation with phenotype. SNPs are the markers that are the focus of the genome wide association study (GWAS), a search across the genome, examining each genetic variant (or region) one by one to see if there is a statistical relationship (association) between SNPs and a phenotype. The genetic variants that are isolated from these GWASs are then often used to engage in either further statistical or downstream biological analysis. For this thesis, SNPs are the genetic markers of choice because they are heritable and abundantly distributed across the genome.[58]

The minor allele frequency (MAF), also referred to as the variant allele frequency, refers to the frequency of the less common allele at a variable site. In the literature, polymorphisms are distinguished by their MAF and categorized as common (MAF > 0.05), low-frequency (0.01 < MAF <0.05), or rare (MAF <0.01) variants.[58]

The phenomenon of polygenicity implies that no single genetic variant determines or is associated with a trait, but rather that it is often hundreds and thousands of genetic variants that each have a small influence on a trait. Such phenotypes are called "complex" since they have a multifactorial genetic basis. This is often related to the common disease-common variant (CD-

CV) hypothesis, which holds that common disease-associated alleles will be found in all populations that manifest a given common disease. In the most extreme omnigenic model, each variant on each gene is assumed to influence a complex trait and will have a small additive or multiplicative effect on the phenotype. [58] The most important recent developments in statistical genetics surround the discovery of ubiquitous polygenicity in most traits that we study. An intuitive implication of polygenicity is that the effect sizes of individual SNPs are smaller than if only a few SNPs would be associated with an outcome. Small effects are harder to discover given a fixed statistical measure of the certainty of the discovery. Polygenicity therefore explains the disappointingly small effects of discovered variants as well as the small number of robustly identified variants.[58]

Recall that polymorphic refers to the presence of more than one allele at a specific locus. A locus (plural loci) is a location on the genome, which could be the location of a gene or marker. A locus or position may refer to the part of the genome that codes for a gene or to the position of a nucleotide in the genome. When an individual has two of the same alleles, regardless of whether it is dominant or recessive, they are called homozygous. Heterozygous refers to having one of each of the different alleles. A person is heterozygous at a gene locus when their cells contain two different alleles. Heterozygosity thus refers to a specific genotype. Homozygous wildtype, on the other hand, refers to the state of having two copies of the more common allele. This distinction also explains the difference between dominant traits, which is when only one allele of a gene is necessary to express the trait versus recessive traits, where both alleles of a gene must be identical to express the trait. For dominant traits we use two capital letters (e.g., AA) and for a recessive trait we use two lowercase letters (e.g., aa). Dominance in melanin deposits results in freckles, for instance. A homozygous freckled person would have the FF genotype while someone without freckles with the homozygous gene would be represented by ff. [58]

A quantitative trait locus (QTL) is a region of DNA which is associated with a particular phenotypic trait, which varies in degree, and which can be attributed to polygenic effects, i.e., the product of two or more genes, and their environment. These QTLs are often found on different chromosomes. The number of QTLs which explain variation in the phenotypic trait indicates the genetic architecture of a trait. Typically, QTLs underlie continuous traits (those traits which vary continuously, e.g., height) as opposed to discrete traits (traits that have two or several character values, e.g., red hair in humans, a recessive trait). Moreover, a single

phenotypic trait is usually determined by many genes. Consequently, many QTLs are associated with a single trait. Another use of QTLs is to identify candidate genes underlying a trait. The DNA sequence of any genes in this region can then be compared to a database of DNA for genes whose function is already known, this task being fundamental for marker-assisted crop improvement. [62]

## 1.3 Genomic Medicine

Genomic medicine is an interdisciplinary medical specialty involving the use of genomic information that has rapidly grown since the completion of the Human Genome Project (HGP) more than a decade ago. The genome is the complete set of information in an organism's DNA.[48] The HGP allowed the investigation of basic genome physiology, and the identification of approximately 10 million common DNA variants. These projects were the first to postulate the possibility of a better understanding of disease pathobiology and pathophysiology via catering the identification and characterization of small variations in the genome, termed single nucleotide polymorphism (SNPs). The venerable field of genetics studies single genes, whereas the emerging field of genomics studies all of a person's genes.[47]

Some key precision medicine applications lie within the realm of cancer diagnosis and potential treatment, for example, the identification of a prostate-specific antigen (a single-strand glycoprotein) which is now routinely used for clinical diagnosis of prostate cancer (PCa). Other pioneering examples relate to the diagnosis of rare diseases. For example, the CFTR gene has been identified as a causal gene for cystic fibrosis, an autosomal, recessive disease. Such advances have led to routine clinical use of both biomarker panels as well as whole exome and genome sequencing both for the case of cancer as well as rare diseases. It has been long been recognized that there is a significant variability in drug response with respect to the efficacy, optimal dose, and adverse drug reactions, with the prevalence of medication-related adverse events among inpatients in the Western world estimated to affect 19% of patients. Genome technology allows for the screening and identification of the right drugs for the right patients, forming the so-called pharmacogenomics field, a key component of the personalized medicine vision.[55]

The prior probability of any variant discovered through genome sequencing being the cause of a patient's rare condition is exceedingly low. Attempts to catalog human genetic diversity have revealed that a typical human genome differs from the reference human genome at 4.1-5

million locations. Most of these variations will be completely benign, while some may have a minor impact on the risk of several common diseases, and only a tiny number may have the ability to cause significant disease in an individual or their children (potentially in combination with variants inherited from their partner). The bulk of these variants are identified through genome sequencing and require thorough filtering to obtain a relevant output. The human genome sequence will someday change many elements of healthcare practice. It will improve our understanding of disease mechanisms and lead to the development of new medications and treatments. In the short term, molecular phenotyping based on genetic and genomic information will enable earlier and more accurate disease prediction and diagnosis, as well as disease progression. The focus of medicine will shift away from late-stage illness cures and toward disease prevention.[49]

A classical biomarker, also known as a biological marker, is any trait that can be used to evaluate a certain disease condition or physiological function. Biomarkers might be correlational (just related to disease) or functional (that is, they have an identified mechanism of action related to disease). Biomarkers can be measured individually or in groups to infer risk, diagnosis, prognosis, and therapy response. Biomarkers include DNA, RNA, proteins, metabolites, host cells, and microbes. Biomarkers can be detected in a wide range of biological materials, including blood, organ tissue, stool, saliva, and urine.[47]

A genomic biomarker is "a DNA or RNA characteristic that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other intervention".[47] It is a DNA sequence that causes disease or is associated with susceptibility to disease. It generally represents the expression, function, or regulation of a gene and can be used clinically to diagnose and monitor disease. A genomic biomarker represents the expression, function, or regulation of a gene. The definition of a genomic biomarker does not include the measurement and characterization of proteins or low molecular weight metabolites. Robust, reproducible, and accessible genomic biomarkers are of diagnostic value and may lead to the identification of causal factors. They can therefore be used clinically to screen for diagnoses, to monitor the activity of diseases, and may also be useful to guide molecularly targeted therapy and personalized regimens or to assess therapeutic response.[47]

### 1.3.1 Advances in Genomic Medicine

Healthcare is becoming more personalized as a result of genomic applications. One of the important approaches for precision medicine is stratifying individual genetic susceptibility based on inherited DNA variation. Disease susceptibility and risk can now be quantified and predicted during birth using "stable genomics," or DNA-based assessments that do not change over a person's lifetime.[3] The ultimate goal of precision medicine is that medicine will be informed by a genetic understanding of the disease rather than a "one size fits all" approach. Precision medicine involves not only researching DNA but also taking into account aspects such as where a person lives, what they do, and their family's health history. Instead of depending on tactics that are the same for everyone, the idea is to develop personalized prevention or treatment approaches to help specific individuals stay healthy or get better.[47] The application of knowledge gained from sequencing human genomes is critical for precision medicine, allowing patients to be matched to the best therapy, so that a patient is treated with the appropriate drug at the right dose at the right time, or changing treatment due to resistance or adaptability through disease evolution.[4] Patients with the same signs and symptoms of cancer often have different outcomes. The precision medicine approach provides a research strategy to develop biomarkers that can be used to classify patients with the same cancer into finer taxa (subclass 1 versus subclass 2) by biomarkers that predict prognoses derived from the synthesis of large amounts of data to identify discriminating biomarkers. For example, patients in subclass 1 who have a worse prognosis (that is, have biomarkers that are associated with poor survival) may be given a more aggressive treatment (treatment X) versus those in subclass 1 who have a better prognosis (that is, have biomarkers that are associated with good outcome) and require a less aggressive therapy (treatment Y). Additionally, the converse may be true where individuals with a worse prognosis are provided less aggressive therapy if no benefit from aggressive treatment has been observed for this subclass.[47]

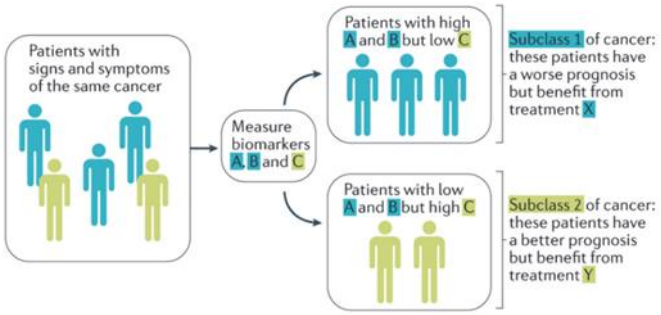


**Figure 1.2**

**Classifying patients into new, specific taxa [47]**

Genetic testing is a type of medical test that detects chromosomal, gene, or protein changes. The majority of the time, testing is used to detect changes associated with inherited disorders. A genetic test's results can confirm or rule out a suspected genetic condition, or they can help determine a person's risk of developing or passing on a genetic disorder. There are now two types of genetic testing available: clinical and direct-to-consumer (DTC). A clinical genetic test is typically performed in a clinical setting with access to trained medical professionals, such as genetic counselors, to assist patients in interpreting the results, which can be extremely difficult to misinterpret. For example, sometimes the results of a genetic test can give a false sense of assurance or sound a false alarm, but a conversation with a genetic counselor can help put the test results into context or explain their implications. By contrast, DTC genetic testing is done at home after ordering a simple test kit online. However, since, the validity of some DTC genetic tests is questionable, and their results are not usually interpreted by a qualified medical professional, their purpose may be more recreational than medical. Privacy and security of genetic data are not completely guaranteed with any DTC genetic test, but consumers can take control of their privacy by knowing and reading the privacy policy and opting out of consent so that their information is not shared with a third party. [47]

Gene therapy is a type of treatment that involves inserting healthy foreign genetic material into a person's cells in order to cure a rare condition or disease. Gene therapy, rather than just treating symptoms, aims to correct the underlying genetic cause of the disease and thus serve as a one-time cure. While gene therapies are becoming more widely available, they are still out of reach for the general public. Luxturna, which is used to treat a rare type of vision loss, has a $850,000 list price.

Understanding why the same genetic condition can manifest so differently in different people is often at an early stage, which makes genetic counselling difficult, particularly in the prenatal setting. It is becoming possible to provide more personalized risk estimates for some genetic conditions by combining knowledge of a person's genetic diagnosis with analysis of other factors that may influence their risk. In general, risk personalization has relied on clinically obvious characteristics: for example, men with pathogenic BRCA variants have a lower risk of developing breast cancer than women with pathogenic BRCA variants. Recently, genetic testing has been developed to supplement 'key' genetic test results in order to provide a more refined personal risk assessment. For example, using a polygenic risk score based on breast and ovarian cancer susceptibility SNPs identified through population GWAS revealed significant differences in absolute cancer risks between women with pathogenic BRCA variants and higher versus lower polygenic risk score values. This has yet to be implemented in routine clinical practice, but it has the potential to help women with pathogenic BRCA variants make better decisions about how and when to manage their cancer risk. [49]

## 1.4 LD-HWE

This section presents two controls that are very important for processing genetic data.

### 1.4.1 Linkage disequilibrium (LD)

Linkage disequilibrium (LD), which refers to the fact that alleles are not randomly associated at different loci. Polymorphisms are inherited together through what is called linkage disequilibrium (LD), which is the nonrandom occurrence in members of a population of the combinations of 2 or more linked genomic loci. In other words, linkage disequilibrium is defined as an association of the alleles present at each of two positions in a genome. For instance, if a T at one SNP locus is generally observed with a G at another SNP locus, these two SNPs are said to be in linkage disequilibrium. Their co-occurrence is more correlated than we would expect by random (equilibrium) conditions. Two alleles (i.e., that are variants of polymorphisms) which are located at different positions at the same chromosome are in LD if they are not inherited independently from one another. In general, alleles which are located close together at the same chromosome will have stronger LD. Conversely, when two SNPs are inherited randomly (i.e., unlinked), they are said to be in equilibrium. The hypothesis of interest is whether the gene is involved in the disease's causal pathway. In this case, the SNP loci chosen within the gene may not be functional; that is, they may not directly cause the disease. These

sites, however, are likely to be associated with disease because they are in linkage disequilibrium (LD) with the functional variant.[50] High LD thus means that two SNPs are linked, which is measured $r^2$.

This measure is based on Pearson's $\chi^2$-statistic for the test of no association between the rows and columns of an r × c contingency table. Specifically, $r^2$ is defined as

$$r^2 = \chi_1^2/N \tag{1.1}$$

Pearson's $\chi^2$-test statistic is given by

$$\chi_1^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{1.2}$$

where I = 1, 2, . . ., r, j = 1, 2, . . ., c and $O_{ij}$ and $E_{ij}$ are respectively the observed and expected cell counts for the I, j cell of an r × c table.[50] The $r^2$ measure is a statistical measure of shared information between two markers and is commonly used to determine how well one SNP can act as a proxy for another. [58]

### 1.4.2 Hardy-Weinberg equilibrium (HWE)

The Hardy-Weinberg equilibrium (HWE) is a theoretical mathematical model describing the probability and distribution of genotype frequencies in a population. The main purpose of the HWE is to express the principle that the amount of genetic variation (allele and genotype frequencies) in a population will remain constant from one generation to the next in the absence of evolutionary influences. HWE denotes independence of alleles at a single site between two homologous chromosomes. Consider for example the simple case of one biallelic SNP with genotypes AA, Aa and aa. HWE implies that the probability of an allele occurring on one homolog does not depend on which allele is present on the second homolog. The HWE is used to model and predict genotype frequencies in large, stable populations. It is an important tool for understanding population structure. If certain assumptions are met, genotype and allele frequencies can be estimated from one generation to the next. In genetic association studies, HWE principles have been applied to detect genotyping error and disease susceptibility loci. When a population is in HWE for a gene, it is not evolving, and allele frequencies will remain the same across generations. The HWE dictates that the frequencies and relative proportions of genotypes remain stable-or in other words in equilibrium-over time if all assumptions of the

HWE are met. The proportions will remain constant at this equilibrium if these five assumptions hold:

1. There is no natural selection (i.e., all genotypes have equal fitness). Natural selection is the increase or decrease of particular genetic traits as a function of the differential fitness and the reproductive success of individuals. In other words, natural selection operates when particular genetic variants render the individuals who bear them more likely to survive. Consequently, those genetic variants increase in frequency in the next generation. Natural selection is said to drive adaptive evolution to select for traits that are beneficial to a particular population within an environment. One way to think about selection is that it is a filter that removes suboptimal alleles from a population so that it is better adapted to its environment. Fitness-also sometimes referred to as longer reproducing, they are considered as no longer evolutionary fit.

2. There is no genetic drift. Genetic drift is a change in allele frequencies oner time in a population of finite size due to random transmission of parental alleles from parents to offspring and due to the fact that some individuals randomly produce more offspring than others, irrespective of their genotype.

3. A closed population (there is no significant migration in or out of the population)

4. Mutation does not occur

5. There is no assortative mating. In genetic research refers to a mating structure in which pairs of individuals that are genetically similar to each other mate with a higher probability than expected under random mating.

If all of these assumptions are met, then four important conclusions can be drawn from the HWE theorem: (1) allele frequencies do not change from one generation to the next, (2) genotype frequencies can be inferred from allele frequencies, (3) only one generation is required to go from non-equilibrium to equilibrium, and (4) once the system is in HWE, it stays in HWE. These assumptions thus entail that the population structure is not from two or more subpopulations, there is no inbreeding (i.e., mating without one or more common ancestors), males and females have similar allele frequencies, all members of the population have equal reproductive success and the population is infinitely large. If the basic assumptions are not met for a particular gene, the population may evolve. Or in other words, genotype frequencies might change. In practice, violation of the HWE may also point to measurement error in genetic data.

Testing the HWE is therefore a crucial part of the quality control process in handling genetic data.[58]

Tests of HWE include Pearson's $\chi^2$-test and Fisher's exact test. The $\chi^2$-test is computationally advantageous but relies on asymptotic theory. Thus, when more than 20% of the expected counts are less than five, Fisher's exact test is preferable. Consider the 2 × 2 table of genotypes at a single locus given in Figure 1.3. Here $n_{11}$ and $n_{22}$ are the number of individuals with genotypes AA and aa, respectively, and these counts are observed. Notably, the genotypes Aa and aA are indistinguishable in population-based investigations, and thus we only observe the sum $n_{12}^* = n_{21} + n_{12}$ and not the individual cell counts, $n_{21}$ and $n_{12}$. The expected counts corresponding to these three observed counts, $n_{11}$, $n_{12}^*$, $n_{22}$, are given respectively by $E_{11} = Np_A^2$, $E_{12} = 2Np_A(1 - p_A)$ and $E_{22} = N(1 - p_A)^2$, where $p_A$ is the probability of A and is estimated based on the observed allele count. That is, we let $p_A = (2n_{11} + n_{12}^*)/(2N)$. The $\chi^2$-test statistic is

$$\chi^2 = \sum_{(i,j)\epsilon C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_1^2 \tag{1.3}$$

where now the summation is over the set C of three observed cells. This statistic is compared with the appropriate quantile of a $\chi_1^2$-distribution to determine whether to reject the null hypothesis of HWE.

| | | Homolog 2 $A$ | Homolog 2 $a$ | |
|---|---|---|---|---|
| Homolog 1 | $A$ | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| | $a$ | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| | | $n_{.1}$ | $n_{.2}$ | $n$ |

**Figure 1.3**

**Genotype counts for two homologous chromosomes**

A statistically significant test of HWE suggests that the SNP under investigation is in Hardy-Weinberg disequilibrium (HWD).[50]

The p-value from Fisher's exact test is based on summing the exact probabilities of seeing the observed count data or something more extreme in the direction of the alternative hypothesis. Fisher showed that the exact probability from a contingency table such as Figure 1.3 is given by

$$p_A = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{22}}}{\binom{N}{n_{.1}}} = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{N!n_{11}!n_{12}!n_{21}!n_{22}!} \tag{1.4}$$

In practice, a test of HWE is used to assess whether either population admixture or stratification is present. While admixture and stratification represent two different phenomena—the former describes in-breeding while the latter implies the presence of multiple subpopulations in which there is no inbreeding—the manifestation of both is a violation of the HWE assumption.[50]

# CHAPTER 2

# HERITABILITY

## 2.1 Introduction

Heritability on the liability scale, $h^2$, quantifies the proportion of variance of liability to disease attributable to inherited genetic factors. Is a measure of how well differences in people's genes account for differences in their traits. [18] All individuals in the population carry some genetic risk variants and likely experience some nongenetic risk factors, but most individuals in the population are not affected– disease status results when the cumulative load exceeds a burden of risk threshold.[18] Heritability is formally defined as a ratio of variances.[17] Two types of heritability can be estimated. The broad-sense heritability ($H^2$) evaluates the proportion of phenotypic variance explained by all genetic factors, including additive effects, dominant effects, and epistasis effects and the narrow-sense heritability ($h^2$), on the other hand, evaluates the proportion of phenotypic variance explained by additive genetic effects.[39] It measures with a single number the fraction of variation between individuals in a population that is due to their genotypes. Because individuals transmit only one copy of each gene to their offspring, most relatives share only single or no copies that are identical by descent (IBD) (the most important exceptions are identical twins and full siblings (sibs)), and dominance and other non-additive genetic effects that are based on sharing two copies do not contribute to their phenotypic resemblance. This is why the selection response and correlation of most relatives depend on $h^2$ and not $H^2$, and why $h^2$ is the usual parameter.[17] Identical by descent (IBD) is a term used in genetic genealogy to describe a matching segment of DNA shared by two or more people that has been inherited from a common ancestor without any intervening recombination. The segments are considered to match if all the alleles on a paternal or maternal chromosome are identical (barring rare mutations and genotyping errors) and if the minimum threshold conditions set by the testing company have been met. Everyone has two copies of each chromosome – one chromosome inherited from their father and one chromosome inherited from their mother. Matching segments can be on half-identical regions (HIRs) (matches on the paternal or maternal chromosome) or fully identical regions (FIRs) (matches on both the paternal and maternal chromosome). FIRs are generally only seen in full siblings

and double cousins but are sometimes found in more distant relatives if the individual comes from an endogamous (intermarrying) population. [63]

## 2.2 Definition of heritability

Because heritability is a ratio of variances, both the numerator and denominator need close scrutiny. The denominator contains the total observed variation, usually excluding variation that is due to known fixed factors and covariates such as sex, age and cohort. The numerator of $h^2$ contains variation that is due to additive genetic values in the population. These values, called 'breeding values' in the literature, are defined as the sum of the average effects of parents' genes that give rise to the mean genotypic value of their progeny. Breeding values can be measured even when the average effects of individual genes cannot. A consequence of the definition of heritability is that it depends on the population, because both the variation in additive and non-additive genetic factors, and the environmental variance, are population specific. Genetic variance depends on segregation in a population of the alleles that influence the trait, the allele frequencies, the effect sizes of the variants and the mode of gene actions. All these variables can differ across populations. Similarly, environmental variance can vary across populations. Therefore, the heritability in one population does not, in theory, predict the heritability of the same trait in another population. In practice, heritabilities of similar traits are often remarkably similar in other populations of the same species, or even across species. Heritability can also differ between sexes, and heritability of the same trait can differ early and late in life.[17]

Observed phenotypes (P) of a trait of interest can be partitioned, according to biologically plausible nature–nurture models, into a statistical model representing the contribution of the unobserved genotype (G) and unobserved environmental factors (E):

$$\text{Phenotype (P)} = \text{Genotype (G)} + \text{Environment (E)} \tag{2.1}$$

The variance of the observable phenotypes ($\sigma_P^2$) can be expressed as a sum of unobserved underlying variances ($\sigma_G^2$ and $\sigma_E^2$ ):

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 \tag{2.2}$$

Heritability is defined as a ratio of variances, by expressing the proportion of the phenotypic variance that can be attributed to variance of genotypic values:

$$\text{Heritability (broad sense)} = H^2 = \frac{\sigma_G^2}{\sigma_P^2} \quad (2.3)$$

The genetic variance can be partitioned into the variance of additive genetic effects (breeding values; $\sigma_A^2$), of dominance (interactions between alleles at the same locus) genetic effects ($\sigma_D^2$), and of epistatic (interactions between alleles at different loci) genetic effects ($\sigma_I^2$):

$$\sigma_G^2=\sigma_A^2 + \sigma_D^2 + \sigma_I^2 \quad (2.4)$$

and

$$\text{heritability (narrow or strict sense)} = h^2 = \frac{\sigma_A^2}{\sigma_P^2} \quad (2.5)$$

## 2.3 Heritability Estimation

Estimates of heritability on the liability scale depend on knowledge of baseline risk of disease in the population from which the twin and family cohorts are drawn and estimates of baseline risk are often surprisingly difficult to pin down. They may also vary between populations, across ages and may depend on whether nongenetic factors have been recorded and included in the analysis. Hence, in reality heritability estimates should be viewed as pragmatic benchmarks representing evidence for low, moderate or high contributions of genetic effects.[18]

Accurate estimation of heritability can show the degree to which genetic factors influence phenotypes and improve our understanding of the genetic basis of disease and disease-related complex traits. Indeed, heritability plays an important role across a range of genetic applications: it is a key for understanding the evolutionary forces underlying natural selection; it determines how a population will respond to selection; it predicts, at least in part, gene mapping power in genome-wide association studies; it can estimate, quite accurately in some cases, the phenotypic value of an individual and thus facilitate genomic selection via predicted breeding values; and it provides an upper limit for the genetic prediction of phenotypes.[39]

### 2.3.1 Heritability with unknown pedigrees

Genetic markers can help to estimate heritability in novel ways. When phenotypes are collected on a sample of individuals whose relatedness is partially or wholly unknown, genetic markers can be used to infer relatedness between pairs of individuals, because related individuals tend to share more marker alleles than unrelated individuals. The inferred relatedness can then be correlated with phenotypic similarity, and quantitative genetic parameters, including

heritability, can be estimated. This method has been applied in evolutionary studies to estimate heritability for quantitative traits when phenotypes and DNA samples are available, but pedigree information is not, for example in fish and plants. A disadvantage of this method is that many polymorphic markers, typically hundreds, are needed to estimate relatedness accurately, for distant relatives in particular. Generally, the closer the relatives the fewer markers are needed. Offspring–parent pairs can be easily identified with only a few polymorphic markers because they always share at least one allele at all marker loci. Funding, rather than the availability of large numbers of polymorphic markers, could be the only limiting factor in the near future, given the rapid discovery of new markers in many species and the development and application of high-density array technology.

### 2.3.2 Exploiting variation in relatedness

Genetic markers can also be used to estimate heritability when the pedigree is known, by estimating the actual or realized relationship between relatives. Apart from offspring–parent pairs (who always share 50% of their genes IBD) and monozygotic twins (who share 100% IBD), the proportion of the genome that is shared IBD varies around its expectation for pairs of relatives because of the stochastic nature of segregation and recombination. A parent has one chromosome from each of its parents, and which parts of these two grandparental chromosomes are passed on to an offspring is a chance event (random segregation). This segregation causes variation in the actual number of alleles shared IBD between relatives. For example, for full sib pairs at a single locus, 25% of all the sib pairs share no alleles IBD (they are 'unrelated' at that locus), 25% share two alleles IBD (they are 'clones' at that locus) and the remaining 50% share one allele IBD. Recombination events during the formation of gametes reduce the variation in the proportion of a chromosome (or genome) that is shared IBD between relatives, by creating more segregating segments: the larger the number of segregating units, the smaller the variation in the proportion of segments shared. The measuring of multiple genetic markers in relatives allows the estimation of the total proportion of the genome that is shared. The amount of variation around the expectation is modest, but measurable. For example, for sib pairs the average proportion shared is 50%, with a standard deviation of approximately 4%. For half sibs the mean and standard deviation are 25% and 3%, respectively. The significance of this new approach is that heritability can be estimated without strong assumptions about the causes of family resemblance, because it is estimated from data within families. In the future, with sufficient data, this will allow unbiased estimation of heritability of contentious phenotypes

such as IQ in humans, unbiased estimation of the genetic contribution to concordance for disease in relatives, and unbiased estimation of additive and non-additive variance that is not affected by confounding factors.[17]

A high heritability means that most of the variation that is observed in the present population is caused by variation in genotypes. It means that, in the current population, the phenotype of an individual is a good predictor of the genotype. However, it does not mean that the phenotype is determined once we know the genotype, because the environment can change or can be manipulated to alter the phenotype. A low heritability means that of all observed variation, a small proportion is caused by variation in genotypes. It does not mean that the additive genetic variance is small. This difference matters because the response to natural or artificial selection depends on the amount of genetic variation in the population. Many phenotypes relating to fitness in natural populations have a large amount of additive genetic variation relative to the mean [17]

## 2.4 Heritability in Genomic Area

GWAS identify SNPs that are statistically correlated with phenotypes of interest. After nearly a century of twin and family studies consistently demonstrating relatively high heritability of traits, there was some expectation that early GWAS would find a few genes of large effect. GWAS results in this respect have been disappointing. While twin studies indicate that roughly 50% of the total variance in cognitive ability is explained by genetic differences, individual SNPs associated with cognitive ability typically explain less than .04% total variance. [64]

One simple method of increasing the amount of phenotypic variance accounted for by GWAS (sometimes referred to as ' $h^2_{GWAS}$') is to sum the total effects of genome-wide significant SNP hits. The first efforts to do this were met with disappointing results as well. Weedon et al. (2008) conducted a GWAS of height which identified 20 SNPs with a combined effect of 3%. These meager results inspired an impetus in the GWAS community to conduct bigger and better studies that would be required to power the small effect sizes of individual SNPs. Bigger data meant more SNPs, which meant more variance explained and, consequently, increased $h^2_{GWAS}$. Even when summing the small effects of hundreds of genome-wide significant SNPs, variance explained by GWAS results are still quite small. [64]

To increase the amount of phenotypic variance accounted for by GWAS results, has been developed the 'SNP heritability' (or $h^2_{SNP}$). SNP-based heritability only measures the variance explained by additive effects of the genotyped or imputed SNPs. First, instead of limiting analysis to SNPs who meet the strict p-value GWAS significance threshold, SNP heritability is derived by analyzing the complete set of SNPs for each participant sample - even those that are not associated with the trait of interest. To this end, SNP heritability is biologically non-obvious. Second, SNP heritability assumes linear additivity of SNP effects. Third, SNP heritability represents the current limit on the total variance that could be explained by SNPs, for any given phenotype. So, for example, if a polygenic score were maximally predictive, it would be equivalent to SNP heritability. Fourth, SNP heritability is consistently lower than traditional heritability. This gap between traditional heritability and SNP heritability is sometimes referred to as the missing heritability problem. [39]

### 2.4.1 SNP Heritability estimation

Accurate estimation of SNP heritability can help us better understand the degree to which measured genetic variants influence phenotypes. A common method to estimate SNP heritability based on summary-statistics is LD Score regression (LDSC). For each SNP, LDSC first computes its LD score, $\ell_j = \sum_k r^2_{jk}$, which captures approximately the number of genetic variants tagged by this SNP. LD score cannot be computed exactly due to the large number of genome-wide SNPs. Instead, it is typically estimated based on SNPs. After obtaining LD score, LDSC regresses the $\chi^2$ test statistic from GWAS on the per-SNP LD scores. Under a polygenic model, in which effect sizes are drawn independently from distributions with variance proportional to 1/ (p(1 – p)), where p is the minor allele frequency (MAF), the expected $\chi^2$ statistic of variant j is:

$$E[\chi^2 | \ell_j] = n\,\ell_j \cdot \frac{h^2_g}{M} + na + 1 \qquad (2.6)$$

where a measure the confounding bias due to potential population stratification and cryptic relatedness, n is the sample size, M is the number of SNPs, such that $\frac{h^2_g}{M}$ is the average heritability explained per SNP. Here, population stratification refers to the presence of a systematic difference in allele frequencies between subpopulations in the data possibly due to different ancestry. Cryptic relatedness occurs when individuals in the study are more closely related to another than thought. Both population stratification and cryptic relatedness, if uncontrolled, can

lead to upward biased SNP heritability estimation. By controlling for population stratification and cryptic relatedness using the parameter a, LDSC can mitigate their influence for SNP heritability estimation. Thus, regressing the $\chi_j^2$ statistics from GWAS against per-SNP LD score $\ell_j$ allows for estimation of $h_g^2$. By modeling summary statistics, LDSC is not only applied to many data sets that previously cannot be analyzed for SNP heritability estimation, it also substantially improves computational speed and makes SNP heritability scalable to large data sets.[39]

## 2.5 Missing Heritability

The development of alternative methods of estimating heritability (e.g., GWAS and SNP heritability) has given rise to what is commonly referred to as the 'missing heritability problem'. The missing heritability arises out of a numerical gap between the heritability measured using pedigree information and the measure through GWAS of the same trait. For example, traditional heritability estimates for IQ obtained using twin and family studies range between .5 and .7 while SNP-based heritability estimates of IQ are currently no greater than .25. Missing heritability is greatest among complex, behavioral traits.

Below pointed out some explanations for missing heritability. Epistatic interactions between SNPs entail that the effect of a given SNP will be modified (enhanced or diminished, for example) in the presence of another SNP, which violates the additivity assumption of SNP heritability. Similarly, gene-environment interaction and epigenesis have been proposed as explanations for missing heritability.[64] The GWAS analyses were not powered enough to capture all the genetic variants involved in disease susceptibility and that a lot of variants with small effects were missed. It is also possible the causal variants are not in complete linkage disequilibrium (LD) with the genotyped SNPs. Genomic heritability estimates could, therefore, be improved by taking into account all the genetic data to incorporate smaller effects that did not reach significance but could, however, significantly contribute to phenotype variability. End of another explanation for the missing heritability is that rare variants that are not captured by SNP-chips could be major contributors of common disease susceptibility. [40]

# CHAPTER 3

# GWAS

## 3.1 Introduction

Genome-wide association studies (GWAS) aim to identify associations of genotypes with phenotypes by testing for differences in the allele frequency of genetic variants between individuals who are ancestrally similar but differ phenotypically as well as genetic associations that may differ across ancestries, complicating direct comparisons between groups of individuals. The typical goal of the GWAS is to identify loci that harbor causative variants (hoping to implicate genes near these loci, thus leading to a better understanding of a disease and novel therapeutics) and to use genetic risk factors to make predictions about who is at risk and to identify the biological underpinnings of disease susceptibility for developing new prevention and treatment strategies. The most commonly studied genetic variants in GWAS are single-nucleotide polymorphisms (SNPs). [41] GWAS are more likely to provide insights into disease pathogenesis than useful information on personalized risk assessment. [21] Previous GWAS have shown that most traits are influenced by thousands of causal variants that individually confer very little risk, are often associated with many other traits and are correlated with causal and non-causal variants that are physically close as a result of linkage disequilibrium, making direct biological, causal inferences complicated.[41] GWASs test millions of separate regression models for associations between genetic variants and a phenotype. Phenotypes can be monogenic traits, strongly influenced by variation within a single gene, but many are polygenic complex traits, which are the result of variation within multiple genes and their interaction with behavioral and environmental factors. The results of a GWAS show the association of each individual SNP with a particular trait or phenotype across all genotyped regions. Since many traits are complex and linked to multiple genetic loci (i.e., polygenic), a GWAS often identifies many genetic variants that each have a small influence on a phenotype. Due to small effect sizes, very large data sources are required and the GWAS discovery typically culminates in many GWAS analyses conducted on multiple data sources and then combined into one meta-analysis.

## 3.2 Study Design

A well-planned study can avoid systematic bias in analysis results while also providing enough statistical power to detect association signals. A trait can be a categorical disease status, such as affected or unaffected, or a disease-related biomarker, such as low-density lipoprotein cholesterol levels. In both population-based and family-based studies, quantitative phenotypes have been shown to have higher power than categorical traits. The choice of trait implies the appropriate statistical tests - logistic regression can be used to model disease susceptibility if the trait is categorical, whereas linear regression can be used for continuous phenotypes. A retrospective case-control study begins with the selection of subjects based on disease status, followed by the collection of genetic and environmental data. This design benefits from a low-cost and convenient sample collection process, and subject recall bias is not an issue when the variables being evaluated are genetic markers. To increase study power, the case-control genetic study can benefit from combining working data with external samples, though ideally the same protocol of sample storage, DNA collection, and genotyping should be followed to avoid systematic bias. A prospective cohort study establishes a cohort and performs baseline genotyping on all subjects, who are then followed up on and disease development is observed. The prospective study nested with a case-control design is a more cost-effective version in which a cohort is followed up on and those who develop disease are collected as cases while a subset of disease-free subjects is selected as controls. The study of genetic association using subjects collected in families is referred to as a family-based association study, and the simplest design consists of affected offspring with one or both parents. This type of study has the advantage of being resistant to population admixture, which is a common problem in case-control studies, because it tests for the disequilibrium of alleles transmitted and not transmitted to affected offspring. According to studies, the family-based design is better suited for disorders with low prevalence (1%), and it has lower power when the disease has a polygenic genetic architecture or unscreened control subjects are used, compared to the unrelated case-control design. [43] The choice of data resource and study design for a GWAS depends on the required sample size, the experimental question and the availability of pre-existing data or the ease with which new data can be collected. There are several excellent public resources available that provide access to large cohorts with both genotypic and phenotypic information, and the majority of GWAS are conducted using these pre-existing resources.[41]

## 3.3 Quality Control (QC) Of Genetic Data

The analysis of genetic data to conduct a GWAS entails an understanding of statistical inference in this setting but also numerous quality checks-referred to as quality control (QC). QC is one of the central aspects of working with genetic data. Low-quality samples and genotype-calling errors will result in an increased number of false-positive and false-negative findings. Quality control (QC) procedures can be used to exclude low-quality samples or markers and prevent misleading associations in subsequent analysis. Because the impact of removing a biomarker is possibly bigger than the impact of removing one subject, the QC should be performed first on a 'per-individual' basis and then on a 'per-biomarker' basis to reduce the likelihood of incorrectly removing a causal association.[43]

### 3.3.1 Sample QC

The first step is to ensure that the individuals included in the sample have high-quality data.

1. Sex inconsistencies occur if discrepancies are found (e.g., an individual is recorded being female but appears homozygous for every X chromosome marker). This can be checked by estimating the homozygosity rate on the X-chromosome. Because males have only one X-chromosome, the expected homozygosity rate is 1 for males and < 0.2 for females[ Homozygous, as related to genetics, refers to having inherited the same versions (alleles) of a genomic marker from each biological parent. Thus, an individual who is homozygous for a genomic marker has two identical versions of that marker]. If a subject's homozygosity rate deviates from the expected homozygosity rate based on the determined sex information, it indicates the likelihood of a sample mix-up or incorrect subject information. If differences are discovered (for example, an individual is listed as female yet seems homozygous for every X chromosome marker), any available study questionnaires should be checked to determine whether there was a sample-handling error that resulted in a sample mix-up. Checking X chromosome heterozygosity may also reveal sex chromosome anomalies such as Turner syndrome (females having karyotype XO), Kleinfelter syndrome (males having karyotype XXY), mosaic individuals (XX/XO, XX/XXY), or females with large stretches of loss-of-heterozygosity on the X chromosome who are otherwise phenotypically normal.[43] The intensity plot depicts the intensity of the X and Y probes (Figure 3.1). Females are intended to have low Y intensity and high X intensity (bottom right corner), while men should have similar X and Y intensities (top left corner). Subjects with inconsistent sex information should be deleted if the discrepancy cannot be addressed by consulting the clinical record. However,

depending on the study's objectives, these people are frequently not excluded from the study just because of sex chromosome anomalies. [42]

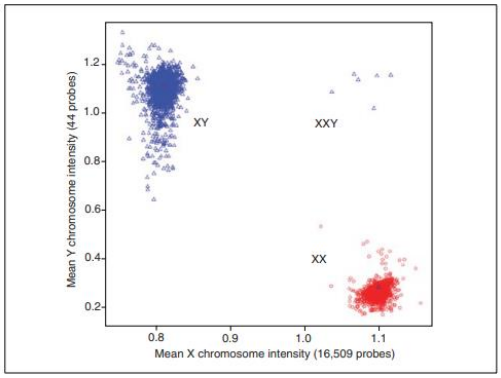


**Figure 3.1**
**Intensity plot [42]**

2. Sample relatedness or duplicate subjects. It is critical to discover unexpected relatives in data when using association tests that assume subject independence. Duplicates and related individuals can significantly introduce bias in association analysis. To related samples in the dataset would result in increased type I and type II errors. This step can be performed by calculating the estimated identity-by-descent (IBD) that measures the proportion of the loci where two individuals sharing 0, 1, or 2 alleles inherited from a common ancestor. The IBD is calculated on the autosomes using SNPs in low linkage disequilibrium ($r^2 < 0.2$). Individuals who share two alleles at each locus are considered duplicated samples or monozygotic twins, and their IBD = 1. IBD = 0.5 for parent-offspring or full siblings, 0.25 for second degree relatives, and 0 for unrelated persons. The observed IBD may vary due to genotyping errors. Thus, if the IBD is greater than 0.98, a pair of subjects may be termed duplicates and, in this case, one subject from the pair should be removed from the data. When relatedness among the subjects is observed, methods that control for kinship relationships can be applied.[43] Using these data, the proportion of loci sharing one allele IBD (Z1) can be plotted by the proportion of loci where individuals share zero alleles IBD (Z0) and points color coded by the relationship type. For clarity, this plot can be restricted to points where the overall kinship coefficient is $\geq 0.05$, as most of the individuals where kinship $\leq 0.05$ will be unrelated. This will produce a plot as shown in Figure 3.2. If it is believed that pedigree records obtained through the original data are accurate, then a point out of place (e.g., points colored as unrelated showing up where most of the parent-offspring pairs cluster) would be indicative of either nonpaternity, adoption,

sample mix-up, or duplicate processing of a single individual. Further investigation employing the original data can be used to attempt to identify the problem. It is also worth noting in studies where datasets from multiple sites are combined that it is possible that the same participant is present in more than one study. These two data points would appear genetically identical across sites. In addition to potentially discovering sample-handling issues, visualizing sample relatedness as shown in Figure 3.2 also reveals any cryptic relatedness that may be present in the study sample. Figure 3.2 shows that many individuals who indicated that they were unrelated (black points) or distantly related (blue points) line up along the diagonal in this plot. These individuals represent second-, third-, fourth-, and fifth-degree relatives. If treated as independent samples in the downstream analyses, having many such as mixed-model regression must be used in place of simple linear or logistic regression. [41]

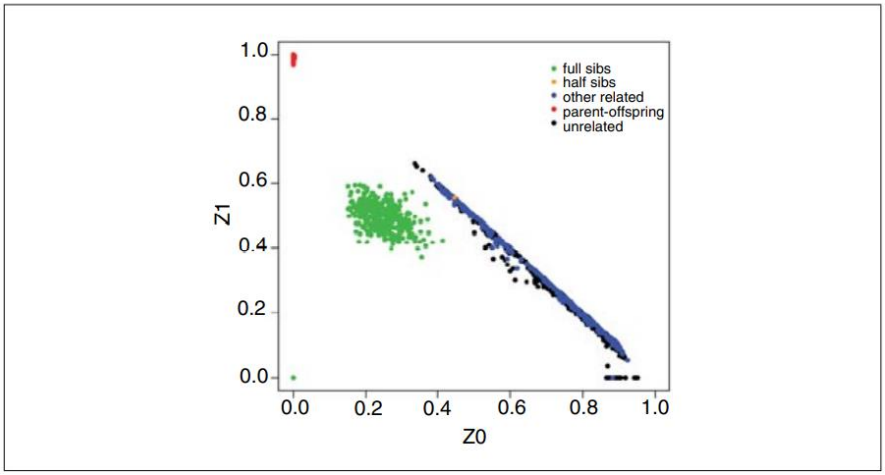


**Figure 3.2**
**IBD presentation[42]**

3. Population substructure. Population structure refers to the patterns found in the genetic data that allow us to determine an individual's ancestry. There is population structure when mating is more likely to occur between some subsets of the population than between others, typically due to geographical structure. Individuals located in geographical proximity to each other are more likely to mate. Population structure is also used to describe a population in which allele frequencies differ between different geographic regions, so a SNP that is common in one population may be rare in another one or even show no variation at all. [58]

The presence of multiple subpopulations (e.g., individuals with different ancestral background) in a study called population stratification. Because allele frequencies can differ between subpopulations, population stratification can lead to false positive associations and/or mask true associations. Restricting analysis to an ancestry group, such as individuals with European ancestry, does not protect us from the risk of including bias in the analysis due to population stratification.[58] The apparent associations would be attributable to ancestral differences rather than an actual association of genes to disease. As a result, it is necessary to examine the study samples for population stratification and use this information to inform subsequent analyses [42]. Principal component analysis (PCA) is frequently used to detect this issue. The idea behind this approach is to provide a low-dimensional representation of the data that captures information on the variability between individuals across SNPs. The aim of PCA is to identify k(k<p) linear combinations of the data, commonly referred to as principal components, that capture overall variability, where p is the number of variables, or SNPs. [50] In PCA, the genotype matrix is normalized and transformed through a linear combination of the input SNPs. The first vector of the converted matrix is called the first principal component (PC), which explains the most variation in the genotype data, followed by the second PC and so on. Finally, the top 10-20 PCs can be incorporated as covariates in a generalized linear model to analyze the effect of a SNP.

The genomic control method computes a variance inflation factor or genomic inflation factor λ, obtained from the robust estimate: $\hat{\lambda}$= median$(X_1^2, X_2^2, \dots, X_p^2)/0{,}456$, where each $X^2$ is a chi-squared distributed statistic calculated from the genome-wide scan of p SNPs. The test statistic $Y^2$, adjusted for the genomic inflation, can be used for the association test: $Y^2 = X^2/\hat{\lambda}$, which follows the chi-square distribution under the null hypothesis. A number of λ close to 1 indicates that data have been properly corrected for population substructure. If the value is more than 1.2, stratification is present. By dividing all of the test data by the value of lambda, it is often possible to adjust for population stratification. [43]

4. Samples with a low genotyping efficiency, or call rate, should be excluded from further analysis. A sample with low DNA concentration will result in a poor genotyping call rate, influenced by where the sample is collected and the amount of sample collected. A sample with more than the usual number of missing genotypes indicates poor sample quality, and the subject should be removed from the data set. Individuals who have missing genotype data across more

than a pre-defined percentage of the typed SNP excluded. A common genotype missing threshold is 5%. If a 98% threshold is applied, this is interpreted as individuals missing genotype data for more than 2% of the standard SNPs being removed. The exact threshold may differ from study to study. The threshold should be set with the purpose of striking a compromise between minimizing the number of samples dropped and maximizing genotyping efficiency. Figure 3.3 shows the proportion of samples (red and blue lines) or SNPs (green line) remaining at different call rate thresholds.[42]

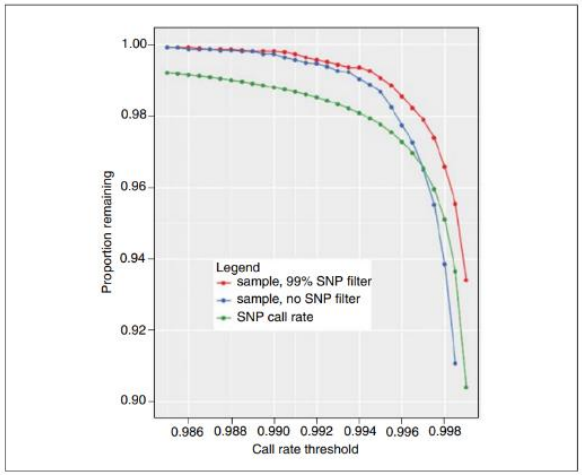


**Figure 3.3**

**Proportion of SNPs or samples remaining as call rate threshold increases [42]**

## 3.3.2 Marker (SNPs) QC

A second set of quality control analyses focuses on the data quality of variants. Several steps are taken sequentially to remove low-quality variants that might introduce bias in the study.

1. Marker genotyping efficiency/call rate. The call rate for a given SNP is defined as the proportion of individuals in the study for which the corresponding SNP information is not missing. Marker genotyping efficiency is a good indicator of marker quality. Excludes SNPs that are missing in a large proportion of the subjects. A recommended threshold for removing SNPs with low call rate is approximately 98% to 99%, although this threshold may vary from study to study. Using a call rate of 98%, meaning that it retains SNPs for which there is less than 2% missing data. Turner *et al.* recommend removing poor-quality SNPs before running the sample genotyping efficiency check discussed above, so that fewer samples will be dropped from the analysis simply because they were genotyped with SNP assays that had poor performance (Figure 3.3).[42]

2. Minor allele frequency. It is particularly critical to filter SNPs based on minor allele frequency (MAF) because SNPs with a low MAF are infrequent, and hence power for discovering SNP phenotype correlations is limited. These SNPs are also more susceptible to genotyping errors. A high degree of homogeneity across research participants at a given SNP often results in insufficient power to infer a statistically meaningful association between the SNP and the characteristic under consideration. This can happen when the MAF is so minimal that the vast majority of people have two copies of the main allele. The MAF threshold should be determined by the sample size. Lower MAF thresholds can be used with larger samples. Figure 1.19.7 illustrates that for uncommon SNPs (1% frequency), the power to identify an association in a large dataset (n = 10,000) with a relatively high effect (odds ratio between 1.3 and 1.7) is quite low. In addition to having reduced power for SNPs with low MAF, these SNPs may also result in misleading relationships due to genotyping errors or population stratification. Turner et al. recommend removing any extremely rare SNPs (including any monomorphic SNPs). The threshold chosen depends on the size of the study and the effect sizes expected. However, in studies with very large sample sizes, it may be beneficial to avoid removing these rare SNPs.[42]

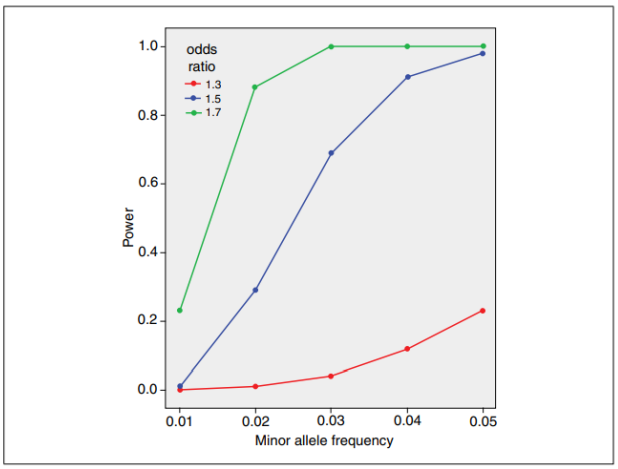


**Figure 1.19.7** This shows the power to detect an association at genome-wide significance ($p < 5\times10^{-8}$), assuming the actual causal SNP is genotyped in a case-control study consisting of 5000 cases and 5000 controls of a common disease with 10% prevalence under an additive model at several different odds ratios. Note that when the MAF is low, power is extremely low even for very large effects (odds ratio = 1.7).

**Figure 3.4 [42]**

3. Hardy-Weinberg equilibrium. Hardy-Weinberg assumptions (see Chapter 1) hold that allele and genotype frequencies remain constant throughout generations. If p represents the frequency of one allele (A) and q represents the frequency of an alternative allele (a), then the genotype frequencies for AA, Aa, and aa are $p^2$, pq, and $q^2$, respectively. The Chi-square goodness-of-fit test can be used to calculate the deviation of observed frequencies from HWE.

The Pearson test is easy to compute, but the $\chi^2$ approximation can be poor when there are low genotype counts, and it is better to use a Fisher exact test, which does not rely on the $\chi^2$ approximation. [see section1.4.2] [43] Deviation from this equilibrium can indicate potential genotyping errors, population stratification, or even true relationship with the trait under study. A real relationship might also cause disequilibrium. SNPs that are significantly out of HWE should not be excluded from the study, but rather identified for further investigation when the association analyses are completed. It is occasionally recommended that HWE be estimated just within the control cohort to avoid omitting potentially meaningful departures (caused by illness connection). The significance level for rejecting based on HWE differs by study and spans from $10^{-5}$ to $10^{-7}$. If various ethnicities are employed in the same study, HWE must be tested independently for each group. [42]

## 3.4 Association Analysis

The proper association test is determined by a number of parameters, including the type of phenotype, the need to adjust for clinical covariates and population structure. Before using association tests, the genotype data must be coded according to the genetic model that has been established. If two SNP alleles are A and a, a dominant model for A will translate the genotypes (AA, Aa, aa) to (1, 1, 0), implying that the presence of the A allele increases the risk of disease by the same amount for AA and Aa genotypes relative to the baseline risk for aa. An additive or co-dominant model will code (AA, Aa, aa) as (2, 1, 0), suggesting that each extra copy of the A allele additively raises disease risk (or an appropriate function of disease risk, such as the log odds of disease). A recessive genetic model for A codes (AA, Aa, aa) as (1, 0, 0), implying that two copies of the A allele are needed to express the phenotypic trait associated with this allele. In GWAS analysis, it is typical to begin with the co-dominant genetic model to search the genome, and then, after related markers are identified, one may opt to do association tests under all scenarios.[43]

### 3.4.1 Single Locus Tests

By focusing on one SNP at a time, a single locus statistical test compares genotype and phenotype. The genotype at a given SNP in the human genetic setting has three levels: homozygous wildtype(AA), heterozygous(Aa), and homozygous rare(aa). If the result is binary, the data can be represented by the two×three contingency table shown in Figure 3.5. In these tests, the null hypothesis states that there is no relationship between genotype and phenotypic

status (the null hypothesis of no association between rows and columns of the 2 × 3 matrix).[43] The odds ratio (OR) is a commonly used measure of association, defined as the ratio of the odds of disease among the exposed to the odds of disease among the unexposed.[50]

| | | Genotype $aa$ | $Aa$ | $AA$ | |
|---|---|---|---|---|---|
| Disease | + | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{1.}$ |
| | − | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{2.}$ |
| | | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | $n$ |

**Figure 3.5**

**2 × 3 contingency table for genotype–disease association[50]**

The OR is written

$$OR = \frac{P(D^+|E^+)/[1-P(D^+|E^+)]}{P(D^+|E^-)/[1-P(D^+|E^-)]} \quad (3.1)$$

For example, the genotype can be set as an indicator for the presence of any other variant. If the three possible genotypes are AA, Aa and aa. Then a dichotomized genotype report could be defined as $E^+$ = (Aa or aa) and $E^-$ = (AA). The corresponding count data is now given in Figure 3.6. In this case, a 2 × 2 contingency matrix results and the OR is equal to

$$\widehat{OR} = \frac{(n_{11}/n_{.1})/(n_{21}/n_{.1})}{(n_{12}/n_{.2})/(n_{22}/n_{.2})} = \frac{n_{11}n_{22}}{n_{21}n_{12}} \quad (3.2)$$

| | | Genotype ($Aa$ or $aa$) | $AA$ | |
|---|---|---|---|---|
| Disease (D) | + | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| | − | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| | | $n_{.1}$ | $n_{.2}$ | $n$ |

**Figure 3.6**

**2× 2 contingency table for genotype–disease association[50]**

A formal test of association between a categorical exposure (genotype) and categorical disease variable (trait) is conducted using Pearson's $\chi^2$-test or Fisher's exact test. In the context of a 2 × 2 table, a test of no association between the rows and columns is equivalent to a test of the single null hypothesis, $H_0$: OR = 1. Pearson's $\chi^2$-test involves first determining the expected cell counts of a corresponding contingency table under the assumption of independence between the genotype and trait. The expected count for the (I,j)-cell is given by $E_{ij} = n_{i.}n_{.j}/n$ for I = 1, 2 and j = 1, 2, 3. Letting the corresponding observed cell counts be denoted $O_{ij}$, Pearson's $\chi^2$-statistic is given by

$$\chi^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)} \qquad (3.3)$$

This statistic has a $\chi^2$-distribution with (r−1)(c−1) degrees of freedom, where r = 2 and c = 3 are the number of rows and columns, respectively.[50]

Fisher's exact test is preferable when at least 20% of the expected cell counts are small ($E_{ij}$< 5). The exact p-value is given by the probability of seeing something as extreme or more extreme in the direction of the alternative hypothesis than what is observed. Fisher derived this probability for the 2 ×2 table of Figure3.6, and it is defined explicitly in Section 1.2.2 for testing HWE.[50]

The following analytical methods shall be used to characterize the association between a genotype and a quantitative trait. Genotype can be defined as an M-level factor and in the simplest case reduces to a binary indicator, for example, for the presence of at least one variant allele at a given SNP locus. Specifically, the t-test is a test of the null hypothesis that the mean is the same in two populations, written $H_0$: μ1 = μ2, where populations are defined by genotype. For example, defining μ1 as the population mean for individuals with genotype AA and μ2 as the population mean for individuals with genotype Aa or aa. The two-sample t-test statistic, assuming equal variances, is given by

$$t = \frac{\overline{y_1}-\overline{y_2}}{\sqrt{s_p^2[1/n_1+1/n_2]}} \sim T_{n_1+n_2-2} \qquad (3.4)$$

where $\overline{y_1}$ and $\overline{y_2}$ are the sample means of the quantitative trait for genotype groups 1 and 2, $s_p^2$ is the pooled estimate of the variance, and n1 and n2 are the respective sample sizes. Under the

null, this statistic has a T-distribution with degrees of freedom equal to $n_1 + n_2 - 2$. The Wilcoxon rank-sum test (also called the Mann-Whitney U -test) is a non-parametric analog to the two-sample t-test and is more appropriate than the t-test if the trait is not normally distributed and the sample size is small. The Wilcoxon rank-sum test is a rank-based test and is used to test the null hypothesis that the medians of a quantitative trait in each of two populations are equal.[50]

If a priori dichotomization of the genotype variables is not desirable, we can perform an analysis of variance (ANOVA) or the non-parametric analog, the Kruskal-Wallis test, to characterize association with a quantitative trait. ANOVA is an extension of the two-sample t-test to the M-sample setting and is based on an F-test for a full model with M −1 genotype indicators (dummy variables) compared with the reduced model with an overall mean. The null hypothesis of an ANOVA using a single SNP is that there is no difference between the trait means of any genotype group. The assumptions of GLM and ANOVA are 1) the trait is normally distributed; 2) the trait variance within each group is the same (the groups are homoskedastic); 3) the groups are independent. The Kruskal-Wallis test similarly extends the Wilcoxon rank-sum test. A Kruskal-Wallis (K-W) test can also be applied and is more appropriate in small-sample settings in which the assumption of normality may not be reasonable.[50]

Because these single-marker tests are special examples of regression models with one predictor variable, simple linear regression produces the same result as the t-test and logistic regression produces the same result as the odds ratio test when testing a single locus. Logistic regression is an extension of linear regression where the outcome of a linear model is transformed using a logistic function that predicts the probability of having case status given a genotype class. Logistic regression is often the preferred approach because it allows for adjustment for clinical covariates (and other factors), and can provide adjusted odds ratios as a measure of effect size. In either case, tests require the trait to be approximately normally distributed for each genotype, with a common variance. If normality does not hold, a transformation (for example, log) of the original trait values might lead to approximate normality. [43]

### 3.4.2 Generalized Linear Models for Covariate Control

Consideration of additional variables in the context of analysis will depend on the scientific question at hand, the biological pathways to disease, and the overarching goal of the analysis. For example, if the aim of a study is to identify the best predictive model (that is, to determine the model that can give the most accurate and precise prediction of cholesterol level for a new individual), then it is generally a good idea to include variables previously associated with the outcome in the model. If the goal is to characterize the association between a given gene and the outcome, then including additional variables, for example, self-reported race, may also be warranted if these variables are associated with both the genotype and the outcome. This phenomenon is typically referred to as "confounding". On the other hand, if a variable such as smoking status is in the causal pathway to disease (that is, the gene under investigation influences the smoking status of an individual, which in turn tends to increase cholesterol levels), then inclusion of smoking status in the analysis may not be appropriate. In this text, the term "covariate" is used loosely to refer to any explanatory variables that are not of specific independent interest in the present investigation. Covariates are also commonly referred to as independent or predictor variables.[50]

A confounding factor is defined as a variable that: (1) is related to the exposure variable, (2) is independently related to the outcome variable, and (3) is not involved in the causal pathway between exposure and disease. For example, determining whether high alcohol consumption (the exposure) is associated with a total cholesterol level (the outcome). Because smoking is associated with heavy alcohol use and also with cholesterol levels in non-heavy alcohol users, smoking status is a possible cause in the aforementioned relationship. In population-based genetic settings, we are generally interested in the association between genotype, as defined by one or more SNPs, and a trait. In this case, a confounding factor is defined as a clinical or demographic variable that is associated with both the genotype and the trait under consideration.

The generalized linear model (GLM) can be used to control potential confounding variables such as age, gender, and medication. Multivariate models have the main advantage that they allow multiple possible confounders and effect modifiers to be properly accounted for. The generalized linear model (GLM) should not be confused with the general linear model for multivariate data. The generalized linear model is a modeling framework that is applicable to a variety of dependent variables, including both quantitative and binary traits, as well as count

data. In this section, we discussed the linear regression model for quantitative traits and then the logistic model for binary outcomes. Both represent special cases of the generalized linear model. The GLM is given in matrix notation by the equation

$$g(E[y]) = X\beta \tag{3.5}$$

where E[Y] = μ denotes the expectation of Y, g() is a link function that performs a monotone transformation on the mean of response variable and X is the design matrix. In the case of a quantitative trait, we let g() be the identity link, and Equation (3.5) reduces to the ordinary linear regression model. The multivariable linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{m} a_j z_{ij} + \varepsilon_i \tag{3.6}$$

in which additional variables are included. The measure of association between the genotype and trait is given by $\beta_1$. Now, however, estimation and testing of this parameter takes into account the additional variables in the model. These additional variables may be confounders or may help to explain the variability in our trait. The inclusion of confounding variables is important for drawing valid conclusions about the effect of genotype on the trait. Adding non-confounding variables to the model will not change our genotype effect estimate substantially. However, by reducing the unexplained variability in our model, including these variables may increase our power to detect the association of our primary independent variable. [50]

As described above, the generalized linear model can also be applied to a binary trait. In this setting, g() is commonly defined as the logit function, reducing Equation (3.5) to the logistic regression model. Logistic regression models provide a setting for modeling dichotomous outcomes based on multiple categorical or continuous predictors. The general form of a univariate logistic model in scalar form is given by

$$logit(\pi_i) = \beta_0 + \beta_1 x_i \tag{3.7}$$

where $\pi_i$= P ($y_i = 1|x_i$) and logit ($\pi_i$) = ln [$\pi_i$/ (1 − $\pi_i$)].

For example, suppose y is an indicator for disease status. The β parameter is then interpreted as the effect of one unit increase in $x$ on the log odds of disease. If x is again a binary variable for the presence of a variant allele, then β is the log odds of disease for individuals with this variant versus those that are homozygous wildtype. In this case, we have OR = exp [$\beta_1$ ]. Again, additional variables can be added to this model to account for potential confounding and effect

modification. Estimation of the parameters is achieved using maximum likelihood methods. Tests of these parameters can be carried out based on the Wald statistic. An advantage of multivariable models is that they provide a natural setting for inclusion of multiple independent variables. This allows for consideration of many predictors of disease as well as providing a means for assessing the potential confounding or mediating role of additional clinical and demographic factors. [50]

### 3.4.3 Linear Mixed Models for Complicated Data Structures

A linear mixed model (LMM) is an extension of the standard linear regression model, wherein the variables are divided into two groups: fixed effects and random effects. Fixed effects are modelled as parameters, i.e., fixed, but unknown, quantities, while random effects are modelled as being drawn from a random distribution – typically a Gaussian distribution with mean zero and an unknown variance. Intuitively, this formulation allows accounting for the random effects, while not specifically estimating the value of each random effect. The linear mixed model (LMM) is an effective method for controlling covariation caused by complex correlation structures. The LMM treats the SNP under test as a fixed effect variable, as well as the clinical and environmental covariates, and the remaining genotypes as random effect variables. The model has the following form:

$$Y = \sum_j jX_j + uG + \sum_k v_k Z_k + \varepsilon \tag{3.7}$$

where G is the test SNP, the variables $X_j$ are fixed effect covariates, and $Z_k$ are genotypes with random effects $v_k$. J is the index of fixed effect covariates and k is the index of random effect genotypes. $v_k$is assumed to be independently drawn from a normal distribution with mean 0 and variance τ, i.e., $v_k \sim$ N(0, τ). E is the error term with $\varepsilon \sim N(0, \sigma^2 I_n)$. $I_n$ is the identity matrix for n individuals and $\sigma^2$ is the variance of the error term. An LMM decomposes the variance associated with phenotype y into the sum of a linear additive genetic and residual component. The variance of Y can be written as:

$$Var(Y) = \tau K + \sigma^2 I_n \tag{3.8}$$

where K is the genetic relationship matrix, or the kinship matrix for related subjects. Because the LMM can account for extra covariance caused by subject relatedness, it can be utilized for association testing in family data as well as data with population stratification. In fact, the LMM is a generalization of the PCA approach in which all PCs are included by default. In short, the

GLM simply corrects population structure; the LMM corrects both population structure and individual kinship relationships.[43]

### 3.4.4 Correcting for Multiple Testing in a GWAS

In a single statistical test, the type I error rate $\alpha$ is the probability of falsely rejecting the null hypothesis when it is true. The p-value for a given hypothesis is determined based on a sample of data and is defined as the probability of observing something as extreme or more extreme, given that the null hypothesis is true. If the p-value is less than $\alpha$ (typically 0.05), then we reject the null hypothesis in favor of the alternative. Formally, for a given null hypothesis denoted $H_0$

$$\text{Type-1 error rate} = P(\text{reject } H_0 \mid H_0 \text{ is true}) \leq \alpha \tag{3.9}$$

In the case where it is desired to test K null hypotheses, given by $H_{0k}$, for k = 1, . . .,K the family-wise error under the complete null (FWEC) is defined as the probability of rejecting one or more of these null hypotheses given that all of them are true. If each test is independent and tested at level a, then

$$\begin{aligned} \text{FWEC} &= P(\text{reject at least one } H_{0k} \mid H_{0k} \text{ is true for all k}) \\ &= 1 - P(\text{reject no } H_{0k} \mid H_{0k} \text{ is true for all k}) \leq 1 - (1-\alpha)^K \end{aligned} \tag{3.9}$$

This ceiling is increasing rapidly. For example, for K = 10 independent trials, FWEC $\leq$ 0.401. That is, if ten independent trials are conducted, each at level a, then the probability of a type 1 error is 40.1%. This phenomenon is referred to as the inflation of the type 1 error rate and is a serious concern in the context of analyzing associations between a large number of SNPs and a trait. [50]

The Bonferroni adjustment for multiple comparisons is perhaps the simplest adjustment that can be applied to address this problem. It simply involves using $\alpha = \alpha/m$ in place of $\alpha$ for the level of each test, where m is the number of tests to be performed. For example, if m = 10 hypothesis tests are performed at a total level of $\alpha = 0.05$, then let $\alpha' = 0.05/10 = 0.005$, so FWEC $\leq 1 - (1 - 0.005)^{10} = 1 - 0.951 = 0.049$. This technique is overly conservative since it presupposes test independence, which is not true for SNPs which are in linkage disequilibrium. Despite this restriction, the Bonferroni adjustment is still a popular GWAS benchmark.[43]

## 3.5 GWAS Results Presentation

The primary output of a GWAS analysis is a list of p-values, effect sizes and their directions generated from the association tests of all tested genetic variants with a phenotype of interest. These data are routinely visualized using Manhattan plots and quantile–quantile plots. Further analysis is then needed to interpret this list of p-values, determining the most likely causal variants, their functional interpretation and possible convergence in meaningful biological pathways.[41]

Association results from GWAS are in the form of lists of summary statistics for millions of genetic variants. Results are "clustered" in blocks corresponding to genetic loci with high LD. For this reason, it is impossible to open the file and browse the results by looking at the summary statistics. The most widely used, recognizable visual tool to explore genome-wide association statistics is the Manhattan plot, which is a type of scatterplot that plots the negative logarithm (base 10) of the association p-value for each genetic variant (y axis) and the chromosome position for each SNP tasted (x axis), where each circle represents a SNP. Manhattan plots are used to visualize GWA significance level by chromosome location. The height of the points in a Manhattan plot is thus inversely related to their p-values. The SNPs shown in the figure are markers, and many will not be the actual causal variant but rather a "tag." In other words, they are tags since nearby variants might actually be driving the association.[58] Due to linkage disequilibrium, typical true signals arise in stacks, formed by neighboring loci in high LD with the causal marker. Figure 3.7 shows a Manhattan plot from a GWAS of 1583 nasopharyngeal carcinoma (NPC) cases and 1894 controls of Chinese descent. The tall stack on chromosome 6 is in the HLA region that has been extensively studied and is known to in duce immune response to the EVB virus infection, which is a major risk factor for NPC.[43]

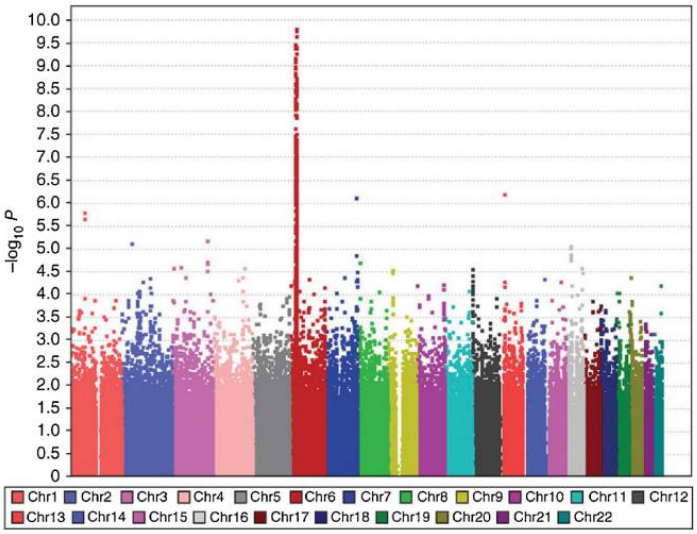


**Figure 3.7**

**Manhattan Plot For GWAS of Nasopharyngeal Carcinoma [43]**

Manhattan plots show the statistical associations of all genetic variants but conceal a considerable amount of important information. Regional associations plots provides additional information regarding chromosome position, genes, recombination rate and linkage disequilibrium levels in a specific genomic region. The x-axis and the y-axis are the same as in a Manhattan plot (genomic position and negative logarithm of association p-value). Figure3.8 is an example of a LocusZoom plot in which the top signal rs1412829 in gene CDNK2A/2B on Chromosome 9 is surrounded by neighboring loci in high LD. The green segments indicate gene regions.

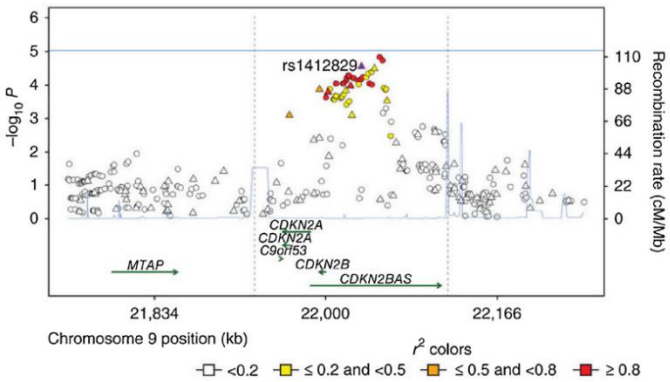


**Figure 3.8**

**LocusZoom Plot[58]**

Another typical figure associated with a GWAS is the Quantile-Quantile (Q-Q) plot, which is examined together with the λ (lambda) statistic. Q-Q plots show the link between the expected and observed distributions of SNP-level test statistics. It is a tool that is used to visualize the appropriate control of population substructure and the presence of an association. Despite meticulous study design and sample collection from a homogeneous cohort, various degrees of population stratification may exist. A blend of unknown or unmanageable ethnicity groups would result in allele frequency differences, inflated variances, and enhanced false relationships. Post-analysis population stratification can be found by visually inspecting the Quantile-Quantile (Q-Q) plot, which compares observed test statistics (or some function of the p-values) to the values that would be obtained from a theoretical distribution. Deviation from the diagonal line suggests the possibility of population stratification and an increase in spurious correlations. The degree of deviation from the line is formally measured by the λ-statistic(genomic control) [43]

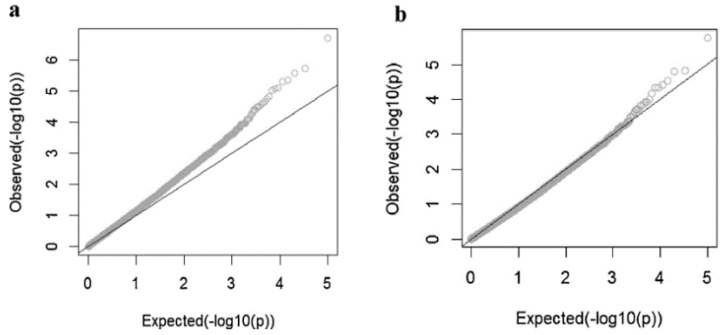


**Figure 3.9**

**a)Q-Q plot of GWAS showing obvious population stratification (λ = 1.14) b) Q-Q plot of normal GWAS corrected for population stratification[41]**

## 3.6 Post GWAS issues

### 3.6.1 Statistical fine-mapping

Because of linkage disequilibrium, many non-causal variations are strongly related with a trait of interest; whether they approach the significance threshold relies on their level of correlation with and intensity of association with the causative variant. As a result, the output of GWAS is grouped in risk loci — groups of correlated variants that all exhibit a statistically significant correlation with the trait of interest — and linkage disequilibrium

often hinders pinpointing causal variations without additional research. Based on observed patterns of linkage disequilibrium and association statistics, fine mapping is an in silico procedure that prioritizes the set of variants most likely to be causal to the target phenotype within each of the genetic loci discovered by GWAS. The most basic fine-mapping analysis is a conditional association analysis of the regional variants (a genetic association analysis that includes fixed effects of genetic variants), which adjusts the regional association signals according to the set of variants in the locus by including the lead variant as a covariate in genotype-phenotype regression models. When there are several association signals, it is typical to apply forward stepwise selection until no relationships remain. This stepwise conditional analysis method is confined to searching all combinatory patterns of potentially credible variations. This is due to the fact that the variant search pattern in each iterative step is heavily dependent on the previously selected variant sets, and the lead initial step frequently includes the lead variant. [41] In Figure 3.10 fine-mapping is applied to identify a set of variants that are likely to include the causal variant (blue box) as well as the most likely causal variant (rs12345; blue dot).

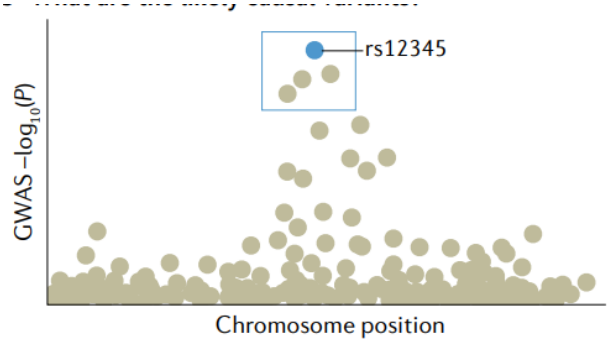


**Figure 3.10 [41]**

## 3.6.2 GWAS Meta-Analysis

Meta-analysis is the statistical synthesis of information from multiple independent studies that increases power and subsequently reduces the risk of false-positive findings. GWAS meta-analyses use what is called summary data, which provide regression coefficients, standard errors, and so on for each genetic marker in a population following a prespecified analysis plan. It is thus not individual-level data but the aggregated summary results. Summary statistics, in a GWAS, is the results obtained after conducting a GWAS, including information on chromosome number, position of the SNP, MAF, effect size (odds ratio/beta), standard error,

and p-value. These statistics are used, for example, to create polygenic scores. Quality Control (QC) is required before conducting a GWAS such as removing variants with low allele frequency, low imputation quality, allele frequency that diverges substantially from a reference sample, or results driven by a specific study that are not replicated elsewhere. An important and time-consuming step in the GWAS meta-analysis is a second set of quality control, which is basically harmonizing the results across studies. Despite providing a unified analysis plan, this cleaning process might take the longest time in an initial project, since analysts might use different software or there are other inconsistencies in the results.[58] Meta-analyses can be performed using a fixed effect model — which assumes error variances are equal across cohorts — or a random effect model to test for heterogeneity in the results; for example, testing whether one or two cohorts clearly deviate from the rest. Combining the contributions of all cohorts allows for a more precise estimation of effect sizes and the significance of effects in GWAS by weighting each individual cohort's results by their sample size or by using the inverse variance method. Sequencing data sets can identify rare variants, although current sequencing data sets are typically too underpowered to test their effects on a phenotype individually; instead, their effects are usually measured in aggregate, such as in genes or gene sets through rare variant burden testing.

## 3.7 GWAS Studies

### 3.7.1 Study 1

Widmer *et al.,* examined improvements to the linear mixed model (LMM) that better correct for population structure and familial relatedness in genome-wide association studies (GWAS). LMMs are based on the estimation of a genetic similarity matrix (GSM), which encodes the pairwise similarity between any two individuals in a cohort. These similarities are estimated from single nucleotide polymorphisms (SNPs) or other genetic variants. Traditionally, when an LMM is used for GWAS, its GSM is estimated from all available SNPs. In this paper, Widmer *et al.* evaluate possible improvements to this approach. They discovered that modifying this approach improves GWAS performance as measured by type I error checking and power. Specifically, when there is only population structure, a GSM constructed from SNPs that predict the phenotype well in combination with principal components as covariates controls type I error and yields higher power than the traditional LMM. In any setting, with or without population structure or family relatedness, a GSM consists of a mixture of two GSM components, one

constructed from all SNPs and the other constructed from SNPs that predict the phenotype well. This again controls for type I error and yields more power than the traditional LMM.[50]

Synthetic SNPs and phenotypes: To reveal the weaknesses of the different models tested, they varied the degree of population structure, family structure, number of causal SNPs, and signal strength across a wide range of possible parameters, including those that yield strong confounding by population structure and family relatedness. Each data set was created with M = 50,000 SNPs and N = 4,000 individuals, as is typical of many GWASs.[50]

The models used to carry out the comparison are presented in the table below.

| Name of model | Model description |
|---|---|
| Linreg | Linear regression |
| LMM(all) | LMM with GSM based on all SNPs |
| LMM(select) | LMM with GSM based on SNP selection |
| LMM(select) + PCs | LMM(select) with PCs added as fixed effects |
| LMM(all + select) | LMM with a mixture of two GSMs |

**Figure 3.11**

**Models Table[50]**

Where the LMM(all + select) model is a new LMM model having a GSM made up of a mixture of two GSMs ($(1-\pi) K_0 + \pi K_1$), one based on all SNPs ($K_0$) and one based on SNP selection ($K_1$).

No population or family relatedness [Three data sets were generated for each possible combination of parameters (number of SNPs and narrow-sense heritability for causal signal), totaling 90 data sets.] SNPs were chosen to maximize phenotypic prediction accuracy. SNPs were identified in particular by searching through multiple sets of SNPs to find those that maximized out-of-sample prediction accuracy as measured by the log likelihood of the phenotype under the LMM. It measured the empirical type I error rate (the proportion of non-causal SNPs deemed significant) as a function of P-value threshold and the empirical power (the proportion of causal SNPs deemed significant) as a function of empirical type I error for each of the three models. The results are shown for various numbers of causal SNPs. Type I error was well controlled by all models. Furthermore, LMM (select) had the greatest power, particularly when the number of causal SNPs was small (and thus the effect sizes were large). It is not surprising that LMM(select) had greater power than Linreg when considering the LMM as linear regression with selected SNPs as covariates. Conditioning on specific SNPs, in other

words, reduces noise in the phenotype. It is also expected that LMM(select) had more power than LMM(all) when there were few causal SNPs, as the use of all SNPs in the GSM obfuscated the true causal signal, a phenomenon known as "dilution."[50]

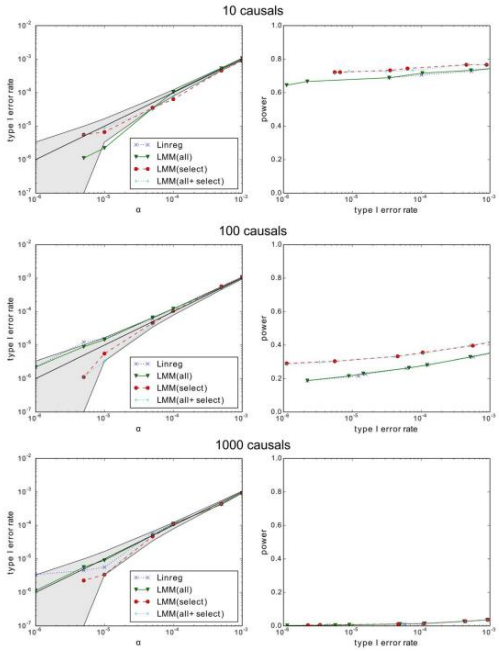


**Figure 3.12**

**Empirical type I error rate and power for no population or family relatedness with purely synthetic data. [50]**

Population structure but no family relatedness [Parameter values used in these simulations were as follows: Number of causal SNPs, Narrow-sense heritability, Degree of population structure Three data sets for each possible combination of parameters were generated, yielding 360 data sets]

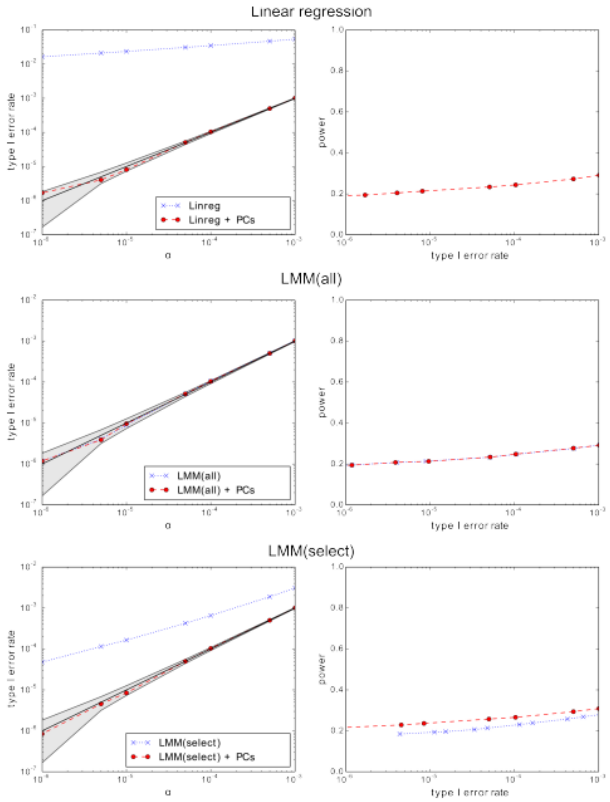


**Figure 3.13**

**Empirical type I error rate and power for population structure but no family relatedness with purely synthetic data.[50]**

The inclusion of PCs had differing effects on the performance of the models (Figure 3.13). LMM(all) controlled type I error well, whether or not PCs were included as fixed effects, and inclusion did not affect power. In contrast, for Linreg, inclusion of PCs led to control of type I error and had little effect on power. Furthermore, the inclusion of PCs led to control of type I error and improved power for LMM(select), as was recently reported in an independent investigation.

Population structure and family relatedness [The parameter values used in these simulations were as follows: Narrow-sense heritability, population structure, family relatedness, and the fraction of individuals belonging to a family. Three data sets for each possible combination of parameters were generated, yielding 1800 data sets in all. No two sets of SNPs were the same]

The model LMM(all + select), which performed best for the setting of family relatedness without population structure, also performed best here (Figure 3.14). These results indicate that the inclusion of all SNPs as part of the mixture GSM led to good control of type I error for both

forms of confounding structure, consistent with our findings for family relatedness alone and population structure alone. Furthermore, the inclusion of selected SNPs as part of the mixture GSM led to improved power, again most notably so when there were a small number of causal SNPs with large effect size (Figure 3.14b). Also on purely synthetic data with population structure but no family relatedness, it was found that LMM(select) yielded better GWAS performance than LMM(all), but only when PCs were used as covariates.[50]

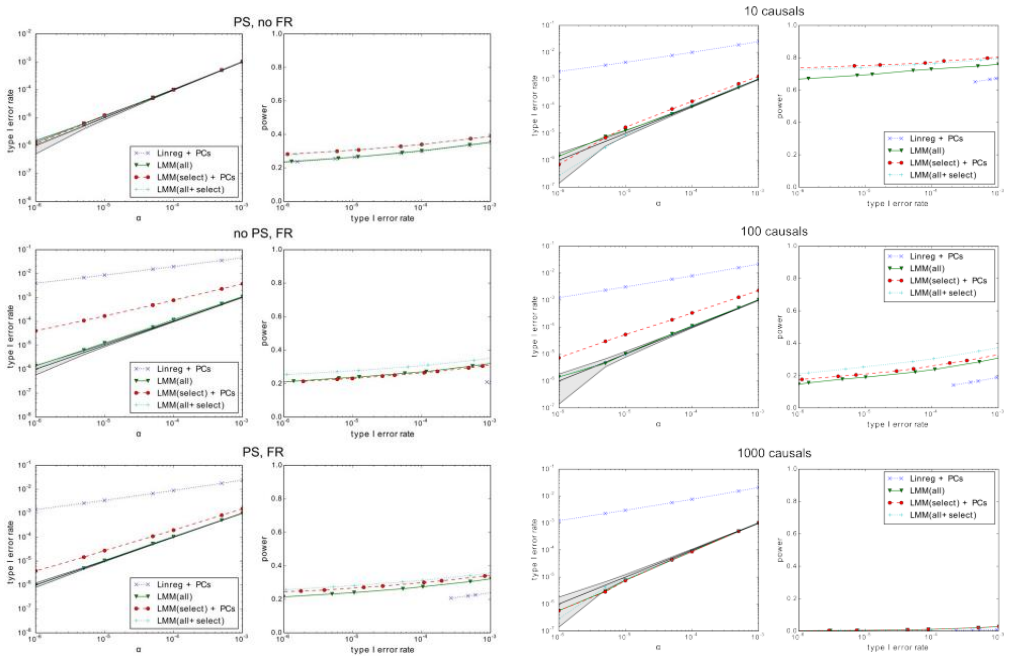


**Figure 3.14**

**a) Empirical type I error rate and power with and without population structure (PS) and family relatedness (FR) b) Empirical type I error rate and power for both family relatedness and population structure with purely synthetic data [50]**

In summary, one potential improvement, building a GSM based on selected SNPs that well predict the phenotype failed rather dramatically. In particular, when population structure, family relatedness, or both were present, this approach failed to control for type I error. Nonetheless, when SNP selection was used in combination with other improvements, it proved useful. Specifically, in the presence of population structure alone, SNP selection in combination with PCs used as covariates-controlled type I error and also yielded more power than the traditional approach. In all settings, with or without population structure or family relatedness, a mixture of two GSMs, one constructed from all SNPs and another constructed from SNPs identified by SNP selection both controlled type I error and yielded more power than the

traditional LMM. Furthermore, the improvements to power afforded by SNP selection were the strongest when some SNPs had a large effect size. Interestingly, we found that a GSM based on all SNPs (or LDsampled SNPs) could account for population structure just as well as PCs. Consequently, if SNP selection picks all SNPs, then there is no need to add PCs to the LMM.[50]

### 3.7.2 Study 2

Recently, genome-wide analysis has identified variants in five chromosomal regions that are significantly associated with a risk of prostate cancer. These variants occur in three independent regions at 8q244-7 and in one region at 17q12 and another at 17q24.3.8 These five regions probably harbor genes that confer susceptibility to prostate cancer or regulate factors affecting critical genes, but the specific genes in these regions have not been identified. Individually, single-nucleotide polymorphisms (SNPs) in each of the five chromosomal regions were shown to have only a moderate association with prostate cancer in previous studies. In this study, Zheng, S. L. *et al.*, investigated whether a combination of SNPs would have a stronger association with prostate cancer than any individual SNP. For this purpose, they assessed the joint associations of SNPs in the five chromosomal regions with prostate cancer in a large-scale study of Swedish men.[57]

Among 3648 identified subjects with prostate cancer, 3161 (87%) agreed to participate. DNA samples from blood, tumor–node–metastasis (TNM) stage, Gleason grade (as determined by biopsy), and levels of prostate-specific antigen (PSA) at diagnosis were available for 2893 subjects (92%). Case subjects were classified as having advanced disease if they met any of the following criteria: a grade 3 or 4 tumor, spread to nearby lymph nodes and metastasis, a Gleason score of 8 or more, or a PSA level of more than 50 ng per milliliter; otherwise, subjects were classified as having localized disease. Control subjects, who were recruited concurrently with case subjects, were randomly selected from the Swedish Population Registry and matched according to the expected age distribution of cases (groups of 5-year intervals) and geographic region. A total of 2149 of 3153 control subjects (68%) who were invited subsequently agreed to participate in the study. DNA samples from blood were available for 1781 control subjects (83%). Serum PSA levels were measured for all control subjects but were not used as an exclusionary variable. A history of prostate cancer among first-degree relatives was obtained from a questionnaire for both case subjects and control subjects.[57]

They selected 16 SNPs from five chromosomal regions (three at 8q24 and one each at 17q12 and 17q24.3) that have been reported to be associated with prostate cancer. Tests for Hardy–Weinberg equilibrium($p>0.05$) were performed for each SNP separately among case subjects and control subjects with the use of Fisher's exact test. Pairwise linkage disequilibrium was tested for SNPs within each of the five chromosomal regions in control subjects. Differences in

allele frequencies between case subjects and control subjects were tested for each SNP with the use of a chi-square test with 1 degree of freedom. For genotypes, a series of tests assuming an additive, dominant, or recessive genetic model were performed for each of the 16 SNPs with the use of unconditional logistic regression with adjustment for age and geographic region; the model that had the highest likelihood was considered to be the best-fitting genetic model for the respective SNP.[57]

They tested the independent effect of each of the five previously implicated regions by including the most significant SNP from each of the five regions in a logistic-regression model with the use of a backward-selection procedure. Multiplicative interactions were tested for each pair of SNPs by including both main effects and an interaction term (a product of two main effects) in a logistic regression model. They tested the cumulative effects of the five SNPs on prostate cancer by counting the number of genotypes associated with prostate cancer (on the basis of the best-fitting genetic model from single-SNP analysis) for these five SNPs in each subject. The odds ratio for prostate cancer for men carrying any combination of one, two, three, or four or more genotypes associated with prostate cancer was estimated by comparing them with men carrying none of the prostate cancer–associated genotypes with the use of logistic regression analysis. They also performed tests for the cumulative effect on prostate-cancer association, which included five SNPs and family history.

Associations of these five SNPs with aggressiveness of prostate cancer (advanced or localized), and family history (yes or no) were tested only among case subjects with the use of a chi-square test of a 2×K table, in which K is the number of possible categories within each variable. Associations of SNPs with the mean age at diagnosis were tested only among case subjects with the use of a two-sample t-test. Because serum PSA levels were not normally distributed, a nonparametric analysis (Wilcoxon rank sum test) was used to assess the association between SNPs and preoperative serum PSA levels in case subjects or PSA levels at the time of sampling in control subjects. All reported p-values are based on a two-sided test.[57]

Results: Significantly different frequencies ($p<0.05$) between case and control subjects were observed for SNPs in each of the five chromosomal regions. At 17q12, SNP rs4430796 had the strongest association with prostate cancer; the frequency of allele T (SNP rs4430796) was 0.61 in case subjects and 0.56 in control subjects($p=6\times10^{-7}$). Of the four SNPs at 17q24.3, three were associated with prostate cancer, but only rs1859962 had a highly significant association

($P=2.1\times10^{-4}$). For SNPs at 8q24, significant associations with prostate cancer were found for all SNPs examined across the three independent regions at 8q24. Of the 16 SNPs, 13 remained significant at $p<0.05$ after adjustment for 16 tests with the use of a Bonferroni correction.

| SNP | Chromosomal Region | Position† | Alternative Alleles | Allelic Tests | | | | | Best-Fitting Genetic Model‡ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Associated Allele§ | Frequency | | Odds Ratio (95% CI)¶ | P Value | Model | Genotype‖ | | Odds Ratio (95% CI) | P Value** |
| | | | | | case subjects | control subjects | | | | reference | associated | | |
| rs4430796 | 17q12 | 33,172,153 | T, C | T | 0.61 | 0.56 | 1.24 (1.14–1.36) | $6.0\times10^{-7}$ | Recessive | CC or TC | TT | 1.40 (1.23–1.59) | $2.68\times10^{-7}$ |
| rs7501939 | 17q12 | 33,175,269 | G, A | G | 0.66 | 0.62 | 1.22 (1.12–1.33) | $9.0\times10^{-6}$ | Recessive | AA or GA | GG | 1.33 (1.17–1.50) | $5.54\times10^{-6}$ |
| rs3760511 | 17q12 | 33,180,426 | A, C | C | 0.41 | 0.38 | 1.17 (1.07–1.27) | $5.0\times10^{-4}$ | Recessive | AA or CA | CC | 1.42 (1.20–1.68) | $4.47\times10^{-5}$ |
| rs1859962 | 17q24.3 | 66,620,348 | G, T | G | 0.54 | 0.50 | 1.17 (1.08–1.28) | $2.1\times10^{-4}$ | Recessive | GT or TT | GG | 1.28 (1.12–1.46) | $3.54\times10^{-4}$ |
| rs7214479 | 17q24.3 | 66,702,544 | C, T | T | 0.50 | 0.48 | 1.08 (0.99–1.18) | 0.07 | Recessive | CC or CT | TT | 1.15 (1.00–1.32) | 0.06 |
| rs6501455 | 17q24.3 | 66,713,406 | A, G | A | 0.56 | 0.54 | 1.09 (1.00–1.19) | 0.05 | Recessive | AG or GG | AA | 1.13 (0.99–1.29) | 0.06 |
| rs983085 | 17q24.3 | 66,723,656 | A, G | A | 0.57 | 0.55 | 1.07 (0.98–1.16) | 0.13 | Recessive | GA or GG | AA | 1.11 (0.97–1.26) | 0.12 |
| rs6983561 | 8q24 (region 2) | 128,176,062 | A, C | C | 0.06 | 0.03 | 1.65 (1.33–2.05) | $4.2\times10^{-6}$ | Dominant | AA | CA or CC | 1.60 (1.28–2.00) | $2.14\times10^{-5}$ |
| rs16901979 | 8q24 (region 2) | 128,194,098 | C, A | A | 0.06 | 0.03 | 1.65 (1.33–2.05) | $4.3\times10^{-6}$ | Dominant | CC | AA or CA | 1.60 (1.28–2.01) | $2.14\times10^{-5}$ |
| rs6983267 | 8q24 (region 3) | 128,482,487 | G, T | G | 0.56 | 0.51 | 1.22 (1.12–1.33) | $3.9\times10^{-6}$ | Dominant | TT | GT or GG | 1.38 (1.19–1.59) | $1.74\times10^{-5}$ |
| rs7000448 | 8q24 (region 3) | 128,510,352 | C, T | T | 0.43 | 0.40 | 1.15 (1.06–1.25) | $1.4\times10^{-3}$ | Dominant | CC | CT or TT | 1.18 (1.04–1.33) | $1.21\times10^{-2}$ |
| rs1447295 | 8q24 (region 1) | 128,554,220 | C, A | A | 0.17 | 0.14 | 1.21 (1.07–1.36) | $1.6\times10^{-3}$ | Dominant | CC | CA or AA | 1.26 (1.10–1.44) | $8.27\times10^{-4}$ |
| rs4242382 | 8q24 (region 1) | 128,586,755 | G, A | A | 0.16 | 0.14 | 1.24 (1.10–1.39) | $5.3\times10^{-4}$ | Dominant | GG | AG or AA | 1.29 (1.12–1.47) | $2.53\times10^{-4}$ |
| rs7017300 | 8q24 (region 1) | 128,594,450 | A, C | C | 0.20 | 0.18 | 1.15 (1.03–1.28) | 0.01 | Dominant | AA | CA or CC | 1.20 (1.05–1.36) | $6.20\times10^{-3}$ |
| rs10090154 | 8q24 (region 1) | 128,601,319 | C, T | T | 0.16 | 0.13 | 1.26 (1.11–1.42) | $2.0\times10^{-4}$ | Dominant | CC | CT or TT | 1.31 (1.14–1.50) | $1.03\times10^{-4}$ |
| rs7837688 | 8q24 (region 1) | 128,608,542 | G, T | T | 0.15 | 0.13 | 1.17 (1.04–1.13) | $9.6\times10^{-3}$ | Dominant | GG | GT or TT | 1.21 (1.06–1.39) | $5.87\times10^{-3}$ |

* CI denotes confidence interval, and SNP single-nucleotide polymorphism.
† The position is based on the National Center for Biotechnology Information database, build 35.
‡ The best-fitting model for each SNP was determined after testing associations of a series of genetic models, including dominant and recessive models, with prostate cancer.
§ These alleles were reported to be associated with prostate cancer in studies published previously.[4-8,10]
¶ Allelic odds ratios are based on the multiplicative model.
‖ Reference genotypes and those associated with prostate cancer for each SNP were defined on the basis of the best-fitting genetic model.
** P values are two-sided and were calculated by the likelihood-ratio test with one degree of freedom, adjusted for age and geographic region.

**Figure 3.15**

**Association of SNPs at Five Chromosomal Regions with Prostate Cancer[57]**

Strong genetic dependence (linkage disequilibrium) among SNPs within each region allowed for a combined analysis in which we were able to select one SNP (the most significant SNP from single SNP analysis) to represent each of the five regions in tests for an independent association with prostate cancer (Figure 3.16). When these five SNPs were included in a multivariate logistic regression model, each of the five remained significantly associated with prostate cancer after adjustment for other SNPs, and each continued to be highly significant when family history was included in the model.

| Variable or SNP† | Chromosomal Region | Alternative Alleles | Reference | | Frequency of Associated Factors‡ | | Regression Coefficient | Odds Ratio (95% CI) | P Value§ | PAR |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Case Subjects | Control Subjects | | | | % |
| Age | | | | | | | 0.01 | 1.01 (1.00–1.02) | 0.02 | |
| Geographic region | | | | | | | −0.77 | 0.46 (0.39–0.54) | <0.001 | |
| Family history | | | No | Yes | 0.19 | 0.09 | 0.80 | 2.22 (1.83–2.68) | $1.15\times10^{-17}$ | 9.89 |
| rs4430796 | 17q12 | T, C | CC/TC | TT | 0.38 | 0.30 | 0.32 | 1.38 (1.21–1.57) | $1.62\times10^{-6}$ | 10.23 |
| rs1859962 | 17q24.3 | G, T | GT/TT | GG | 0.30 | 0.25 | 0.24 | 1.28 (1.11–1.47) | $5.49\times10^{-4}$ | 6.54 |
| rs16901979 | 8q24 (region 2) | C, A | CC | AA/CA | 0.10 | 0.07 | 0.42 | 1.53 (1.22–1.92) | $1.83\times10^{-4}$ | 3.58 |
| rs6983267 | 8q24 (region 3) | G, T | TT | GT/GG | 0.82 | 0.77 | 0.32 | 1.37 (1.18–1.59) | $3.44\times10^{-5}$ | 22.17 |
| rs1447295 | 8q24 (region 1) | C, A | CC | CA/AA | 0.31 | 0.26 | 0.19 | 1.22 (1.06–1.40) | $5.31\times10^{-3}$ | 5.41 |
| All five SNPs | | | | | | | | | | 40.45 |
| All five SNPs and family history | | | | | | | | | | 46.34 |

* CI denotes confidence interval, PAR population attributable risk, and SNP single-nucleotide polymorphism.
† A family history of prostate cancer and five SNPs were included in the multivariate logistic-regression model with adjustment for age and geographic region.
‡ For SNPs, the reference genotype and those associated with prostate cancer at each SNP were determined on the basis of the best-fitting model after testing associations of a series of genetic models with prostate cancer.
§ P values were calculated by the likelihood-ratio test.

**Table 3.16**

**Adjusted Odds Ratios for Representative SNPs at Five Chromosomal Regions and Family History [57]**

When multiplicative interaction was tested for each possible pair of these five SNPs with the use of an interaction term in logistic regression, none were significant at $P<0.05$. However, the five SNPs appeared to have a cumulative association with prostate cancer, after adjustment for age, geographic region, and family history (Figure 3.17). Men who carried one, two, three, or four or more of the five SNPs had an increasing likelihood of having prostate cancer, as compared with men who did not carry any of the five SNPs (p-value for trend, $6.75\times10^{-27}$). When family history was included as another risk factor (coded as 0 or 1) for a total of six possible prostate-cancer associated factors, they observed a stronger cumulative effect after adjustment for age and geographic region (p-value for trend, $4.78\times10^{-28}$). For example, men who carried any five or more of these six factors had an odds ratio of 9.46 (95% confidence interval [CI], 3.62 to 24.72) for prostate cancer, as compared with men who carried none of the six factors ($P=1.29\times10^{-8}$).[57]

| Variable | Case Subjects | Control Subjects | Regression Coefficient | Odds Ratio (95% CI) | P Value† | P Value for Trend‡ |
|---|---|---|---|---|---|---|
| | *no. of subjects (%)* | | | | | |
| **Genotypes at five SNPs§** | | | | | | |
| Age | | | 0.01 | 1.01 (1.00–1.02) | 0.02 | |
| Geographic region | | | −0.76 | 0.46 (0.40–0.55) | <0.001 | |
| Family history | | | 0.8 | 2.22 (1.83–2.68) | $7.73\times10^{-18}$ | |
| No. of associated genotypes¶ | | | | | | |
| 0 | 162 (5.6) | 173 (10.1) | NA | 1.00 | | |
| 1 | 883 (30.8) | 631 (36.8) | 0.41 | 1.50 (1.18–1.92) | $9.46\times10^{-4}$ | |
| 2 | 1123 (39.1) | 618 (36.0) | 0.67 | 1.96 (1.54–2.49) | $4.19\times10^{-8}$ | |
| 3 | 548 (19.1) | 255 (14.9) | 0.79 | 2.21 (1.70–2.89) | $4.33\times10^{-9}$ | |
| ≥4 | 154 (5.4) | 38 (2.2) | 1.5 | 4.47 (2.93–6.80) | $1.20\times10^{-13}$ | $6.75\times10^{-27}$ |
| **Genotypes at five SNPs and family history‖** | | | | | | |
| Age | | | 0.01 | 1.01 (1.00–1.02) | 0.02 | |
| Geographic region | | | −0.75 | 0.47 (0.40–0.55) | <0.001 | |
| No. of associated factors** | | | | | | |
| 0 | 144 (5.0) | 174 (10.1) | NA | 1.00 | | |
| 1 | 778 (26.9) | 581 (33.6) | 0.48 | 1.62 (1.27–2.08) | $1.27\times10^{-4}$ | |
| 2 | 1053 (36.4) | 622 (36.0) | 0.73 | 2.07 (1.62–2.64) | $5.86\times10^{-9}$ | |
| 3 | 642 (22.2) | 286 (16.6) | 0.99 | 2.71 (2.08–3.53) | $9.54\times10^{-14}$ | |
| 4 | 236 (8.2) | 60 (3.5) | 1.56 | 4.76 (3.31–6.84) | $9.17\times10^{-19}$ | |
| ≥5 | 40 (1.4) | 5 (0.3) | 2.24 | 9.46 (3.62–24.72) | $1.29\times10^{-8}$ | $4.78\times10^{-28}$ |

**Figure 3.17**

**Cumulative Effect of Associated Factors on the Risk of Prostate Cancer[57]**

They calculated the specificity and sensitivity of the regression model by constructing receiver operating-characteristic (ROC) curves and calculated statistics for the area under the curve (AUC) to estimate the ability of each of three models to distinguish case subjects from control subjects. The AUC was 57.7 (95% CI, 56.0 to 59.3) for model 1 (age and region alone), 60.8 (95% CI, 59.1 to 62.4) for model 2 (age, region, and family history), and 63.3 (95% CI, 61.7 to 65.0) for model 3 (age, region, family history, and the number of genotypes associated with prostate cancer at the five SNPs). The AUC was significantly higher for model 3 than for model 2 ($P=6.12\times10^{-6}$).[57]

Zheng, S. L. *et al.*, found that the presence of the five prostate-cancer–associated SNPs was independent of PSA levels in both case subjects and control subjects, which suggests that some men with low PSA levels may have an increased risk of prostate cancer if they carry one or more of the prostate-cancer–associated genotypes described here. However, this proposition also requires testing in a prospective trial, particularly one that uses PSA in combination with the associated SNPs and family history. They do not know the mechanism by which the SNPs we analyzed could affect the risk of prostate cancer. Other than SNP rs4430796, which is located within the TCF2 gene, the specific genes that are affected by the rest of the SNPs have not been identified. Since the five SNPs in our study appear to be associated with a risk of prostate cancer in general, rather than with a more or less aggressive form, we suspect that the genetic variants act at an early stage of carcinogenesis. This study is only a first step toward

defining a genetic association with prostate cancer in populations. Future investigations will need to test the value of these findings in assessing the risk of prostate cancer in individual men.[57]

# CHAPTER 4

# POLYGENIC RISK SCORE (PRS)

## 4.1 Introduction

GWAS have made clear that only a very small proportion of the total genetic contribution can be unambiguously attributed to variation in particular loci of the genome. Most such genetic contributions are thus spread across the huge landscape of the genome, with many loci each contributing a small, almost undetectable effect on the phenotypes. To date, GWAS have identified thousands of loci that are associated with a range of complex human traits and diseases, including cardiovascular diseases, cancers, obesity and Alzheimer's disease. These data have provided numerous insights into the genes and pathways that cause disease, but more recently the use of these data for disease risk prediction has gained interest. Many common, complex diseases now have numerous, well-established risk loci and likely harbor many genetic determinants with effects too small to be detected at genome-wide levels of statistical significance. A simple and intuitive approach for converting genetic data to a predictive measure of disease susceptibility is to aggregate the effects of these loci into a single measure, the polygenetic risk score (PRS). The genetic architecture of most phenotypes and health conditions is polygenic in nature. With the growth of genome-wide association studies (GWASs) and larger samples, PRSs have increasingly emerged as a major tool in several areas of quantitative genetic research. [58] This approach is particularly valuable for complex traits that lack common risk variants of large effect, including schizophrenia and height. The predictive power of a PRS is limited by the number of SNPs tested and a trait's heritability and prevalence, but theoretically can be high. In practice, in some situations, a PRS may identify high-risk individuals and can identify risk classes that could inform a range of treatment options. [1]

A polygenic risk score (PRS) is a numeric summary of the relationship between multiple genetic loci and a phenotype. A PRS estimates the genetic risk of an individual for some disease or trait, calculated by aggregating the effects of many common variants in the genome, each of which can have a small effect on a person's genetic risk for a given disease or condition [4]. PRS analyses aim to provide insight into the genetic architecture using evidence for association from variants that do not pass the stringent threshold of association. As the threshold of sample

dataset p-value increases, the number of SNPs included in the PRS also increases, and hence the ratio of false/true positives increases. [18] PRS can also be used to determine the presence of a genetic signal in underpowered studies, to infer the genetic architecture of a trait, for screening in clinical trials, and as a biomarker.[58]

## 4.2 PRS theory

Polygenic scores are derived directly from the genome-wide associations in GWASs. Using the summary statistics from these to construct an estimate of how SNPs combine to explain the trait of interest. With the increasing availability of genetic data in large cohort studies such as the UK Biobank, inclusion of this genetic risk as a covariate in statistical analyses is becoming more widespread [4]. The purpose of risk scores is twofold: (1) to predict the likelihood of an individual developing disease, a reaction to a drug or a particular outcome of interest based on some amount of available information, usually genetic, clinical, demographic, or a combination, and (2) to estimate the level of predictive power that is captured by associated variants. Predicting a greater proportion of the "risk" for the outcome of interest indicates the level of success of predictors included in the risk score. A PRS may estimate the overall likelihood, or risk, that an individual has of developing an outcome of interest based on the genotypes and variants identified as being associated with that outcome. Because an individual's genetic profile is set at birth, and therefore because risk for disease could theoretically be determined prior to (most) environmental exposures, a great deal of hope has been invested in developing these models as an advancement of precision medicine. Family history is typically seen as a good proxy for genetic risk as it reflects shared genetic and environmental factors and thus is incorporated into clinical history, when possible, for genetic diseases. Furthermore, a positive family history reflects a certain level of disease risk, while a negative family history does not imply the opposite. One goal of implementing the PRS is to improve upon these factors for a more comprehensive and accurate assessment of disease risk beyond what family history can estimate. A PRS can be based solely on available genetic data or can incorporate environmental, phenotypic, and/or demographic information. [1] As genetic factors capture only the genetic contribution to risk and as PGSs capture only part of the genetic risk, PGSs cannot be diagnostically accurate risk predictors. Nonetheless, for many common complex genetic disorders, such as cancers and heart disease, there is increasing interest in evaluating PGSs for early disease detection, prevention, and intervention [8]

PRS analyses can be characterized by the two key input data sets that they require: (i) base data (GWAS), which consists of summary statistics (e.g., betas and p-values) of genotype-phenotype associations at genetic variants (SNPs) genome-wide, and (ii) target data, which consists of genotypes and, in most cases, phenotypes, and should be independent of the GWAS sample to avoid additional bias and overfitting. It is in the target sample that the PRS analyses were performed, which may involve merely computing PRSs in all the target individuals, conducting association testing between the PRSs and phenotypes or outcomes of interest, or predicting individuals' risk of disease or medication side effects in clinical settings. It is important to distinguish between base and target data to avoid overfitting. Overfitting can be defined as fitting a model too closely to one set of data, greatly limiting its predictive ability in external data. Often, an overfit model will reflect effects beyond true biological effects, such as random noise or population-specific effects. PRS weights are generated in base, or training, data. The standard approach to choosing weights involves using GWAS summary statistics. The natural logarithm of the odds ratio ($\ln [OR] = \beta$) is the common selection and is considered as the $\beta$ for each copy of each SNP. These effect estimates can carry either risk ($OR > 1$, $\beta > 0$) or a protective effect ($OR < 1$, $\beta < 0$). Test, or target data is the genomic data to which the weights from the base data are applied. Only SNPs that are included in both the base and target data with strand agreement will be considered in generation of the PRS. In general, target and base data should not include any of the same participants, in order to, again, avoid overfitting by biasing the sets used to generate the weights.[53] It is also important to consider overlap in participants between the base and target data. This must take into account not only individuals that may be present in both the base and target data, but also the potential existence of close relatives between the datasets. Because of this, it is ideal to remove these close relatives or choose alternative base or target data to avoid such relationships. Failure to do so can lead to "overfitting" of the PRS model by capturing environmental or behavioral effects that are due to similar upbringing and exposures among relatives. If the parameters of the PRS calculation have not been previously optimized, then the target sample can be used both for this optimization and for the analysis, as long as careful cross-validation or permutation procedures are applied. Ideally, analysis is also performed on an independent validation sample to ensure the generalizability of results. [5] Using an independent validation dataset allows unbiased estimation of the predictive performance, avoiding optimism due to overfitting. Generally, once predictive performance plateaus or declines in the validation set, the optimal trade-off of signal

and noise has been reached.[52] To find an appropriate dataset, ensure that the data is phenotypically relevant. It is important to note that the target sample, should not be too different from the base sample. The base sample is the sample or the collection of studies that have been used to calculate the original summary statistics of the GWAS. Also, ancestry composition should not differ too much between the base and the target sample. If allele frequencies of the SNPs used in the score differ too much between the two samples, this will result in a very imprecise score that cannot be used for any further analysis, even for highly heritable traits. GWAS genotypes in a PRS discovery sample may not be representative of those in the validation or application set leading to attenuated performance of the PRS. [58] In practice, PRSs are linear combinations of the phenotype-associated alleles across the genome, typically weighted by GWAS effect sizes. It is thus a single quantitative measure that can be interpreted as a measure of an individual's genetic propensity toward a phenotype relative to a population. [58] In general, PRS for an individual defined as the weighted sum of a person's genotypes at M loci. A PRS for individual *i* can be calculated as the sum of the allele counts $\alpha_{ij}$ (0, 1, or 2) for each SNP $j$ = l, ... $M$, multiplied by a weight $\beta_j$

$$\text{PRS}_i = \sum_{j=1}^{M} a_{ij} \tag{5.1}$$

where the weights (or effect sizes : the increase in the trait value (usually reported as a beta) or disease risk (usually reported as an OR) associated with each additional copy of the risk allele) $\beta_j$ are transformations of GWAS coefficients [58] (the log odds ratio or the estimated regression coefficient from a linear or logistic regression) [13] It is also important to underline that in calculating PRSs on a binary (e.g., case/control) phenotype, the effect sizes used as weights are typically reported as log Odds Ratios (log (ORs)). Assuming that relative risks on a disease accumulate on a multiplicative rather than an additive scale, then PRSs should be computed as a summation of log (OR)-weighted genotypes. PRS values are computed in relation to a hypothetical individual with the non-effect allele at every SNP, and, thus, they provide only a relative (compared to other individuals) estimate of risk (or trait effect) rather than an absolute estimate.[5] Equation 1.1 shows that it is a linear combination of the effects of multiple SNPs on phenotype. The underlying model in a PRS is also usually additive, since it is measured the number of "risk alleles" for each SNP included in the score. However, recessive or dominant models can also be used in the construction of a PRS. Typically, these scores include hundreds-to-thousands of SNPs, motivated by theory and data showing that many diseases are polygenic.

In this way, PRS aggregate the contribution of an individual's germline genome into a single number proportional to the risk for a given disease.[52]

The resulting score is approximately normally distributed in the general population, with higher scores indicating higher risk.[4] The central limit theorem dictates that if a PRS is based on a sum of independent variables (here, SNPs) with identical distributions, then the PRS of a sample should approximate the normal (Gaussian) distribution. This is true even if the PRS has extremely low predictive accuracy, since the sum of random numbers is approximately normally distributed, and so a normally distributed PRS in a sample should not be considered as validation of the accuracy of a PRS. However, strong violations of these assumptions, such as the use of many correlated SNPs or a sample of heterogenous ancestry (thus, SNPs with markedly different genotype distributions), can lead to non-normal PRS distributions. Thus, inspection of PRS distributions may highlight calculation errors or problems of population stratification in the target sample for which researchers did not adequately control.[5] An additional assumption is the absence of gene-gene interactions (or epistasis) since SNP effects are assumed to be independent. In order to create a PRS, required summary statistics that are calculated from a GWAS of the trait of interest and the individual-level genotype data in which you would like to apply your PRS. The GWAS summary statistics should not include the same data that are used for calculating the PRS, which would introduce additional bias leading to overfitting. [58]

There is a benefit to adding PRS to existing clinical risk scores, the unique characteristics of PRS open up possibilities for earlier prevention. Indeed, a study to predict the development of T1D in high-risk children (family history of T1D) found that a PRS was only predictive of progression to T1D before any metabolic abnormalities were present (high DPT-1 score), indicating the value of a T1D PRS for predicting those likely to progress to disease. For cardiovascular disease, traditional risk factors are typically not measured early in life and can have substantial temporal variation. In contrast, individuals can be genotyped early in life, and have their PRS for a wide range of complex diseases. For those at substantially increased lifetime risk of disease, but without elevated traditional risk factors, targeted lifestyle interventions could be used to reduce their risk, for example, by more frequent follow-ups or more stringent targets for traditional risk factors (e.g., cholesterol) [52]

## 4.3 Quality Control

The power and validity of PRS analyses are dependent on the quality of the base and target data. Both data sets must be subjected to QC to at least the standards used in GWAS studies (as described in chapter 3), while numerous QC issues unique to PRS analyses require special attention and are summarized below:

Heritability check. A critical factor in the accuracy and predictive power of PRSs is the power of the base (GWAS) data, and so to avoid reaching misleading conclusions from the application of PRSs, it is recommended to perform PRS analyses only that use GWAS data with a $h^2_{SNP} > 0.05$. [5]

Effect allele. Some GWAS results files do not make clear which allele is the effect allele and which is the non-effect allele. If an incorrect assumption is made in computing the PRS, then the effect of the PRS in the target data will be in the wrong direction, and so to avoid misleading conclusions, it is critical that the effect allele from the base (GWAS) data is known. [5]

Target sample size. It is recommended to perform PRS analyses that involve association testing on target sample sizes of ≥100 individuals and caution against analyses that utilize base data with low $h^2_{SNP}$ and small target sample size. This is to minimize the generation of misleading results due to the less stringent QC feasible on small samples.

File transfer. Since most base GWAS data are downloaded online, and base/ target data transferred internally, one should ensure that files have not been corrupted during transfer. Corrupt files can generate PRS calculation errors. [5]

Genome build. Ensure that the base and target data SNPs have genomic positions assigned on the same genome build.

Ambiguous SNPs: : If the base and target data were generated using different genotyping chips and the chromosome strand (+/-) that was used for either is unknown, then it is not possible to pair up the alleles of ambiguous SNPs (i.e., those with complementary alleles, either C/G or A/T SNPs) across the data sets, because it will be unknown whether the base and target data are referring to the same allele or not. While allele frequencies could be used to infer which alleles are on the same strand, the accuracy of this could be low for SNPs with MAF close to 50% or when the base and target data are from different populations. Therefore, we recommend removing all ambiguous SNPs to avoid introducing this potential source of systematic error. [5]

Duplicate SNPs: Ensure that there are no duplicated SNPs in either the base or target data since this may cause errors in PRS calculation unless the code/software used specifically checks for duplicated SNPs.[5]

Sex-check: It is standard in GWAS QC to remove individuals for whom there is a difference between reported sex and that indicated by the sex chromosomes. While these may be due to differences in sex and gender identity, they could also reflect mislabeling of samples or misreporting and are, thus, considered potentially unreliable data. In addition to this check, if the aim of an analysis is to model autosomal genetics only, then we recommend that all X and Y chromosome SNPs are removed from the base and target data to eliminate the possibility of non-autosomal sex effects influencing results. [5]

Sample overlap: Sample overlap between the base and target data can result in substantial inflation of the association between the PRS and trait tested in the target data and so must be eliminated.[5]

Relatedness: A high degree of relatedness among individuals between the base and target data can also generate inflation of the association between the PRS and target phenotype. Assuming that the results of the study are intended to reflect those of the general population without close relatedness between the base and target samples, then relatives should be excluded. If genetic data from the relevant base data samples can be accessed, then any closely related individuals (eg. 1st/2nd degree relatives) across base and target samples should be removed. If this is not an option, then every effort should be made to select base and target data that are very unlikely to contain highly related individuals.[5]

Choi *et al.* recommend the below QC criteria for standard analyses: genotyping rate >0.99, sample missingness<0.02, Hardy-Weinberg Equilibrium $p > 1 \times 10^{-6}$, heterozygosity within 3 standard deviations of the mean, minor allele frequency (MAF) >1% (MAF >5% if target sample N <1000)[5] In Figure 4.1 summarized the fundamental features of a PRS analysis.

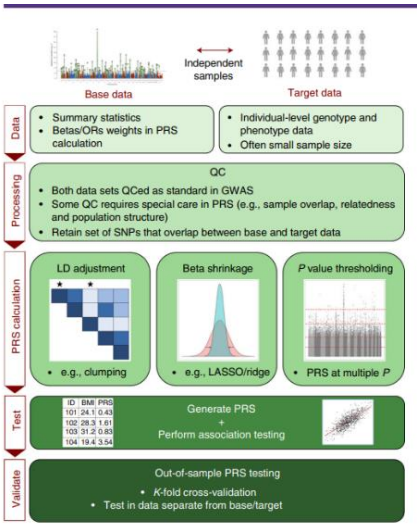


**Figure 4.1**

**The fundamental features of a PRS analysis [5]**

## 4.4 PRS Performing

There are now many methods to calculate PRSs, which differ in terms of two key criteria: which SNPs to include and what weights to allocate to them. The optimal method may differ among traits depending on the sample size of the sample data set and on the genetic architecture of the trait (the number, frequencies, and effect sizes of causal variants), particularly given the linkage disequilibrium (LD) correlation structure between SNPs. Often, when new PGS methods are introduced, comparisons are made between a limited set of methods using simulated data and their application to some real-data examples. [8]

### 4.4.1 Shrinkage of GWAS Effect Size Estimates

Given that SNP effects are estimated with uncertainty and since not all SNPs influence the trait under study, the use of unadjusted effect size estimates of all SNPs could generate poorly estimated PRSs with high standard error. To address this, two broad shrinkage strategies have been adopted: [5]

(1) PRS methods that perform shrinkage of all SNPs generally exploit commonly used statistical shrinkage/regularization techniques, such as LASSO or ridge regression, or Bayesian approaches that perform shrinkage via prior distribution specification. Under different approaches or parameter settings, varying forms of shrinkage can be achieved: e.g., LASSO

regression reduces small effects to zero, while ridge regression shrinks the largest effects more than LASSO but does not reduce any effects to zero. The most appropriate shrinkage to apply is dependent on the underlying mixture of null and true effect size distributions, which is probably a complex mixture of distributions that vary by trait. Since the optimal shrinkage parameters are unknown a priori, PRS prediction is typically optimized across a range of possible parameter values, which in the case of LDpred, for example, includes a parameter for the fraction of causal variants.[5]

(2) In the classic PRS calculation method, only those SNPs with a GWAS association p-value below a certain threshold (e.g., $P < 1 \times 10^{-5}$) are included in the calculation of the PRS, while all other SNPs are excluded. This approach effectively shrinks all excluded SNPs to an effect size estimate of zero and performs no shrinkage on the effect size estimates of those SNPs included. Since the optimal p-value threshold is unknown a priori, PRSs are typically calculated over a range of thresholds, association with the target trait is tested for each, and the prediction is optimized accordingly. This process is analogous to tuning parameter optimization in the formal shrinkage methods.[5]

### 4.4.2 Controlling LD

If genetic association testing is performed using joint models of multiple SNPs, then independent genetic effects can be estimated despite the presence of LD. However, association tests in GWASs are typically performed one SNP at a time, which, combined with the strong correlation structure across the genome, makes estimating the independent genetic effect extremely challenging. If independent effects were estimated in the GWAS or by subsequent fine-mapping, then PRS calculation can be a simple summation of those effects. If, instead, the investigator is using a GWAS based on one-SNP-at-a-time testing, then there are two main options for approximating the PRS that would be obtained from independent effect estimates: (1) SNPs are clumped (i.e., thinned, prioritizing SNPs at the locus with the smallest GWAS P value) so that the retained SNPs are largely independent of each other and, thus, their effects can be summed up, assuming additivity; and (2) all SNPs are included, accounting for the LD between them. In the classic PRS calculation method, option (1) is combined with p-value thresholding and is called the C+T (clumping + thresholding) method, while option (2) is generally favored in methods that implement traditional shrinkage techniques. The relatively similar performance of the classic approach to more sophisticated methods may be due to the

clumping process capturing conditionally independent effects well; note that clumping does not merely thin SNPs by LD at random (like pruning) but preferentially selects SNPs most associated with the trait under study and retains multiple SNPs in the same genomic region if there are multiple independent effects there: clumping does not simply retain only the most-associated SNP in a region. A criticism of clumping, however, is that researchers typically select an arbitrarily chosen correlation threshold for the removal of SNPs in LD. Both clumping and LD modeling require estimation of the LD between SNPs. Assuming that LD values derived from the base data are unavailable, then those from a reference sample of the same ancestry should be used to approximate these. If there are no reference samples well matched to the population composition of the base data, then the target data can be used to estimate the LD instead. However, if base and target samples are drawn from different populations, then the base data LD may be poorly approximated and the PRS accuracy reduced accordingly.[5]

### 4.4.3 Population stratification

When selecting the analysis data for the survey, it is important to be aware of the potential for PGS inflation in the target sample due to population stratification. A major concern in GWAS and PRS studies is that their results may be affected by confounding due to population genetic structure. Since environmental risk factors also tend to be geographically structured, this creates the potential for associations between many genetic variants and the tested trait that are confounded by, for example, location. Uncorrected, this can lead to false positive genotype-phenotype associations and consequently inflated estimates of PRS prediction. PRS prediction can also be inflated by a household effect, whereby the genetics of an individual are correlated with their household environment when created by parents (or siblings) with shared genetic tendencies (e.g., of diet, books or exercise). A key difference between these sources of PRS inflation is that the genetic variants leading to inflation due to population genetic structure are typically non-causal of the outcome, being incidentally associated with location and environmental risk factors, whereas those creating the household effect are (indirectly) causal. Stringent adjustment of effects via genetic principal components (PCs) or the use of mixed models should be applied to both the base and target samples to minimize inflation due to population structure, but the possibility of complex structure causing residual confounding cannot be ruled out. On genome-wide association studies (GWASs) most discoveries to date have been conducted on European ancestry populations. European ancestry-based polygenic scores derived from GWASs cannot be directly used for prediction in non-European ancestry

populations due to differences in linkage disequilibrium (LD), allele frequencies, and genetic architecture. The frequencies of the SNPs used for PRSs contain a strong population component even without applying any PRS weighting. [58]

### 4.4.4 Clumping and thresholding method (C+T)

This method can be used to calculate a score based on any number of genetic variants, including all SNPs. It considers the LD structure of the data by selecting independent SNPs to avoid oversampling of more densely genotyped SNPs.[58] Briefly, C+T uses the GWAS effect size estimates as SNP weights and includes independent SNPs with association p-values lower than a threshold (chosen after application in a tuning sample). It is the most commonly used method. [8] For prediction purposes, including less significant SNPs (than the GWAS p-value threshold) can substantially improve predictive performance. Therefore, when using C+T, one has to choose a p-value threshold that balances between removing informative variants when using a stringent p-value threshold and adding too much noise in the score by including too many variants with no effect.[58] Generally, it selects the p-value threshold that achieves the highest correlation/association with the phenotypes in a validation dataset that contains a measure of the phenotype under study. This approach, however, becomes less useful if the phenotype is not available in the target dataset.[11] The clumping step aims at removing redundancy in included effects that is simply due to linkage disequilibrium (LD) between variants. Yet, clumping may as well remove independently predictive variants in nearby regions; to balance this, C+T uses as hyper-parameter a threshold on correlation between variants included, therefore the correct choice must be made for the hyper-parameters, so to maximize predictive performance of the polygenic score derived. Most of the time, people use default values for these parameters, except for the p-value threshold, for which they look at different values and choose the one maximizing predictive ability in a training set.[9] A PRS is defined as the sum of allele counts of the remaining SNPs weighted by the corresponding regression coefficients.[9]

It is important to avoid double-counting causal variants. Two main approaches can be used to select independent SNPs:

1. LD pruning is the process of genetic marker selection based on their LD. LD pruning is a statistical procedure used to remove redundant SNPs or, in other words, pairs of correlated SNPs. This method selects only one representative SNP from each LD block in the genotype

data. For LD pruning, the pairwise correlation between the markers in a specific range of the genome is calculated. This region is then scanned and if for any pair of markers, the correlation is greater than the specified threshold, the marker with the smallest minor allele frequency (MAF) is discarded, otherwise both markers are retained. In the event that both markers have the same MAF, the one in the latter position is pruned. The process continues until the whole genome has been scanned.The aim is for the final set of markers to contain those that are nearly uncorrelated. [14]

2. Clumping, instead, selects the SNP with the lowest p-value association in each LD block. Clumping looks at the most significant SNP first, computes correlation between this index SNP and nearby SNPs and removes all the nearby SNPs that are correlated with this index SNP beyond a particular threshold (e.g., $r^2 = 0.2$). The clumping step aims at removing redundancy in included effects that is simply due to linkage disequilibrium (LD) between variants. Clumping is preferred since it selects the most statistically significant variant in the locus. [58] While clumping retains one SNP per LD block, pruning can end up with multiple SNPs or no SNPs at all for a region.[14]

A common approach to selecting SNPs for PRS's calculation is based on the p-value of the association within the summary statistics. [58] Thresholding consists of removing SNPs with a p-value larger than a p-value threshold in order to reduce noise in the score.[9] Generally, several p-value thresholds are tested to maximize prediction.[9] Generally speaking, stricter p-value thresholds are more suitable for traits that are not polygenic while more lenient thresholds perform the best for polygenic traits. The aim of the research will also shape the decision. If the goal is to maximize prediction, having more SNPs would be the better choice. However, the more variants that are included in the calculation, the greater the risk that it includes unnecessary "noise" in the PRS. Both steps, clumping and thresholding, represent a statistical compromise between signal and noise.[7] The clumping step prunes redundant correlated effects caused by linkage disequilibrium (LD) between variants. Similarly, thresholding must balance between including truly predictive variants and reducing noise in the score by excluding null effects.[7] Next comes the calculation of PRS as described below. The gold-standard strategy for guarding against generating overfit prediction models and results is to perform out-of-sample prediction. First, parameters are optimized using a training sample, and then the optimized model is tested in a test or validation data set to assess performance. In the PRS setting involving base and target data sets, it would be incorrect to believe that out-of-sample

prediction has already been performed, because polygenic scoring involves two different data sets; in fact, the training is performed on the target data set, meaning that a third data set is required for out-of-sample prediction.[5]

### 4.4.5 LDpred Method

Polygenic risk scores have shown great promise in predicting complex disease risk and will become more accurate as training sample sizes increase. As discussed above, the standard approach to calculating P+T risk scores is likely to discard information and may reduce forecast accuracy. Another, more advanced class of PRS methods is based on approaches typically used either to perform regression with correlated data and/or to select an optimal subset of predictors in a regression model. Unlike the P+T model, these approaches attempt to model the effects of all markers jointly. In the Bayesian statistical framework, a prior probability distribution for the parameters of interest is combined with data to produce a refined posterior distribution, from which inference is made. These prior distributions are based on prior knowledge of how genetic effects are distributed. This prior knowledge is combined with the data to yield a posterior distribution, from which inference is made. LD information is incorporated via appropriate reference populations to account for correlation between effects. Each of these methods aims to shrink effect sizes of non-causal SNPs to increase sparsity, thereby increasing predictive accuracy and computational tractability. In general, these models apply shrinkage to marker effects (i.e., summary statistics) that incorporate LD information from a reference panel.[10]. Vilhjálmsson *et al.* introduce LDpred, a method that infers the posterior mean effect size of each marker by using a prior on effect sizes and LD information from an external reference panel. By using a point-normal mixture prior to the marker effects, LDpred can be applied to traits and diseases with a wide range of genetic architectures. Unlike P+T, LDpred has the desirable property that its prediction accuracy converges to the heritability explained by the SNPs as sample size grows. [12] In practice, LDpred is a different way to estimate the weights using a Bayesian approach. The method assumes a point-normal mixture prior to the distribution of effect sizes and takes into account the correlation structure of SNPs by estimating the LD patterns from a reference sample of unrelated individuals.[10]

The method: LDpred calculates the posterior mean effects from GWAS summary statistics by conditioning on a genetic architecture prior and LD information from a reference panel. The inner product of these re-weighted and the test-sample genotypes is the posterior mean

phenotype and thus, under the model assumptions and available data, an optimal (minimum variance and unbiased) predictor. The prior for the effect sizes is a point-normal mixture distribution, which allows for non-infinitesimal genetic architectures. The prior has two parameters: the heritability explained by the genotypes and the fraction of causal markers (i.e., the fraction of markers with non-zero effects). The heritability parameter is estimated from GWAS summary statistics and accounts for sampling noise and LD. The fraction of causal markers is allowed to vary and can be optimized with respect to prediction accuracy in a validation dataset, analogous to how P+T is applied in practice. Hence, similar to P+T (where p-value thresholds are varied and multiple PRSs are calculated), multiple LDpred risk scores are calculated with the use of priors with varying fractions of markers with non-zero effects. The value that optimizes prediction accuracy can then be determined in an independent validation dataset. LD is approximated using data from a reference panel (e.g., independent validation data). The posterior mean effect sizes are estimated via the Markov chain Monte Carlo (MCMC) method and applied to the validation data to obtain the PRS. In the special case of no LD, posterior mean effect sizes with a point normal prior can be viewed as a soft threshold and can be computed analytically. A key feature of LDpred is that it relies on GWAS summary statistics, which are often available even when raw genotypes are not. For this reason, the main approaches that Vilhjálmsson *et al.* compare with LDpred are: PRS based on all markers (unadjusted PRS), P+T, and LDpred specialized to an infinitesimal prior (LDpred-inf). [12]

Phenotype Model: For phenotype model, Y be a N×1 phenotype vector and X be a N×M genotype matrix, where N is the number of individuals, and M is the number of genetic variants. Vilhjálmsson *et al.* assumed throughout that the phenotype Y and individual genetic variants $X_i$ have been mean centered and standardized to have variance 1. They model the phenotype as a linear combination of M genetic effects and an independent environmental effect ε, i.e., $Y = \sum_{i=1}^{M} X_i\beta_i + \varepsilon$, where $X_i$ denotes the $i^{th}$ genetic variant, $\beta_i$ is its true effect, and ε is the environmental and noise contribution. In this setting, the (marginal) least-squares estimate of an individual marker effect is $\widehat{\beta_i} = \frac{X_i'Y}{N}$. For clarity, they implicitly assume that they have the standardized effect estimates available as summary statistics. In practice, they usually have other summary statistics, including the p value and direction of the effect estimates, from which they infer the standardized effect estimates.[12]

First, they exclude all markers with ambiguous effect directions, i.e., A/T and G/C SNPs. Second, from the p-values, they obtain Z scores and multiply them by the sign of the effects (obtained from the effect estimates or effect direction). Finally, they approximate the least-squares estimate for the effect by $\widehat{\beta_i} = si\left(\frac{zi}{\sqrt{N}}\right)$, where $s_i$ is the sign of the effect, $z_i$ is the z-score obtained from the p-value and N is the sample size. If the trait is a case-control trait, this transformation from the p value to the effect size can be thought of as being an effect estimate for an underlying quantitative liability or risk trait.[12]

Bpred: Bayesian Approach in the Special Case of No LD. Under the assumption that the phenotype has an additive genetic architecture and is linear, then estimating the posterior mean phenotype boils down to estimating the posterior mean effects of each SNP and then summing their contribution into a risk score. Under a model, the optimal linear prediction given some statistic is the posterior mean prediction. This prediction is optimal in the sense that it minimizes the prediction error variance. Under the linear model described above, the posterior mean phenotype (or the optimal predictor of the trait $Y$ [14]) given GWAS summary statistics and LD is

$$\mathrm{E}\,(\mathrm{Y}|\hat{\beta}, \widehat{D}) = \sum_{i=1}^{M} X_i' \, E\left(\beta_i \middle| \hat{\beta}, \widehat{D}\right) \tag{1.2}$$

Here, $\hat{\beta}$ denotes a vector of marginally estimated least-squares estimates obtained from the GWAS summary statistics and $\widehat{D}$ refers to the observed genome-wide LD matrix in the training data, i.e., the samples for which the effect estimates are calculated. Hence, the quantity of interest is the posterior mean marker effect given LD information from the GWAS sample and the GWAS summary statistics. In practice, we might not have this information available to us and are forced to estimate the LD from a reference panel. In most of our analyses, we estimated the local LD structure in the training data from the independent validation data. Although this choice of LD reference panel can lead to a small bias when one estimates individual prediction accuracy, this choice is valid whenever the aim is to calculate accurate PRSs for a cohort without knowing the case-control status a priori. In other words, it is an unbiased estimate of the PRS accuracy when the validation data is used as an LD reference, which we recommend in practice.[12]

If all samples are independent and all markers are unlinked and have effects drawn from a Gaussian distribution, i.e., $\beta_i \sim_{iid} N(0, (h_g^2/M))$, then this is an infinitesimal model, which

represents a genetic architecture where all genetic variants are causal. Under this model, the posterior mean can be derived analytically:[12]

$$\mathrm{E}\left(\beta_i \middle| \tilde{\beta}\right) = \mathrm{E}(\beta_i|\tilde{\beta}_l) = \frac{h_g^2}{h_g^2 + \frac{M}{N}} \tilde{\beta}_l \tag{1.3}$$

An arguably more reasonable prior for the effect sizes is a non-infinitesimal model, where only a fraction of the markers is causal and affect the trait. We can model non-infinitesimal genetic architectures by using mixture distributions with a mixture parameter p, which denotes the fraction of causal markers. Consider the following Gaussian mixture prior to this. Assume that the effects are drawn from a mixture distribution as follows: [12]

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) with\ probability\ p \\ 0\ with\ probability\ (1-p), \end{cases}$$

Where p is the probability that a marker is drawn from a Gaussian distribution, i.e., the fraction of causal markers. Under this model, the posterior mean (which is a shrinkage of the original GWAS effect ($\tilde{\beta}_i$)[14]) can be derived as:[12]

$$\mathrm{E}(\beta_i|\tilde{\beta}_i) = \left(\frac{h_g^2}{h_g^2 + \frac{Mp}{N}}\right) \overline{p_l} \tilde{\beta}_i, \tag{1.5}$$

Where $\overline{p_l}$ is the posterior probability that the i[th] marker (SNP) is causal and can be calculated analytically.[12] For more accurate predictions, the authors recommend that the user specify a range of different fraction values $p$, which will be optimized on independent testing data.[14] Vilhjálmsson *et al.* refer to this Bayesian shrink without LD as Bpred.[12]

LDpred: Bayesian Approach in the Presence of LD. If we allow for loci to be linked, then we can derive posterior mean effects analytically under a Gaussian infinitesimal prior (described above). We call the resulting method LDpred-inf, and it represents a computationally efficient special case of LDpred. If we assume that distant markers are unlinked, the posterior mean for the effect sizes within a small region $l$ under an infinitesimal model is well approximated by [12]

$$\mathrm{E}\left(\beta^l | \tilde{\beta}^l, D\right) \approx \left(\frac{M}{Nh_g^2} I + D_l\right)^{-1} \tilde{\beta}^l \tag{1.6}$$

Here, $D_l$ denotes the regional LD matrix within the region of LD, and $\tilde{\beta}^l$ denotes the least-squares-estimated effects for SNPs within that region. The approximation assumes that the heritability explained by the region is small and that LD with SNPs outside of the region is negligible. Interestingly, under these assumptions, the resulting effects approximate the standard mixed-model genomic BLUP effects. LDpred-inf is therefore a natural extension of the genomic BLUP to summary statistics. In practice, we do not know the LD pattern in the training data, and we need to estimate it by using LD in a reference panel. Deriving an analytical expression for the posterior mean under a non-infinitesimal Gaussian mixture prior is difficult, and thus LDpred approximates it numerically by using an approximate MCMC Gibbs sampler. [12]

In general, the LD matrix is given by the following relation: $D = \frac{XX'}{N}$ and for the locus-LD the relation becomes: $D_l = \frac{X^{(l)}X^{(l)\prime}}{N}$ .

Estimation of the Heritability Parameter: In the absence of population structure and assuming independent and identically distributed mean-zero SNP effects, the following equation has been shown to hold:

$$\mathrm{E}\,(x_j^2)=1+\frac{N h_j^2 l_j}{M} \tag{1.7}$$

where $x_j^2$ is the $x^2$-distributed test statistic at the j$^{th}$ SNP, and $l_j = \sum_k \left[r^2(j,k) - \left(1 - \frac{r^2(j,k)}{N} - 2\right)\right]$ summing over k neighboring SNPs in LD, is the LD score for the j$^{th}$ SNP. Taking the average of both sides over SNPs and rearranging them, we obtain a heritability estimate: $\widetilde{h_j^2} = \frac{\overline{(x^2}-1)M}{\bar{l}N}$ ,where $\overline{x^2} = \sum_j(\frac{x_j^2}{M})$ and $\bar{l} = \sum_j(\frac{l_j}{M})$ . [12]

When LDpred is applied to real data, two parameters need to be specified beforehand. The first parameter is the LD radius, i.e., the number of SNPs that we adjust for on each side of a given SNP. There is a trade-off when we decide on the LD radius. If the LD radius is too large, then errors in LD estimates can lead to apparent LD between unlinked loci, which can lead to worse effect estimates and poor convergence. If the LD radius is too small, then we risk not accounting for LD between linked loci. Vilhjálmsson *et al.* found that an LD radius of approximately M/3,000 (the default value in LDpred), where M is the total number of SNPs used in the analysis, works well in practice. Regarding the choice of the LD panel, its LD

structure should ideally be similar to the training data for which the summary statistics are calculated. In simulations, Vilhjálmsson *et al.* found that the LD reference panel should contain at least 1,000 individuals. The second parameter is the fraction p of non-zero effects in the prior. This parameter is analogous to the p value threshold used in P+T. Vilhjálmsson *et al.* recommendation is to try a range of values for p (e.g., 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 3E-4, 1E-4, 3E-5, 1E-5; these are default values in LDpred). This will generate 11 sets of SNP weights, which can be used for calculating polygenic scores. One can then use independent validation data to optimize the parameter, analogous to how the p value threshold is optimized in the P+T method.[12]

When using LDpred, Vilhjálmsson *et al.* recommend that SNP weights (posterior mean effect sizes) are calculated for exactly the SNPs used in the validation data. This ensures that all SNPs with non-zero weights are in the validation dataset. In practice, we use the intersection of SNPs present in the summary-statistics dataset, the LD reference genotypes, and the validation genotypes. If the validation cohort contains more than 1,000 individuals, with the same ancestry as the individuals used for the GWAS summary statistics, then we suggest using the validation cohort as the LD reference as well. These steps are implemented in the LDpred software package.[12]

LDpred-inf is a special case of LDpred when all variants are assumed to be causal (i.e., p = 1). Under this infinitesimal model, the posterior mean effect sizes are closely approximated by

$$\mathrm{E}\,(\beta|\tilde{\beta}, \widehat{D}) \approx \left(\frac{M}{N g_h^2} I + D\right)^{-1} \tilde{\beta} \qquad (1.8)$$

where D denotes the LD matrix between the markers in the training data, and $\tilde{\beta}$ denotes the marginally estimated marker effects. [12]

Although LDpred is a substantial improvement over existing methods for using summary statistics to conduct polygenic prediction, it still has limitations. First, the method's reliance on LD information from a reference panel requires that the reference panel be a good match for the population from which summary statistics were obtained; in the case of a mismatch, prediction accuracy might be compromised. Second, the point-normal mixture prior distribution used by LDpred might not accurately model the true genetic architecture, and it is possible that other prior distributions might perform better in some settings. Third, in those instances where raw genotypes are available, fitting all markers simultaneously might achieve higher accuracy

than methods based on marginal summary statistics. Fourth, LD reference panels are likely to be inadequate for rare variants, motivating future work on how to treat rare variants in PRSs. Despite these limitations, LDpred is likely to be broadly useful in leveraging summary-statistics datasets for polygenic prediction of both quantitative and case-control traits. As sample sizes increase and polygenic predictions become more accurate, their value increases, both in clinical settings and for understanding genetics. LDpred represents substantial progress, but more work remains to be done. One future direction would be to develop methods that combine different sources of information. In addition, using different prior distributions across genomic regions or functional annotation classes could further improve the prediction.[12]

### 4.4.6 LASSOSUM

As there is no inherent information on linkage disequilibrium (LD) in summary statistics, a pertinent question is how we can use LD information available elsewhere to supplement such analyses. To answer this question, Mak T.S.H *et al.*, propose a method for constructing PRS using summary statistics and a reference panel in a penalized regression framework, called lassosum. They also propose a general method for choosing the value of the tuning parameter in the absence of validation data. Lassosum use non-Bayesian strategies to consider large numbers of markers jointly applies least absolute shrinkage and selection operator (LASSO) regression to downweight, and perhaps omit altogether, effects of correlated markers. Two important parameters for lassosum, which may require optimizing using external data, are λ, which determines the fraction of effects shrunk to 0, and s, the shrinkage parameter.[10]

The method: Given a linear regression problem $y = X\beta + \varepsilon$, where X denotes an $n$-by-$p$ data matrix, and y a vector of observed outcomes, the LASSO(least absolute shrinkage and selection operator) is a popular method for deriving estimates of β and predictors of (future observations of) y, especially in the case where $p$ (the number of predictors/columns in X) is large and when it is reasonable to assume that many β are 0. LASSO obtains estimates of β (weights in the linear combination of X) given y and X by minimizing the objective function

$$f(\beta) = (y - X\beta)^T(y - X\beta) + 2\lambda\|\beta\|_1^1 \quad (1.9)$$

$$= y^T y + \beta^T X^T X\beta - 2\beta^T X^T y + 2\lambda\|\beta\|_1^1 \quad (1.10)$$

where $\|\beta\|_1^1 = \sum_i \beta_i$ denote the $L_1$ norm of $\boldsymbol{\beta}$, for a particular fixed value of $\lambda$. In general, depending on $\lambda$, a proportion of the $\beta_i$ are given the estimate of 0. It is also a specific instance

of penalized regression where the usual least square formulation of the linear regression problem is augmented by a penalty, in this case $2\lambda\|\beta\|_1^1$. LASSO lends itself to being used for estimation of $\boldsymbol{\beta}$ in the event where only summary statistics are available, because if $\boldsymbol{X}$ represent standardized genotype data and $\boldsymbol{y}$ standardized phenotype, divided by $\sqrt{n}$, then Equation (1.10) can be written as

$$f(\beta) = y^T y + \beta^T R\beta - 2\beta^T r + 2\lambda\|\beta\|_1^1 \quad (1.11)$$

where r= $X^T$ y represents the SNP-wise correlation between the SNPs and the phenotype, and R=$X^T X$ is the LD matrix, a matrix of correlations between SNPs. Equation (1.11) suggests a method for deriving PGS weights as estimates of $\boldsymbol{\beta}$ by minimizing $f(\boldsymbol{\beta})$. Estimates of $\boldsymbol{r}$ can be obtained from summary statistics databases that are publicly available for major diseases/phenotypes and estimates of LD ($\boldsymbol{R}$) from publicly available genotypes such as the 1000 Genome database. Equation (1.11) suggests a method for deriving PGS weights as estimates of $\boldsymbol{\beta}$ by minimizing $f(\boldsymbol{\beta})$. [11]

An issue that surfaces when we substitute $\boldsymbol{R}$ and $\boldsymbol{r}$ with the estimates derived from publicly available data is that the genotype $\boldsymbol{X}$ used to estimate $\boldsymbol{R}$ and $\boldsymbol{r}$ will in general be different. In particular, it will be more appropriate to write $\boldsymbol{R} = X_r^T X_r$ to indicate that the genotype used to derive estimates of LD ($X_r$) will not in general be the same as the genotype that gave rise to the correlations $\boldsymbol{r}$. Writing Equation (1.11) as

$$f(\beta) = y^T y + \beta^T X_r^T X_r \beta - 2\beta^T X^T y + 2\lambda\|\beta\|_1^1 \quad (1.12)$$

however, would imply that (1.12) is no longer a LASSO problem, because it is no longer a penalized least squares problem. A minimum to (1.12) can still be sought, although the solutions would often be unstable and nonunique, since $y^T y + \beta^T X_r^T X_r \beta - 2\beta^T X^T y$ will not generally have a finite minimum.

A natural solution to this problem is to regularize Equation (1.12). In particular, if we replace $X_r^T X_r$ with $R_s = (1-s)X_r^T X_{r_{sI}}$, for some 0 <$s$< 1, then

$$f(\beta) = y^T y + \beta^T R_s \beta - 2\beta^T r + 2\lambda\|\beta\|_1^1 \quad (1.13)$$

will be equivalent to a LASSO problem. First, we note that $y^T y$ is a constant and thus replacing it with any other constant will not change the solution. $R_s$ is necessarily positive definite for 0 <$s$< 1. This means that there always exists W and v such that

$$W^T W = R_s \text{ and } W^T v = r \tag{1.14}$$

Substituting (1.14) into (1.13) and replacing $y^T y$ with $v^T v$, we see that (1.13) can be written in a form such as (1.9) and is therefore a LASSO problem. Expanding (1.13) into

$$f(\beta) = y^T y + (1-s)\beta^T X_r^T X_r \beta - 2\beta^T r + s\beta^T \beta + 2\lambda \|\beta\|_1^1 \tag{1.15}$$

Equation (1.15) encompasses a number of submodels as special cases. For example, when $\lambda = 0$ and $s = 1$, the estimated PGS becomes equivalent to simply using the entire set of correlation estimates without shrinkage or subset selection.[11]

Tuning parameters selection: $\lambda$ and $s$ need to be chosen. Generally, in the presence of a validation dataset, $\lambda$ can be chosen by maximizing the correlation of the PGS with the validation phenotype data, just as it has been done in the choice of a p-value cutoff points in standard PGS calculation method(C+T). In principle, this method can be used to choose a suitable value for $s$ also, although repeating the estimation over different values of $s$ is much more time-consuming. Thus, in this paper, Mak T.S.H *et al.* set $s$ to a few chosen values and examined whether they are sufficient in arriving at a PGS with reasonable prediction accuracy. A more pressing problem is that validation phenotypes are not often available. And here Mak T.S.H *et al.* try to simulate this procedure in the following manner, which they refer to as pseudovalidation in this paper and can be applied to any PGS method that requires a tuning parameter. The analysis of this method is beyond the scope of this thesis and is therefore not discussed further. The reader is referred to the source [11].

| | *P*-value thresholding w/o clumping | *Standard approach:* Clumping + thresholding (C+T) | Penalised Regression | Bayesian Shrinkage |
|---|---|---|---|---|
| Shrinkage strategy | *P*-value threshold | *P*-value threshold | LASSO, Elastic Net, penalty parameters | Prior distribution, e.g. fraction of causal SNPs |
| Handling Linkage Disequilibrium | N/A | Clumping | LD matrix is integral to algorithm | Shrink effect sizes with respect to LD |
| Example software | PLINK | PRSice [12] | Lassosum [19] | LDpred [38] |

**Figure 4.2**

**Comparison of different approaches for performing PRS analyses [10]**

When performing a PRS analysis, it is important to consider which approach and software may be best suited for handling the research question. The primary decisions to make are how to decide which SNPs to include and how to modify effect size estimates. Depending on the genetic architecture of the trait or disease in question, each of these approaches could have merit. Beyond the method, the choice of approach and software could be influenced by the specific research question, data availability and type of data, goodness of fit metrics, and computation speed.[53]

## 4.5 Validation and prediction

Validation of the PRS underpins its usefulness. If incorrect decisions or conclusions are drawn, the PRS may lack precision and accuracy. Validation is also inherently intertwined with prediction. Prediction is the estimation of $R^2$, which is the proportion of variance explained by the regression model. Generally, it is important to understand the amount of variability that can be explained by including a particular PRS in a model, namely the increase in the $R^2$ when a PRS is entered into a model compared to the baseline model. The baseline model is the simplest possible prediction, which is used as a starting point against which additional variables are added. Then, it also generally included the population stratification variables (e.g., the first 10 or 20 PCAs) and other relevant covariates. Typically, a regression is performed on the target sample, with the PRS as a predictor of the target trait or experimental outcome, and covariates are included as appropriate. Dudbridge also showed that the power of PRS association testing is optimized using equal-sized base and target sample sizes, while individual-level predictive accuracy is optimized by maximizing base sample size. For binary traits, Nagelkerke $R^2$ used to measure more generally how much of the variation in the observed outcomes can be explained by the model's predictions [5]

A typical PRS study involves testing evidence for an association between a PRS and a trait(s) in the target data. The association between PRS and outcome can be measured with standard association or goodness-of-fit metrics, such as the p-value derived in testing a null hypothesis of no association, phenotypic variance explained ($R^2$) or effect size estimate (beta or OR) per unit of PRS or between specific strata (e.g., high versus low-risk individuals), and with measures of discrimination in disease prediction, such as area under the receiver operator curve (AUC) or area under the precision recall curve. The most frequent way to describe how much variance is explained by a PRS is to run a regression model with the phenotype as the dependent

variable and the PRS as an independent variable and calculate the $R^2$. The interpretation of $R^2$ is that it is the proportion of variance explained by the regression model. When using covariates, it is common to estimate the gain in $R^2$ in two steps. First, a regression model with covariates but without a PRS is estimated. In the second step, the PRS was added to the models and estimated the differences in the $R^2$ of the two models. Since the statistical distribution of the $R^2$ is not a standard one, it is not possible to estimate a confidence interval unless we use nonparametric statistical techniques. $R^2$ has a theoretical upper bound equal to the heritability of the phenotype; however, SNP heritability ($h^2_{SNP}$), the heritability using all available SNPs in the data, will be the upper bound for PRS. For binary traits, the approach is very similar, but instead of estimating a series of linear regression models, we estimate logistic regression models and report the gain in pseudo-$R^2$. Alternatively, it is possible for binary traits to estimate the area under the curve as a measure of the accuracy of the PRS to explain the phenotype.[58] The area under the curve (AUC) measures the predictive ability of a receiver operating characteristic (ROC) generated based on the PRS for a sample of individuals. The AUC is a function of the ability of the risk score to correctly identify the presence (sensitivity) or absence (specificity) of the outcome of interest.[58]

The AUC compares the rates of true positives (sensitivity) and false positives (1 –specificity) and indicates the overall performance of predictive models. Sensitivity, the probability of correctly classifying an affected individual as affected, indicates the ability of the model to correctly predict individuals with the outcome of interest; specificity, the probability of correctly classifying an unaffected individual as unaffected, indicates the ability of the model to accurately screen out individuals without the outcome of interest. The AUC of the ROC curve is a measure of overall performance of the model, and ranges from 0 to 1. Model performance based on AUC may differ depending on the phenotype being measured but, in general, an AUC of 0.5 is considered null (no better than chance), and an AUC of at least 0.8 is considered to be very good, especially for a complex trait. An AUC less than 0.5 likely indicates a data error or that the model is predicting the wrong outcome. Diagnostic tests, by contrast, tend to have AUCs at 0.95 or higher for clinical use.[53] Models are expected to have an AUC >0.75 for informative screening of individuals who are at increased disease risk and a very high AUC (as high as 0.99) for a diagnostic test. The higher the AUC, the more precise the prediction and, thus, the greater the clinical utility of the combination of factors included in the model.[1] Other than the ROC curve, there are a few options to visualize goodness of fit

metrics of a PRS model. The first of these includes a common boxplot (binary traits or categorized quantitative traits) or scatter plot (quantitative traits) for plotting the PRSs against the trait of interest. These visually demonstrate the ability of the model to discriminate the outcome. To graphically depict fit of a model against other models, one option is to plot the $R^2$ values across multiple thresholds in a bar plot.[53]

## 4.6 PRS Studies

### 4.6.1 Study 1

About simulations: Vilhjálmsson *et al.*, performed three types of simulations: (1) simulated traits and simulated genotypes; (2) simulated traits, simulated summary statistics and simulated validation genotypes; and (3) simulated traits based on real genotypes. They used the point-normal model for effect sizes, for most of the simulations:

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) with\ probability\ p \\ 0\ with\ probability\ (1-p), \end{cases}$$

Furthermore, for all of the simulations, they used four different values for p (the fraction of causal loci). They simulated genotypes with the adjacent squared correlation between SNPs set to 0 (unlinked SNPs) and 0.9 (SNPs in LD). In order to compare the performance of the LDpred method at large sample sizes, they simulated summary statistics that they used as training data for the PRSs. They also simulated two smaller samples (2,000 individuals) representing independent validation data and a LD reference panel.

When there is no LD, the least-squares effect estimates (summary statistics) are sampled from a Gaussian distribution, $\widehat{\beta_i}|\beta_i \sim_{iid} N(\beta_i, (\frac{1}{N}))$ , where $\beta_i$ are the true effects. To simulate marginal effect estimates without genotypes in the presence of LD, they first estimate the LD pattern empirically by simulating 100 linked SNPs for 1,000 individuals for a given value and average over 1,000 simulations. This matrix captures the LD pattern in the validation data given that they simulate it by using the same procedure. Using this LD matrix D, we then sample the marginal least-squares estimates within a region of LD (SNP chunk) as $\hat{\beta}|\beta \sim_{iid} N(D\beta, (\frac{D}{N}))$, where D is the LD matrix. When simulating traits by using the Wellcome Trust Case Control Consortium (WTCCC) genotypes (Figure 5.6), they performed simulations under four different scenarios representing different number of chromosomes: (1) all chromosomes, (2)

chromosomes 1–4, (3) chromosomes 1 and 2, and (4) chromosome 1. They used 16,179 individuals in the WTCCC data and 376,901 SNPs that passed quality control (QC). Also, they used 3-fold cross-validation, whereby 1/3 of the data was validation data and 2/3 were training data.

Results: Vilhjálmsson *et al.*, first considered simulations with simulated genotypes. They assessed accuracy by using squared correlation (prediction $R^2$) between observed and predicted phenotypes. The Bayesian shrink imposed by LDpred generally performed well in simulations without LD. In Figure 4.3 the four subfigures a-d correspond to different genetic architectures where they vary p, the fraction of variants with (non-zero) effects drawn from a Gaussian distribution. Bpred denotes the analytical solution to LDpred, which can be derived in the absence of LD. As expected, Bpred outperforms P-value thresholding in the absence of LD, although not by much.

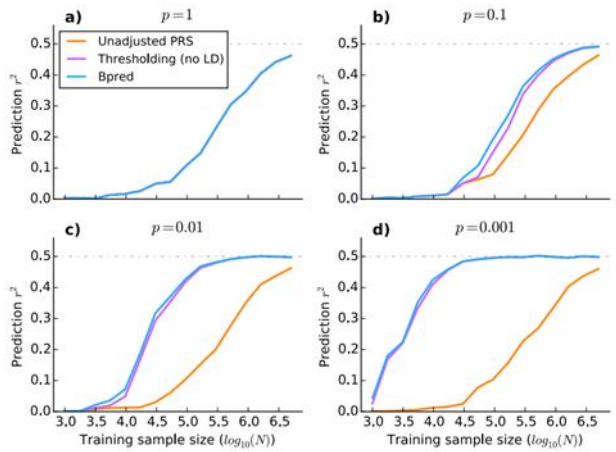


**Figure 4.3**

**Comparison of methods using simulated genotypes without LD[12]**

However, LDpred performed particularly well in simulations with LD (Figure 4.4); the larger improvement (e.g., versus P+T) in this case indicates that the main advantage of LDpred is in its explicit modeling of LD. The four subfigures a-d correspond to different genetic architectures where p varies, which represents the fraction of variants with (non-zero) effects drawn from a Gaussian distribution. Note that when p=0.001, the chance of two causal variants being in LD is very small (~1%), and thus the improvement from accounting for LD in LDpred is negligible compared to P+T.

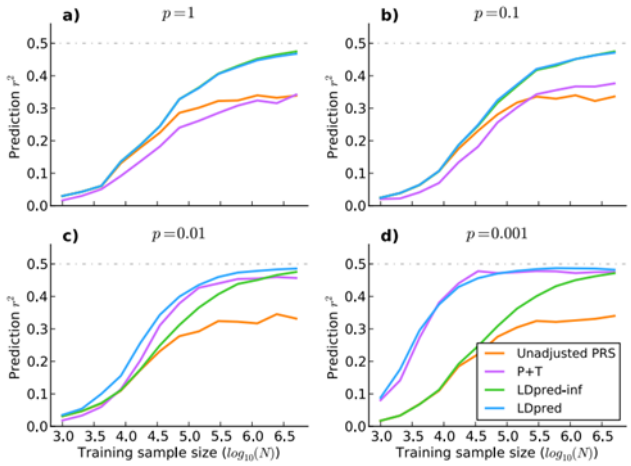


**Figure 4.4**

**Comparison of methods using simulated genotypes with LD[12]**

They also evaluated the prediction accuracy as a function of the sample size of the LD reference panel (Figure 4.5). LDpred performs best with an LD reference panel of at least 1,000 individuals. These results also highlight the importance of using an LD reference population with LD patterns similar to the training sample, given that an inaccurate reference sample will have effects similar to those of a small reference sample.

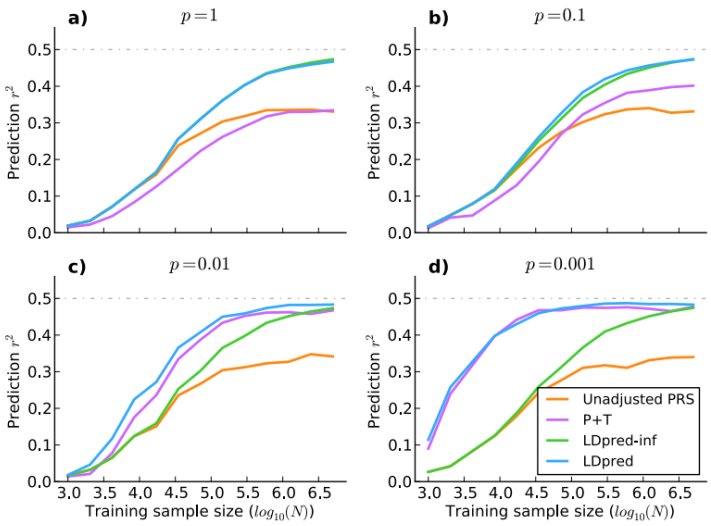


**Figure 4.5**

**Prediction accuracy for methods as a function of LD reference sample size[12]**

Using real WTCCC genotypes (15,835 samples and 376,901 markers after QC), Vilhjálmsson *et al.* simulated infinitesimal traits with the heritability set to 0.5. They

extrapolated results for larger sample sizes ($N_{eff}$) by restricting the simulations to a subset of the genome (smaller M), leading to larger N/M. Results are displayed in Figure 4.6A. LDpred-inf and LDpred (which are expected to be equivalent in the infinitesimal case) performed well in these simulations—particularly at large values of $N_{eff}$, consistent with the intuition from Equation 1.8 that the LD adjustment arising from the reference-panel LD matrix (D) is more important when $Nh_g^2/M$ is large. On the other hand, P+T performed less well, consistent with the intuition that pruning markers loses information. The four subfigures correspond to p = 1 (A), p =0.1 (B), p = 0.01 (C), and p = 0.001 (D) for the fraction of simulated causal markers with (non-zero) effect sizes sampled from a Gaussian distribution. To aid interpretation of the results, they plot the accuracy against the effective sample size, defined as $N_{eff} = (N/M_{sim})M$ , where N = 10,786 is the training sample size, M = 376,901 is the total number of SNPs, and $M_{sim}$ is the actual number of SNPs used in each simulation: 376,901 (all chromosomes), 112,185 (chromosomes 1–4), 61,689 (chromosomes 1 and 2), and 30,004 (chromosome 1). The effective sample size is the sample size that maintains the same N/M ratio if all SNPs are used.

They next simulated non-infinitesimal traits by using real WTCCC genotypes and varying the proportion p of causal markers. Results are displayed in Figures 4.6(B–D). LDpred outperformed all other approaches, including P+T, particularly at large values of N/M. For p =0.01 and p= 0.001, the methods that do not account for non-infinitesimal architectures (unadjusted PRSs and LDpred-inf) performed poorly, and P+T was second best among these methods.

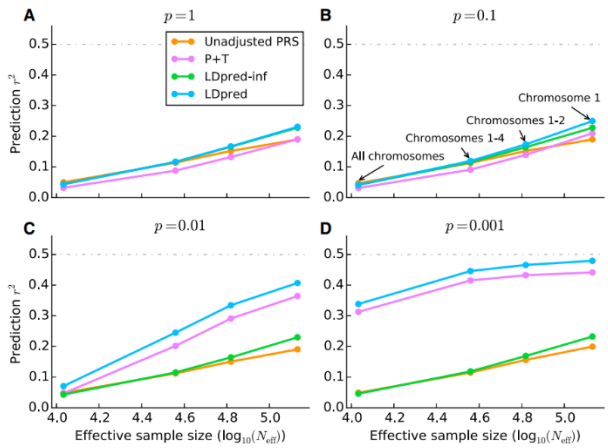


**Figure 4.6**

**Comparison of Four Prediction Methods Applied to Simulated Traits[12]**

In this section, we present the results of a study carried out by Vilhjálmsson et al. using real data. WTCCC Genotype Data were initially used, and a quality check was performed on them. In particular pruning variants with missing rates above 1%, and removing individuals with genetic relatedness coefficients above 0.05, so they left 15,835 individuals genotyped for 376,901 SNPs, including 1,819 individuals with bipolar disease (BD), 1,862 individuals with coronary artery disease (CAD), 1,687 individuals with Crohn disease (CD), 1,907 individuals with hypertension (HT), 1,831 individuals with rheumatoid arthritis (RA), 1,953 individuals with type 1 diabetes (T1D), and 1,909 individuals with type 2 diabetes (T2D). For each of the seven diseases, was performed 5-fold cross-validation, whereby 1/5 of the data was validation data and 4/5 were training data, on affected individuals and 2,867 control individuals. For each of these analyses, they used the validation data as the LD reference data when using LDpred and when performing LD pruning. In Figure 4.7, represents the comparison of LDpred to other summary-statistics-based methods across the seven WTCCC disease datasets. It plotted the prediction accuracy of the different methods as estimated from 5-fold-cross-validation. The Nagelkerke prediction $R^2$ is shown on the y axis. LDpred significantly improved the prediction accuracy for the immune-related diseases T1D, RA, and CD. LDpred attained significant improvement in prediction accuracy over P+T for T1D (p-value= 4.4E-15), RA (p-value = 1.2E-5), and CD (p-value = 2.7E-8). For the other diseases with more-complex genetic architectures, the prediction accuracy of LDpred was similar to that of P+T, potentially because the training sample size was not sufficiently large enough for modeling LD to have a sizeable impact. [12]

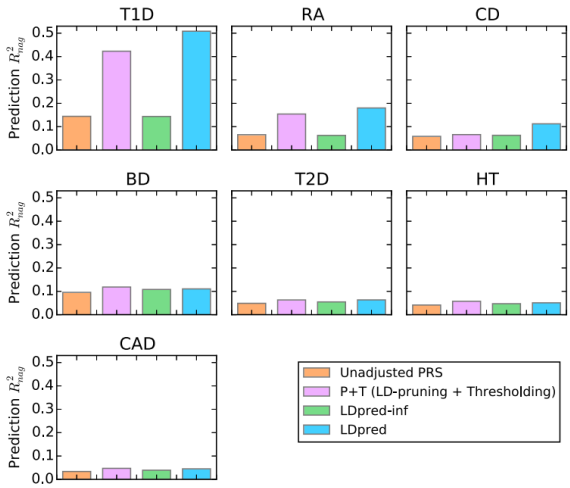


**Figure 4.7**

**Comparison of Methods Applied to Seven WTCCC Disease Datasets[12]**

Finally, six large summary-statistics datasets were analyzed in this study. For all of the validation datasets, it was used the chip genotypes and filter individuals with more than 10% of genotype calls missing and filtered SNPs that had a missing rate more than 1% and a minor allele frequency (MAF) greater than 1%. In addition, SNPs that had ambiguous nucleotides, i.e., A/T and G/C removed. They matched the SNPs between the validation and GWAS summary-statistics datasets on the basis of the SNP rsID and excluded triplets, SNPs for which one nucleotide was unknown, and SNPs that had different nucleotides in different datasets. This was Vilhjálmsson *et al.* QC procedure for all large summary-statistics datasets that they analyzed. For all six of these traits, they used the validation dataset as the LD reference data when using LDpred and when performing LD pruning. By using the validation dataset as LD reference data, they were only required to coordinate two different datasets, i.e., the GWAS summary statistics and the validation dataset. They calculated P+T risk scores for different p-value thresholds by using grid values (1E-8, 1E-6, 1E-5, 3E-5, 1E-4, 3E-4, 1E-3, 3E-3, 0.01, 0.03, 0.1, 0.3, 1), and for LDpred they used the mixture probability (fraction of causal markers) values (1E-4, 3E-4, 1E-3, 3E-3, 0.01, 0.03, 0.1, 0.3, 1). They are used to find the optimal prediction value from a validation dataset for LDpred and P+T.

The Psychiatric Genomics Consortium 2 (PGC2) SCZ summary statistics consisted of 34,241 affected and 45,604 control individuals. The ISC (International Schizophrenia Consortium) cohorts and the MGS (Molecular Genetics of Schizophrenia) cohorts used as validation datasets. After the QC the ISC cohort consisted of 1,562 affected and 1,994 control

individuals genotyped on ~518,000 SNPs that overlapped with the GWAS summary statistics. The MGS dataset consisted of 2,681 affected and 2,653 control individuals after QC and had ~549,000 SNPs that overlapped with the GWAS summary statistics.

The International Multiple Sclerosis Genetics Consortium summary statistics used for multiple sclerosis (MS). These were calculated with 9,772 affected and 17,376 control individuals (27,148 individuals in total) for ~465,000 SNPs. As an independent validation dataset, used the BWH/MIGEN chip genotypes with 821 affected and 2,705 control individuals. After QC, the overlap between the validation genotypes and the summary statistics only consisted of ~114,000 SNPs, which used for the analysis.

For breast cancer (BC), the Genetic Associations and Mechanisms in Oncology (GAME-ON) BC GWAS summary statistics used, consisting of 16,003 affected and 41,335 control individuals. As validation genotypes, we combined genotypes from five different datasets. None of these 307 affected or 560 control individuals were included in the GWAS summary-statistics analysis, and they thus represent an independent validation dataset. We used the chip genotypes that overlapped the GWAS summary statistics, which resulted in ~444,000 genotypes after QC.

For CAD, we used the transatlantic Coronary Artery Disease Genome-wide Replication and Meta-analysis (CARDIoGRAM) consortium GWAS summary statistics. These were calculated with 22,233 affected and 64,762 control individuals (86,995 individuals in total) for 2.4 million SNPs. For T2D, we used the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium GWAS summary statistics. These were calculated with 12,171 affected and 56,862 control individuals (69,033 individuals in total) for 2.5 million SNPs. For both CAD and T2D, we used the Women's Genomes Health Study (WGHS) dataset as validation data, where we randomly down-sampled the control individuals. For CAD, we validated the predictions in 923 individuals with cardiovascular disease and 1,428 control individuals, and for T2D we used 1,673 affected and 1,434 control individuals. We used the genotyped SNPs that overlapped the GWAS summary statistics, which amounted to about ~290,000 SNPs for both CAD and T2D after QC.

Prediction accuracies for LDpred and other methods are reported in Figure 4.8 (Nagelkerke $R^2$). For all five traits, LDpred provided significantly better predictions than other approaches (for the improvement over P+T, the p values were 6.3E-47 for SCZ, 2.0E-14 for MS, 0.020 for

BC, 0.004 for T2D and 0.017 for CAD). The relative increase in Nagelkerke $R^2$ over other approaches ranged from 11% for T2D to 25% for SCZ. This is consistent with the fact that the simulations showed larger improvements for highly polygenic traits, such as SCZ. Noted that for both CAD and T2D, the accuracy attained with >60,000 training samples from large meta-analyses (Figure 4.8) is actually lower than the accuracy attained with <5,000 training samples from the WTCCC (Figure 4.7).

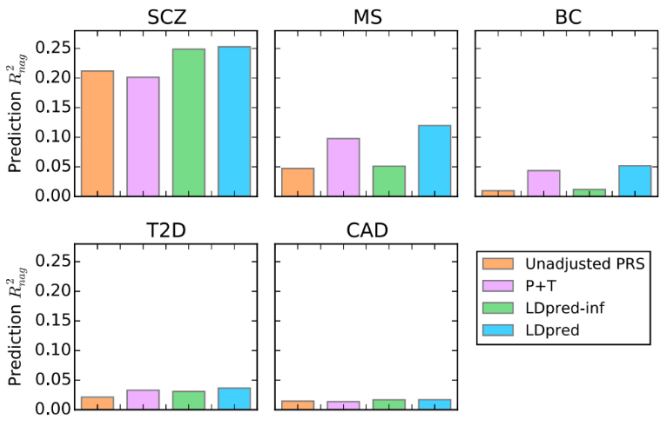


**Figure 4.8**

**Comparison of Methods Training on Large GWAS Summary Statistics for Five Different Diseases [12]**

PRSs are likely to become clinically useful as GWAS sample sizes continue to grow. However, unless LD is appropriately modeled, their predictive accuracy will fall short of their maximal potential. Our results show that LDpred is able to address this problem—even when only summary statistics are available—by estimating posterior mean effect sizes by using a point normal prior and LD information from a reference panel. Intuitively there are two reasons for the relative gain in prediction accuracy of LDpred PRSs over P+T. First, LD pruning discards informative markers and thereby limits the overall heritability explained by the markers. Second, LDpred accounts for the effects of linked markers, which can otherwise lead to biased estimates. These limitations hinder P+T regardless of the LD pruning and p-value thresholds used.[12]

Clarifications: In this study the unadjusted PRS is simply the sum of all the estimated marker effects for each allele, i.e., the standard unadjusted polygenic score for the $I^{th}$ individual is $\sum_{j=1}^{M} X_{ij}\widehat{\beta_j}$, where $X_{ij}$ denotes the genotype for the $i^{th}$ individual and the $j^{th}$ genetic variant.

P+T refers to the strategy of first applying informed LD pruning with $r^2$ threshold 0.2 and subsequently applying p-value thresholding, where the p-value threshold is optimized over a grid with respect to prediction accuracy in the validation data.

### 4.6.2 Study 2

Mak T.S.H et al, performed a number of simulation studies to assess the performance of their proposed method, lassosum. In their first simulation study, they used the Wellcome Trust Case Control Consortium (WTCCC) Phase 1 data for seven diseases. They filtered variants and participants using the following QC criteria: genotype rate >0.99, minor allele frequency >0.01, missing genotype per individual<0.01. SNP rsID included in the 1000 Genome project (Phase 3, release May 2013) genotype data, with matching reference and alternative alleles, on top of the QC done by the original researchers (Wellcome Trust Case Control Consortium, 2007). This resulted in 358,179 SNPs and 15,603 individuals, of which 2,859 were controls. [11]

They randomly chose two 1,000 samples as two test datasets. In the first dataset $\boldsymbol{X}^{(1)}$, validation and pseudovalidation were performed to determine the optimal value of $\lambda$. This choice of $\lambda$ and/or $s$ was applied in the other test dataset $\boldsymbol{X}^{(2)}$ in the assessment of prediction accuracy. Prediction accuracy was assessed by the correlation of the PGS with the true predictor $X^{(2)}\beta$. Except when assessing the performance of using different reference panels, they used the first test dataset $\boldsymbol{X}^{(1)}$ as the reference panel also. In assessing the impact of using different reference panels, they let the 1000 Genome East Asian (EAS) subpopulation ($n = 503$) be their test dataset. They compared the performance of using four different reference panels: (1) the original sample that generated the summary statistics, (2) a sample of 1,000 from the WTCCC, (3) the EUR subpopulation from the 1000 Genome project, and (4) the EAS subpopulation from the 1000 Genome project. The above simulations were repeated 10 times and were compared with the approach of p-value thresholding (with and without clumping) and LDpred. For clumping, they used an $R^2$ of {0.1, 0.2, 0.5, 0.8}. (As mentioned before, Clumping is a method for selectively clumping together SNPs in a LD region. Each region is tagged by a lead SNP. In the method implemented in this paper, they start with the most significant SNP. All SNPs that are correlated with this SNP by an $R^2$ of greater than a certain threshold (e.g. 0.2) within a certain region are clumped together with this SNP. They continue this process with the second and third most significant SNP, until all SNPs are clumped into a region). For p-value thresholding, they used the set of p-values {$5e{-}8$, $1e{-}5$, $1e{-}4$, $1e{-}3$, 0.0015,0.002,0.0025, …,

0.995, 1} as possible p-value thresholds. For LDpred, they used the set of proportion of causal SNPs {0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1}. For p-value thresholding and LDpred, they used a validation dataset as well as pseudovalidation to select the best threshold and proportion of causal SNPs, respectively.[11]

Results: WTCCC simulations were run in this study with summary statistics sample sizes of 10,000, 50,000, and 250,000, respectively. They used two different values for P (causal), which represents the expected proportion of causal SNPs: 0.1 and 0.01. P (causal)=0.01 represents a scenario with fewer causal SNPs and larger effect sizes, whereas P (causal) = 0.1, represents a scenario in which causal SNPs have smaller effect sizes and are more evenly distributed across the genome. Figure 4.9 displays the performance of lassosum with different values of $\lambda$ for one of the simulations. It can be seen that in all the simulation scenarios, the general pattern is that predictive performance increases with $\lambda$ up to a point and then decreases, often rapidly. Using a validation dataset or alternatively pseudovalidation is usually effective in helping us select a value of $\lambda$ that is close to the optimal (Circles are values of λ chosen with a validation dataset and triangles are values of λ chosen with pseudovalidation). Comparing different values of $s$, the shrinkage parameter, they see that the maximum attainable correlation is generally lower for $s = 1$, the scenario where lassosum reduces to soft thresholding, that is, where information on LD is ignored, except when $n = 10{,}000$ and $P$ (causal)=0.1. In addition, $s = 0.5$ and $s = 0.2$ usually gives better performance than $s = 0.9$.[11]

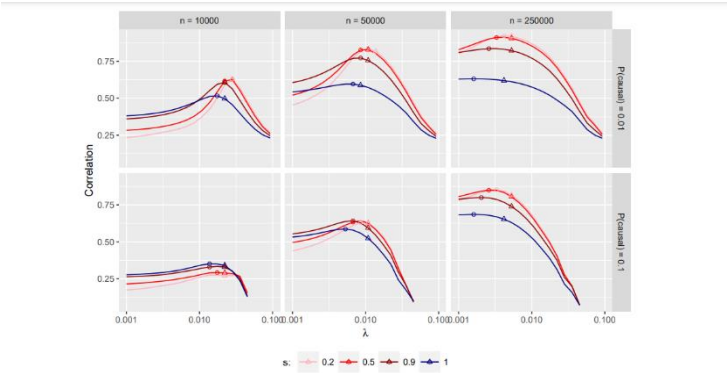


**Figure 4.9**

**The predictive performance of lassosum with respect to λ[11]**

Figure 4.10, shows the average prediction performance over 10 simulations, comparing the use of pseudovalidation and a validation dataset with phenotype data as well as using the

minimum $\lambda$ value of 0.001. They use $\lambda = 0.001$ for comparison because it is shown in Figure 4.9 that in general the prediction performance of lassosum approaches a constant as $\lambda$ tends to 0, whereas when $\lambda$ approaches 1, the performance drops sharply. Thus, using $\lambda$ close to 0 represents a conservative, safe option. When $s = 0.2$ or 0.5, the performance of pseudovalidation was very similar to using a real validation phenotype. Both approaches were clearly superior to the conservative option of setting $\lambda = 0.001$. When $s = 0.9$ or $s = 1$, pseudovalidation was still clearly superior to setting $\lambda = 0.001$ for $n = 10,000$ and $n = 50,000$ and $P$ (causal)=0.01. In all simulations, the performance of p-value thresholding was similar to the use of lassosum with $s = 1$. It is also observed that lassosum with $s = 0.2$ or $s = 0.5$ tended to give the best performance overall. Thus, it is reasonable to maximize over $s$ also using either a validation phenotype or pseudovalidation when using lassosum. [11]

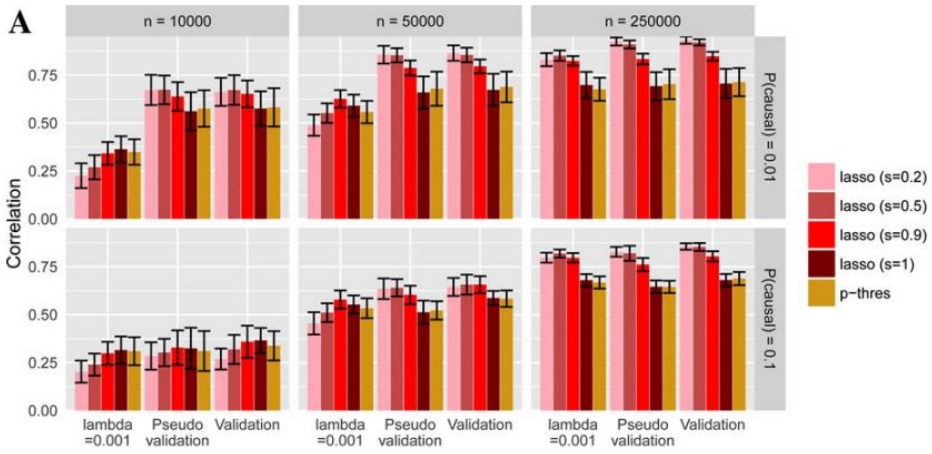


**Figure 4.10**

**Comparing the use of a validation dataset with phenotype data and pseudovalidation in selecting the tuning parameter λ[11]**

Figure 4.11 shows the comparison of the performance of lassosum with clumping and p-value thresholding, as well as with LDpred. For lassosum, there has been optimization over both $\lambda$ and $s = \{0.2, 0.5, 0.9, 1\}$. For comparison, there has been optimization over p-value thresholds and clumping $R^2 = \{0.1, 0.2, 0.5, 0.8,$ no clumping$\}$. Finally, there has been optimization for LDpred over $P$ (causal) = {0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1}. For p-value thresholding, clumping led to a noticeable increase in prediction accuracy, except when $P$ (causal)=0.1 and $n = 10,000$. However, in all scenarios, lassosum was superior to clumping and thresholding. The result was similar whether the method was optimized using a validation

dataset or pseudovalidation. LDpred did not appear to have the claimed advantage over p-value thresholding in these simulations. Which may be due to the fact that the size of the reference sample used was only 1,000, smaller than the recommended size of at least 2,000 in the paper. However, they found that the performance of LDpred did not improve even when the sample size of the reference panel (and test panels) was set to 5,000 (Figure 4.12). [11]

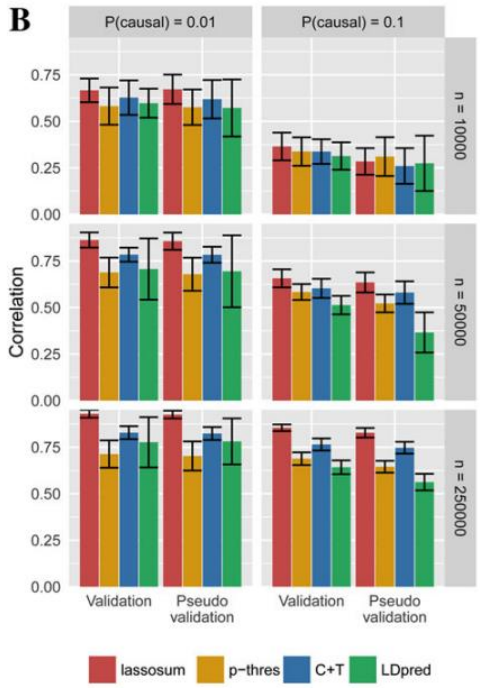


**Figure 4.11**

**Comparing the performance lassosum, (p-thres), C + T and LDpred[11]**

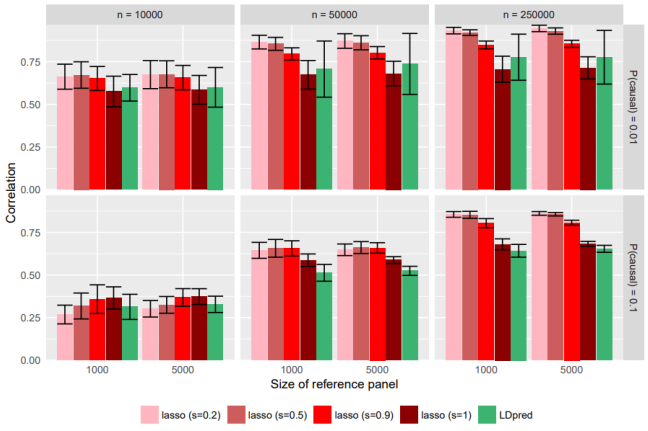


**Figure 4.12**

**Examining the impact of the size of the reference panel in the prediction accuracy of lassosum[11]**

The effect of using different reference panels when using lassosum was investigated in Figure 4.13. There has been generation of the summary statistics using the entire WTCCC sample and using four different reference panels for their LD information: (1) the original WTCCC sample that generated the summary statistics, (2) a sample of 1,000 from the WTCCC, (3) the EUR subpopulation from the 1000 Genome project, and (4) the EAS subpopulation from the 1000 Genome project, which also served as the test sample. It was found that for the small sample size ($n = 10,000$) scenario the use of the different reference panels made relatively little difference to predictive performance. However, as sample size increased, using the true sample that generated the summary statistics led to noticeably improved predictive performance. For many scenarios, using the 1000 Genome EUR sample as the reference panel led to a similar performance as using the original summary statistic sample. A clear advantage for using the summary statistics sample was only shown in the scenario with the most power ($n = 250,000$ and $P$ (causal)=0.01). Using the wrong (EAS) reference sample was clearly inferior when the sample size was above 50,000, but it was still better than simple p-value thresholding. [11]

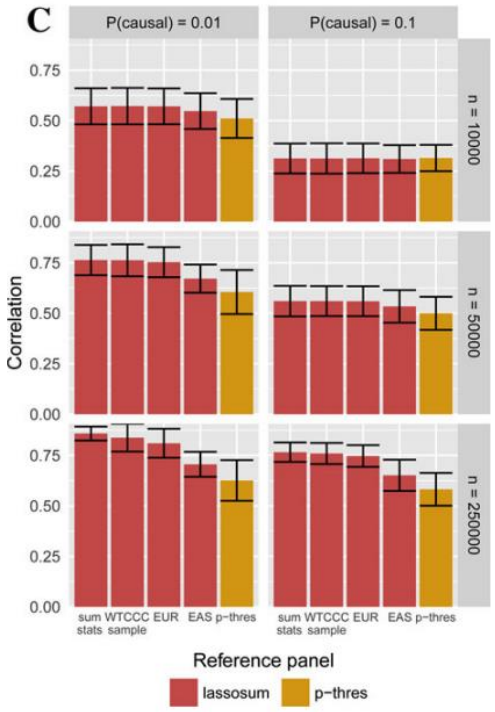


**Figure 4.13**

**The effect of using different reference panels on lassosum [11]**

Next, it was examined the performance of lassosum in a larger simulated dataset with around 8 million SNPs, with a focus on clumping, to see whether prefiltering by clumping can be an effective method in reducing the number of SNPs in the analysis. The sample size for the summary statistics was set at 200,000. Six levels of clumping ($R^2$ = 0.01, 0.05, 0.1,0.2, 0.5, and 0.8) were applied to the data, resulting in around 190.000, 330.000, 430.000, 610.000, 1.170.000, and 1.940.000 SNPs respectively. (The actual number depends on the simulations) It was not perform LDpred for $R^2 > 0.2$ because it was too time and memory intensive. In Figure 4.14A, the results from this simulation appear. As shown, clumping was beneficial in improving prediction performance for p-value thresholding, and the best performance was achieved with an $R^2$ of 0.5 or 0.8. For lassosum, performance decreased with increasing level of clumping (decreasing $R^2$). lassosum with no clumping gave the best performance overall. LDpred performed poorly in this simulation, likely because the reference panel size was too small. [11]

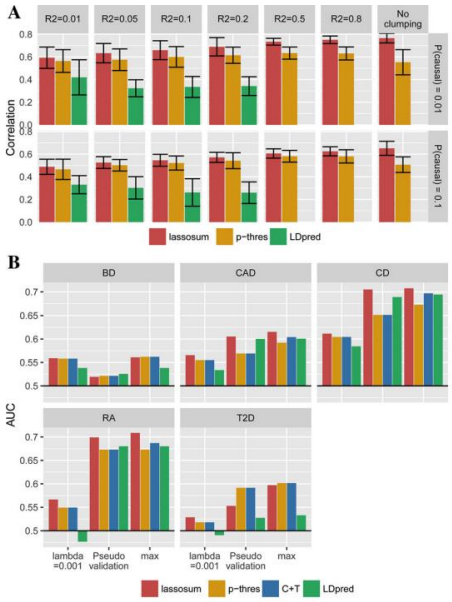


**Figure 4.14**

**(A) Performance of lassosum in a large simulated dataset. (B) Performance of lassosum vs. other methods when using real summary statistics data from meta-analyses[11]**

In Figure 4.14B, the results were presented for using real summary statistics from five large meta-analyses to predict phenotypes in the WTCCC data. [Bipolar disorder:$n$(cases)=7, 481, $n$(controls) = 9, 250, coronary artery disease:$n$(cases) = 60, 801, $n$(controls) = 123, 504, Crohn's disease:$n$(cases) = 22, 575, $n$(controls) = 46, 693, RA:$n$(cases) = 14, 361, $n$(controls)

= 43, 923, Type 2 diabetes:$n$(cases) = 26, 488, $n$(controls) = 83, 964]. Because all these meta-analyses included the WTCCC as one of the studies, PGS derived using these summary statistics directly would overfit the data. To overcome this problem, we attempted to isolate the non-WTCCC components of the summary statistics by reversing the fixed-effects meta-analysis equations: $\beta_{meta} = \frac{\beta_S/\sigma_S^2+\beta_{\bar{S}}/\sigma_{\bar{S}}^2}{1/\sigma_S^2+1/\sigma_{\bar{S}}^2}$, $\frac{1}{\sigma_{meta}^2} = \frac{1}{\sigma_S^2} + \frac{1}{\sigma_{\bar{S}}^2}$,

where, $\beta_S$ and $\sigma_S$, denote the log odds ratio and standard error from the WTCCC study and $\beta_{\bar{S}}$ and $\sigma_{\bar{S}}$ the contribution to the meta-analysis apart from WTCCC. SNPs with negative $\sigma_{\bar{S}}^2$ were set to have zero effect size. Prediction accuracy of the summary statistics-derived PGS were assessed by the area under the ROC curve (AUC) statistic when used to predict disease status in the WTCCC dataset with the relevant disease and the 2,859 controls. The testing sample was also used as the reference panel.[11]

In all cases, the use of pseudovalidation resulted in a PGS that is close to the maximum AUC across all tuning parameters and was clearly superior to using $\lambda$ =0.001. For BD, CAD, CD, and RA, the performance of lassosum, LDpred, and clumping and thresholding were similar, although a slightly higher AUC was observed for lassosum. For T2D, the maximum AUC was surprisingly achieved by p-value thresholding without clumping. [11]

In summary, Mak T.S.H *et al.* have proposed the calculation of PGS using a penalized regression approach using summary statistics and examined its performance in simulation experiments. Their proposed approach, lassosum, in general appeared to give better prediction than p-value thresholding with or without clumping as well as the LDpred, for which they failed to demonstrate the claimed superior performance over p-value thresholding. Clumping was beneficial for p-value thresholding in most scenarios but not for lassosum. In some scenarios, clumping actually decreases the predictive power of p-value thresholding, such as in the simulations with $P$(causal)=0.1 and $n$ = 10,000. Also, increasing the sample size of the reference panel will generally increase prediction accuracy as well, although this comes at a cost of exponentially increasing running times. In simulations, Mak T.S.H found that gains in prediction accuracy from a larger reference panel were usually modest.[11]

Moreover, they showed that pseudovalidation method is effective in selecting a parameter value that is close to the optimum. Having a validation dataset with phenotype data generally provides an even more reliable method for selecting the tuning parameter. However, in the

event where this is unavailable, pseudovalidation offers an alternative and can be applied to any PGS method that requires a tuning parameter. Mak T.S.H *et al.* have focused on the performance of lassosum as a method, they note that it is more generally an instance of penalized regression. Potentially, other penalties can be used, which can lead to even better prediction. They chose the LASSO penalty because of its simplicity. Some limitations of the present study are worth bearing in mind when considering these results. For example, summary statistics may be inflated due to population stratification in the data where they are generated. As summary statistics are often derived from meta-analyses, it is also possible that there is underlying heterogeneity in effect sizes. How these impact PRS calculation is currently unknown. The simplicity of lassosum makes it an ideal framework from which more complex methods can be developed. [11]

### 4.6.3 Study 3

Type 2 diabetes (T2D) is a global public health problem. Identifying individuals at high risk for T2D for early targeted detection, prevention and intervention is of great public health importance. In addition to known behavioral and environmental factors, T2D has been shown to have a strong genetic component. Genome-wide association studies (GWAS) have successfully identified many common genetic variants that confer susceptibility to T2D. However, all of these common genetic variants discovered by GWAS may only be able to explain a small proportion of the overall heritability and therefore result in low predictive power. The polygenic risk score (PRS) that aggregates the information of many common single nucleotide polymorphisms (SNPs), weighted by the effect size resulting from large-scale GWAS discovery, has been used to predict T2D risk. PRS is expected to have better predictive power and the potential to improve performance in T2D risk assessment. In this study, Liu *et al.,* to further explore the prediction capability of the PRS model in identifying high-risk individuals for T2D, proposed a new strategy to construct PRS model by the following three-step filtering procedure to consider a statistical compromise between signal and noise. First, rather than including SNPs across the whole genome, it selected a subset of SNPs by a lenient significance threshold ($p \leq 5 \times 10^{-2}$) from a huge number of SNPs included in large-scale GWASs. Second, it set $r^2$ equal to 0.2, 0.4, 0.6, and 0.8 as candidate LD pruning thresholds. Third, it set p-value thresholds as $5 \times 10^{-2}$, $5 \times 10^{-4}$, $5 \times 10^{-6}$, and $5 \times 10^{-8}$. After applying the above thresholds to the GWAS summary data, a total of 16 candidate PRS models were then generated. Testing was conducted using the UKB testing dataset (n = 182,422) to avoid the

model overfitting issue. Finally, the best predictive PRS model among a set of candidate PRS models was chosen and evaluated in the UKB validation dataset (n = 262,751). They also considered non-genetic risk factors, including sex, age, physical measurements, and clinical factors, to further increase prediction accuracy. [54]

The study was conducted based on the UKB project, one of the largest prospective cohort studies. A total of 487,409 individuals with available genotyping array and a total of 625,394 variants were originally collected from UKB. Subsequently, SNPs and individuals with very high levels of missingness were filtered out. Based on a relaxed threshold of 0.2 (>20%), 89,752 variants and 30,855 subjects removed. There were also 262,751 SNPs removed with minor allele frequency $< 1 \times 10^{-6}$. Finally, 456,451 individuals and 271,687 variants passed QC and were considered in the following analysis. Liu *et al.*, further imputed the inevitably missing values of T2D-related risk factors, including sex, age, physical measures [e.g., BMI, waist circumference] and clinical factors [e.g., high-density lipoprotein (HDL), low-density lipoprotein (LDL)] by their means. To analyze individuals with a relatively homogeneous ancestry, the population was constructed centrally based on a combination of self-reported ancestry and genetically confirmed ancestry using the first 10 principal components (i.e., PC1, ..., PC10). To construct, test, and further validate the robustness of the polygenic predictor of T2D, they randomly divided the overall data into two parts, i.e., the testing and validation dataset. The data were split over two datasets, a testing dataset and a validation dataset. Liu *et al.*, assigned 40% of all individuals to the UKB testing dataset (n = 182,422) and the remaining 60% to the UKB validation dataset (n = 274,029). Other ratios were also tried to divide the testing and validation datasets, i.e., 30–70%, 50–50%, 60–40%, and 70–30%. There were nearly 5.494% (n = 10,023) participants who were cases in the testing dataset and 5.575% (n = 15,277) in the validation dataset. Individuals in the UKB validation dataset were distinct from those in the UKB testing dataset. The details of the study design are described in Figure 4.15. For PRS model construction, summary statistics from a T2D GWAS conducted among 60,786 participants with 12,056,346 SNPs of European ancestry were used. The UKB samples did not overlap with the samples from the discovery GWAS. From these summary statistics, SNPs were selected according to the association p-values ($p \leq 5 \times 10^{-2}$) obtained from the above GWAS, and 50,224 SNPs remained. Liu et al., then considered multiple $r^2$ thresholds (0.2, 0.4, 0.6, and 0.8) and p-value thresholds ($5 \times 10^{-2}$,$5 \times 10^{-4}$,$5 \times 10^{-6}$, and $5 \times 10^{-8}$). [54]

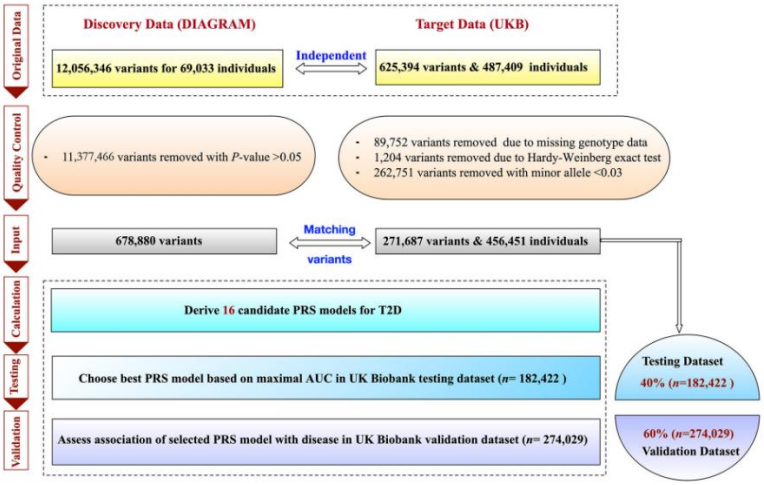


**Figure 4.15**

**Flowchart for the polygenic risk score (PRS) model for type 2 diabetes[54]**

A total of 16 candidate PRS models were generated to obtain an optimal PRS model. The performance of these 16 PRS models was evaluated in the UKB testing dataset and they chose the best one for further validation analysis. The AUCs of these 16 candidate PRS models ranged from 0.691 to 0.792 (Figure 4.16). The best PRS model was selected with the highest AUC [AUC = 0.792, 95% CI: (0.787, 0.796)] based on 25,454 SNPs when $p \leq 5 \times 10^{-2}$ and $r^2 < 0.2$. The AUCs of different ratios of the testing and validation datasets are shown in Figure 5.17. The AUCs of different ratios were very close to each other, ranging from 0.791 to 0.795. The AUC of the 40–60% ratio had the best performance in the validation dataset [AUC = 0.795, 95% CI: (0.790, 0.800)].

| Tuning parameter | SNP number | AUC (95% CI) |
|---|---|---|
| $p \leq 5 \times 10^{-8}$ and $r^2 < 0.2$ | 363 | 0.706 (0.701–0.711) |
| $p \leq 5 \times 10^{-8}$ and $r^2 < 0.4$ | 486 | 0.702 (0.697–0.707) |
| $p \leq 5 \times 10^{-8}$ and $r^2 < 0.6$ | 670 | 0.696 (0.691–0.701) |
| $p \leq 5 \times 10^{-8}$ and $r^2 < 0.8$ | 957 | 0.691 (0.686–0.697) |
| $p \leq 5 \times 10^{-6}$ and $r^2 < 0.2$ | 750 | 0.715 (0.710–0.720) |
| $p \leq 5 \times 10^{-6}$ and $r^2 < 0.4$ | 1,013 | 0.709 (0.704–0.714) |
| $p \leq 5 \times 10^{-6}$ and $r^2 < 0.6$ | 1,335 | 0.701 (0.696–0.706) |
| $p \leq 5 \times 10^{-6}$ and $r^2 < 0.8$ | 1,853 | 0.696 (0.691–0.701) |
| $p \leq 5 \times 10^{-4}$ and $r^2 < 0.2$ | 2,616 | 0.736 (0.732–0.741) |
| $p \leq 5 \times 10^{-4}$ and $r^2 < 0.4$ | 3,394 | 0.726 (0.721–0.731) |
| $p \leq 5 \times 10^{-4}$ and $r^2 < 0.6$ | 4,299 | 0.715 (0.710–0.720) |
| $p \leq 5 \times 10^{-4}$ and $r^2 < 0.8$ | 5,690 | 0.708 (0.703–0.713) |
| $p \leq 5 \times 10^{-2}$ and $r^2 < 0.2$ | **25,454** | **0.792 (0.787–0.796)** |
| $p \leq 5 \times 10^{-2}$ and $r^2 < 0.4$ | 32,600 | 0.782 (0.777–0.787) |
| $p \leq 5 \times 10^{-2}$ and $r^2 < 0.6$ | 40,001 | 0.771 (0.766–0.776) |
| $p \leq 5 \times 10^{-2}$ and $r^2 < 0.8$ | 50,224 | 0.760 (0.755–0.765) |

*AUC was determined using a logistic regression model adjusted for sex, age, and the first 10 principal components of ancestry. The highest AUC is denoted by the bold values.*

**Figure 4.16**

**The predictive power of candidate polygenic risk score (PRS) models for T2D[54]**

| Dataset | 30–70% | 40–60% | 50–50% | 60–40% | 70–30% |
|---|---|---|---|---|---|
| Testing | 0.791 | 0.792 | 0.794 | 0.795 | 0.794 |
| | (0.781–0.791) | (0.787–0.796) | (0.790–0.800) | (0.791–0.799) | (0.790–0.799) |
| Validation | 0.794 | 0.795 | 0.793 | 0.792 | 0.791 |
| | (0.790–0.799) | (0.790–0.800) | (0.789–0.797) | (0.787–0.796) | (0.781–0.791) |

*AUC was determined using a logistic regression model adjusted for sex, age, and first 10 principal components of ancestry.*

**Figure 4.17**

**AUCs of different ratios of the testing and validation dataset[54]**

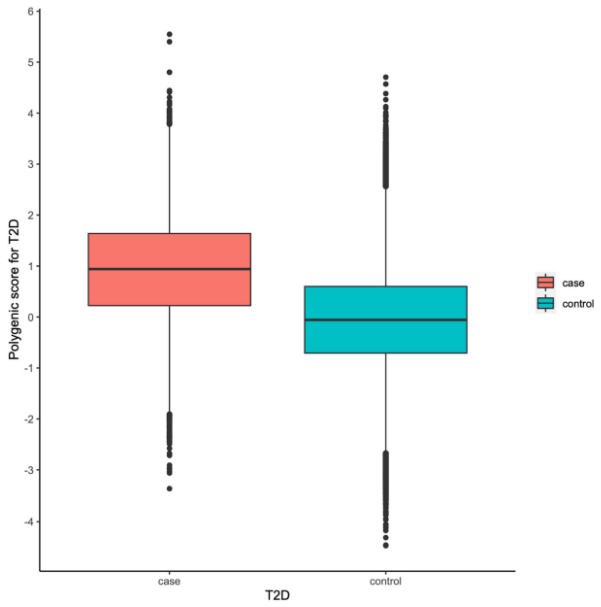


**Figure 4.18**

**PRS among type 2 diabetes (T2D) cases versus controls in the UKB validation dataset[54]**

To facilitate interpretation, they scaled PRS to have a zero mean and one standard deviation. Liu *et al.,* investigated whether their PRS model could identify individuals at high T2D risk. Figure 4.18 showed that the median of the standardized PRS was 0.941 for individuals with T2D versus -0.056 for individuals without T2D, a difference of 0.997 ($p < 0.00001$). The standardized PRS approximated a normal distribution across the population, with the empirical risk of T2D rising sharply in the right tail of the distribution (Figure 4.19). The PRS model identified nearly 30% of the population at greater than or equal to fivefold risk, 12% of the population at greater than or equal to sixfold risk, and the top 7% of the population at greater than or equal to sevenfold increased risk for T2D, as shown in Figure 4.19A. Then, they stratified the population according to the percentiles of the PRS and defined the top 10 percentiles as the "high risk" group while the bottom 10 percentiles were the "low risk" group. The odds ratio was assessed in a logistic regression model adjusting for sex, age, and the first 10 principal components of ancestry. Figure 4.19B shows the prevalence of T2D increases with the percentiles of the PRS model. There were 5,642 (18.698%) cases in the "high risk" group among 30,174 individuals, while only 282 (0.935%) cases in the "low risk" group, corresponding to a nearly 20-fold increase in the risk of T2D comparing the top 10 percentiles versus the bottom 10 percentiles.[54]

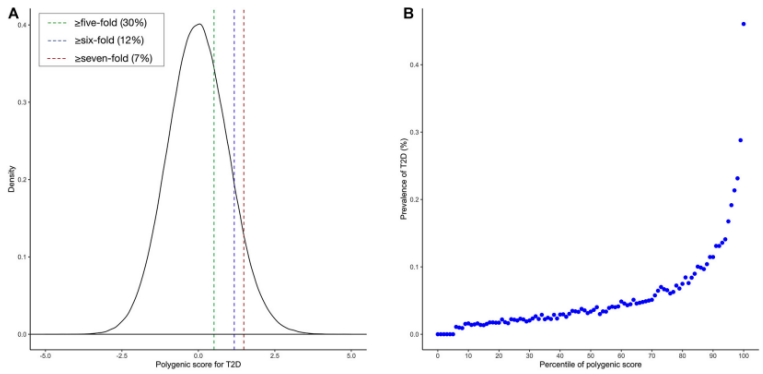


**Figure 4.19 | Risk for type 2 diabetes (T2D) according to polygenic risk score (PRS)**

**(A) Distribution of PRS for T2D in the UKB validation dataset (n = 301,736) (B) Prevalence of T2D according to 100 groups of the UKB validation dataset stratified according to the percentile of the PRS for T2D[54]**

Figure 4.20 showed that the AUCs of $model_3$, which only included PRS into the prediction model without adjusting for any other covariates, was 0.749 [95% CI: (0.744,0.754)] in the testing dataset and 0.755 [95% CI: (0.752, 0.755)] in the validation dataset. Interestingly, if only considering sex, age, and the first 10 principal components of ancestry into the model, the AUC was 0.667 [95% CI: (0.663, 0.672)]. After adding PRS, the AUC reached 0.795 [95% CI: (0.790, 0.800)], which increased about 13% than $model_2$. The AUC of $model_4$ (i.e., considering sex, age, PC, BMI, WC, DBP, SBP, GL, CL, HDL, LDL, and TL simultaneously) was 0.880 [95% CI: (0.878, 0.888)] and raised to 0.901 [95% CI: (0.897, 0.904)] in the validation dataset when adding PRS into the model. In brief, the polygenic score indeed helps to identify high-risk individuals for T2D, while the role of T2D-related covariates could also help increase prediction accuracy. As showed in Figure 5.22, PRS, sex, age, physical measurements, and most clinical factors were all significantly associated with T2D ($p < 0.0001$).[54]

| Dataset | Mean | $model_2$ | $model_3$ | $model_1$ | $model_4$ | $model_5$ |
|---|---|---|---|---|---|---|
| Testing | −0.003 | 0.671 (0.666–0.676) | 0.749 (0.744–0.754) | 0.792 (0.787–0.796) | 0.886 (0.882–0.889) | 0.902 (0.899–0.905) |
| Validation | −0.003 | 0.667 (0.663–0.672) | 0.755 (0.752–0.755) | 0.795 (0.790–0.800) | 0.882 (0.878–0.888) | 0.901 (0.897–0.904) |

*$model_1$: AUC was determined using a logistic regression model adjusted for sex, age, and the first 10 principal components of ancestry. $model_2$: AUC was determined using a logistic regression model only considering sex and age. $model_3$ : AUC was determined using a logistic regression model only considering genome-wide polygenic score. $model_4$: AUC was determined using a logistic regression model considering demographic factors, physical measurements, and clinical factors. $model_5$ : AUC was determined using a logistic regression model adjusted for sex, age, body mass index, waist circumference, diastolic blood pressure, systolic blood pressure, glucose level, cholesterol level, high-density lipoprotein, low-density lipoprotein, triglyceride level, and the first 10 principal components of ancestry.*

**Figure 4.20 | AUC of different models in the testing and validation dataset[54]**

$$\mathrm{T2D} \sim \mathrm{PRS} + \mathrm{sex} + \mathrm{age} + \boldsymbol{PC}$$

$$model_2 : \mathrm{T2D} \sim \mathrm{sex} + \mathrm{age} + \boldsymbol{PC};$$

$$model_3 : \mathrm{T2D} \sim \mathrm{PRS};$$

$$model_4 : \mathrm{T2D} \sim \mathrm{sex} + \mathrm{age} + \boldsymbol{PC} + \mathrm{BMI} + \mathrm{GL} + \mathrm{CL} + \mathrm{HDL} + \mathrm{LDL} + \mathrm{TL} + \mathrm{WC} + \mathrm{DBP} + \mathrm{SBP};$$

$$model_5 : \mathrm{T2D} \sim \mathrm{PRS} + \mathrm{sex} + \mathrm{age} + \boldsymbol{PC} + \mathrm{BMI} + \mathrm{GL} + \mathrm{CL} + \mathrm{HDL} + \mathrm{LDL} + \mathrm{TL} + \mathrm{WC} + \mathrm{DBP} + \mathrm{SBP}.$$

**Figure 4.21**

**Prediction models[54]**

In summary, about 30% of participants were at greater than or equal to fivefold increased risk of developing T2D, 12%were at greater than or equal to sixfold risk, and the top 7% were at greater than or equal to sevenfold increased risk. Particularly, the stratified PRS according to their percentiles showed that the “high-risk” group is strongly associated with the risk of T2D. The above results suggest that our PRS model can be used as a powerful tool in identifying individuals at high risk of T2D. Although the present study has made important contributions in identifying individuals with increased risk of developing T2D; however, there exists one major limitation. Individuals in the UKB dataset are primarily European ancestry; the specific PRS calculated here may not have optimal predictive power for other ethnic groups because the allele frequencies, LD patterns, and effect sizes of common SNPs may be different across populations with different ethnic backgrounds.[54]

| Variables | Estimate beta | Stand error | Z | p-value |
|---|---|---|---|---|
| (Intercept) | 24.500 | 0.495 | 49.474 | $< 2 \times 10^{-16}$ |
| PRS | 12370.000 | 167.400 | 73.943 | $< 2 \times 10^{-16}$ |
| CL | −0.591 | 0.057 | −10.377 | $< 2 \times 10^{-16}$ |
| HDL | 0.051 | 0.063 | 0.876 | 0.381 |
| LDL | 0.010 | 0.068 | 0.140 | 0.888 |
| TL | 0.285 | 0.013 | 21.826 | $< 2 \times 10^{-16}$ |
| Sex | −0.214 | 0.028 | −7.731 | $1.070 \times 10^{-14}$ |
| WC | 0.045 | 0.002 | 28.356 | $< 2 \times 10^{-16}$ |
| BMI | 0.036 | 0.004 | 9.325 | $< 2 \times 10^{-16}$ |
| Age | 0.060 | 0.002 | 38.401 | $< 2 \times 10^{-16}$ |
| DBP | −0.018 | 0.001 | −13.928 | $< 2 \times 10^{-16}$ |
| SBP | 0.005 | 0.001 | 7.626 | $2.410 \times 10^{-16}$ |
| GL | 0.449 | 0.006 | 69.917 | $< 2 \times 10^{-16}$ |
| PC10 | 0.020 | 0.004 | 4.726 | $2.280 \times 10^{-16}$ |

*BMI, body mass index; CL, cholesterol level; DBP, diastolic blood pressure; GL, glucose level; PRS, genome-wide polygenic score; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SBP, systolic blood pressure; TL, triglyceride level; WC, waist circumference.*

**Figure 4.22**

**Parameter estimations under model5 in validation dataset][54]**

## 4.7 Conclusion

PRS is a useful tool that can be used to summarize genetic information into a single variable for statistical analysis. However, PRSs are only now beginning to make the transition from association analyses in research settings to utility in clinical settings, and there are a number of technical, practical, and ethical issues that must be addressed before widespread clinical adoption. As the body of literature surrounding PRS increases, so too will the ability to use PRSs to predict relative disease risk earlier in life. This relies upon the notion that an individual's personal genome is effectively unchanging from birth to death, so genetic risk should remain constant. Although there is much work to be done to make this a reality, it may eventually be possible that clinicians will be able to determine risk for a wide array of diseases based on a single PRS-optimized genotyping chip. Informing disease risk for a myriad of outcomes early in life will help improve individualized prevention efforts, screening, and life planning.

The utilities of PRS have been explored in many common diseases, such as cancer, coronary artery disease, obesity, and diabetes, and in various non-disease traits, such as clinical biomarkers. These applications demonstrated that PRS could identify a high-risk subgroup of these diseases as a predictive biomarker and provide information on modifiable risk factors driving health outcomes. On the other hand, there are several limitations to implementing PRSs in clinical practice, such as biased sensitivity for the ethnic background of PRS calculation and

geographical differences even in the same population groups. Also, it remains unclear which method is the most suitable for the prediction with high accuracy among numerous PRS methods developed so far. Although further improvements of its comprehensiveness and generalizability will be needed for its clinical implementation in the future, PRS will be a powerful tool for therapeutic interventions and lifestyle recommendations in a wide range of diseases.[56]

In summary, this thesis has reviewed common methods to construct and evaluate PRSs. In developing and performing a PRS analysis, there are many options to consider depending on the underlying goals of the study. Careful use of data and interpretation of results are a necessity in order not to overstate the current clinical importance of PRSs. Nevertheless, the potential for disease prediction using PRSs should not be ignored and, with increasing sample sizes, their use should increase if limitations are appropriately identified.[53] Predictive diagnosis or risk profiling should provide opportunities for environmental modification (such as smoking cessation), early therapy (for example, administering statins for individuals at risk of cardiovascular disease) or targeted cancer screening (for example, the use of colonoscopy in families or individuals at genetic risk of colorectal cancer). Diagnostic medicine will become increasingly important as our understanding of disease susceptibility and progression markers improves and as the tools for rapid and effective disease prediction and monitoring are developed.[2]

# Bibliography

[1] R. P. Igo Jr, T. G. Kinzy, and J. N. Cooke Bailey, 'Genetic risk scores', *Current protocols in human genetics*, vol. 104, no. 1, p. e95, 2019.

[2] S. Song, W. Jiang, L. Hou, and H. Zhao, 'Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies', *PLOS Computational Biology*, vol. 16, no. 2, p. e1007565, 2020.

[4] J. A. Collister, X. Liu, and L. Clifton, 'Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists', *Frontiers in Genetics*, vol. 13, 2022.

[5] S. W. Choi, T. S.-H. Mak, and P. F. O'Reilly, 'Tutorial: a guide to performing polygenic risk score analyses', *Nature protocols*, vol. 15, no. 9, pp. 2759–2772, 2020.

[6] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, 'New approaches to population stratification in genome-wide association studies', *Nature reviews genetics*, vol. 11, no. 7, pp. 459–463, 2010.

[7] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G. Blum, 'Making the most of clumping and thresholding for polygenic scores', *The American Journal of Human Genetics*, vol. 105, no. 6, pp. 1213–1221, 2019.

[8] G. Ni *et al.*, 'A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts', *Biological psychiatry*, vol. 90, no. 9, pp. 611–620, 2021.

[9] F. Privé, 'Genetic risk score based on statistical learning', 2019.

[10] R. P. Igo Jr, T. G. Kinzy, and J. N. Cooke Bailey, 'Genetic risk scores', *Current protocols in human genetics*, vol. 104, no. 1, p. e95, 2019.

[11] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham, 'Polygenic scores via penalized regression on summary statistics', *Genetic epidemiology*, vol. 41, no. 6, pp. 469–480, 2017.

[12] B. J. Vilhjálmsson *et al.*, 'Modeling linkage disequilibrium increases accuracy of polygenic risk scores', *The american journal of human genetics*, vol. 97, no. 4, pp. 576–592, 2015.

[13] H.-C. So and P. C. Sham, 'Improving polygenic risk prediction from summary statistics by an empirical Bayes approach', *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[14] D. Chasioti, J. Yan, K. Nho, and A. J. Saykin, 'Progress in polygenic composite scores in Alzheimer's and other complex diseases', *Trends in Genetics*, vol. 35, no. 5, pp. 371–382, 2019.
[15] T. Ge, C.-Y. Chen, Y. Ni, Y.-C. A. Feng, and J. W. Smoller, 'Polygenic prediction via Bayesian regression and continuous shrinkage priors', *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.
[16] L. D. Giacco and C. Cattaneo, 'Introduction to genomics', *Molecular Profiling*, pp. 79–88, 2012.
[17] P. M. Visscher, W. G. Hill, and N. R. Wray, 'Heritability in the genomics era—concepts and misconceptions', *Nature reviews genetics*, vol. 9, no. 4, pp. 255–266, 2008.
[18] N. R. Wray, S. H. Lee, D. Mehta, A. A. Vinkhuyzen, F. Dudbridge, and C. M. Middeldorp, 'Research review: polygenic methods and their application to psychiatric traits', *Journal of child psychology and psychiatry*, vol. 55, no. 10, pp. 1068–1087, 2014.
[19] A. J. Vargas and C. C. Harris, 'Biomarker development in the precision medicine era: lung cancer as a case study', *Nature Reviews Cancer*, vol. 16, no. 8, pp. 525–537, 2016.
[20] G. Novelli, C. Ciccacci, P. Borgiani, M. P. Amati, and E. Abadie, 'Genetic tests and genomic biomarkers: regulation, qualification and validation', *Clinical cases in mineral and bone metabolism*, vol. 5, no. 2, p. 149, 2008.
[21] R. Simon, 'Genomic biomarkers in predictive medicine. An interim analysis', *EMBO molecular medicine*, vol. 3, no. 8, pp. 429–435, 2011.
[22] D. G. Covell, 'Data mining approaches for genomic biomarker development: applications using drug screening data from the cancer genome project and the cancer cell line encyclopedia', *PLoS One*, vol. 10, no. 7, p. e0127433, 2015.
[23] K. Strimbu and J. A. Tavel, 'What are biomarkers?', *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010.
[24] J. K. Aronson and R. E. Ferner, 'Biomarkers—a general review', *Current protocols in pharmacology*, vol. 76, no. 1, pp. 9–23, 2017.
[25] J. J. Douglas and J. A. Roussel, 'The role of genomics to identify biomarkers and signaling molecules during severe sepsis.', *Minerva Anestesiologica*, vol. 82, no. 3, pp. 343–358, 2015.
[26] S. Matsui, 'Genomic biomarkers for personalized medicine: development and validation in clinical studies', *Computational and mathematical methods in medicine*, vol. 2013, 2013.

[27] M. Hassan *et al.*, 'Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review', *International Journal of Molecular Sciences*, vol. 23, no. 9, p. 4645, 2022.

[28] J. P. Ioannidis, P. Castaldi, and E. Evangelou, 'A compendium of genome-wide associations for cancer: critical synopsis and reappraisal', *JNCI: Journal of the National Cancer Institute*, vol. 102, no. 12, pp. 846–858, 2010.

[29] B. K. Bulik-Sullivan *et al.*, 'LD Score regression distinguishes confounding from polygenicity in genome-wide association studies', *Nature genetics*, vol. 47, no. 3, pp. 291–295, 2015.

[30] J. H. Moore and M. D. Ritchie, 'The challenges of whole-genome approaches to common diseases', *Jama*, vol. 291, no. 13, pp. 1642–1643, 2004.

[31] R. Tibshirani, 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[32] F. Gunes, 'Penalized regression methods for linear models in SAS/STAT®', presented at the Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc. http://support. sas. com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels. pdf, 2015.

[33] I. H. T. Guideline, 'Definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories e15', 2007.

[34] G. Novelli, C. Ciccacci, P. Borgiani, M. P. Amati, and E. Abadie, 'Genetic tests and genomic biomarkers: regulation, qualification and validation', *Clinical cases in mineral and bone metabolism*, vol. 5, no. 2, p. 149, 2008.

[35] D. Golan, S. Rosset, and D.-Y. Lin, 'Mixed models for case-control genome-wide association studies: major challenges and partial solutions', in *Handbook of Statistical Methods for Case-Control Studies*, Chapman and Hall/CRC, 2018, pp. 495–514.

[36] C. Dandine-Roulland and H. Perdry, 'The use of the linear mixed model in human genetics', *Human heredity*, vol. 80, no. 4, pp. 196–206, 2015.

[37] A. S. Kaler, J. D. Gillman, T. Beissinger, and L. C. Purcell, 'Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize', *Frontiers in plant science*, p. 1794, 2020.

[38] B. Hayes, 'Overview of statistical methods for genome-wide association studies (GWAS)', *Genome-wide association studies and genomic prediction*, pp. 149–169, 2013.

[39] H. Zhu and X. Zhou, 'Statistical methods for SNP heritability estimation and partition: A review', *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1557–1568, 2020.
[40] E. Génin, 'Missing heritability of complex diseases: case solved?', *Human Genetics*, vol. 139, no. 1, pp. 103–113, 2020.
[41] E. Uffelmann *et al.*, 'Genome-wide association studies', *Nature Reviews Methods Primers*, vol. 1, no. 1, pp. 1–21, 2021.
[42] S. Turner *et al.*, 'Quality control procedures for genome-wide association studies', *Current protocols in human genetics*, vol. 68, no. 1, pp. 1–19, 2011.
[43] M. H. Wang, H. J. Cordell, and K. Van Steen, 'Statistical methods for genome-wide association studies', 2019, vol. 55, pp. 53–60.
[44] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes, 'A guide to genome-wide association analysis and post-analytic interrogation', *Statistics in medicine*, vol. 34, no. 28, pp. 3769–3792, 2015.
[45] C. Li, Y. Fu, R. Sun, Y. Wang, and Q. Wang, 'Single-locus and multi-locus genome-wide association studies in the genetic dissection of fiber quality traits in upland cotton (Gossypium hirsutum L.)', *Frontiers in plant science*, vol. 9, p. 1083, 2018.
[46] N. Mkize, A. Maiwashe, K. Dzama, B. Dube, and N. Mapholi, 'Suitability of GWAS as a Tool to Discover SNPs Associated with Tick Resistance in Cattle: A Review', *Pathogens*, vol. 10, no. 12, p. 1604, 2021.
[47] A. J. Vargas and C. C. Harris, 'Biomarker development in the precision medicine era: lung cancer as a case study', *Nature Reviews Cancer*, vol. 16, no. 8, pp. 525–537, 2016.
[48] S. C. Roth, 'What is genomic medicine?', *Journal of the Medical Library Association: JMLA*, vol. 107, no. 3, p. 442, 2019.
[49] R. H. Horton and A. M. Lucassen, 'Recent developments in genetic/genomic medicine', *Clinical Science*, vol. 133, no. 5, pp. 697–708, 2019.
[50] Widmer, C. *et al.* Further improvements to linear mixed models for genome-wide association studies. *Scientific reports* **4**, 1–13 (2014).
[51] Stephan, J., Stegle, O. & Beyer, A. A random forest approach to capture genetic effects in the presence of population structure. *Nature communications* **6**, 1–10 (2015).
[52] Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Human molecular genetics* **28**, R133–R142 (2019).

[53] Osterman, M. D., Kinzy, T. G. & Bailey, J. N. C. Polygenic risk scores. *Current Protocols* **1**, e126 (2021).

[54] Liu, W., Zhuang, Z., Wang, W., Huang, T. & Liu, Z. An improved genome-wide polygenic score model for predicting the risk of type 2 diabetes. *Frontiers in genetics* **12**, 632385 (2021).

[55] Bravo-Merodio, L. *et al.* Translational biomarkers in the era of precision medicine. *Advances in clinical chemistry* **102**, 191–232 (2021).

[56] Konuma, T. & Okada, Y. Statistical genetics and polygenic risk score for precision medicine. *Inflammation and Regeneration* **41**, 1–5 (2021).

[57] Zheng, S. L. *et al.* Cumulative association of five genetic variants with prostate cancer. *New England Journal of Medicine* **358**, 910–919 (2008).

## Books

[58] Mills, M. C., Barban, N. & Tropf, F. C. *An introduction to statistical genetic data analysis*. (Mit Press, 2020).

[59] Foulkes, A. Applied statistical genetics with R-for population-based association studies [e-Book]. (2019).

[60] Blum, A., Hopcroft, J. & Kannan, R. *Foundations of data science*. (Cambridge University Press, 2020).

[61] Laird, N. M. & Lange, C. *The fundamentals of modern statistical genetics*. (Springer, 2011).

## Internet Sites

[62] Wikipedia contributors, Quantitative trait locus — Wikipedia, the free encyclopedia, https://en.wikipedia.org/w/index.php?title= Quantitative_trait_locus&oldid=1112557834, [Online; accessed 27- September-2022] (2022).

[63] https://isogg.org/wiki/Identical_by_descent

[64] S. M. Downes, L. Matthews, Heritability, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Spring 2020 Edition, Metaphysics Research Lab, Stanford University, 2020.