Development and Application of Methodology for Genome-Wide Association Studies of Age of Disease Onset in Homogeneous and Admixed Populations

Thesis submitted in accordance
with the requirements of the University of Liverpool
for the degree of Doctor in Philosophy
by

Odessica Nicole Hughes

August 2020

Abstract

In the ensuing "post genome era", genome-wide association studies (GWAS) emerged as a powerful and eventually the standard tool for investigating the genetic architecture of common human diseases. Despite the success of the GWAS design, there are inherent limitations that initially limited its widespread use, particularly in populations with diverse genetic backgrounds due to issues of confounding resulting from population structure and admixture. Current GWAS methodology is limited in its ability to distinguish the genetic variants associated with the age-of-onset (AOO) of disease independently of overall risk variants. However, genetic risk scores (GRS), which simultaneously utilize information from multiple single nucleotide polymorphisms (SNPs), have the potential to identify individuals at risk of early-age-onset (EAO) disease because they are expected to have a greater genetic burden. The focus of the research in this thesis is to evaluate methods for detecting associations of AOO of disease with SNPs and GRS in the presence of population structure and admixture.

First, a simulation study was undertaken to evaluate the power to detect association between SNPs and AOO of disease in an admixed population under an additive genetic model in a time-to-event (TTE) framework. Investigations compared the performance of the Cox proportional hazards (PH) model and the general Weibull model. The simulation study evaluated the impact of population admixture on statistical power. Results demonstrated that the power of the general Weibull model was largely consistent with that of the Cox PH model. The presence of an inflated type I error rate due to confounding with ancestry was evident in the testing of association of AOO of disease with a tag SNP for the causal variant, which was addressed with the inclusion of ancestry as a covariate in the model. Additionally, results also suggested that greater levels of admixture have the potential to result in reduced power, particularly in situations where the risk allele frequency (RAF) is very different between the ancestral populations when testing is based on the tag SNP, rather than the causal variant, which is the more common occurrence in GWAS.

Second, an investigation of the utility of GRS to detect an association with AOO of disease in ancestrally homogenous populations was undertaken, which featured a component of both simulation and real data application. The work entailed the application of GRS to investigate the association of AOO of type 2 diabetes (T2D) in two independent GWAS originating from the Northwestern University Gene (NUgene) Banking Project and the Wellcome Trust Case Control Consortium (WTCCC). Investigations were centred around three modelling approaches: Cox PH, proportional odds, and logistic regression models. The Cox PH model, which considered both cases and controls in a TTE framework assessed AOO at the end of the study period, while controls were censored at their current age. However, within the proportional odds model framework, AOO was viewed as an ordinal outcome which distinguishes between controls, late-age-onset (LAO) and EAO cases. Within the binary logistic regression framework, contrast was made between cases and controls, irrespective of AOO. Results formulated on the basis of the P-value, as the measure of strength of evidence, indicated that the performance of the T2D GRS within the logistic modelling framework, which assessed T2D status, was substantially better when compared to the TTE modelling framework (Cox PH model), which assessed AOO of T2D. Consideration was also given to the proportional odds modelling framework which encompassed LAO and EAO, however, the proportional odds assumption was found not to be valid in both the NUgene

and WTCCC datasets. Further analysis based on GRS simulation studies indicated that when there are high levels of censoring, i.e. low proportion of individuals affected by the disease, there is little advantage in applying the Cox PH model over the logistic models. However, low levels of censoring seem to favour the Cox PH model, particularly when the magnitude of the SNP effect is small, and the RAF is low.

Third, an investigation was undertaken to compare the performance of T2D GRS to detect association with AOO of the disease across individuals of European, Asian, and African ancestry in the UK Biobank. Overall, the results based on the P-value, which represents a measure of the strength of evidence, indicated that the T2D GRS was more strongly associated with AOO of T2D in the Cox PH model based on cases and controls (censored at their current age) relative to the logistic model, which assessed T2D status.

In conclusion, the research in this thesis has demonstrated that GRS have the potential to advance common disease genetic research in relation to AOO of complex diseases. Additionally, improvements to methods developed to detect and account for population structure is paramount for GWAS discoveries as sample sizes continue to grow and for the clinical implementation of risk prediction models based on GRS. The application of GRS in ancestrally diverse or admixed populations is key to the realization of the vision of personalized medicine or personalized healthcare for all.

Acknowledgements

I consider myself fortunate to have had the opportunity to undertake this research and as such owe an enormous debt to the following individuals and organizations who have contributed towards the shaping of this PhD thesis. Without their input the completion of this thesis would not have been possible.

At the outset, I would like to express my appreciation to my supervisors Professor Andrew Morris and Professor Andrea Jorgensen for their continued guidance, advice, and encouragement throughout my doctoral journey. Their immense knowledge, motivation, enthusiasm have been invaluable, I could not have imagined having better advisors and mentors for my research.

I am also grateful to members of the Statistical Genetics and Pharmacogenomics Research Group as well as the PhD students and staff of the Department of Biostatistics who have supported me throughout my doctoral research endeavour. It has been an immense privilege to know and work with you all.

I gratefully acknowledge the funding received from the Commonwealth Scholarship Commission (CSC) in the United Kingdom through the award of a Commonwealth Scholarship to complete this doctoral programme. And my present employer, the Government of Anguilla (GoA) who granted me study leave and therefore the time to complete this programme of study.

Finally, I wish to thank my family and friends for their continuous support and encouragement throughout, I will forever be grateful.

Statement of contribution

The work contained in this thesis was undertaken by the candidate under the supervision of Professor Andrew P. Morris (Primary Supervisor) and Professor Andrea L. Jorgensen (Secondary Supervisor), Department of Biostatistics, Statistical Genetics & Pharmacogenomics Group, University of Liverpool. The research aims and objectives pertaining to the thesis, which is composed of simulation studies and T2D genotype samples pertaining to AOO of disease in a GWAS setting, were formulated jointly with supervisors.

The candidate developed the R syntax used to conduct the population structure simulation studies with guidance and support from supervisors. In relation to the GRS AOO simulation studies, the R script developed by Professor Andrew P. Morris were adapted and used to perform the GRS AOO simulation analysis. All data analysis of the simulated data was undertaken by the candidate including interpretation of results and production of tables and figures. Strategic direction and support, including critical review of simulation findings was provided by supervisors which included avenues for improvement.

The aspect of the thesis relating to AOO of T2D were based on three T2D GWAS SNP genotype datasets. These three genotype samples which originated from Northwestern University Gene (NUgene) Banking Project, Wellcome Trust Case Control Consortium (WTCCC) and UK Biobank were made available by Professor Andrew P. Morris. Additionally, all phasing and/or, imputation, and ancestry inference via principal components analysis (PCA) pertaining to the three genotype samples were also processed or facilitated by Professor Andrew P. Morris. Data analysis of the genotype samples was undertaken by the candidate including interpretation of results and production of tables and figures. Strategic direction and support, along with critical review of the reported findings was provided by supervisors.

Table of contents

List of tables

List of figures

Abbreviations

AFR	Africa
AFT	Accelerated failure time
AIM(s)	Ancestry informative marker(s)
AMD	Age-related macular degeneration
AOO	Age-of-onset
AWclust	Allele-sharing distance and Ward's minimum variance hierarchical clustering (AWclust)
BMI	Body mass index
CFH	Complement factor H gene
CI	Confidence interval
DIAGRAM	DIAbetes Genetics Replication And Meta-analysis
DM	Diabetes mellitus
EA	Effect allele
EAF	Effect allele frequency
EAO	Early age onset
EM	Expectation-maximization
ES	Effect size
EUR	Europe
GLM	Generalized linear modelling
GRM	Genetic relationship matrix
GRS (s)	Genetic risk score (s)
GWAS (s)	Genome-wide association study (studies)
HR	Hazard rate /ratio
HRC	Haplotype Reference Consortium
HWE	Hardy Weinberg equilibrium
IBD	Identity by descent
IBS	Identity by state
IDF	International Diabetes Federation
LAO	Late age onset
LD	Linkage disequilibrium
LE	Linkage equilibrium
LMM	Linear mixed models

MAF	Minor allele frequency
MCMC	Markov chain Monte Carlo
MDS	Multidimensional scaling
MENA	Middle East and North Africa
MLE	Maximum likelihood estimation
NAC	North America and the Caribbean
NCP	Non-centrality parameter
NEA	Alternative allele
NUgene	Northwestern University Gene
OLS	Ordinary least squares
OR	Odds ratio
PCA	Principal component analysis
PCs	Principal components
PH	Proportional Hazards
PRS (s)	Polygenic risk score (s)
RAF	Risk allele frequency
SACA	South and Central America
SE	Standard error
SEA	South East Asia
SHIPS	Spectral Hierarchical clustering for the Inference of Population Structure
SNP (s)	Single nucleotide polymorphism (s)
T2D	Type 2 diabetes
TOPMed	Trans-Omics for Precision Medicine
TTE	Time-to-event
UK	United Kingdom
USA	United States of America
WHO	World Health Organization
WP	Western Pacific
WTCCC	Wellcome Trust Case Control Consortium

Publications and presentations of work in this thesis

PRESENTATIONS

Work in this thesis presented at departmental and university events or meetings at the University of Liverpool.

Poster Presentations

Odessica N. Hughes, Andrea L. Jorgensen, Andrew P. Morris***;*** Department of Biostatistics, Statistical Genetics & Pharmacogenomics Group. ***Investigating polygenic contribution to age-at-onset of type 2 diabetes;*** **Institute of Translational Medicine Research Poster Day;** University of Liverpool, November 2018.

Odessica N. Hughes, Andrea L. Jorgensen, Andrew P. Morris; Department of Biostatistics, Statistical Genetics & Pharmacogenomics Group. ***Investigating polygenic contribution to age-at-onset of type 2 diabetes*****. Faculty of Health and Life Sciences Postgraduate Students Faculty Poster Day,** University of Liverpool, March 2019.

Oral Presentations

Odessica N. Hughes, Andrea L. Jorgensen, Andrew P. Morris***. Methodology - Investigating methods to account for admixture in genome-wide association studies of time-to-event outcomes. Biostatistics End of Year Ph.D. Student Talks,*** University of Liverpool, May 2017.

Odessica N. Hughes, Andrea L. Jorgensen, Andrew P. Morris***. Results - Investigating methods to account for admixture in genome-wide association studies of time-to-event outcomes. Biostatistics Research Day,*** University of Liverpool, May 2018.

Odessica N. Hughes, Andrea L. Jorgensen, Andrew P. Morris. **Investigating methods for the identification of polygenic contribution to risk and age-at-onset of type 2 Diabetes.** ***Statistical Genetics & Pharmacogenomics Group meeting,*** University of Liverpool, April 2019.

Chapter 1: Introduction

Chapter Outline

This chapter introduces the genome-wide association study (GWAS) design, which forms the basis of the research undertaken in this thesis. Methodological challenges commonly encountered in the implementation of the GWAS design are discussed. The chapter focuses on issues affecting the determinants of statistical power and likely sources of false positive findings in a GWAS setting attributable to population structure. The chapter also highlights key issues in assessing the relative statistical power of various analysis approaches that form part of the thesis research as well as issues surrounding age-of-onset (AOO) GWAS based on single nucleotide polymorphism (SNP) and genetic risk score (GRS) association approaches.

1.1. | Introduction to GWAS

In the ensuing "post genome era", GWAS emerged as a powerful and eventually the standard tool for investigating the genetic architecture of common human diseases and complex traits [1-3]. GWAS have evolved from small-scale studies consisting of less than 100 cases of disease [4] of primarily single ancestry populations to large-scale international consortium-based studies consisting of tens of thousands of cases of disease that are often ascertained from multi-ancestry populations [5]. With the availability of large-scale population biobanks and advances in genotyping technologies, this trend towards larger and larger sample sizes is likely to continue. Owing to the polygenic (disease resulting from the combined action of two or more genes) nature of most common diseases requiring large samples to detect the moderate effects of associated variants, increasingly large samples have become a feature of common disease GWAS [6, 7].

Since the first application of the GWAS it has experienced remarkable success, though not in the way that was originally envisioned. Immediate identification of the causal variant and heritability, which refers to the fraction of phenotypic variance explained by genetic variation [8], fully accounted for were among the initial expectations of GWAS. However, many identified variants have no known biological effects and associated variants were found to account for less than 5% to 10% of heritability implied by family (and twin) studies [9]. Even with this initial setback, as of 2017, approximately 10,000 robust associations for a wide range of complex traits

have been attributed to GWAS [10]. There are also several known or candidate drugs linked to genes from GWAS signals of common complex diseases that includes type 2 diabetes (T2D), rheumatoid arthritis, osteoporosis, and schizophrenia [10]. Despite the success of the GWAS, however, there are inherent limitations that initially limited its widespread use, particularly in populations from diverse genetic backgrounds [11]. Originally, GWAS were undertaken in homogenous populations due to concerns relating to geographical confounding between the disease and SNPs, which can result in inflated type I error rates if not accounted for in the association analysis [12]. Adding to this was the likelihood of reduced power to detect association due to potential heterogeneity in allelic effects on the disease across diverse genetic backgrounds [13].

To effectively tackle global health challenges, particularly as they relate to common complex diseases like T2D, it is essential to investigate the properties (power and type I error rates) of statistical methods for GWAS analysis in the presence of population structure. Population structure, which refers to the state where sub-populations are distinguishable by observed genotypes [14], is especially problematic for common disease genetics as the multiple SNP influences are usually of modest genetic effect and as a result can be dwarfed by confounding [15].

The impact of genetics in public health is presently limited in the sense that accurate prediction of disease risk at the individual level is still lacking for most common diseases and across different ancestry groups [16, 17]. However, to realise the vision of personalized medicine or personalized healthcare, inclusion of diverse populations in genetic research is essential, given the potential to use genetics to treat common diseases through better targeted intervention strategies. Exploring the link between genes, environment and lifestyle has the potential to provide new insight and a clearer understanding of the mechanisms associated not only with disease risk but also AOO of disease and disease progression. Identifying individuals who may have inherited a genetic predisposition to common diseases along with genetic factors that are likely to influence the AOO of disease, particularly for common complex diseases, is an important consideration in medical research. Identification of individuals with a likely earlier onset of disease and utilizing better targeted screening strategies, based on knowledge of AOO of the disease, has the potential to improve patient survival and reduce treatment costs.

1.2. | Mapping strategies in genetic and genomic research

In human genetics and genomic research, several approaches have been developed to identify disease risk variants (or risk variants for other health related outcomes). These approaches to mapping disease variants are formulated on the foundation of linkage disequilibrium (LD). With the emergence of the GWAS design, alternative mapping strategies applicable to admixed populations have also emerged.

1.2.1. | Gene mapping and patterns of linkage disequilibrium

Mapping strategies in genetics takes advantage of an important property of human genomic structure, which involves the processes of linkage and allelic association [18]. Underpinning disease gene mapping strategies is linkage disequilibrium (LD). LD refers to the phenomenon where two or more alleles at different SNPs are genetically linked and are correlated. Based on this principle, an individual with one particular allele at a genetic locus (location of a variant on the genome) are likely to have a specific allele at a second genetic locus [3]. The extent of the LD in a population is influenced by a number of factors including rate of recombination, rate of mutation, genetic drift, non-random mating, and population structure.

A genetic marker that facilitates association analysis refers to any region of the genome whose location on a chromosome or chromosomal DNA can be identified and displays sequence variation between individuals of a population [19]. This genetically determined sequence variation is usually readily identifiable by direct observation [20]. Chromosomal aberration (any abnormal chromosome complement resulting from an alteration in chromosome structure or number [19]), a gene, or a locus containing a DNA polymorphism (presence of discreetly different forms of a gene or a character [21]) are examples of different forms of genetic markers [19]. The SNP is the most common form of variation in the human genome and is commonly used as a marker in genetic association analysis. The SNP is defined as a single base change in a DNA sequence. To date, around 325 million SNPs have been identified in the human genome, 15 million of which are present at frequencies of 1% or higher across different populations worldwide [22, 23].

LD eliminates the need to genotype every single SNP in the human genome whilst still accomplishing near complete coverage of common variation of the human genome. The LD property of the human genome enables a carefully selected representative set of SNPs to determine the status of other SNPs [24]. These SNPs are usually referred to as tag SNPs. In association mapping, a tag SNP could be associated with a disease, not because it is biologically causal, but because it is statistically correlated with a causal variant. The most common measure used to assess LD is the squared correlation coefficient ($\mathbf{r^2}$), which measures the level of correlation between two markers or SNPs. An $\mathbf{r^2}$ of 1 indicates that the two markers are perfectly correlated while an $\mathbf{r^2}$ of 0 indicates that the two markers are completely independent [25].

Genetic investigations in diverse populations require careful consideration regarding the structure of LD within different ancestry groups as well as admixed populations. Ancestral populations differ in their degree of LD, therefore the number of marker SNPs needed for complete genomic coverage may vary. As an example, populations of African ancestry are older and therefore have shorter stretches of LD, as they have experienced more generations for LD to decay, compared to non-African ancestry populations [26]. Furthermore, population bottleneck events (sharp reduction in population size), such as the "out of Africa" exodus, resulted in lost haplotypes (reduced genetic diversity), which then results in increased LD in non-African ancestry populations [27]. Therefore, in African ancestry populations, which has lower levels of LD relative to European populations, more SNPs is needed to achieve near complete genomic coverage. This property of LD has strong implications for GWAS design [28].

1.2.2. | Mapping strategies

Before the GWAS era: Prior to 2000, the family-based disease mapping approach, linkage analysis, the primary method of investigation at the time, proved highly successful for single gene diseases (mendelian diseases) but was not as successful with common and complex diseases [29]. Linkage refers to the tendency for disease causing genes and other genetic markers to be inherited together because of their location near one another on the same chromosome [30]. The linkage analysis approach was found in general to be more powerful for detecting rare genes with large effects [31, 32]. As a result of the limitations of the linkage analysis approach, genetic association analysis, which aims to identify a genetic variant that

influences a disease or trait at the population level [33], has emerged as the central tool applied to common, complex diseases. The way in which this gene mapping approach is applied depends on two different strategies. One depends on prior biological knowledge that points to a particular polymorphism in candidate genes or regions that are applied in candidate gene association studies [34]. The other approach requires a very high density of genetic markers (described in secton1.2.3) in genomic regions which are used to investigate the entire genome and applied to GWAS. Prior to the GWAS era, the candidate gene approach had become the standard tool for common complex disease investigations. However, the candidate gene approach was not as successful as the GWAS approach. The main difficulties with the candidate gene approach related to the fact that: (i) investigations were limited to protein-coding regions of genes and therefore the scope of investigations did not consider the impact of neighbouring genes; (ii) at the time, sample sizes were often very small, and thus underpowered; and (iii) Data quality and analytical protocols at the time were not adequate to address confounding due to population structure which resulted in false-positive findings that failed to be replicated [35, 36].

Alternative mapping strategies: In the context of GWAS, an admixed population refers to individuals formed of two or more genetically distinguishable and previously isolated ancestral populations [37]. Admixture mapping is the most common alternative gene mapping strategy for admixed populations in GWAS. It is applied to identify disease susceptibility variants in an admixed population resulting from a recent mixture of two or more ancestrally distinct populations [38]. This approach is most beneficial when the disease susceptibility variant has different allele frequencies in the ancestral (parental) populations because of drift or selection. Admixture mapping requires a genotyping panel that can differentiate chromosomal segments in admixed individuals by their ancestral origins. Additionally, polymorphic markers must differ in frequency in the ancestral populations, and there must be at least 10% admixture [39]. Examples of diseases that show a difference in incidence between the two "ancestral" populations includes multiple sclerosis in Africans vs. Europeans (population relative risk 0.50), or hypertension in Africans vs. Europeans (population relative risk 2.61) [39].

1.2.3. | Emergence of the GWAS design

The success of GWAS in identifying genetic risk factors for diseases and genetic differences in drug response *(efficacy and adverse reactions to common drugs)* – a field known as pharmacogenomics - has been gradually paving the way for personalized medicine [40]. The completion of the International HapMap Project in 2003 provided the platform for the advancement of GWAS in biomedical research. Key findings from the project facilitated a better understanding of the patterns of human sequence variation, which when coupled with advances in genotyping technology, made it feasible to conduct population based GWAS. This has had significant impact on public health strategies as it has allowed the switch in focus from mendelian or single gene disease (that is diseases influenced by a single gene) with large effects that affect a few families, to multiple gene or complex diseases (that is diseases influenced by multiple genes) that affect many in the general population [3].

Among the first success stories of the GWAS era was a 2005 study of age-related macular degeneration (AMD) where the study identified two SNPs in the complement factor H gene (CFH) strongly associated with AMD [4]. The 2007 publication by the Wellcome Trust Case Control Consortium (WTCCC) marked a pivotal stage in the advancement of GWAS for complex diseases. The study identified 24 association signals across seven diseases: type I diabetes, T2D, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis and inflammatory bowel disease [41]. This study was also instrumental in the development of protocols for quality control and association analysis in GWAS that are still widely used today.

The advancement of GWAS methodologies provides the opportunity for genes associated with complex diseases to be identified although there are still many challenges that have the potential to affect the validity of GWAS findings going forward. The challenge posed by the presence of unaccounted population structure, which can result in false positives in GWAS analysis, is the focus of this research.

1.3. | Key stages of the GWAS approach

This section of the thesis gives a general overview of the key stages of a typical GWAS. These stages include the initial discovery stage, which is often followed by replication and validation. Other follow-up GWAS approaches often undertaken include meta-analysis and fine mapping and are therefore also highlighted here.

1.3.1. | General process of initial GWAS discovery

GWAS, which have been formulated on the premise of the "common disease common variant hypothesis", maintains that common diseases are likely influenced by genetic variants that are also common in the population [42]. GWAS can be characterized as a study of genetic variation across the human genome that is designed to identify genetic associations with human diseases or other health related outcomes. It is in essence, a non-candidate-driven approach ("hypothesis free"), which involves rapidly scanning several hundreds of thousands (or millions) of genetic markers (most commonly SNPs) across genomes of many people to find genetic variations associated with a particular disease or health related outcome [43]. This approach, which is primarily a population-based observational approach, involves the comparison of the SNPs alleles of individuals with the disease (cases) with similar individuals who have not develop the disease (controls). If a SNP allele is more frequent in individuals with the disease, the SNP risk allele is deemed "associated" with the disease (see Figure 1.1). Additionally, if a SNP allele is found to be more frequent in controls when compared to case, the associated allele may be deemed "a protective allele" from the disease under consideration. The difference in allele frequency between the cases and controls is usually considered significant at the genome-wide significance p-value threshold of 5×10^{-8} (this standard threshold was originally based on the effective number of independent common SNPs across the human genome [~ 1 million with MAF ≥ 5%], which was determined on the basis of the LD block structure in European ancestry populations [44, 45]). The observed associated SNP serves as a marker for the genomic region responsible for the disease or outcome of interest. However, further investigations of the variants within the candidate region are required to identify the causal variant [24, 32].

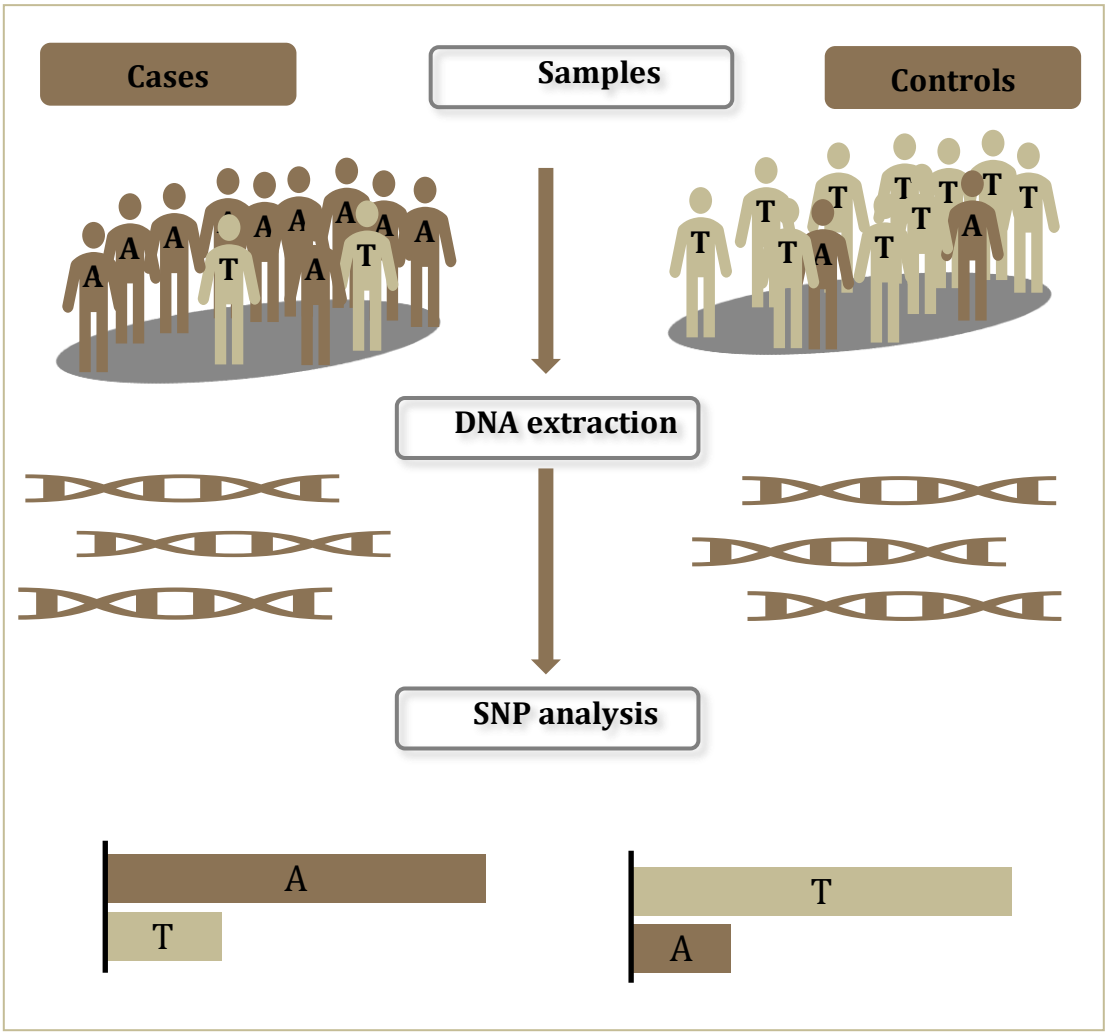


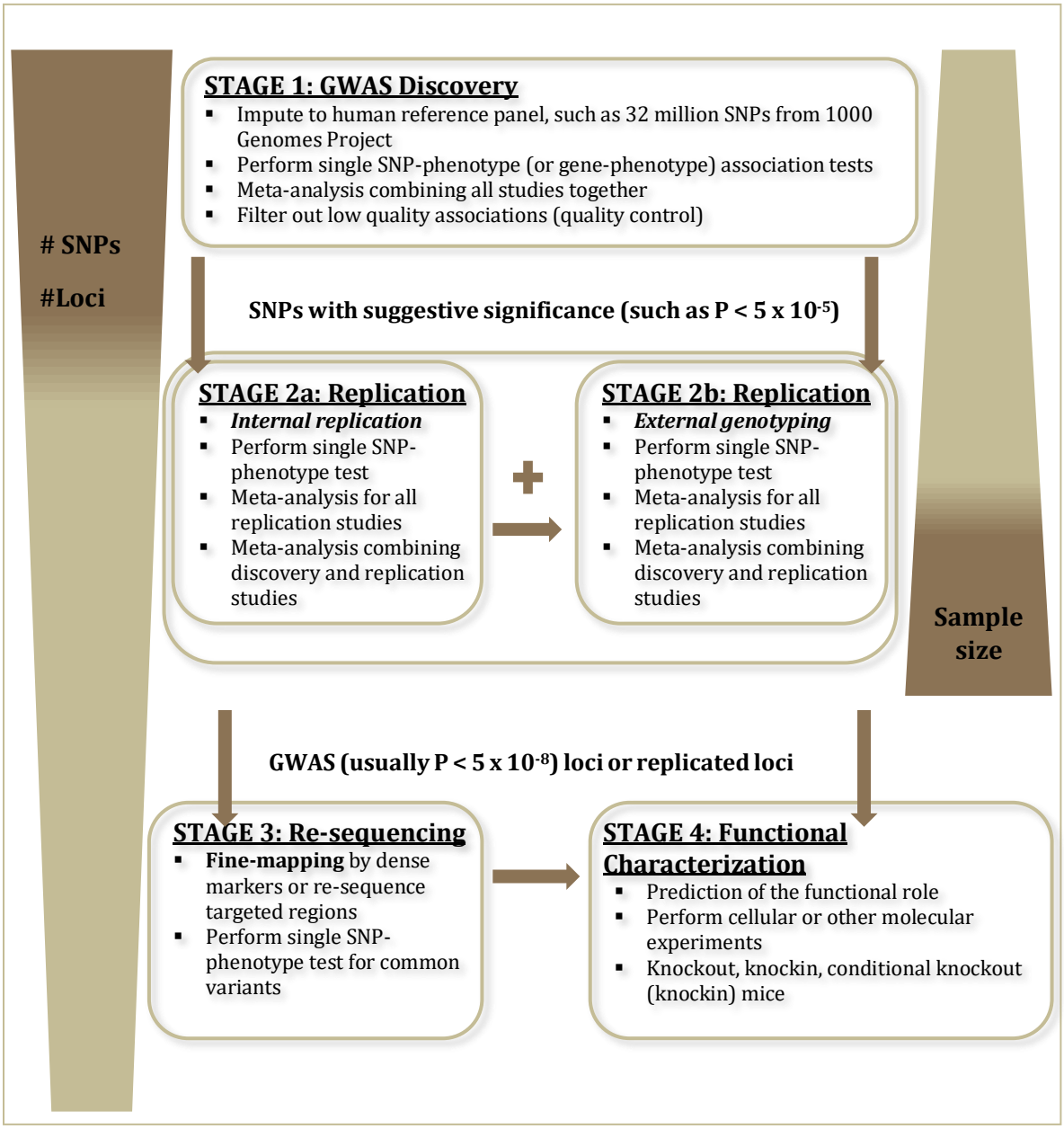
Figure 1. 1 - Diagram to show typical allele distribution which GWAS seek to identity

1.3.2. | Replication and validation GWAS

Due to the inherent limitations of the GWAS methodology a process of validation is required for new discoveries. GWAS is particularly vulnerable to false positive associations introduced through population differences resulting from population structure and/or genetic admixture. This is compounded by the requirement for larger sample sizes needed to optimize statistical power as the risk of confounding increases in relation to increasing sample sizes[46]. Replication of initial GWAS findings in an additional independent sample drawn from the same population is the gold standard for validation of GWAS findings [3]. A study that is well-

powered is also a crucial requirement for undertaking replication GWAS based on the findings of a discovery GWAS to facilitate the identification of false-positive findings. Requirement for a larger sample size relative to the initial GWAS is also important to address other potential issues including over-estimation of effect size. Investigations in other geographical populations are also conducted after findings have been validated in a replication sample to determine if associated SNPs have an ethnic-specific effect. Replicated loci or genomic regions are determined on the basis of a genetic effect that is consistent in terms of magnitude and direction across both the original discovery dataset and the validation dataset. All SNPs in high LD with the tested SNP are potential replication candidates [3], however, there are researchers who believe that replication should only be claimed if the same variant, phenotype and genetic model is involved [47].

Measures to streamline all elements of the GWAS validation and replication process have been ongoing, however, there remains areas of concern that are fundamental to GWAS long-term success. Validation and replication in GWAS have suffered inconsistencies in a number of areas, particularly as it relates to; (1) an established criteria for identifying associated SNPs to use in replication studies; and standard definition for a proper replication study and criterion for refuting the finding based on the replication results [48]. These issues have become even more pertinent with the availability of large biobanks, as independent external replication cohorts may not be available or possible. Additionally, the rationale for selecting replication SNPs often varies among studies. Regarding studies that include diverse or admixed populations, there is a lack of clarity in terms of the significance threshold applied, because of the underlying LD structure which may be quite different to European ancestry populations. For GWAS based on admixed populations thresholds ranging from $p < 1 \times 10^{-8}$ to $p < 1 \times 10^{-9}$ based on method of genotype ascertainment, genetic and variant frequency have been proposed [49].



Source ([50])

Figure 1. 2 - General design and workflow of GWAS

1.3.3. | Meta-analysis of GWAS

The realization that sample size is critical to GWAS success saw the formation of major global consortia to tackle the genetic basis of many common diseases [7]. The requirement for large sample sizes is often beyond the capacity of a single GWAS, as a result meta-analysis is usually employed to combine data from multiple studies of relatively small sample sizes, with the expectation to detect genes underlying susceptibility loci with greater power (Figure 1.2). By combining summary statistics from individual GWAS, more precise estimates of genetic effects are produced and hence provide more convincing conclusions. In genetic research, privacy, data access issues and use of different genotyping platforms often result in limitations on researchers to directly combine individual datasets [51]. Furthermore, combining and analysing raw data from all studies may prove to be very laborious and at the same time offer no gains in efficiency when compared to meta-analysis based on summary statistics [1, 52]. Therefore, meta-analysis of summary statistics provides an avenue to bring into context the overall genetic evidence pertaining to a disease outcome in a cost-effective manner [53].

1.3.4. | Fine mapping in GWAS

After the initial discovery GWAS, further investigations of the variants within the associated regions are often required to identify the causal variant (Figure 1.2). Discovery GWAS primarily identifies a tag SNP, which is often not the causal SNP, but rather a SNP in LD with the true functional SNP. With numerous variants now known to be associated with many common diseases, this creates a shift in focus in terms of establishing effective strategies for identifying the causal gene and thereby the biological mechanisms that underlie these diseases. With the advent of next-generation sequencing, deep sequencing and functional studies provide an additional avenue to ascertain the biological mechanisms associated with the causal SNPs. However, fine mapping aided by high-density genotype imputation is often undertaken in an effort to identify the causal gene. Through genotype imputation, which estimates the value of untyped or unknown alleles, low-density genotyped SNPs, ($\sim 10^5 - 10^6$ SNPs) are increased to a level of high-density ($\sim 10^7 - 10^8$ SNPs) [54]. The end goal is to establish causality, functionality and determining effects which can inform strategies for disease diagnosis, prognosis, prevention and treatment [55].

1.4. | Methodological challenges of the GWAS design

This section highlights the main methodological challenges commonly encountered in the conduct of GWAS and its impact on statistical power and the false positive error rate. The different forms of population structure are described and their impact on GWAS is discussed in some detail. The determinants of statistical power are also outlined, as well as potential sources of reduced power or false positives that can arise throughout the various stages of the GWAS process, are also discussed. Special attention is given to false positive findings that may be attributable to the presence of unaccounted for population structure.

1.4.1. | Population structure and its impact in GWAS

Population structure, distant cryptic relatedness (3rd -9th degree relatives: individuals are closely related, but this shared ancestry is unknown) and family structure (1st and 2nd degree relatives [56]), are the most prominent confounding issues to consider in the design and analysis of genetic or genomic association studies. As indicated in section 1.1, population structure describes the state where populations are distinguishable by observed genotypes, which in genetic studies have the potential to result in different forms of confounding. Large samples (N > 5,000) from populations or population cohorts are expected to contain individuals who have ancestry originating from different geographical populations [57]. Even in relatively genetically homogenous populations different levels of fine-scale substructure have been observed [58, 59]. It has been demonstrated in past research that population structure has the potential to create spurious results, particularly when methods rely on large numbers of small effects such as polygenic scores [60], which have been applied to disease risk prediction.

The three main types of population structure that have been distinguished are discrete structures, admixed populations and hierarchical structures [61]. Discrete structures refer to a population that consists of mutually exclusive distinct subpopulations (assumes a partition of the population into "islands" [62]). An admixed population (described in section 1.2) allows individual-specific proportions of ancestry arising from actual or hypothetical ancestral islands. Hierarchical structures comprise of both the discrete distinct subpopulations and admixed populations.

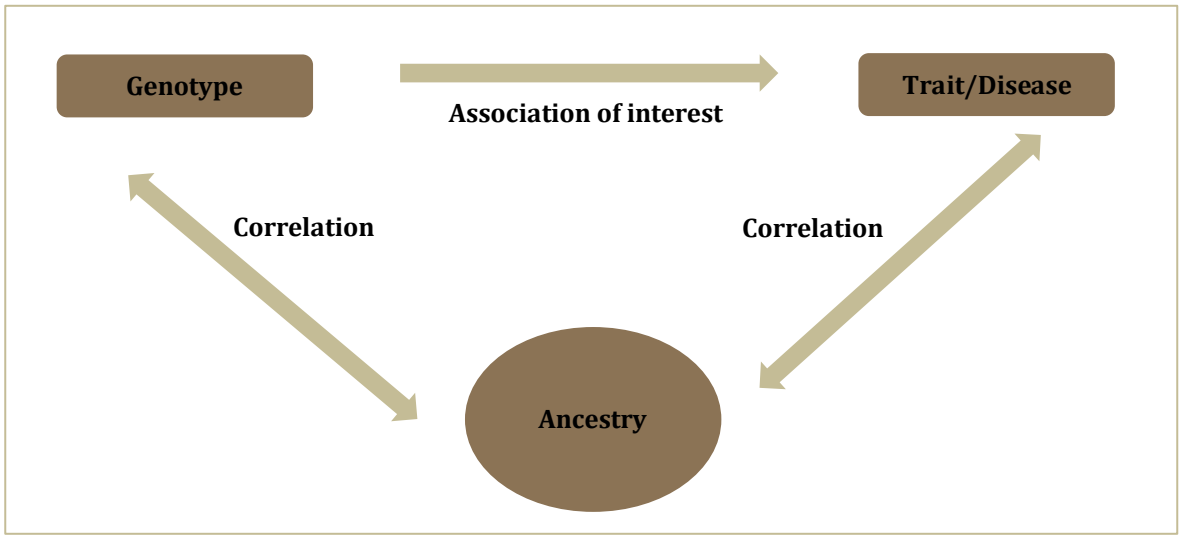


Figure 1. 3 - Confounding due to ancestry

When confounding arises as a result of population structure, ancestry is associated with both genotype and the disease investigated (Figure 1.3). Therefore, associations in GWAS could potentially be the result of the underlying structure of the ancestral population and not the disease associated locus if population structure is not taken into consideration. Population structure is often aligned with geography. At the continental level, the Caribbean and the Americas contains some of the most highly admixed populations [63, 64], South America, in particular, is one of the most ancestrally diverse regions in the world [65]. The ancestry landscape of these admixed populations was shaped by complex admixture events arising during the era of colonization and the Atlantic Slave Trade. Common examples of admixed populations in this region include African Americans, Latin Americans, and African Caribbean populations. The ancestral composition of these admixed populations comprises various combinations of European, Native American, West African, and East Asian ancestry. The Uyghur population of central Asia with ancestral contributions from European and East Asian populations and segments of the South African population who are of African and European ancestry are also examples of admixed populations [66, 67]. Over the years, several statistical methods and tools have been developed to both detect and account for population structure in GWAS. These methods are discussed in section 1.5.

1.4.2. | Determinants of statistical power in GWAS

Statistical hypothesis testing is subject to both type I and type II errors, which have implications for statistical power. The type I error rate refers to the probability of making the error of rejecting null hypothesis ($\mathbf{H_0}$) when it is true (denoted as $\boldsymbol{\alpha}$) while the type II error rate refers to the probability of failing to reject H_0 when it is false (denoted as $\boldsymbol{\beta}$). In a genetic research setting, statistical power ($\mathbf{1-\beta}$) is defined as the probability of detecting an effect, given that the effect is real or alternatively, the probability of correctly rejecting the ($\mathbf{H_0}$) when it is in fact false. The ($\mathbf{H_0}$) typically refers to an effect size that signifies no association (example odds ratio (OR) = 1 or log OR = 0), whereas the alternative hypothesis ($\mathbf{H_1}$) usually refers to an effect size (for a two-sided test) that suggests an association(example OR ≠ 1 or log OR ≠ 0). Adequate statistical power is a requirement for a successful GWAS undertaking. In GWAS, the main determinants of statistical power are sample size, SNP effect size (for binary outcomes, such as disease status, this is commonly measured by the OR), minor allele frequency (MAF), significance threshold and in situations where there is a binary outcome, the ratio of cases to controls. Increasing the sample size results in increased power as the accuracy of the estimated effect size is improved (i.e. smaller standard error of effect size estimate). A larger effect size results in increased power because it increases difference from the ($\mathbf{H_0}$) value. Significance testing in GWAS must adopt stringent significance thresholds to allow for multiple testing, and the more stringent (smaller) the critical p-value for rejecting $\mathbf{H_0}$, the lower the statistical power. Furthermore, in case-control studies, the proportion of cases moving closer to 0.5 increases power because it increases the accuracy of effect size estimation. At the design stage, the non-centrality parameter (NCP):a measure as to the degree to which the null hypothesis is incorrect, is often used as an intermediate to determine the power of a genetic study for a given significance threshold based on available information in respect to these parameters (the expected effect size of the SNPs, MAF, ratio of cases to controls (binary outcomes) and sample size).

The test statistics given by Equation 1.1, under the ($\mathbf{H_0}$) of no association which is distributed as a chi-square distribution with m degrees of freedom **(df)** is a central chi-square. However, under the ($\mathbf{H_1}$) it is a non-central chi-square distribution [52] where $\boldsymbol{\chi^2_{\alpha(\mathbf{df=m})}}$ corresponds to the $\boldsymbol{\chi^2}$ degrees of freedom in respect to a predetermined significance threshold α and the NCP described in Equation 1.2.

$$Z^2 \sim \chi^2_{\alpha(df=m)} \left((\theta/SE_{bin})^2\right)$$

Equation (1.1)

In biomedical research, statistical power of 80% is widely used as a benchmark to avoid false negative associations and to determine a cost-effective sample size [68]. In an effect to acquire the appropriate sample of cases and controls that form part of epidemiological studies, a more complex sampling design than simple random sampling is often required. To account for this added complexity the effective sample size (ESS) is often applied as a measure to determine the minimum sample size that would be required to achieve the same level of precision if the samples were a simple random sample [69]. Here the ESS have been denoted $\mathbf{n\pi(1-\pi)}$, where *n* corresponds to the total number of cases and controls in the overall sample size and π the proportion of cases. The NCP_{bin} of an additive GWAS model for binary case-control data [70, 71] is given by Equation 1.2.

$$\boldsymbol{NCP_{bin} = (\theta/SE_{bin})^2 \approx 2\,f(1-f)\,n\pi(1-\pi)\theta^2}$$

Equation (1.2)

The NCP_{bin} consists of the estimated true effect size denoted θ and its associated standard error denoted SE_{bin}. Where $\boldsymbol{NCP_{bin}}$ denotes the non-centrality parameter for an additive GWAS model for binary case-control data; total number of cases and controls in the overall sample size n; effect size θ, which usually refer to the log OR; MAF f; and proportion of cases π.

1.4.3. | Sources of reduced power or false positives in GWAS

The areas or aspects of the GWAS process that have the potential to give rise to false positive findings or reduce power are outlined in this section. The main aspects of GWAS covered include study design issues relating to common disease GWAS; changes in the main technological tools for genotyping and imputation (process of imputing missing or untyped SNPs genotypes based on observed nearby SNP genotypes in high LD); and established data quality control procedures pertaining to genotyped samples from an individual and SNP perspective. The impact of statistical analysis procedures is also discussed.

1.4.3.1. |Issues at design stage of common diseases GWAS

At the design stage of the GWAS, the approach to sample selection, determination of the disease or phenotype or its classification and the main factors that may influence the disease outcome are of prime consideration.

GWAS samples: Presently, within a public health framework, there is recognition that health outcomes are to a great extent influenced by a range of social, cultural, political, economic, environmental, behavioural and biological (which encompasses genetics) factors [72]. Together, these factors have been termed "the determinants of health". In the context of common and complex disease genetics, determining the extent of the role of these determinants of health in the onset of a disease can be challenging due to the issue of confounding. The underlying demographic characteristics within specific populations can often mask associations in the context of disease and health related outcomes. Ethnicity, which is essentially the product of cultural, geographical, and biological differences between groups of individuals or subpopulations [11], often presents a challenge in the conduct of biomedical research. Additionally, age and gender are often associated with several lifestyle, physical and chemical exposures as well as disease and other health related outcomes. Of these potential confounders, ethnicity is linked to the most critical confounding factor in GWAS, population structure. To account for potential confounders, they are usually included at the analysis stage as covariates in the model. However, to address the issue of population structure, different study design strategies for ascertaining samples have also been explored.

To reduce the effects of confounding due to population structure, matching cases to controls on the basis of ethnicity or ancestry has been regarded as a potential solution. Additionally, GWAS were initially restricted to a single ancestral population. In this setting, individual populations were analysed separately based on self-identified ethnicity or ancestry and combined statistically in a meta-analysis. However, self-identified ethnicity is subject to misclassification and adding to this is the fact that the problem of fine scale structure (i.e. structure within an ethnic group) remains. Additionally, there is also the inability to easily match for the levels of admixture within admixed individuals [73]. As a result, GWAS based exclusively on family data was considered a viable alternative solution. Data on affected cases and their parents are collected in this family-based association design, and a comparison is made between alleles

transmitted to the child, and those that are not. However, this approach has two drawbacks; (1) it is less powerful than population-based GWAS as the two parents are required to form a single matched control; and (2) parental data may not always be available, particularly for late onset diseases, as parents are more likely to be deceased.

Disease classification: Phenotypic misclassification and phenotypic heterogeneity can confound the relationship between SNP and disease outcome. The accuracy of the diagnostic criteria used for disease classification have implications for statistical power. In case-control studies, misclassification reduces substantially the power to detect associations [74]. It has been noted from past studies that utilizing the International Classification of Diseases (ICD) codes alone to define phenotypes via electronic diagnostic code data can result in substantial misclassification effects. Furthermore, the accuracy of inferring disease phenotypes from electronic diagnostic codes can vary widely across diseases [75]. This may be further complicated or compounded by phenotypic heterogeneity (refers to mutations in the same gene resulting in similar but different diseases or variation in the expression of the disease (expressivity)). An example of this is dementia, which is a disease of the brain that is often the result of similar complex disorders, the most common of these includes Alzheimer's disease (AD), vascular dementia, frontotemporal dementia, dementia with Lewy bodies, and Parkinson's disease [76].

Genetic heterogeneity: An added challenge to gene discovery in GWAS is genetic heterogeneity, which refers to the phenomenon where several distinct genetic variants may give rise to the same phenotype [77]. Research pertaining to a wide range of diseases suggest that complex diseases are characterized by remarkable genetic heterogeneity. Furthermore, genetic heterogeneity is known to exist between ancestries for a substantial portion of loci associated with complex diseases [78]. The effects of genetic heterogeneity can cause a reduction in the power of association tests between a single SNP and phenotype [79]. Genetic heterogeneity manifests itself on two levels, allelic and locus heterogeneity. Allelic heterogeneity describes the situation where different mutations within a single gene locus cause the same disease [80]. In such a situation, a single marker may fail to capture all disease variants when they exist on the same gene. Locus heterogeneity describes the situation where genetic variants in completely unrelated gene loci cause a single disease but only one mutant locus is needed for the disease to manifest [80]. However, these causal variants may be on a

single pathway (signalling, regulatory, metabolic) [81]. This suggests that different combinations of multiple genes may independently influence disease risk. However, from a single gene perspective, cases and controls may appear to be the same, which has implications for statistical power. As for any specific causal gene only a subset of cases will contain a variant in that gene, while other cases will have causal variants in other genes in the pathway [81]. It has been suggested that AOO of disease variation and variation in severity of disease may reflect underlying genetic heterogeneity [82] or individuals having a higher genetic load of risk variants [83, 84].

1.4.3.2. |Main technological tools of GWAS

Advances in genotyping and imputation technologies, though advantageous in the long-term, have the potential to impact power. Differences in genomic coverage and the range of SNPs included on genotyping microarrays in terms of MAF are likely to impact power. Issues relating to genotyping and imputation technologies are described in more detail below.

Genotyping technologies: Genotyping by SNP microarray is the prime instrument that has enabled the application of GWAS. However, errors in genotyping resulting from either the genotyping experiment or genotyping calling process is another potential major source of false positive GWAS findings [85]. Adding to this is the use of different microarray platforms, Affymetrix and Illumina being the most commonly used, which could result in different levels of genomic coverage. Genotype SNP arrays have different genomic coverage therefore, the level of power could vary from one SNP array to another.

LD is an important factor in this process and ancestral populations differ in their extent and structure of LD, therefore the number of marker SNPs needed for complete genomic coverage may vary. The extent of genomic coverage of GWAS chips is measured on the basis of the percent of common SNPs from a reference panel having an r^2 of 0.8 or more with at least one SNP on the platform. The different genotype SNP arrays have been shown to have similar power in populations of European ancestry but may vary in populations of non-European ancestry. In populations with weaker LD, such as African ancestry populations, power is increased by applying denser and ancestry specific SNP array platforms.

An additional concern is the level of accuracy in genotype calling which is usually assessed on the basis of call rate, signal quality and call accuracy. Differential genotyping error, which refers to the situation where the genotyping error rates in cases and controls are different, resulting in the inflation of the type I error rate is another concern [86]. The potential for this type of error is more likely in large scale studies where cases and controls may be genotyped at various sites.

Table 1. 1 - Most commonly used public reference panels

Reference panel	**Number of reference samples**	**Number of sites (autosomes + X chromosome)**	**Ancestry distribution**	**Year**
The International HapMap Project phase 3	1,011	1.4 million	Multi-ethnic	2010
1000 Genomes Project phase 1	1,092	28.9 million	Multi-ethnic	2012
1000 Genomes Project 3	2,504	81.7 million	Multi-ethnic	2015
UK10K Project	3,781	42.0 million	European	
Haplotype Reference Consortium (HRC)	32,470	40.4 million	Predominantly European	2016
Trans-Omics for Precision Medicine (TOPMed)	60,039	239.7 million	Multi-ethnic	2017

Source (Genotype Imputation from Large Reference Panels: Annual Review of Genomics and Human Genetics 2018) [87].

Imputation technologies: Genotype imputation, introduced in 2007 [88], is a mechanism that further leverages the correlations between nearby alleles due to LD to predict genotypes at SNPs that have not been directly genotyped on an array, but which are available on a high-density reference panel (such as the 1000 Genomes Project). Imputation, which involves the estimation of unknown alleles based on the observation of nearby alleles in high LD, enables

the inclusion of ungenotyped SNPs in association testing and thereby increases power. Genotype imputation also facilitates GWAS meta-analysis based on different genotyping platforms by generating a common set of variants that can be analysed across all the studies. For example, the Illumina 660k array only contains 20% of the SNPs included in the Affymetix 6.0 SNP array [87]. Additionally, common SNPs across different generations of the same platform or commercial brand can also be generated by genotype imputation. The accuracy of genotype imputation is impacted by several factors including sample size and SNP coverage of the GWAS. In addition, the accuracy of the imputation depends on the reference panels employed in the process. The accuracy of genotype imputation is affected primarily by the density of the genotyping array and sequencing coverage in the reference panel; MAF; haplotype accuracy in both reference and sample being imputed; and the software employed. Some of the more widely used software tools for genotype imputation include FastPHASE, Beagle, Minimac, IMPUTE, MACH and SHAPEIT, [87, 89].

Research investigating imputation accuracy indicates that there will be difficulties when imputing populations for which there is a limited number of reference individuals. Therefore, the choice of reference panel for samples of different ancestral origins has implications for statistical power. The most commonly used public reference panels are listed in Table 1.1.

1.4.3.3. |Data quality procedures in GWAS

Standard data quality control procedures have been implemented as part of the GWAS process to help mediate the negative impact on power from various potential confounding sources. Procedures designed to assess the quality of data in a GWAS from the perspective of individual sample quality are outlined. This is followed by procedures designed to assess the quality of individual SNPs.

Sample quality control: In a standard GWAS, the data quality procedures deployed as part of the assessment of individual samples include individual level missingness, sex discrepancy (which refers to the difference between the assigned sex and the sex determined based on the genotype), heterozygosity rate, cryptic relatedness and ancestry outliers. As high levels of missingness or genotype failure rates within samples is an indication of poor DNA quality or technical problems, the level of missingness among individuals is assessed based on a user -

defined missingness threshold. Individuals with high rates of genotype missingness are removed. It is recommended that this process is performed in two stages based first on a more relaxed threshold (example 0.2) and then a more stringent threshold (example 0.02) after a check of SNP specific missingness (described below) [90]. The final user-defined threshold is typically in the range of 95%-99% completeness, in keeping with the overall required stringency of the quality control [91].

Additionally, checks for discrepancies in reported sex and sex based on the X chromosome are usually undertaken using the heterozygosity/ homozygosity rates for the X chromosome. The X chromosome F statistic provides an indication of the deviation of the observed number of heterozygote variants from that expected under Hardy-Weinberg equilibrium (HWE). Typically, the reduction in heterozygosity is assessed with reference to HWE. HWE represents a state within a given population, where the balance in the relative number of alleles is maintained from generation to generation assuming (1) mating is random; (2) no natural selection; (3) no migration; (4) no mutation: and (5) population is large. For a genetic locus at equilibrium, the Hardy-Weinberg Proportions relating to the genotypes composed of those alleles is expressed by $p^2 + 2pq + q^2 = 1$, where p^2 corresponds to the frequency of AA (homozygous A), 2pq corresponds to frequency of Aa [heterozygous], and q^2 corresponds to the frequency of aa [homozygous a]) [92]. Males are expected to have an X chromosome F statistic value of > 0.8 and females are expected to have an X chromosome F statistic < 0.2 [90].

The overall heterozygosity rate, excluding the X chromosome, for each individual in the sample is also a measure of DNA sample quality. Genome-wide, excessive, or reduced proportion of heterozygote genotypes could potentially be an indication of DNA contamination or inbreeding. It is therefore recommended that individuals who deviate ±3 SD from the overall sample heterozygosity rate mean should be removed [90]. The mean heterozygosity rate for each individual is given by $(\mathbf{J} - \mathbf{O}/\mathbf{J})$, where $\mathbf{J}$ is the number of non-missing genotypes and $\mathbf{O}$ is the number of observed homozygous genotypes for a given individual [25].

To address potential issues with cryptic relatedness, identity by descent (IBD), which is determined by the proportion of the genome shared by a pair of individuals from a common ancestor is commonly applied to remove unknown second-degree relatives or duplicate samples. Conventionally, it is assumed that individuals included in a GWAS are unrelated i.e. no

pair of individuals is closely related, where second-degree relatives are treated as sufficiently unrelated. Inclusion of closely related individuals could result in bias of SNP effect size and standard errors if not accounted for in the analysis. Therefore IBD, a measure of how strongly pairs of individuals included in a sample are genetically related, is commonly applied. Pairwise identity-by-state (IBS), which is based on the average proportion of alleles shared in common at genotyped SNPs (excluding the sex chromosome), can be applied to genome-wide data to estimate the degree of recent shared ancestry for a pair of individuals [25, 91]. IBS is calculated for each pair of individuals in the sample based on SNPs in low LD (typically $r^2 < 0.2$), where the selection of uncorrelated SNPs is referred to as pruning. If pairs of individuals are found to be related, which is usually based on an IBS metric (pi-hat) threshold value > 0.1875 [25], one of the related pair is removed from the sample [25].

To identify individuals who potentially may be considered ancestry outliers principal component analysis (PCA) is usually applied to genotyped data to form a relatedness matrix where eigen decomposition is performed to generate a smaller set of variables through a few linear combinations of the original variables. At the continental level, most of the variation in ancestry is usually explained by the first two PCs, therefore the first two PCs, which can be viewed graphically via a scatter plot, are used to identify "ancestry outliers" who are subsequently removed from the dataset.

SNP quality control: Data quality pertaining to the SNPs included in a standard GWAS are often assessed based on level of missingness, MAF, deviation from HWE and, for studies that undertake genotype imputation, imputation quality. Assessment of the level of missingness within SNPs involves determining the number of samples for which a genotype has not been assigned. To facilitate the identification and removal of low-quality SNPs, exhibiting excessive genotype missingness, different genotype calling rate thresholds are applied after examination of the overall genotype data. It is typical for the final user-defined genotype calling rate threshold to be in the range of 95%-99% [91]. SNPs with low MAF are more prone to genotyping errors and standard GWAS analysis methods are not powered to detect association with low MAF SNPs. Therefore, only SNPs with a MAF above a user-defined (determined after examination of the genotyped data) threshold are included. Generally, SNPs with a MAF of less than 1-5% are excluded, however with advances in imputation this threshold has been decreasing [93]. Furthermore, a more stringent genotype calling rate is sometimes

applied for SNPs with a MAF < 5% [94]. However, the threshold applied may depend on the sample size with larger samples applying lower MAF thresholds. For large (100,000) and moderate (10,000) samples, thresholds of 0.01 and 0.05 are commonly used [90]. However, for large sample studies like the UK Biobank, MAF thresholds as low as 0.1% have been applied [95].

Deviation from HWE is an indication of evolutionary selection, however, it is also a common indicator of genotyping errors. Therefore, the samples of controls included in a GWAS are assessed for deviation from HWE. Cases are typically not included in this assessment as genetic markers are expected to deviate from HWE if there is an association between markers and disease. Differences in call rates between the cases and controls are also assessed (differential genotyping error) as this can result in inflation of the type I error rate. Furthermore, for studies that undertake genotype imputation, assessment of imputation quality is an important standard quality control procedure designed to facilitate the removal of poorly imputed SNPs. Different imputation quality metrics are associated with each genotype imputation software tool. Among the three most applied software tools the r^2, the allelic r^2 metrics are implemented in Minimac and Beagle and the information metric (or info score) is implemented in IMPUTE2 [91, 96, 97]. Although there is no consensus on filtering thresholds for removing poorly imputed SNPs, an info score greater than 0.4 is generally considered acceptable in relation to IMPUTE2 (values range between 0 and 1 where values approaching 1 is an indication of a SNP imputed with high certainty) [98]. For Minimac and Beagle r^2 values greater than 0.3 are typically used for filtering [99].

1.4.3.4. |Primary statistical analysis procedures in GWAS

In GWAS statistical tests of association are developed primarily within the generalized linear modelling (GLM) framework. The GLM enables or allows for adjustment for clinical covariates (and other factors) to be measured and accounted for in the modelling process. As the primary phenotype of common disease GWAS is disease status, where case-control data are often obtained, association testing is usually applied via the logistic regression model. The logistic regression model is an extension of linear regression modelling where the outcome (as case control data is a binary outcome; affected disease cases versus unaffected controls) of the linear model is transformed using a logistic function that predicts the probability of having case status

given a genotype class. The measure of the effect of the genotype SNP on disease outcome, adjusted for confounding factors like ancestry, is usually based on the odds ratio (OR) (features of the logistic model are described further in section 1.7.4).

The genotypes for a SNP can be grouped into genotype classes or models that reflect an assumed relationship between the genotyped SNP and disease outcome. The choice of model applied in the analysis or test can have implications for statistical power. This is because the degrees of freedom for the test may be altered depending on the number of genotype-based groups. Additionally, loss of statistical power may also occur if model assumptions regarding the relationship between the genotyped SNP and disease outcome are wrong. These genetic models include, the dominant, recessive, multiplicative, or additive models of which the additive genetic model is most often applied (see Figure 1.4). The additive genetic model assumes that the risk allele effect is linearly related to the number of risk alleles, while the dominant genetic model assumes that the risk allele effect is related to the presence of the risk allele. Additionally, the recessive model assumes that the risk allele effect is related to the presence of both risk alleles.

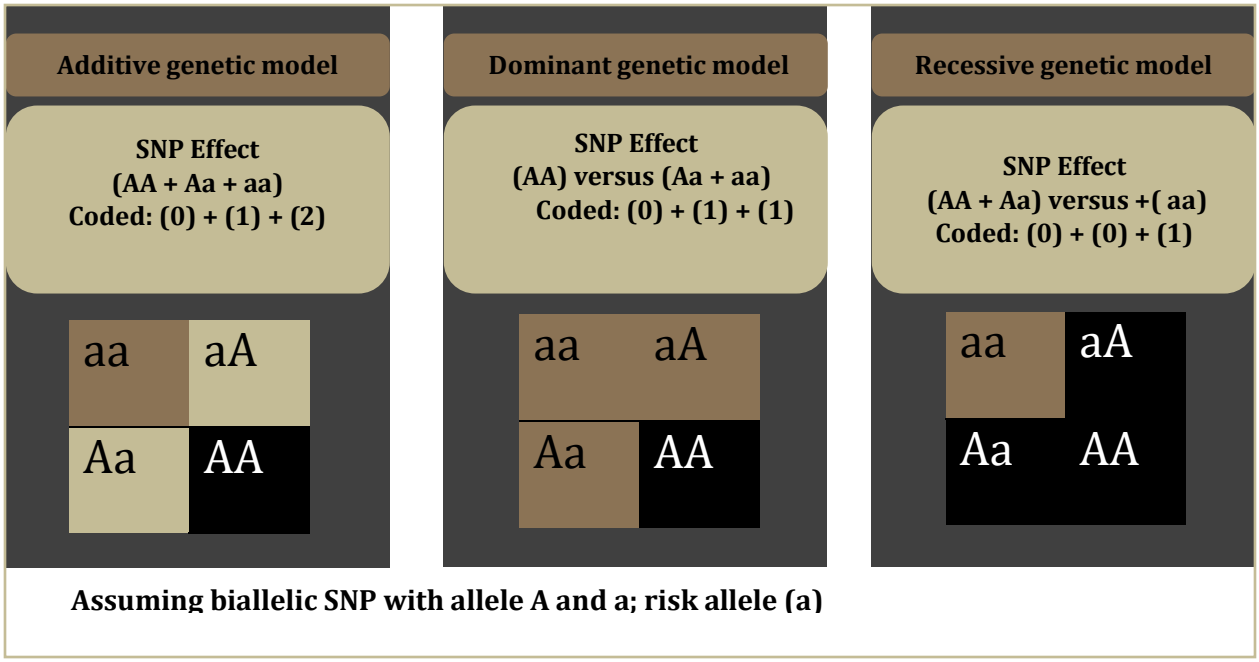


Figure 1. 4 - SNP effect according to genotype phenotypic model

Currently the standard practice in GWAS is to test each SNP that is typed on the GWAS genotyping microarray independently of each other to identify associated signals pertaining to

the trait or disease under consideration. This leads to the problem of multiple testing as the probability of observing a "significant" result purely by chance increases with the number of statistical tests performed. For example, in a test involving 500,000 SNPs: 5,000 expected to be significant at $\alpha < .01$; 500 expected to be significant at $\alpha < .001$ and 0.05 expected to be significant at $\alpha < 10^{-7}$. In association testing, the P-value which is the probability of making a type I error is used as a measure of statistical evidence against the null hypothesis, where a smaller p-value indicates stronger evidence substantiating the alternative hypothesis. The p-value in disease GWAS is an indication of how likely a suspected disease associated variant is due to random chance. In genetic research, different statistical significance p-value thresholds have been applied to differentiate true positives from false positives. The genome-wide significance P-value threshold of 5×10^{-8} has become a standard for common-variant GWAS and is based on patterns of LD in European populations. The three main methods that have been developed to address multiple testing are Bonferroni correction, false discovery rate and permutation testing [90].

This standard genome-wide significance threshold was formulated on the basis of permutation tests applied to International HapMap Consortium (IHC) genotype data in 2005 to estimate the number of independent chromosomal segments with MAF ≥ 5% [45]. Based on the LD structure of European ancestry populations, the effective number of independent SNPs across the genome was approximated by counting 1 SNP per LD block, plus all SNPs outside of blocks (interblock SNPs). On this basis a typical European ancestry population under study has about 1 million independent chromosomal segments with MAF ≥ 5%. The genome-wide significance threshold is equivalent to the Bonferroni correction for the approximately one million independent tests performed in a GWAS [100] (*Bonferroni correction for m tests set significance level to $\alpha = .05/m$).* In European ancestry populations, the genome-wide significance threshold has been shown to adequately control for the number of independent SNPs in the entire genome, regardless of the actual SNP density in the population under investigation [90]. Concurrently, SNP genotyping arrays can genotype up to 4 million markers and imputed genotyped SNPs increases further the number of tested SNPs [90]. Additionally, African ancestry populations are estimated to have around 2 million independent chromosomal segments [101]. Due to the greater level of genetic diversity among those individuals of African ancestry, a more stringent threshold is required (probably close to 1.0×10^{-8}) [90].

1.5. | Ancestry inference and application in biomedical research

This section describes the primary methods applied in genomic research to infer the ancestry of human populations based on genotype data. The section focuses on its application to the issue of population structure. Methods for detecting and describing population structure are outlined and includes an assessment of their strengths, limitations, and areas for improvement essential for the long-term development of the GWAS methodology.

1.5.1. | Ancestry inference in biomedical research

Ancestry inference is an important part of the framework for the analysis of population genetic or genomic data. It has various applications, but in humans it is often applied to account for the effects of population structure in genetic studies of traits and common diseases. Self-reported ancestry is usually uninformative for the purpose of classifying individuals into distinct population groups based on ancestry [102]. This is especially so in populations consisting of admixed individuals, as it is not possible to account for the degree of admixture within individuals based on self-reported ancestry alone. Technological and computational advances in the field of genomics have enabled the development of inference methods based on genotype data. Generally, methods designed to detect or account for population structure have two main approaches to inferring genetic ancestry. Global ancestry inference, the first approach, is geared towards estimating the genome level contribution proportions from each ancestral population, which provides a global view of admixture in the target population [103]. Global ancestry is thus defined as the relative proportion of ancestral blocks from each contributing population across the genome [104]. However, in the second approach, local ancestry inference is made in regard to the number of copies of chromosomes from a particular population are at a given site where local ancestry is defined as the genetic ancestry of an individual at a particular chromosomal location, where an individual can have 0, 1 or 2 copies of an allele derived from each ancestral population [103, 104].

Since 2003, several methods for inferring ancestry have been applied that incorporate both local and global ancestry, or global ancestry only. These methods differ primarily in their modelling approach. On a broad level, these methods can be classified into two main

approaches: parametric and non-parametric. In parametric approaches, global ancestry inference is made on the basis of an ancestry coefficient assuming a specified statistical model. Parametric approaches are based on several genetic assumptions about the SNPs including HWE and linkage equilibrium (LE). The most widely used parametric approach is STRUCTURE (first proposed in 2000), which relies on Bayesian Markov chain Monte Carlo (MCMC) [105]. The method can identify subpopulations from genome-wide genotyped samples through the detection of allele frequency differences within the data which are then used to assign individuals to those discrete sub-populations [106]. The genotype data used to infer ancestry is usually based on a selected set of unlinked or uncorrelated genetic markers (null markers) that are not associated with the disease or trait of interest. This approach has limitations that include difficulties in assigning individuals to subpopulations when they are a continuous mixture of ancestral subpopulations or admixed individuals [107]. Other approaches which have been popularly adapted include ADMIXTURE (first proposed 2009) and FRAPPE (first proposed in 2006) [14], which are based on maximum likelihood estimation (MLE). An expectation-maximization (EM) algorithm is used to optimize the likelihood for both allele frequencies and fractional group memberships in both methods, however, a faster optimization algorithm is utilized by ADMIXTURE. ADMIXTURE, which has many of the same capabilities of STRUCTURE, has the advantage of less computing time while maintaining similar accuracy as STRUCTURE [14]. Furthermore, ADMIXTURE has been shown to be more accurate in estimating global genetic ancestry than FRAPPE [108]

In the non-parametric approaches, which require no underlying modelling assumptions, multivariate analysis techniques are utilized to infer structure in the data. The most commonly applied multivariate analysis techniques applied to ancestry inference include PCA, clustering analysis, multidimensional scaling (MDS) and principal coordinate analysis [102]. On a broad level these approaches can be categorised as: (1) dimension reduction–based methods, as they typically apply a dimension reduction technique to reduce the dimensions of the space of genetic markers before clustering is applied; and (2) distance-based methods that compute pairwise similarities/distances between individuals before applying clustering on the computed allele-sharing distance matrix to infer population structure [105]. Examples of dimension reduction–based methods include PCA, singular value decomposition (SVD) and MDS. Distance-based methods include allele-sharing distance and Ward's minimum variance

hierarchical clustering (AWclust), Spectral Hierarchical clustering for the Inference of Population Structure (SHIPS) and NETVIEW [105].

In relation to population structure, estimation of genetic ancestry has two main areas of application in biomedical research: (1) individual global ancestry can be applied as a genetic background covariate for population structure control; and (2) locus-specific ancestry can be directly used to detect association with disease, which is referred to as admixture mapping (introduced in section 1.2.2) [109]. Historically, panels of ancestry informative markers (AIM, i.e. markers that exhibit marked differences in allele frequencies between two or more populations) were commonly employed to infer ancestry [110] for admixture mapping. In order to apply admixture mapping procedures, genotyping data are required from both the admixed and ancestral populations. As a result, ancestry specific reference panels have been developed that can distinguish between populations with continental differences originating from African, Asian, Native Amerindian and European populations. The first admixture scans were published in 2005, and by 2010 high-density mapping panels were constructed for African Americans, Latino/Hispanics and Uyghurs *populations* [111]. Reference panels designed to capture fine-grained intracontinental admixture are also being developed [109]. Information pertaining to genetic ancestry is also needed to facilitate genotype imputation of untyped genotype SNP in GWAS which usually require matched ancestral reference panels.

1.5.2. | Accounting for population structure

As the application of the GWAS approach continues to widen geographically and given the ever-growing size and complexity of genetic data, the availability of accurate and efficient tools to detect and/or account for the effects of population structure becomes even more paramount. To effectively undertake genetic association analysis in the general population, a clear insight into the underlying genetic population substructure is key, particularly in populations consisting of individuals originating from diverse geographical backgrounds. Historically, global ancestry has been used to control for the effects of population structure in GWAS. However, there is increasing recognition that both the effects of local and global ancestry need to be accounted for. As indicated in section 1.4.1, three main forms of population structure have been described in the literature: discrete structure; admixed population and hierarchical structure (includes both discrete and admixed individuals). Furthermore, development of

computationally efficient and scalable ancestry inference methods is essential for the sustainable development of GWAS methodology.

An array of approaches to address population structure in population-based association studies have been proposed over the past two decades. These proposed methods, which utilize genotype information from a whole genome set of SNPs or from a set of selected AIM encompass or extend across two main strategies: (1) overdispersion of test statistic approach; and (2) genetic ancestry inference approach. The overdispersion of test statistic approach involves direct estimation of the level of inflation in the test statistic owing to population structure. Based on genome-wide association summary statistics, these approaches use an overdispersion model to determine a test statistic appropriate empirical distribution [112, 113]. To correct for the inflation due to population structure a uniform correction factor can be applied to the original test statistics for each SNP [114], For the overdispersion of test statistic approach an overall genome-wide inflation measure can be applied (i) based on a genomic control measure or (ii) based on LD score regression intercept.

The approach formulated on the basis of genetic ancestry inference is designed to minimize potential population structure by distinguishing the most plausible underlying subpopulations within the overall population. Depending on the assumed form of population structure (discrete structure; admixed population and hierarchical structure), some genetic ancestry inference methods may assign individuals to a single ancestry while taking into account the effects of admixture through local ancestry inference, while other approaches assign individuals to a single ancestral population using global ancestry inference without taking into account the effects of admixture. Methods may also differ regarding whether or not they take into account the assumed underlying LD structure of the different ancestral populations. In the context of the genetic ancestry inference approach, when correcting for inflation resulting from population structure, population membership is viewed as an unmeasured covariate or more generally, proportions of ancestry from different populations are seen as unmeasured covariates [115]. For the genetic ancestry inference approach, methods applied entail (i) obtain ancestry-based covariates to account for the effects of ancestry differences resulting in population structure and (ii) linear mixed models for both population structure and cryptic relatedness.

1.5.2.1. |Overdispersion of test statistic approaches

Genomic control approach: The genomic control approach is among the earliest statistical methods proposed to address population structure [107]. In this approach a genome-wide measure is used to adjust the test statistic of all SNPs included in the analysis. This genome-wide measure is termed the inflation factor lambda **(λ)** and is assumed to be the same across the human genome. The value for the inflation factor is arrived at by calculating the chi-squared statistics for a set of unlinked or uncorrelated genetic markers (null markers) not associated with the disease or trait of interest. The empirical median for this set of chi-squared statistics is divided by the median of the chi-squared distribution (with the appropriate degrees of freedom for the statistical test) to obtain the value for the inflation factor that is used to adjust for the effects of population structure. This inflation factor is used to adjust the observed p-values of the candidate SNPs so that the corrected median p-value will be 0.5. Because the vast majority of variants are expected not to be associated with the disease or trait of interest the median observed p-value is expected to be close to 0.5 in the absence of population structure. This correction is applied by dividing the actual association test chi-square statistics by the computed inflation factor value. The value of the inflation factor can be considered also as a measure of the extent of the effects of confounding on the association statistics. A λ value of 1 is an indication of no inflation. However, a λ value of 1.03 or higher suggests that there may be inflation [57]. One of the main limitations of the genomic control measure is its inability to handle the effects of admixed populations.

LDscore regression approach: LD Score regression is one of the more recent developments for addressing the issue of population structure based on GWAS summary statistics [116]. It is a measure designed to ensure that confounding due to population structure does not inflate the number of false positives. The ability to distinguish inflation due to polygenicity from bias is an important issue in GWAS, particularly with the increasing sample sizes of GWAS. LD score regression quantifies the contribution of each SNP by examining the relationship between the test statistics and LD [117]. Estimates have been found to be a more accurate measure of test score inflation when compared to the genome-wide genomic control measure. The LD score regression approach is formulated on the basis that the more genetic variation that a marker tags, the higher the probability that it will tag a causal variant. In contrast, variation due to

population structure or cryptic relatedness is not expected to correlate with LD. The method is implemented by regressing the test statistics from GWAS against LD score. The intercept minus one from this regression is an estimator of the mean contribution of confounding to the inflation of the test statistics [118]. The correction is applied by dividing all the GWAS χ^2 statistics by the intercept value, which is expected to have the effect of restoring the average χ^2 statistic of these null SNPs to the theoretically proper value of unity and thereby bring the Type I error rate close to the targeted level [116].

1.5.2.2. |Genetic ancestry inference approaches

Genetically derived ancestry as covariate approach: In genomic research, population structure is often accounted for by including ancestry derived from genotype data into the model as a covariate. Within this framework, ancestry is usually based on global ancestry inference. Methods developed or applied to ancestry inference have been described in section 1.5.1. Over the past decades, several parametric and non-parametric methods have been proposed that utilize these ancestry inference procedures. The main limitation of parametric approaches relates to their sensitivity to sample size, which can affect model assumptions and have proven to be impractical for large GWAS datasets. Furthermore, parametric approaches are not applicable to highly structured populations due to limits on the number of subpopulations that can be inferred. In contrast, non-parametric approaches have proven more viable owing to the advantage of having more efficient computational costs [105].

Among the non-parametric approaches applied in biomedical research, PCA is the most cited dimensional reduction method [105] used to detect and or mediate the impact of population structure. The PCA approach involves the application of a set of unlinked or uncorrelated genetic markers (null markers) not associated with the disease or trait of interest. Each genotyped SNP is modelled as a quantitative variable in the number of copies of the minor allele (additive genetic model). Based on each genotyped marker included in the analysis, the PCA determines the pattern of genetic variation over the individuals in the samples, which is designed to capture on a broad level the degree of "genetic similarity" among individuals. Each resulting axis explains as much of the genetic variance in the data as possible with the constraint that each component is orthogonal to the preceding components. Ancestry is usually explained by the top PCs. However, it is typical for one to ten PCs to be modelled [119]. The resulting

PCs are included in the statistical analysis to account for the effects of population structure as covariates [107]. One of the main limiting factors affecting the PCA method is when population structure is very complex because too many PCs are needed [120].

Mixed modelling approach: Covariate-based approaches generally assume that the genotyped samples are unrelated and are therefore only designed to address the issue of population structure as it relates to ancestry differences. However, linear mixed models (LMM) are designed to address confounding due to both relatedness and population structure. The level of relatedness is captured in the modelling of the covariance of the genome-wide SNP genotype data between individuals [121]. The linear mixed modelling approach is formulated around three main steps: (1) the modelling process builds a genetic relationship matrix (GRM) which models genome-wide sample structure; (2) estimates the GRM contribution to phenotypic variance using a random effects model with or without additional fixed effects; (3) computes association statistics that account for the GRM component of phenotypic variance [122]. The mixed model approach is among the more recent approaches that have been proposed to address population structure and is currently widely used in GWAS to account for population structure and relatedness for both continuous and binary traits. The increasingly routine application of LMM in GWAS of binary phenotypes is owed to LMM flexibility in being able to account for population structure, as well as, their computational tractability, when compared to logistic mixed models [123]. However, due to concerns regarding population structure resulting in a violation of the LMM constant residual variance assumption, a logistic mixed model approach for GWAS of binary traits has also been proposed to account for population structure [118]. More recently also , protocols relating to the application of linear models with binary phenotypes to prevent loss of power, even in the presence of extreme case-control imbalance have also been proposed [123]. A limitation of LMM when applied to binary phenotypes relates to the inability to directly generate ES estimates on the OR scale. However, methods to acquire approximations on the log OR scale have been developed [123].

1.5.2.3. |Software and tools

As the size of GWAS continues to grow both the effectiveness and efficiency of these ancestry inference methods becomes even more pertinent. Along with the methods proposed for correcting population structure, numerous software tools have also been introduced to support their application [14, 103, 105, 108, 124]. In relation to the methods outlined in this section a summary of the available tools for global ancestry inference is outlined in Table 1.2.

Table 1. 2 - Summary of global ancestry methods and software

Program	Method	Function	Related Publications
Eigensoft	PCA	Calculate PCA from genotype data	1) Patterson, N., Price, A.L. and Reich, D., 2006. Population structure and eigenanalysis. PLoS genet, 2(12), p.e190. 2) Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics, 38(8), pp.904-909. (***Link to software download***: https://reich.hms.harvard.edu/software)
LASER	PCA	Calculate PCA from sequencing data (low pass)	3) LASER 1.0 algorithm: Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., Fulton, R., Wilson, R.K. and Mardis, E.R., 2014. Ancestry estimation and control of population stratification for sequence-based association studies. Nature genetics, 46(4), pp.409-415. 4) LASER 2.0 algorithm: Wang, C., Zhan, X., Liang, L., Abecasis, G.R. and Lin, X., 2015. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. The American Journal of Human Genetics, 96(6), pp.926-937. 5) LASER server: Taliun, D., Chothani, S.P., Schönherr, S., Forer, L., Boehnke, M., Abecasis, G.R. and Wang, C., 2017. LASER server: ancestry tracing with genotypes or sequence reads. Bioinformatics, 33(13), pp.2056-2058. (***Link to software download***: http://laser.sph.umich.edu/)

Program	Method	Function	Related Publications
FlashPCA	PCA	Rapid calculation of PCA	6) version ≥2: Abraham, G., Qiu, Y. and Inouye, M., 2017. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics, 33(17), pp 2776-2778. 7) version ≤1.2.6: Abraham, G. and Inouye, M., 2014. Fast principal component analysis of large-scale genome-wide data. PloS one, 9(4), p.e93766. (***Link to software download***: https://github.com/gabraham/flashpca)
PC-AiR	PCA	PCA in samples that may contain cryptically related participants	8) Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M. and Conomos, M.P., 2019. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics, 35(24), pp.5346-5348. (***Link to software download***: http://bioconductor.org/packages/release/bioc/html/GENESIS.html)
PCAmask	PCA	PCA in highly structured populations	9) Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D. and Kenny, E.E., 2017. Human demographic history impacts genetic risk prediction across diverse populations. The American Journal of Human Genetics, 100(4), pp.635-649. (***Link to software download***: https://github.com/armartin/ancestry_pipeline)
PLINK	MDS	Calculation of multi - dimensional scaling variables from IBD distance matrix	10) Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics, 81(3), pp.559-575. (***Link to software download***: http://zzz.bwh.harvard.edu/plink/)
EMMA	Mixed model	Perform linear mixed model analysis for quantitative traits	11) Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J. and Eskin, E., 2008. Efficient control of population structure in model organism association mapping. Genetics, 178(3), pp.1709-1723. (***Link to software download***: http://mouse.cs.ucla.edu/emma/)

Program	Method	Function	Related Publications
GEMMA	Mixed Model	Perform linear mixed model analysis for quantitative traits	12) Zhou, X. and Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nature genetics, 44(7), p.821. 13) Zhou, X. and Stephens, M., 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods, 11(4), pp.407-409. (***Link to software download***: http://www.xzlab.org/software.html)
EMMAX	Mixed Model	Perform linear mixed model analysis for quantitative traits more quickly than EMMA	14) Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E., 2010. Variance component model to account for sample structure in genome-wide association studies. Nature genetics, 42(4), pp.348-354. (***Link to software download***: http://genetics.cs.ucla.edu/emmax/)
LD score regression	LD score regression	Calculate genomic inflation parameters accounting for LD	15) Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L. and Neale, B.M., 2015. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics, 47(3), pp.291-295. (***Link to software download***: https://github.com/bulik/ldsc)
PC loading regression	PC loading regression	Improved population structure control compared with PCA	16) Bhatia, G., Furlotte, N.A., Loh, P.R., Liu, X., Finucane, H.K., Gusev, A. and Price, A.L., 2016. Correcting subtle stratification in summary association statistics. bioRxiv, p.076133. (***Link to software download***: Not yet available)

Source (Population stratification in genetic association studies, 2017) [14]

1.6. | Common disease GWAS

This section provides an overview of the disease areas that have benefited from common disease GWAS both in terms of the disease risk variants that have been identified and AOO of disease variants. The prospects of undertaking a common disease GRS GWAS is also discussed.

In relation to the GRS, a general overview of some of the most common pseudo R^2 measures that can be used to assess the predictive power and accuracy of GRS or explained variance attributable to the GRS is also included. These pseudo R^2 measures are compared with the view to determining the most appropriate pseudo R^2 metric that can be applied to assess the relative performance of different GRS models.

1.6.1. | Disease risk GWAS

In an epidemiological framework the risk of disease is assessed via longitudinal studies where exposures or risk factors are measured at the start of the study period (Figure 1.5). The probability of developing the disease conditional on exposure or risk factors is assessed based on the disease status of individuals included in the study. Through these epidemiological studies conducted in the general population the presence of disease within a population is often measured on the basis of incidence rates (occurrence of new cases of disease during specified time period); prevalence rates (proportion of cases of disease in a population at a given time); or survivorship rates (proportion of individuals surviving over a specific period). Although the data obtained from prospective cohort studies are of better quality when compared to case-control studies, the case-control design is most commonly applied in GWAS. This is due to the fact that case-control studies are less costly and time-consuming when compared to prospective cohort studies. As case-control GWAS are the most applied, regression analysis of binary traits or disease are usually undertaken within a logistic regression framework (described in section 1.7.4). The primary phenotype of interest is usually disease status; however, the genetic determinants of many chronic complex diseases includes an AOO component.

In medical genetics, rare and common diseases are usually characterized by different levels of penetrance. Penetrance is defined as the percentage of individuals having a mutation or genotype who exhibit clinical signs or phenotype of the associated disorder or genotype [125]. A highly penetrant allele means that the trait it produces will almost always be apparent in an individual carrying the allele (which is often the case for single gene diseases). On the other hand, low penetrant alleles will only occasionally produce the associated trait. Therefore, with low penetrance it is more difficult to distinguish the genetics from environmental factors. Furthermore, penetrance at a given allele may be polygenic (i.e allele is modified by the presence or absence of polymorphic alleles at other gene loci) [126]. The diseases or trait

altering variants discovered by GWAS tend to be common and of low penetrance [127](common alleles having small genetic effects). The proportion of variance of a trait or disease controlled by genes is defined as the heritability. Therefore, heritability is a measure of the genetic contribution to phenotypic variation. If common alleles have small genetic effects (low penetrance), but common diseases show heritability (inheritance in families), then multiple common alleles must influence disease susceptibility [3].

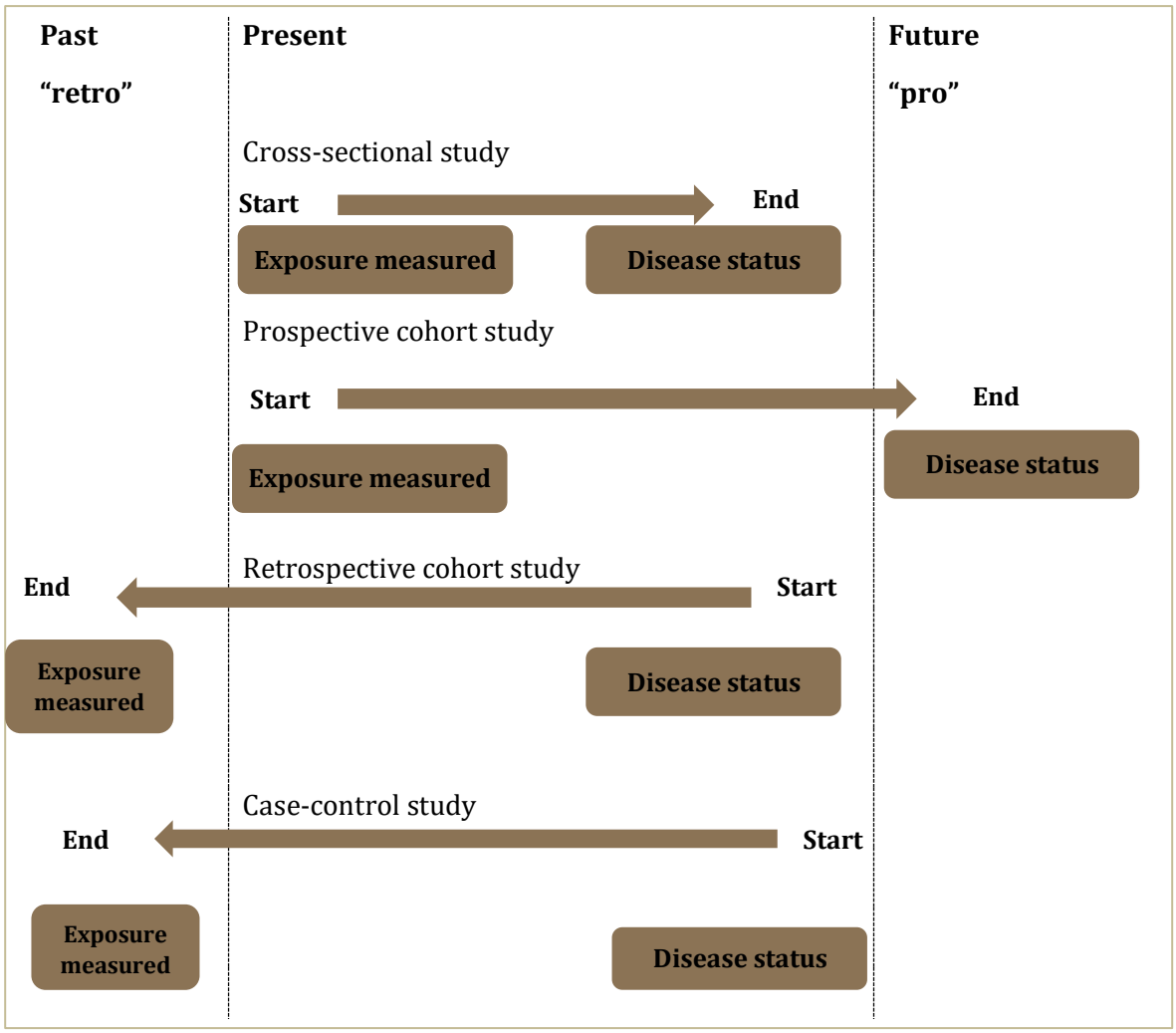


Figure 1. 5 - Epidemiological study designs

As indicated in section 1.4.3.1, genetic heterogeneity presents an additional challenge in common disease GWAS. In a clinical setting genetic heterogeneity refers to the presence of a variety of genetic defects that cause the same disease [128]. This is often because of mutations

at different loci on the same gene. Diseases known to experience this phenomenon includes Alzheimer's disease, cystic fibrosis, and polycystic kidney disease.

Despite these challenges, since the introduction of the GWAS design, numerous genetic variants associated with the risk of many common diseases have been identified. Cardiovascular disease, many different types of cancers, Alzheimer's disease, Parkinson's disease, inflammatory bowel disease and T2D are among the more burdensome diseases that have benefited most from GWAS. Before 2007 there were fewer than 20 genetic variants associated with the risk of common diseases or traits [127], but by 2018 this had changed dramatically. More than 161 genetic risk loci have been associated with coronary artery disease accounting for 15% of the genetic contribution to the disease [129]. Among the four major neurodegenerative diseases which includes Alzheimer's disease and Parkinson's disease over 200 loci have been found to be associated with disease risk [130]. Furthermore, the number of risk loci associated with inflammatory bowel disease are in the region of 240 [131].

1.6.2. | Age-of-onset GWAS

Genetics play a significant role in determining risk to common diseases as well as AOO and for some diseases the effect of AOO may be greater than the genetic effect of risk [132]. For preventive and therapeutic strategies, a better understanding of the biological mechanisms impacting not only disease risk but also AOO of disease is critical. However, current research generally focuses on identifying disease risk variants and to a much lesser extent AOO of disease. This may be because most GWAS are based on the case-control design, therefore it would be more difficult to obtain data on AOO retrospectively [133] and due to cost constraints prospective cohort studies are on the decline. Additionally, GWAS analysis tools for time-to-event (TTE) analysis is not as readily available as tools for logistic regression.

Evidence based on heritability studies seem to indicate that the level of heritability (which is often wide ranging) attributed to a common disease is influenced by AOO [134]. Earlier AOO for some common diseases is associated with increased risk of complication and comorbidities as well as more aggressive progression of the disease or severity. Identification of genetic variants associated with AOO independent of disease risk may provide an opportunity for improved drug therapies. The first Alzheimer's disease GWAS focused on identifying genes

associated with AOO conducted in 2011 [135]. Additionally, a Parkinson's disease study in 2015 found that two loci (GBA and TMEM175/GAK) significantly altered the AOO of Parkinson's disease [136].

Given the limitations of current GWAS methodology in distinguishing the genetic variants associated with the AOO of disease independently of overall risk variants, different approaches have been proposed to distinguish early onset individuals from late onset individuals. This is based on the hypothesis that an increased polygenic burden is expected in individuals with an earlier onset relative to late onset. This is because earlier onset of disease indicates less exposure to lifestyle risk factors, therefore onset of disease is likely to be influence by genetics to a much greater extent.

1.6.3. | GRS GWAS

GWAS discoveries have contributed tremendously to the current knowledge of genetics and its implication for human health. Given the interplay of both genetic susceptibility and environmental risk factors in common diseases, numerous applications of risk prediction models have been explored incorporating both components [137]. The extent to which genetic and non-genetic factors are included depends on the underlying heritability of the disease.

The genetic component of these risk prediction models commonly employs a risk profiling approach that incorporates information from multiple variants [83]. Identification and analysis of contributing genomic variants associated with many common complex diseases have enabled the application of genetically based risk scores. Given the underlying polygenic architecture of most common diseases, single SNP GWAS analysis in isolation lacked the ability to simultaneously assess the overall genomic risk of an individual associated with a disease or trait. Risk scores, derived from genetic data, are designed to capture an individual's overall genomic risk to disease by aggregating the number of risk variant alleles present in an individual into a single numeric measure of risk. In some instances, these risk variant alleles are weighted in terms of their impact usually estimated by their log OR, which are commonly obtained from discovery GWAS summary data.

Applications of risk prediction models based on polygenic scores have implemented various strategies or approaches for capturing the genetic component of disease risk. The most common of these include the GRS and the Polygenic risk score (PRS) [138]. The primary difference between the two approaches is that GRS represents a weighted or unweighted sum of risk alleles for a limited number of robustly associated SNPs, ie SNPs showing strong evidence of association with the disease (usually genome-wide significant SNPs). On the other hand, PRS represent a weighted or unweighted average of SNP effects across the whole genome. SNPs not significantly associated with the disease are incorporated in the measure and is based on the current hypothesis that heritability of disease can be captured by many SNPs which have a small effect on disease [57, 138].

GRS, which is explored further within this thesis, have the potential to identify individuals at risk of early age-onset disease because they are expected to have a greater genetic burden. Reduced exposure to lifestyle risks factors implies greater genetic burden of disease risk variants which would be reflected in higher GRS values. Therefore, the application of GRS is potentially beneficial given the limitation within GWAS in terms of its ability to distinguish the genetic variants associated with the AOO of disease independently of overall risk variants.

1.6.3.1. |Constructing GRS

To undertake risk prediction based on GRS requires information on the SNPs known to be associated with the disease or trait. Information on the associated SNPs as well as their associated summary statistics are usually obtained from the largest published GWAS currently available. This published GWAS is usually termed the "discovery or base GWAS". Construction of the GRS require independent GWAS samples of genotype data. These independent datasets in which the GRS are constructed are often termed the "target GWAS". The formulas that can be used to calculate the GRS for each individual in a target GWAS sample are described in Equation 1.3 for a weighted GRS and Equation 1.4 for an unweighted GRS. The effect size weighting for the weighted GRS is usually based on the OR value provided in the base GWAS.

$$\boldsymbol{GRS_{iw} = \sum_{j=1}^{J}(\beta_{jbase} \times G_{ij})}$$

Equation (1.3)

$$\boldsymbol{GRS_{iu} = \sum_{j=1}^{J} G_{ij}}$$

Equation (1.4)

Where $\boldsymbol{GRS_{iw}}$ is the weighted GRS of the **ith** individual; $\boldsymbol{\beta_{jbase}}$, refers to the effect size (log OR) for the risk allele in the base GWAS; $\boldsymbol{G_{ij}}$ is the genotype dosage (defined as the number of copies of the risk allele at each SNP present in an individual) or expected number of risk alleles present (0,1 or 2) for SNP (j) of individual (i) and $\boldsymbol{GRS_{iu}}$ is the unweighted GRS of the ith individual. $\boldsymbol{J}$ refers to the total number of SNPs included in the calculation of the GRS. The weighted or unweighted genotype dosage values associated with each SNP are added up for each individual to formulate the overall GRS per individual.

1.6.3.2. |Measures of predictive power and accuracy

Common disease risk prediction on an individual level within a genetic research framework, that incorporates an element of genetic prediction based on combined effects of multiple SNPs is increasingly becoming routine. However, measures to assess their predictive power or accuracy are critical before implementation into a clinical setting can be considered. The combined effects of the multiple SNPs included in a risk score can be measured by means of the coefficient of determination, denoted R^2, which is commonly used to quantify the phenotypic variance explained by the combined SNPs. As common diseases are invariably due to the combined interplay of genetics and the environment, the heritability of the disease in question provides a useful quantification of the importance of the genetic component of the disease. As individual SNPs typically have low penetrance, coupled with environmental impact, the estimated heritability provides a useful upper boundary for measuring the relative importance of the combined effects of multiple SNPs. Presently, most common diseases risk scores explain between 10% and 20% of the variance in disease risk [139], while heritability typically ranges between 30% and 60% [7].

In ordinary least squares (OLS) regression analysis, R^2 is the standard statistical measure used to assess the goodness of fit of a model. It is an overall measure of the accuracy of the regression

model [140]. However, OLS R^2 are not appropriate for the class of models used to analyse binary, ordinal and TTE outcomes, which are common in medical research. As a result, these classes of models used to analyse such data are limited in terms of the ability to compare the relative predictive power across models. Several methods akin to the standard R^2 measure used in OLS regression models have been proposed as a measure of predictive accuracy or explained variation, but none have been adapted as standard for models based on maximum likelihood estimation [141]. These measures are often referred to as pseudo R^2 measures as they are not a true R^2 measure in the same sense as the standard OLS R^2 measure, but appear to be on the same scales, as values range from 0 to 1. In general, pseudo R^2 compares the log-likelihood from the null model (model with only an intercept) to the log-likelihood from the full model (model with all the covariates included). Several methods have been proposed for binary and other categorical outcome data. The methods most commonly applied include the Mcfadden [142] pseudo R^2; Cox and Snell [143] pseudo R^2; and Nagelkerke [144] pseudo R^2.

The McFadden's R^2 is defined as:

$$R^2_{McF} = 1 - \ln(L_M) / \ln(L_0)$$

Equation (1.5)

The Cox and Snell R^2 is defined as:

$$R^2_{C\&S} = 1 - (L_0 / L_M)^{2/n}$$

Equation (1.6)

The Nagelkerke R^2 is defined as:

$$R^2_N = R^2_{C\&S} / R^2_{Max}$$

Equation (1.7)

In these equations, L is the estimated likelihood, M refers to the model with predictors (full model); O refers to the model without predictors (intercept model); n refers to the size of the sample. R^2_{Max} *for the* Nagelkerke R^2 is defined by $(1 - (L_0)^{2/n})$. The Nagelkerke measure adjusts the Cox and Snell measure for the maximum value so that 1 can be achieved. Application of the different pseudo R^2 measures depends to a large extent on the primary objective of the assessment, that is: (1) the square of the correlation between predicted and observed values; (2) improvement in fitted model from adding predictors to a null model; (3) proportion of explained variance in the data by the model. (e.g. R^2 can be calculated by subtracting the unexplained variance from one). Unlike normal generalized linear models, the different

definitions of R^2 do not coincide or lead to the same quantity in generalized linear models that are not normal. The McFadden's R^2 is regarded as an approach that mirrors more closely approach 2 and 3, while both the Cox and Snell R^2; and the Nagelkerke R^2 is viewed as an approach that mirrors approach 2. However, with the exception of the Nagelkerke R^2, a maximum value of 1 is unattainable for most pseudo R^2 measures. For this reason, the Nagelkerke R^2 measure has been applied to genetic risk prediction models as a measure of explained variance in genetic research [145, 146]. In this context 1- $\boldsymbol{R^2_N}$ can be interpreted as the proportion of variance unexplained by the genetic variants included in the prediction model [146].

R^2 is a useful measure for comparing different models and also to determine the contribution of a single variable to the overall model. Nested models are useful for assessing the individual contribution of a single variable while controlling for covariates in the model. The effect of adding an additional variable to a model can be assessed by comparing the R^2 of two nested models. The larger model is usually referred to as the complete (or full) model, and the smaller the reduced (or restricted) model. The contribution from this additional term can be obtained from the coefficient of partial determination (partial R^2). The partial R^2 measure is defined as the percentage of dispersion that can be described by the predictors specified in the fuller model but cannot be described in a reduced model [147]. It is important to note that model comparison is limited to: (1) comparing different models using the same dataset and outcome measure; and (2) the same pseudo R^2 must be used across models to facilitate comparison.

1.7. | Statistical analysis of GWAS

In this section consideration is given to different approaches to undertaking data analysis of different outcome measures commonly encountered in biomedical research. A general outline of the three main statistical approaches that are explored throughout this thesis is provided. Included are the descriptions of the various model equations and assumptions.

1.7.1. | Approaches to statistical analysis in GWAS

A key consideration in the design and analysis of GWAS of complex human disease outcomes or pharmacogenetic outcomes is the type of outcome of interest. The spectrum of outcome measures typically includes binary outcomes (e.g. presence or absence of disease; dead or alive), continuous outcomes (e.g. blood pressure; and pain scale) and TTE outcomes (e.g. time to death; AOO of disease; and time from start of drug therapy to first adverse event). However, as indicated in section 1.4.3.4, the most common approach to association testing is the case-control setup where the allele frequency of each SNP for individuals with the disease (cases) and individuals without the disease (controls) is compared. An estimate of SNP effect size is often measured based on the OR via logistic regression. The effect of the SNP is commonly based on the additive genetic model (described in section 1.4.3.4) where it is assumed that genotype effect is linearly related to the number of risk alleles (0, 1 or 2). The logistic regression approach assumes a binary outcome (presence or absence of disease), which is generally less powerful for outcomes which have a time element as is the case for AOO of disease, when compared to other approaches, particularly the TTE analysis approach [148, 149]. The TTE analysis approach is generally more powerful as it incorporates information on follow-up time span and allows for censoring, i.e. it considers both censoring and time [150](the censoring aspect of TTE analysis is discussed further in section 1.7.2). Within this thesis consideration is given to three of these outcome measures, TTE, ordinal and binary outcomes with a view to exploring their relative effectiveness in terms of statistical power in the context of AOO of disease GWAS.

1.7.2. | Analysis of TTE outcomes

Data analysis within a TTE analysis framework comprise a set of statistical procedures designed to interrogate data that have time as a key outcome of interest [151, 152]. A unique aspect of TTE analysis is that the research interest is typically a combination of whether the event has occurred (binary outcome) and when it has occurred (continuous outcome) [153]. Additionally, in TTE analysis, only some individuals will have experienced the event by the end of the study, which gives rise to the phenomenon of censoring (described in further detail below). As a result of censoring, not all individuals will have an event time, but instead a censoring time if the event of interest has not occurred. Censoring is addressed in the modelling process to enable valid

inferences regarding the data being analysed [153]. In undertaking TTE analysis, important methodological considerations include a clearly defined: (1) outcome variable (time until the occurrence of an event); (2) time origin (point as which follow-up begins); (3) time scale (time from the beginning of follow-up); and (4) criteria for exiting the study (for censored observations a criteria for exiting the study is need).

As the focus of this thesis is to investigate methods for the analysis of AOO of common disease in a GWAS setting, the time origin could, for example, refer to age at birth or age at entry into the study and time scale AOO of the disease of interest. Regarding censoring, in longitudinal epidemiological studies of common diseases, not all individuals under observation will experience the occurrence of the disease during the study period. Additionally, some individuals may also be lost to follow-up due to drop out or death due to a cause unrelated to the disease of interest. Both incidents are regarded as right censoring and is the most frequent form of censoring encountered in longitudinal studies. Left censoring, which is less frequent, occurs when the event of interest (in this instance occurrence of disease) has occurred before enrolment. Because of censoring, the AOO of disease is unknown for these censored individuals, but what is known is that their AOO is greater than the age at which they were last observed, their censored age. This event-free period for the censored individuals contributes information that is incorporated into the TTE analysis. Generally, censored individuals at the end of the study period are usually censored at their current age and incorporated into the TTE analysis as censored individuals, or controls.

In some situations, however, it may be appropriate to undertake a case only TTE analysis. In this analysis, only individuals who have been confirmed as cases and with a known AOO, if AOO is the timescale, are included in the TTE model. In the context of genetic research, a case only approach may be undertaken where for some diseases selection of appropriate controls may prove challenging [154].

The family of proportional hazards (PH) models is by far the most widely used regression specification applied in biomedical research to simultaneously assess the effects of potential risk factors and/or covariates on survival or event time (in this context AOO of disease). The hazard function, which describes the instantaneous rate of occurrence over time forms the basis of the PH model approach. In these models the Hazard Ratio (HR), which describes the

improvement in one group over the other in terms of rate at which events occur is the key measure of association for the PH model. These models rely on the fundamental assumption of proportionality of the hazards, which implies that potential risk factors or covariates included in the model have a constant impact on the hazard or risk over time. The Cox PH model, the most popular PH model, is characterised as a semi-parametric model because a parametric assumption is made concerning the effect of the predictors (and/or covariates) on the hazard function [155]. Therefore, the regression component of the model is fully parametric where predictors or covariates in the model are linearly related to the log hazard [155]. However, no assumption is made regarding the hazard function itself, which is left unspecified.

Although the PH framework has been highly successful in identifying and quantifying major risk factors for human disease [156] , nevertheless, there are situations where the PH assumption may not be appropriate. Moreover, there are some circumstances where more accurate estimates can be obtained via parametric approaches [157]. These parametric models are distinguished by their distributional form for the survival and hazard function. Among the most important parametric forms applied in biomedical research is the two-parameter Weibull distribution, which is more flexible than the Cox PH as the underlying hazard rate is not restricted to being constant over time. In this thesis, both TTE models are used. In the context of AOO of disease, the parametric Weibull model assumes AOO has a Weibull distribution. However, the semi-parametric Cox PH do not assume a specific distribution for AOO of disease but does assume a specific relationship between predictor (s) and or covariate(s) and the outcome (in this context disease status).

In the general Weibull model, the hazard function at time t is given by:

$$\boldsymbol{h(t) = \lambda \nu t^{\nu-1} exp^{(\beta X)}}$$

Equation (1.8)

In this model, $\boldsymbol{\lambda}$ is the positive scale parameter and $\boldsymbol{\nu}$ is the shape parameter, which determines whether the hazard rate decreases (ν<1), increases (ν>1) or remains constant (ν=1) over time. The baseline hazard rate is given by $\boldsymbol{\lambda \nu t^{\nu-1}}$, which is scaled by the function of covariates, $\boldsymbol{X}$, and corresponding regression coefficients, $\boldsymbol{\beta}$, via $\boldsymbol{exp^{(\beta X)}}$. In this context, the variable of

interest is the genotype of the causal SNP, coded under an additive model by the number of risk alleles carried (0, 1 or 2), and $\boldsymbol{\beta}$ is the log HR of the risk allele.

The Cox PH model is a special case of the general Weibull model for which the hazard rate is constant over time (i.e. ν=1), such that:

$$\boldsymbol{h(t) = \lambda exp^{(\beta X)}}$$

Equation (1.9)

In the presence of censoring two outcome variables are considered. The first is the observed time **(Y),** where $\mathbf{t_i}$ denotes age at last disease-free observation (censoring age $\mathbf{C_i}$) or AOO of disease ($\mathbf{T_i}$) for the $\boldsymbol{i^{th}}$ individual. The second is the censoring variable **($\boldsymbol{\delta}$)**, where $\boldsymbol{\delta_i}$**=1** indicates occurrence of the disease for the $\boldsymbol{i^{th}}$ individual; and $\boldsymbol{\delta_i}$**= 0** indicates that the disease has not occurred at the last observed $\mathbf{t_i}$ for the $\boldsymbol{i^{th}}$ individual.

With censoring, the joint partial likelihood **($\mathbf{L_p}$)** which is used to estimate the model coefficients **($\boldsymbol{\beta}$)** in given by:

$$\boldsymbol{L_P(\beta) = \prod_{i=1}^{n} \left[\frac{exp^{\beta X}}{\sum_{j \in \mathcal{R}(t_i)} exp^{\beta X}} \right]^{\delta_i}}$$

Equation (1.10)

In this equation, $\boldsymbol{\delta_i = 0}$, if $\boldsymbol{t_i}$ is a censoring time for the $\boldsymbol{i^{th}}$ individual , while $\boldsymbol{\delta_i = 1}$ if $\boldsymbol{t_i}$ is an event time (AOO) for the $\boldsymbol{i^{th}}$ individual. Additionally, $\boldsymbol{n}$ denotes the total number of observations; $\boldsymbol{j}$ denotes the number of individuals who experience occurrence of the disease; and $\boldsymbol{\in \mathcal{R}(t_i)}$ denotes the risk set which represents the set of individuals who are at risk of developing the disease at age $\boldsymbol{t}$.

1.7.3. | Analysis of ordinal outcomes

In the second approach AOO is viewed as an ordinal outcome where the proportional odds model [158, 159] is applied to assess the association of AOO of disease with late and early onset disease. Here the median age of cases is often used to distinguish between late and early onset disease. Unaffected individuals, who were considered as censored observations in the TTE analysis, are also included here as a third category. The ordinal outcome is therefore comprised of unaffected individuals (controls); late age onset (LAO); and early age onset (EAO). The proportional odds model is defined as:

$$\boldsymbol{Logit}\,[\,\boldsymbol{P}\,(\boldsymbol{Y_i}\,\leq\,\boldsymbol{d}\mid\boldsymbol{X})\,]=\boldsymbol{log}(\frac{\boldsymbol{\pi_d(X)}}{\mathbf{1}-\,\boldsymbol{\pi_d(X)}})=\mathfrak{x}_{\boldsymbol{d}}\,-\,\boldsymbol{\beta'}\,\boldsymbol{X},\boldsymbol{d}\,=\mathbf{1}\,,\ldots\,\boldsymbol{D}-\mathbf{1}$$

Equation (1.11)

Where $\boldsymbol{\pi_j(X)}=\mathbf{P}\,(\boldsymbol{Y_i}\leq\,\mathbf{d}\mid\mathbf{X})$ represents the probability of being at or below category $\mathbf{d}$ given a predictor variable. Additionally, $\boldsymbol{Y_i}$ denotes the ordered response variable with possible values **(1, 2,...d)** for the $\boldsymbol{i^{th}}$ individual; **D** number of response categories; $\mathbf{d}$ ordered response value; and $\mathfrak{x}_{\mathbf{d}}=(\mathfrak{x}_{\mathbf{1}},\ldots\,\mathfrak{x}_{\mathbf{D-1}})$ are the cut points or intercepts. The predictor genotype SNP is denoted by $\mathbf{X}$ and $\boldsymbol{\beta}$ is the logit coefficient which corresponds to the predictor. The proportional odds model assumes that the effects due to any predictor variable included in the model are the same across the ordered disease status categories (i.e proportional). As a result, a single logit coefficient is estimated for each predictor variable. Therefore, it is expected that the intercept pertaining to each response or ordered disease status category would be different, but the slope is the same for all categories.

1.7.4. | Analysis of binary outcomes

The binary outcome measure, which in this instance compares individuals affected and unaffected by a disease or trait of interest, are usually applied using the most common approach in GWAS, the logistic regression model [160] which is defined as:

$$\textbf{Logit}\,(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma + \theta X$$

Equation (1.12)

Where $\boldsymbol{\pi_i = P(Y_i = 1)}$ represents the probability of the occurrence of the disease of interest in individual **i** denoted $\mathbf{P(Y_i = 1)}$, while $\mathbf{1 - \pi_i = P(Y_i = 0)}$ corresponds to the probability of not developing the disease of interest for individual **i**. In the logistic model, the logit-transformed probabilities associated with disease outcome is modelled as a linear relationship with the predictor variables. The model predictor which in genetics is most commonly a genotype SNP is denoted by **X** while the corresponding coefficient is denoted by **θ.** The intercept (denoted γ) represents the overall probability of a single case of disease when all predictors are zero, it corresponds to the log of the baseline odds. The model assumes that the probabilities of disease occurrence for an individual $\mathbf{Y_i}$ has a Bernoulli distribution, therefore, the expected values for $\mathbf{Y_i}$ ($\mathbf{E(Y_i)}$) is equal to $\boldsymbol{\pi_i}$.

1.8. | Thesis objective and structure

The focus of the research in this thesis is to evaluate methods for detecting associations of SNPs and GRS with AOO of disease in the presence of population structure due to both substructure and admixture. Chapter 2 focuses on investigating methods to account for admixture in GWAS of TTE outcomes within admixed populations. The power to detect association of AOO of disease and SNPs in a TTE framework is investigated via simulations. Primarily, investigations compared the performance of two TTE models (Cox PH and Weibull models). The simulation study evaluates the impact of admixture on statistical power (which assumed an admixed population).

Chapter 3, which includes a component of both simulation and real data application, focuses on investigating methods based on the polygenic contribution to common diseases in single-ancestry populations. Investigations involves the construction of GRS which entails different versions of weighted and unweighted GRS. These GRS are used to test for association with AOO of T2D using two independent GWAS datasets. The first GWAS originates from the Northwestern University Gene (NUgene) Banking Project and comprises of 1,115 individual samples of which 46% are cases of T2D. The second GWAS originates from the Wellcome Trust

Case Control Consortium (WTCCC) and comprises of 3,810 individual samples with 24% being cases of T2D. A meta-analysis to combine the results of the two datasets and a GRS simulation study to further assess the statistical properties of the different statistical approaches is also undertaken. In the GRS simulation study evaluation is concentrated primarily on evaluating the impact of censoring on relative power between different statistical approaches.

Chapter 4 focuses again on investigating methods based on polygenic contribution to common diseases but in multi-ancestry populations, where different approaches to applying GRS to non-European ancestry populations is included in the investigations. Investigations involves the construction of GRS to test for an association with AOO of T2D using a UK Biobank dataset. The UK Biobank dataset consisted of just under 400,000 individuals of which approximately 17,000 are cases of T2D and comprised individuals of European (96%), Asian (2%) and African (2%) ancestry.

Chapter 5 brings together the major themes of the thesis and contains the general discussions, conclusions and recommendations for further work which are based on key findings of the work undertaken in chapters 2, 3, and 4.

Chapter 2: Investigating methods to account for population structure in association studies of age -of-onset of disease

Chapter Outline

In this chapter, the type I error rate (false positive rate) and power to detect association between age-of-onset (AOO) of disease and single nucleotide polymorphisms (SNPs) within an admixed population is investigated in a time-to-event (TTE) framework via simulations. The simulations compare the performance of the Cox proportional hazards (PH) and general Weibull TTE models. The simulation study evaluates the impact of population admixture on statistical power. Incorporated in the simulations is a comparison between the traditional association analysis approach, which seeks to establish an association between SNP genotype and the AOO, and an "admixture mapping" approach, which is based on local ancestry at a specified locus and seeks to identify genomic regions at which ancestry is associated with AOO.

2.1. | Introduction

From the onset of genome-wide association studies (GWAS), population structure was identified as a fundamental challenge affecting the validity of GWAS findings. As highlighted in Chapter 1, GWAS, if conducted in ancestrally diverse populations, have the potential to result in inflated type I (false positive) error rates due to the mechanisms of geographical confounding between the disease and SNP, if not accounted for in the association analysis [161, 162]. In a push to address population structure, numerous strategies have been explored to both detect and account for population structure. However, their performance depends on the type of population structure present within the population. These strategies are discussed in section 1.5, covering both global (provides global view of admixture) and local (provides locus specific ancestry) ancestry inference methods.

As indicated in section 1.4.1, common examples of recently admixed populations are based in the Caribbean and the Americas. African Caribbean populations are estimated to have ~65–95% West African, ~4–27% European, and ~0–6% Native American ancestry [163]. In relation to the Unites States of America (USA), research has indicated that African Americans are estimated to have on average 73.2% African, 24.0% European, and 0.8% Native American ancestry, while Latino Americans are estimated to have on average 18.0% Native American ancestry, 65.1% European ancestry, and 6.2% African ancestry [63].

In a standard GWAS, localization of the causal gene or region is facilitated by linkage disequilibrium (LD; discussed in section 1.2.1). However, LD is known to differ among ancestral populations [164]. Through the mechanism of LD, genotype SNP microarrays are formulated on the basis of representative SNPs, often referred to as a "tag SNP", which are used as proxies for a group of neighbouring SNPs in high LD that are usually found in haplotype blocks (set of SNPs found on the same chromosome that tend to be inherited together) throughout the human genome. These genotyping microarrays used in GWAS are designed to cover LD blocks, representing up to 80% of all SNPs with minor allele frequencies (MAF) > 5% in the genome of European ancestry populations [3]. Therefore, as the overall LD structure of the human genome in ancestral populations is different, the power to detect genetic associations across ancestral populations is not consistent as genomic coverage is likely to be different. As indicated in Chapter 1 section 1.4.3.2, to increase power in non-European ancestry populations, specific SNP array platforms are often applied. However, within the LD blocks, the pairwise LD between the causal SNP and tag SNP usually measured by the squared correlation coefficient (r^2) also differs among ancestral populations (GWAS genotyping microarray chips designed to capture pairwise LD of $r^2 \geq 0.8$ between SNPs).

In this chapter, simulations were undertaken to assess the impact of population admixture, on the false positive error rate and power of association of a SNP with AOO of disease. The study, which considers an admixed population, was evaluated in a TTE framework under an additive genetic model. Within this TTE framework, the relative performance of the Cox PH and general Weibull TTE models was the primary focus of evaluation. The range of scenarios considered included those where it was assumed that the causal SNP is directly genotyped as well as scenarios where the causal SNP is not directly genotyped, but association is instead tested with a correlated tag SNP due to LD, which varies across ancestral populations. As part of the admixed population simulation study framework, admixture mapping (described in Chapter 1) where ancestry at a specified marker locus forms the basis of analysis was compared to the standard genotype disease association approach [38, 165].

2.2. | Methods

2.2.1. | Description of study of admixed population

The simulation study considered a disease of interest in a sample of individuals ascertained from an admixed population originating from two ancestral populations. The impact of a bi-allelic causal SNP on the AOO of disease is considered where individuals in the study are followed from birth for 50 years, with a record made of the age at which a disease occurs. The impact of lost to follow-up because of drop-out were incorporated. Those that are unaffected at the age of 50 are considered as censored observations. It is assumed that the ancestry of both the maternal and paternal chromosomes in the genomic region flanking the causal SNP is known or correctly inferred.

A range of scenarios were considered with regards to population, genetic and TTE parameters (see Appendix A Figure A.2.1). The population component described: (i) the probability that each chromosome (maternal and paternal chromosome) of a sampled individual in the genomic region flanking the causal SNP belongs to one of two ancestral populations; and (ii) the RAF of the causal SNP in each ancestral population. The genetic component described the log HR of genotypes at the causal SNP, under an additive model in the number of risk alleles, which is assumed to be homogenous across ancestries. Finally, the TTE component described: (i) the TTE model (Cox PH or Weibull); and (ii) the baseline hazard (discussed further in section 2.2.2). As censored observations are not expected to contribute significantly to the level of statistical power, in all simulations, the baseline hazard rate was selected to achieve approximately 5% right-censored observations (i.e. not affected by the disease at the end of the study). The setting for right censoring also allowed censoring due to dropout.

Initially, it was assumed that testing for association of AOO was with the causal SNP. However, in practice, the causal SNP might not be directly tested in the GWAS analysis. To reflect the more common occurrence, in practice [166], testing of a bi-allelic tag SNP in LD with the causal SNP was also considered, parameterized in terms of the squared correlation coefficient (r^2) between them, allowing for the fact that the structure of LD varies between ancestries. The values considered covered the full spectrum ranging from $r^2 = 0$ (SNPs are in complete linkage equilibrium) to $r^2 = 1$ (SNPs are in complete linkage disequilibrium). The specific levels of LD

values examined based on r^2 included 0, 0.05, 0.15, 0.25, 0.5, 0.75, 0.85, 0.95, and 1. It was assumed that local ancestry at the tag SNP was the same as at the causal SNP, and that allele frequencies of the tag SNP and causal SNP were the same within an ancestry, to reduce the space of parameters investigated in the study. The complete list of parameters used in the simulation models is outlined in Table 2.1.

Table 2. 1 - Description of time-to-event models and admixture simulation components

Model	Weibull model	Cox PH model
(1) Single SNP model	$h(t) = \lambda \nu t^{\nu-1} \exp^{(\beta_1 * X_1)}$	$h(t) = \lambda \exp^{(\beta_1 * X_1)}$
(2) Single SNP with ancestry as a covariate	$h(t) = \lambda \nu t^{\nu-1} \exp^{((\beta_1 * X_1) + (\beta_2 * X_2))}$	$h(t) = \lambda \exp^{((\beta_1 * X_1) + (\beta_2 * X_2))}$
Description of parameters in relation to TTE models		
Model parameters: λ = Scale parameter ν = Shape parameter β= Log-hazard ratio due to covariates X X = covariate (s) X_1 Genotype of causal SNP and X_2 Causal SNP locus ancestry in admixed individuals		
Coding of model covariate(s): Genotype of causal SNP which is coded AA=0; Aa =1; aa=2 Ancestry of chromosomes of individuals at a specified locus which is based on the number of alleles originating from ancestry 1 coded; zero copies of ancestry 1 allele=0; one copy of ancestry 1 allele=1; two copy of ancestry 1 alleles=2 Genotype of Tag SNP which is coded BB=0; Bb =1; bb=2		
Simulation set values in relation to TTE component		
Weibull Model $\lambda = \nu t^{\nu-1}$ = Baseline hazard rate ν (value: 2 representing increasing HR **β1 = Log-hazard ratio due to Genotype of causal SNP (common diseases characterized by small to moderate SNP effect sizes)** Range of values: 0 - 0.0875 (0,0.0125,0.025,0.0375,0.05,0.0625,0.075,0.0875) **Cox PH Model** λ = Baseline hazard rate **ν** (value: 1; representing constant HR) **β1 = Log-hazard ratio due to Genotype of causal SNP (common diseases characterized by small to moderate SNP effect sizes)** Range of values: 0 - 0.175 (0,0.025,0.05,0.075,0.10,0.125,0.15,0.175)		
Simulation set values in relation to population and genetic component		
Overall admixed population size (1,000) Number of ancestral populations (2) **Values for ancestry proportion** ancestry 1 (0.1, 0.3, 0.5); ancestry 2 (0.9, 0.7, 0.5) **Values for RAF for causal SNP and tag SNP** ancestry 1 (0.1, 0.2, 0.3); ancestry 2 (0.5, 0.5, 0.5) **Values for level of LD between causal SNP and tag SNP** Based on r^2 (0, 0.05, 0.15, 0.25, 0.5, 0.75, 0.85, 0.95, 1)		

2.2.2. | Simulation models

As part of the simulation study, two TTE models were considered based on the two-parameter Weibull distribution: (i) the general Weibull model; and (ii) the Cox PH model. In the general Weibull model, the hazard function at time t is given by Equation 1.8 (described in section 1.7.2). Simulations were based on shape parameter (ν) values of 2 representing an increasing hazard rate. The independent variable, a simulated single causal SNP, was included as a genotype additive model (coded in the number of risk alleles carried (0, 1 or 2). Additionally, the values for (β), log HR of the risk allele consisted of values in the range 0 - 0.0875 (0, 0.0125, 0.025, 0.0375, 0.05, 0.0625, 0.075, 0.0875) for AOO simulated under the general Weibull model and values in the range of 0 - 0.175 (0, 0.025, 0.05, 0.075, 0.10, 0.125, 0.15, 0.175) for AOO simulated under the Cox PH model (Table 2.1). In the simulations based on the Cox PH model (described in section 1.7.2), the hazard function is given by Equation 1.9. The Cox PH model is a special case for the general Weibull model for which the hazard rate is constant over time (i.e. ν=1).

2.2.3. | Simulation process

The simulations, which were implemented using the R programming language, version 3.3.4 [167, 168], were performed in three steps for each replicate of data, given the three components of the simulation scenario: population, genetic and TTE. Each replicate of data comprised 1,000 individuals (description of the three steps are outlined in Figure 2.1).

In the first step of the admixed population simulations the ancestral origin of each of an individual's two chromosomes (maternal and paternal chromosome) at the causal SNP were simulated, independently, under a binomial distribution (assuming two ancestral populations), given the specified relative frequencies of each ancestry in the population (description of the three steps are outlined in Figure 2.1).

In the second step of the admixed population simulations, conditional on the ancestry of each of the two chromosomes, their associated allele at the causal SNP was simulated under a binomial distribution, given the specified RAF in each ancestral population, and under the assumption of Hardy-Weinberg equilibrium (HWE). The two alleles pertaining to the two

chromosomes of individuals in the sample were then merged to form the genotype for the causal SNP.

In the third and last step, conditional on the genotype of the individual at the causal SNP, the AOO of the disease was simulated under either the general Weibull model or Cox PH model, given the specified log HR of the risk allele.

For simulations including a tag SNP, genotypes at the tag SNP were simulated by first calculating haplotype frequencies across the causal SNP and tag SNP in each ancestral population for the admixed population simulations, according to the r^2. The scope of the scenarios allowed for the possibility of different LD between the causal SNP and tag SNP in the different ancestral populations but assumed the allele frequency of the two SNPs to be the same to reduce the simulation parameter space.

In the context of an admixed population, alleles at the tag SNP were simulated under a binomial distribution for each chromosome separately, given the haplotype frequencies in the ancestral populations and conditional on the ancestry of the individual at the causal SNP. HWE (see Appendix A Table A.1.1 -A.1.4) was assumed at the tag SNP.

2.2.3.1. |Simulating genotype data in an admixed population

Generating ancestry of chromosomes: The admixed population simulations was centred around simulating the local ancestry of a causal SNP in admixed individuals assumed to originate from two ancestral populations. The process begun by randomly generating 1,000 datasets consisting of 1,000 individuals of mixed ancestry. The ancestry of the maternal and paternal chromosomes was simulated independently based on the binomial distribution. The R syntax used to generate the ancestry of each chromosome is outlined in Appendix D.1.1 and D.1.2. The key steps involved in generating the ancestry of the chromosomes include; (1) specification of the ancestry proportion for each ancestral population; (2) simulation of the maternal chromosome ancestry based on the binomial distribution for an admixed population founded on two ancestral populations (facilitated by the R rbinom() function); (3) simulation of the paternal chromosome ancestry based on the binomial distribution for an admixed population founded on two ancestral populations; (4) the ancestry of the maternal and paternal

chromosomes columns were merged which gives the number of chromosomes originating from each ancestry.

Generating genotype of causal SNP: The allele at the causal SNP for each chromosome was simulated based on the RAF specific to its ancestral population. The R syntax used to generate the allele of each chromosome is outlined in Appendix D.2.1. The main steps entailed; (1) specification of the RAFs associated with each ancestral population; (2) simulation of the allele associated with the maternal chromosome based on the binomial distribution given its ancestry at the causal SNP (facilitated by the R rbinom() function); (3) simulation of the allele associated with the paternal chromosome based on the binomial distribution given its ancestry at the causal SNP; (4) the allele of the maternal and paternal chromosomes columns were merged to create the genotype of the causal SNP column.

Generating genotype of tag SNP: For those simulations that considered the situation where the causal SNP was not directly genotyped, the genotype of the tag SNP correlated with the causal SNP was also simulated. The alleles of the tag SNP were simulated to reflect a range of LD (measured using the r^2) levels between zero and one (Appendix D.3.1). The process involved; (1) calculation of the allele frequencies of the tag SNP based on the haplotype frequencies derived from causal SNP allele frequencies and specified LD values; (2) simulation of the tag SNP maternal chromosome allele in each ancestral population based on the binomial distribution (facilitated by the R rbinom() function); and (3) simulation of the tag SNP paternal chromosome allele in each ancestral population based on the binomial distribution. (4) the alleles of the maternal and paternal chromosomes columns were merged to create the genotype of the tag SNP column.

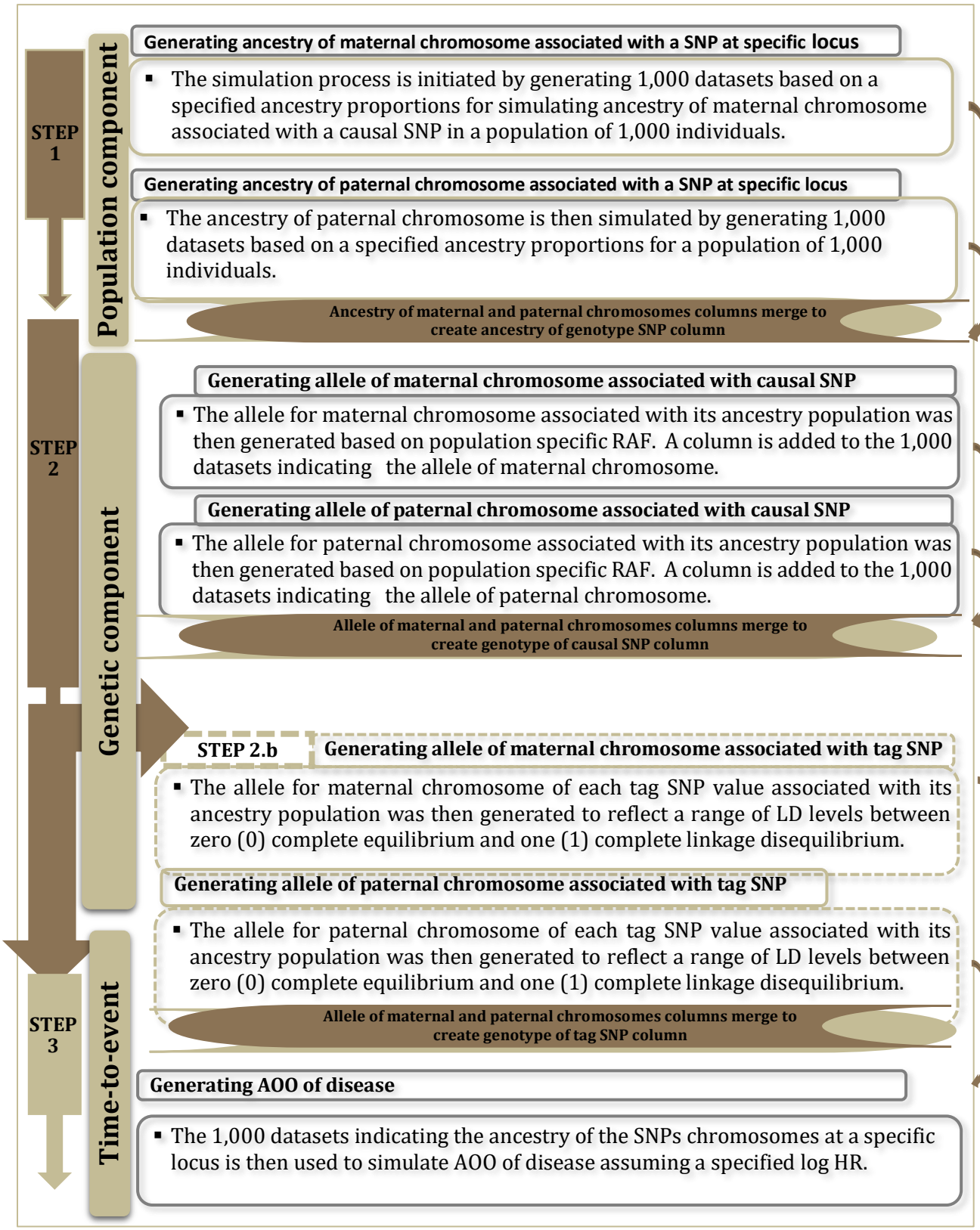


Figure 2. 1 - Description of data generating process in an admixed population

2.2.3.2. | Simulating AOO of disease conditional on the causal SNP genotype

Following the simulation of the ancestry of the maternal and paternal chromosomes at the causal SNP and genotype of the causal SNP, the AOO of disease conditional on the causal SNP genotype under an additive model was simulated assuming a range of log hazard ratios (HRs) (Appendix D.4). The process entailed: (1) specification of the shape and scale parameter values; (2) specification of the log HRs; (3) causal SNP entered into the TTE model as a continuous explanatory variable indicative of the genetic additive model (AA genotype coded as 0; Aa=1; aa=2); (4) specification of the baseline hazard rate; and (5) specification of the censoring rate due to dropout based on the exponential distribution.

The settings used to simulate AOO in the admixed population simulations assuming a 5% right censoring rate under the Cox PH model has been outlined in Appendix D.4.1 and under the Weibull model Appendix D.4.2. The R rweibull() function with a shape parameter of one, where the HR was assumed constant was used to simulate AOO under the Cox PH model. The settings also allowed censoring due to dropout, which was facilitated by the R rexp() function.

2.2.4. | Association analysis

Analysis of each simulated data set was performed using both the Cox PH model and the Weibull model, irrespective of the TTE model used for simulation. The analysis considered the inclusion of the causal SNP genotype (or tag SNP genotype, as appropriate) as the independent variable in the regression model, with or without adjustment for ancestry. A description of the models fitted are outlined in Table 2.1. To assess the impact of the admixture mapping approach, analysis using ancestry as the single independent variable in the model was also considered. Here ancestry represented the ancestry of the causal SNPs locus. As noted from Chapter 1 admixture mapping is formulated on the basis of whether an individual at a specified locus has 0, 1, or 2 copies of a population specific allele [104]. Additionally, the false positive error rate and power to detect an association was assessed at a nominal 5% level of significance, given by the proportion of replicates for which the association p-value was less than 0.05.

The survival package [169]was used to perform both the Cox PH and Weibull TTE analysis via the coxph and survreg functions respectively. An excerpt of the R syntax used to undertake the Cox PH analysis is outlined in Appendix E.1, while R syntax used for the Weibull analysis is outlined in Appendix E.2.

2.3. | Results

This simulation study consisted of an admixed population consisting of two ancestral (parental) populations based on AOO of disease simulated under a Cox PH model or general Weibull model. The simulations focused primarily on evaluating the power to detect an association with AOO of disease and a causal SNP or tag SNP. The simulations first assessed the impact of the causal SNP HR on power, the scope of which incorporated two primary population specific characteristics; (1) RAF within each ancestral population; and (2) the ancestry proportion or percentage contribution from each ancestral population which comprise the admixed individual. Secondly, the simulations also assessed the impact of the ancestry proportions within the admixed populations in greater detail, where different levels of ancestry proportions and RAF were assessed. A more detailed evaluation was also conducted in relation to the impact of the relative RAF between ancestral populations within the admixed population, here RAF was fixed in one ancestral population while it was varied in the second ancestral population. Simulations were also undertaken that considered the impact of the LD between the causal SNP and a tested tag SNP, which covered both: (i) the situation where the level of LD between the causal SNP and tag SNP was the same in the ancestral populations; and (ii) the situation where LD levels between the causal SNP and tag SNP were different among the ancestral populations which formed the admixed population. Additionally, the two primary population specific characteristics described above, were also incorporated into the assessment (RAF and ancestry proportion). To assess the presence of inflation in the type I error rate, models of both the causal SNP and tag SNP adjusted by the locus specific ancestry of the causal SNP were also incorporated as part of the assessment.

2.3.1. | AOO simulated under Cox PH model

Impact of causal SNP HR on power: Figure 2.2 presents the power to detect association of the causal SNP with AOO of disease as a function of log HR, based on analyses with both the Cox PH and Weibull models. The nine plots present power across different parameter settings for the RAF in the two ancestral populations and their ancestry proportions within the admixed population. The analysis which illustrates the relative performance of the Cox PH and Weibull models indicated that there was no notable difference in power between the two analysis methods. Additionally, there was no indication that inflation in the type I error rate was an

issue. (Table 2.2). The mean type I error rate for the causal SNP based on the Cox PH model without and with adjustment for ancestry were 5.1% (CI: 4.4% - 5.7%) and 4.9% (CI: 4.4% - 5.3%) respectively.

Table 2. 2 - Type I error rate associated with causal SNP HR simulated under a Cox PH model in an admixed population

Model	log HR	Ancestry proportion (P1=0.1,P2=0.9)	Ancestry proportion (P1=0.3,P2=0.7)	Ancestry proportion (P1=0.5,P2=0.5)
RAF (P1=0.1, P2=0.5)				
Causal SNP Cox PH	0	4.6%	5.9%	5.8%
Causal SNP + ancestry Cox PH	0	4.4%	5.2%	4.9%
Causal SNP Weibull	0	4.6%	6.0%	6.1%
Causal SNP + ancestry Weibull	0	4.3%	5.2%	4.9%
RAF (P1=0.2, P2=0.5)				
Causal SNP Cox PH	0	3.7%	5.2%	6.0%
Causal SNP + ancestry Cox PH	0	4.3%	4.6%	5.8%
Causal SNP Weibull	0	3.8%	5.1%	6.2%
Causal SNP + ancestry Weibull	0	4.6%	4.5%	6.1%
RAF (P1=0.3, P2=0.5)				
Causal SNP Cox PH	0	4.2%	5.9%	4.3%
Causal SNP + ancestry Cox PH	0	4.5%	5.8%	4.5%
Causal SNP Weibull	0	4.5%	5.8%	4.8%
Causal SNP + ancestry Weibull	0	4.5%	5.7%	4.4%
Summary				
	MEAN	**SE**	**Lower 95% CI**	**Upper 95% CI**
Causal SNP Cox PH	5.1%	0.3%	4.4%	5.7%
Causal SNP + ancestry Cox PH	4.9%	0.2%	4.4%	5.3%
Causal SNP Weibull	5.2%	0.3%	4.6%	5.9%
Causal SNP + ancestry Weibull	4.9%	0.2%	4.4%	5.4%

Descriptions*: Log HR: log hazard ratio; SE: standard error; CI: confidence interval*

In general, there was a small reduction in power to detect association of the causal SNP with AOO with adjustment for ancestry, which indicated that the association of AOO was described by the causal SNP genotypes, independently of ancestry. The small reduction in power was most apparent in situations of equal or near equal admixture (i.e. equal ancestral proportions) and marked differences in RAF as illustrated in Figure 2.2 Plot 4 and Plot 7.

As anticipated, for TTE data simulated under a Cox PH model, analysis with the Cox PH model with the causal SNP genotype as the independent variable was more powerful for detecting association with AOO of disease than using locus specific ancestry of the causal SNP as the independent variable in the model. In this approach admixed individuals were assumed for example, at a specified causal locus to have 0, 1, or 2 copies of an allele which originated from ancestral population 1.

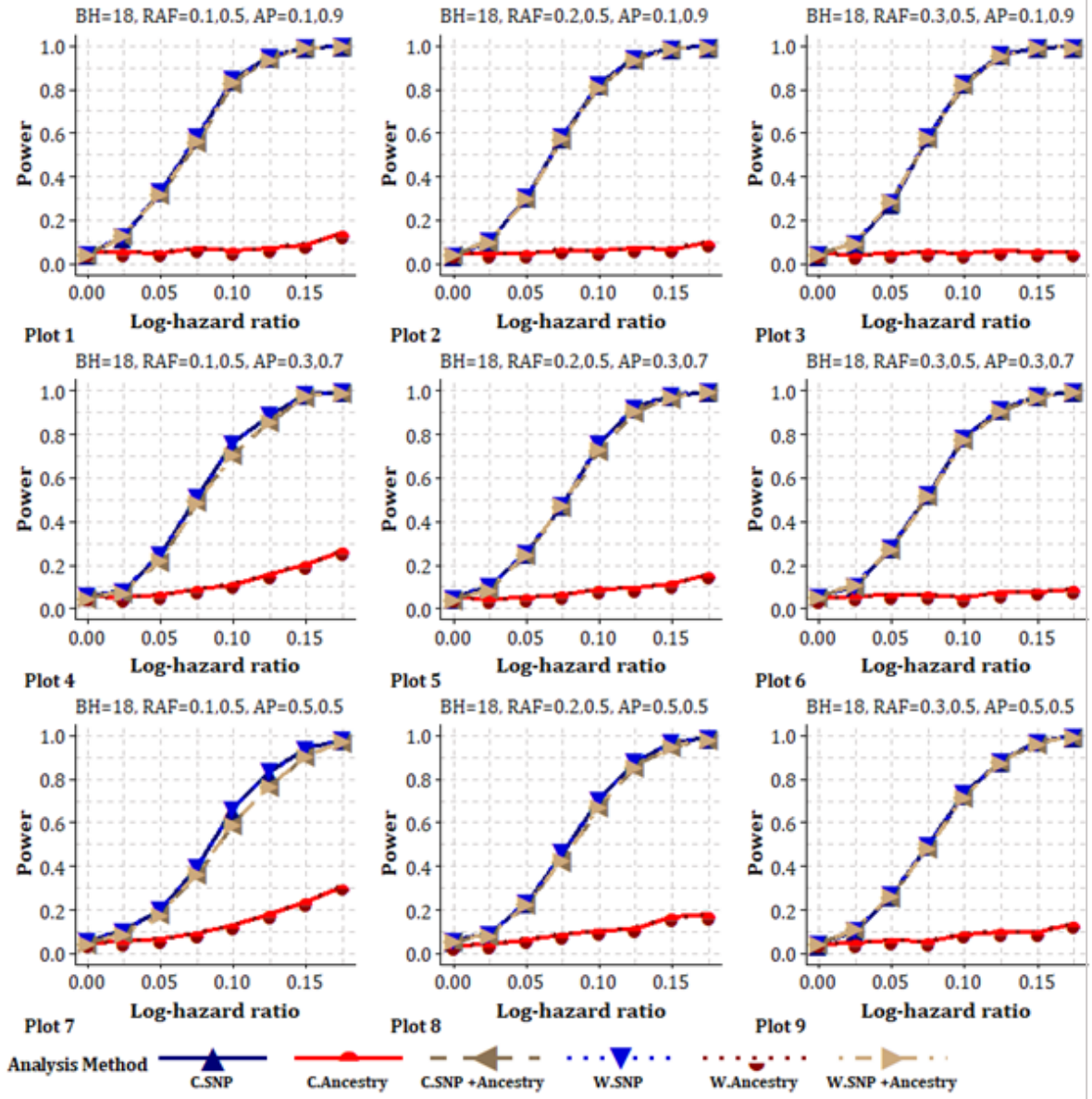


Figure 2. 2 - Power to detect association of a causal SNP with AOO of disease (simulated under a Cox PH model) as a function of log HR assuming admixed population originating from two ancestries

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and log HR on the x axis for each TTE model analysed. Cox PH model with SNP as the single explanatory variable (navy blue); Weibull model with SNP as the single explanatory variable (blue); Cox PH model with ancestry as the single explanatory variable (red); Weibull model with ancestry as the single explanatory variable (dark red); Cox PH model with SNP as explanatory variable and ancestry as covariate (burlywood brown); Weibull model with SNP as explanatory variable and ancestry as covariate (light burlywood brown).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: ancestry proportion; C.SNP: Cox PH model with SNP variable; C.Ancestry: Cox PH model with ancestry variable; C.SNP +Ancestry: Cox PH model with SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.Ancestry: Weibull model with ancestry variable; W.SNP +Ancestry: Weibull model with SNP variable and ancestry covariate.

In the models based on locus-specific ancestry the power to detect an association with AOO of disease was influenced by the relative difference in RAF between the ancestral populations and the relative ancestry proportions within the admixed population. The more similar the relative ancestry proportions in the admixed population power to detect an association was increased. This is because greater levels of admixture between the two mixing ancestral populations maximizes information pertaining to ancestry and disease risk, therefore power is positively impacted. Additionally, the wider the gap in terms of the difference in RAF between the ancestries power was also increased. With RAF that are distinguishable between the two ancestral populations aids the locus-specific ancestry-based model to detect an association with AOO of disease.

Impact of ancestry proportion on power: Figure 2.3 presents the power to detect association of the causal SNP with AOO of disease as a function of ancestry proportion. The four plots present power across different parameter settings for the RAF in the two ancestral populations. The simulations assumed a log HR of 0.1 (HR=1.11) for the risk allele at the causal SNP. It was noted that with a combination of equal admixture and a marked difference in RAF between ancestries, the difference in power to detect association of the causal SNP with AOO of disease between the unadjusted casual SNP model and the causal SNP model adjusted for ancestry was most apparent as illustrated in Figure 2.3 Plot 1(RAF: 10% compared to 50%). Furthermore, observations from the locus-specific ancestry-based model seemed to indicate that maximal power maybe attained when there is a marked difference in RAF between ancestries coupled with equal or near equal admixture (i.e. equal ancestral proportions). However, with an assumed log HR of 0.1 for the risk allele of the causal SNP, power was relatively low when compared to the models based on the causal SNP, with or without adjustment for ancestry.

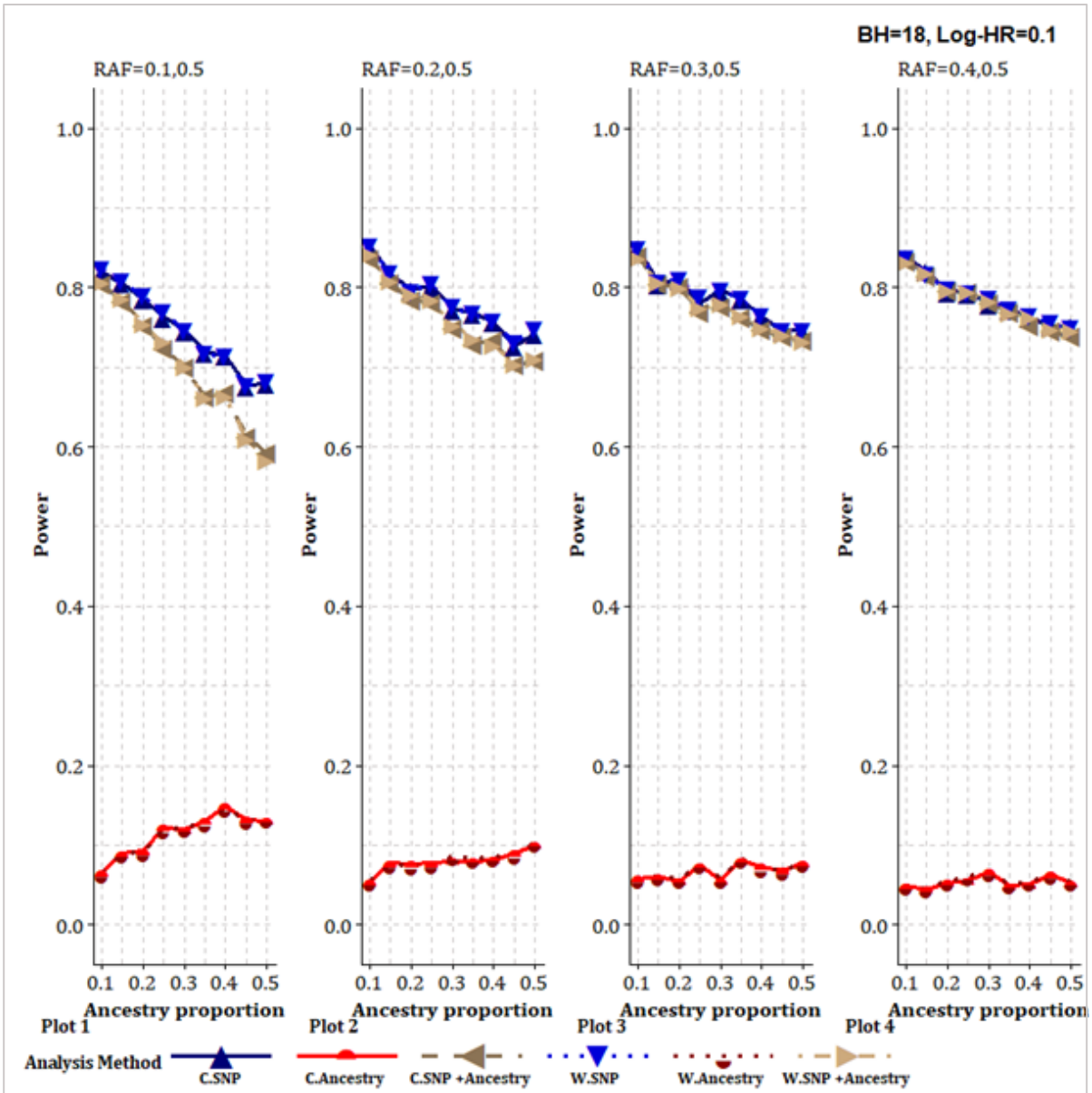


Figure 2. 3 - Effect of ancestry proportion on power to detect an association with AOO of disease (simulated under a Cox PH model) assuming admixed population originating from two ancestries

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and log HR on the x axis for each TTE model analysed. Cox PH model with SNP as the single explanatory variable (navy blue); Weibull model with SNP as the single explanatory variable (blue); Cox PH model with ancestry as the single explanatory variable (red); Weibull model with ancestry as the single explanatory variable (dark red); Cox PH model with SNP as explanatory variable and ancestry as covariate (burlywood brown); Weibull model with SNP as explanatory variable and ancestry as covariate (light burlywood brown).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: ancestry proportion; C.SNP: Cox PH model with SNP variable; C.Ancestry: Cox PH model with ancestry variable; C.SNP +Ancestry: Cox PH model with SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.Ancestry: Weibull model with ancestry variable; W.SNP +Ancestry: Weibull model with SNP variable and ancestry covariate.

Impact of RAF on power: Figure 2.4 presents the power to detect association of the causal SNP with AOO of disease as a function of ancestry specific RAF. The simulations assumed a log HR of 0.1 (HR=1.11) for the risk allele at the causal SNP. Additionally, the ancestry proportions were assumed to be 0.5 in both ancestral populations. The individual nine plots illustrate the impact on power to detect association with AOO of disease when the levels of RAF vary between the ancestral populations. In each scenario, RAF was varied in one ancestral population while it was held fixed in the other ancestral population. Evaluation of the role of ancestry specific RAF on power to detect association of the causal SNP indicates that maximal power, as expected, was attained when the risk allele has frequency of 50% in both ancestries. However, reduction in power was most apparent when there are marked differences in ancestry specific RAF as illustrated in Figure 2.4 Plot 1 (where RAF is 0.1 in ancestry 2) and Plot 9 (where RAF is 0.9 in ancestry 2).

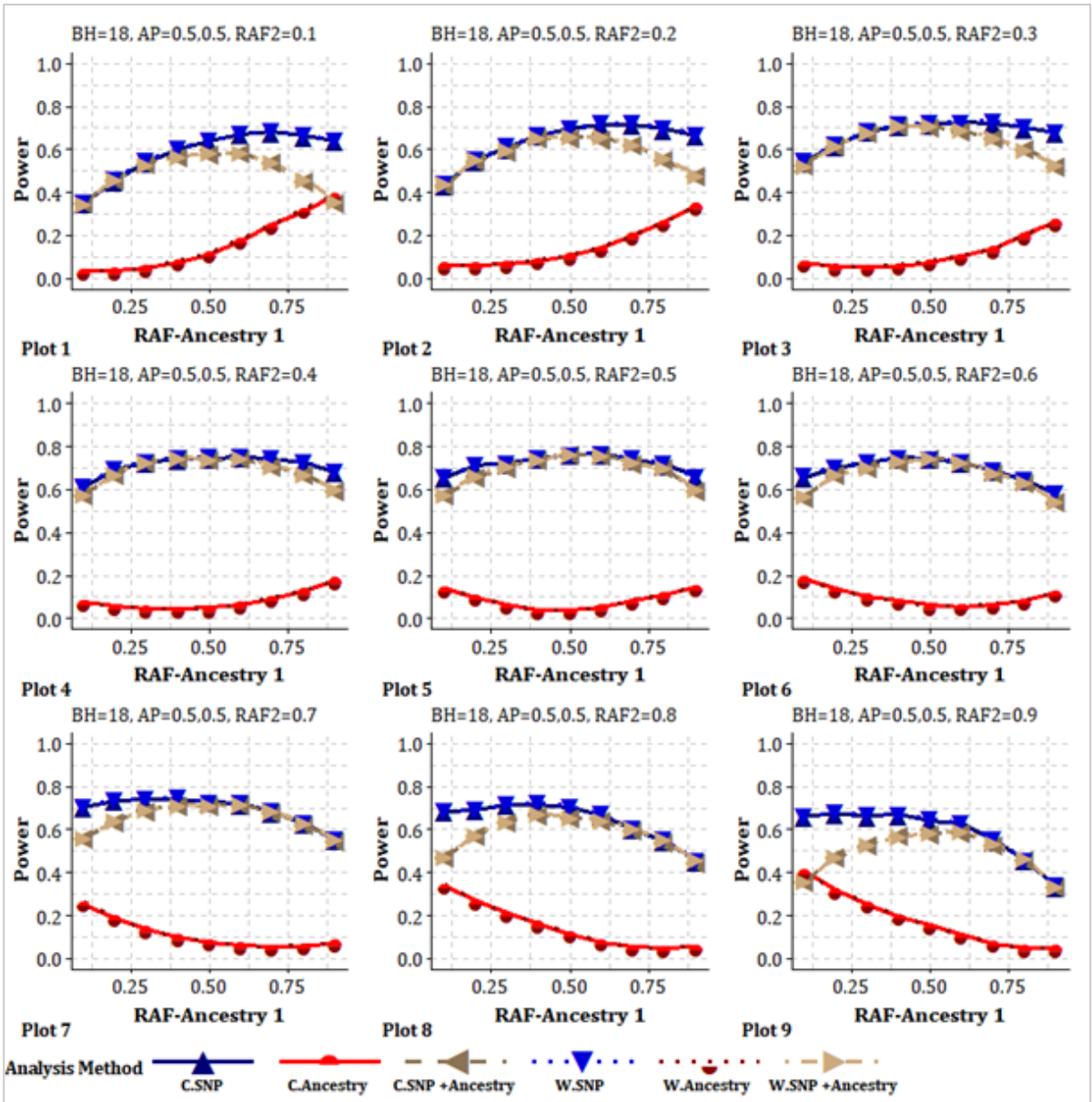


Figure 2. 4 - Effect of ancestry RAF on power to detect an association with AOO of disease (simulated under a Cox PH model assuming a log HR of 0.1 and an AP of 0.5 in both ancestral populations)

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and RAF in ancestry 1 on the x axis for each TTE model analysed. Cox PH model with SNP as the single explanatory variable (navy blue); Weibull model with SNP as the single explanatory variable (blue); Cox PH model with ancestry as the single explanatory variable (red); Weibull model with ancestry as the single explanatory variable (dark red); Cox PH model with SNP as explanatory variable and ancestry as covariate (burlywood brown); Weibull model with SNP as explanatory variable and ancestry as covariate (light burlywood brown).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: ancestry proportion; C.SNP: Cox PH model with SNP variable; C.Ancestry: Cox PH model with ancestry variable; C.SNP +Ancestry: Cox PH model with SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.Ancestry: Weibull model with ancestry variable; W.SNP +Ancestry: Weibull model with SNP variable and ancestry covariate.

Impact of LD with a tag SNP on power: Illustrated in Figure 2.5 is the power to detect association of AOO of disease with a tested tag SNP as a function of LD with the causal SNP within an admixed population. Simulations assumed a log HR of 0.5 (HR=1.65) for the risk allele at the causal SNP. The nine plots present power across different parameter settings for RAF in the two ancestral populations and the proportions of the two ancestral population within the admixed population. First, the scenario in which LD between the causal SNP and tag SNP was assumed to be the same in both ancestral populations was considered. Figure 2.5 illustrates the power of tag SNP association models, with and without adjustment for ancestry. The effect of confounding due to ancestry was particularly evident when the RAF was substantially different in the two ancestral populations (10% compared to 50%) (see Figure 2.5 plot 7 and plot 4) and the ancestry proportions were the same or similar within the admixed population. Furthermore, the effects of confounding are of particular concern when the ancestry proportions are the same within the admixed population. When LD was zero and the ancestry proportion was 50% for both ancestral populations power based on the Cox PH model was 22%, 11.3% and 8.4% for RAF 0.1,0.5; 0.2,0.5; and 0.3,0.5 respectively. However, by including ancestry as a covariate, in addition to the tag SNP reduces the type I error to the correct levels (5% power expected).

After adjustment for the effects of ancestry, the results seem to indicate that for the tag SNP to be comparable with the causal SNP, higher levels of LD are required between the causal SNP and tag SNP in both ancestral populations, when the ancestry proportions are the same for each ancestral population. For instance, when there was a difference in the relative ancestry proportions in the admixed population (10% compared to 90%; 30% compared to 70%), LD of at least r^2 0.5 was required to achieve the same power for the analysis of the tag SNP to be comparable with the causal SNP. However, when the ancestry proportion was 50% for both ancestral populations LD of at least r^2 0.75 was required for the tag SNP model to be comparable with the causal SNP.

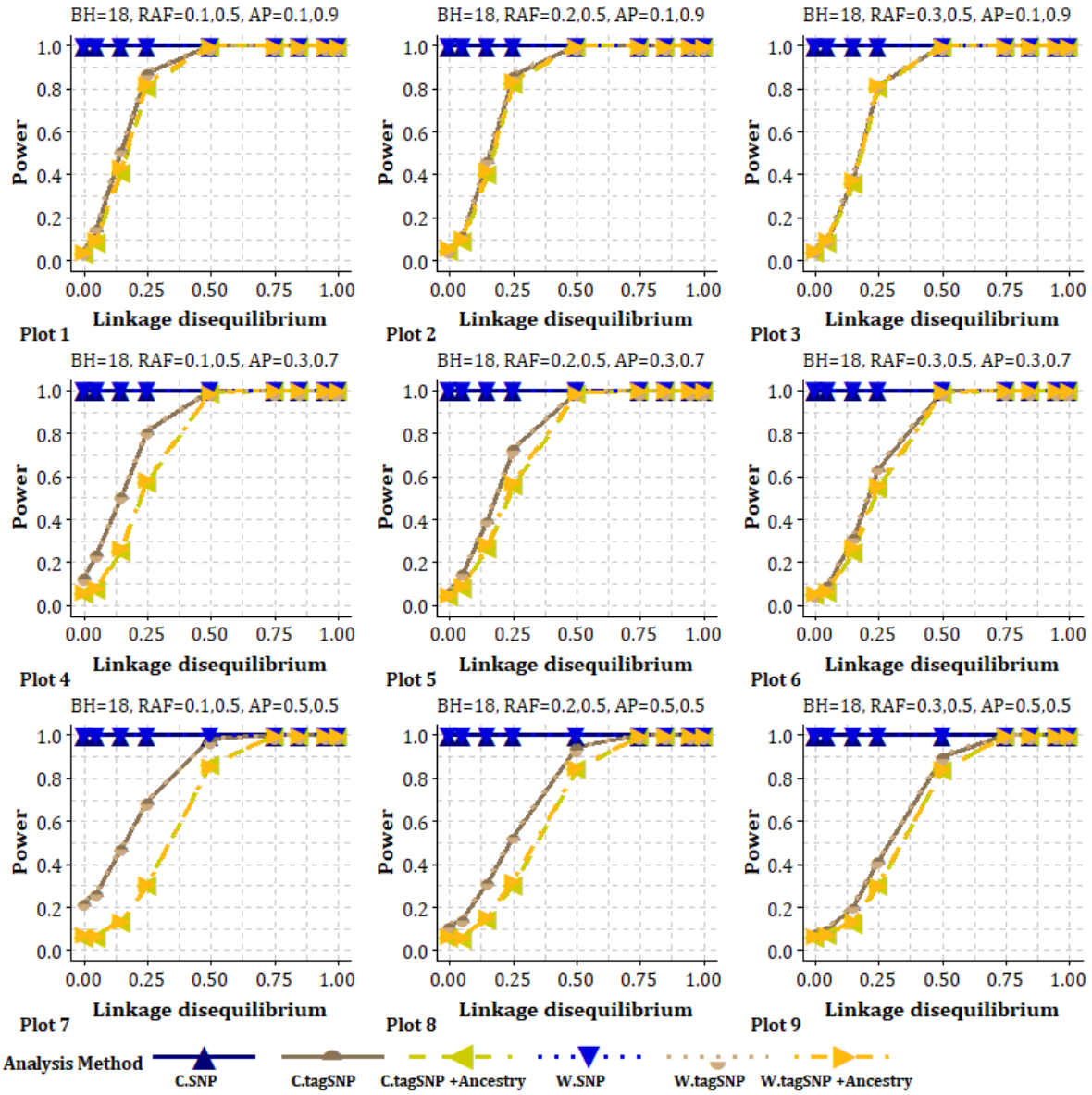


Figure 2. 5 - Effect of LD on power to detect an association with AOO of disease assuming levels of LD between tag SNP and causal SNP are the same in the ancestral populations and a log HR of 0.5

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and linkage disequilibrium in both ancestries on the x axis, for each TTE model analysed. Cox PH model with causal SNP as the single explanatory variable (navy blue); Weibull model with causal SNP as the single explanatory variable (blue); Cox PH model with tag SNP as the single explanatory variable (burlywood brown); Weibull model with tag SNP as the single explanatory variable (light burlywood brown); Cox PH model with tag SNP as explanatory variable and ancestry as covariate (yellow green); Weibull model with tag SNP as explanatory variable and ancestry as covariate (gold).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: Ancestry proportion; C.SNP: Cox PH model with SNP variable; C.tagSNP: Cox PH model with tag SNP variable; C.tagSNP +Ancestry: Cox PH model with tag SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.tagSNP: Weibull model with tag SNP variable; W.tagSNP +Ancestry: Weibull model with tag SNP variable and ancestry covariate.

Next, a more practical setting in which the level of LD between the causal SNP and a tag SNP were different within the two ancestral populations were also considered. For this scenario, the power to detect association of AOO of disease with the tag SNP is illustrated in Figure 2.6. Here a log HR of 0.5 (HR=1.65) for the risk allele at the causal SNP was assumed and is illustrated as a function of LD with the causal SNP, which was assumed to be different within the two ancestral populations. In Figure 2.6 the nine plots present power across different parameter settings for LD within the ancestral populations. In each scenario LD was varied in one ancestral population while it was held fixed in the other ancestral population. Additionally, the parameter settings in relation to the RAF and the ancestry proportion within the admixed population were fixed at (0.1,0.5) and (0.5,0.5) respectively.

Inflated type I error rates due to admixture evident as the unadjusted model with tag SNP showed power around 20% even when LD in both populations was zero (Figure 2.6). However, the expected levels were observed when ancestry was included in the model in addition to the tag SNP.

These results further illustrate that higher levels of LD are required between the causal SNP and tag SNP when the ancestry proportions are the same for each ancestral population. However, it appears that power is not seriously impacted if the LD between the causal SNP and the tag SNP is not high in one of the two ancestral populations. For instance when the contribution from both populations was the same (50%) and there was a gap in terms of RAF between the populations (0.1 versus 0.5) the adjusted tag SNP model was comparable with the causal SNP in terms of power for all levels of LD in the first population when LD in the second was 0.75.

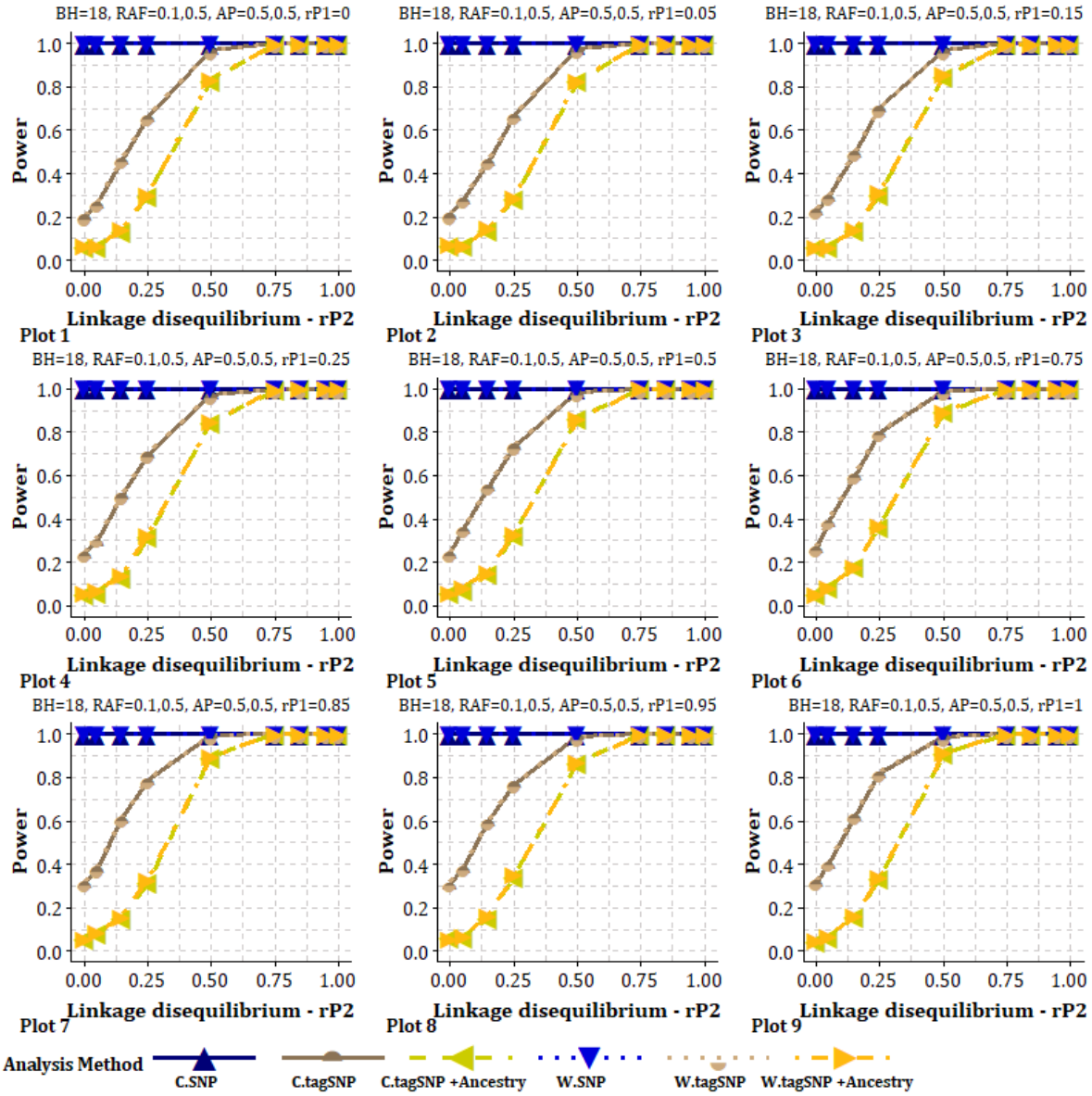


Figure 2. 6 - Effect of LD on power to detect an association with AOO of disease assuming levels of LD between tag SNP and causal SNP are different among ancestral populations and a log HR of 0.5

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and linkage disequilibrium in ancestry 2 (rP2) on the x axis, for each TTE model analysed. Cox PH model with causal SNP as the single explanatory variable (navy blue); Weibull model with causal SNP as the single explanatory variable (blue); Cox PH model with tag SNP as the single explanatory variable (burlywood brown); Weibull model with tag SNP as the single explanatory variable (light burlywood brown); Cox PH model with tag SNP as explanatory variable and ancestry as covariate (yellow green); Weibull model with tag SNP as explanatory variable and ancestry as covariate (gold).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: Ancestry proportion; rP1: linkage disequilibrium in ancestral population 1; rP2: linkage disequilibrium in ancestral population 2; C.SNP: Cox PH model with SNP variable; C.tagSNP: Cox PH model with tag SNP variable; C.tagSNP +Ancestry: Cox PH model with tag SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.tagSNP: Weibull model with tag SNP variable; W.tagSNP +Ancestry: Weibull model with tag SNP variable and ancestry covariate.

The relative performance between the tag SNP and ancestry models were also compared. Findings suggested that ancestry-based models may be useful in situations where LD between the causal SNP and the tag SNP is 0.5 or less. This is particularly so in cases where there is a marked difference in terms of RAF between the two ancestral populations and contribution from both populations is high (30%) or the same (50%). For example, in a situation where the RAF was markedly different between population (10% compared to 50%) and contribution from both populations was high (30%) or the same (50%) modelling with ancestry was more powerful than models based on the tag SNP when LD was 0.4 or less (Appendix A Figure A.3.1 plot 4 and 7). Similarly, when RAF was again markedly different between the populations and contribution from both ancestral populations was 50% the ancestry-based model was more powerful than models based on the tag SNP when LD was 0.5 or less. These findings were not seriously affected if LD levels between the causal SNP and the tag SNP varied in only one of the ancestral populations (Appendix A Figure A.3.2).

2.3.2. | AOO simulated under the Weibull model

Impact of causal SNP HR on power: Illustrated in Figure 2.7 are the power to detect association of the causal SNP with AOO of disease as a function of log HR simulated under the Weibull model assuming an increasing hazard rate with a shape parameter value of 2. The nine plots illustrate power across different parameter settings for the RAF in the two ancestral populations and the ancestry proportion within the admixed population. In terms of power, it was noted that under these conditions no notable difference in performance between the Cox PH and Weibull analysis models were observed. There was also no indication of a problem with inflated type I error rates (Table 2.3). The mean type I error rate for the causal SNP based on the Cox PH and Weibull models without adjustment for ancestry were 4.7% (CI: 4.1% - 5.4%) and 4.8% (CI: 4.2% - 5.4%) respectively.

Table 2. 3 - Type I error rate associated with causal SNP HR simulated under a Weibull model in an admixed population

Model	log HR	Ancestry proportion (P1=0.1,P2=0.9)	Ancestry proportion (P1=0.3,P2=0.7)	Ancestry proportion (P1=0.5,P2=0.5)
RAF (P1=0.1, P2=0.5)				
Causal SNP Cox PH	0	5.5%	3.6%	4.5%
Causal SNP + ancestry Cox PH	0	5.6%	4.6%	4.0%
Causal SNP Weibull	0	5.5%	3.8%	4.6%
Causal SNP + ancestry Weibull	0	5.7%	4.4%	4.3%
RAF (P1=0.2, P2=0.5)				
Causal SNP Cox PH	0	3.5%	5.5%	6.0%
Causal SNP + ancestry Cox PH	0	3.6%	5.9%	5.6%
Causal SNP Weibull	0	3.7%	5.8%	5.9%
Causal SNP + ancestry Weibull	0	3.7%	6.0%	5.5%
RAF (P1=0.3, P2=0.5)				
Causal SNP Cox PH	0	4.5%	5.0%	4.5%
Causal SNP + ancestry Cox PH	0	4.7%	4.7%	4.5%
Causal SNP Weibull	0	4.6%	4.8%	4.5%
Causal SNP + ancestry Weibull	0	5.0%	4.7%	4.6%
Summary				
	MEAN	**SE**	**Lower 95% CI**	**Upper 95% CI**
Causal SNP Cox PH	4.7%	0.3%	4.1%	5.4%
Causal SNP + ancestry Cox PH	4.8%	0.3%	4.2%	5.4%
Causal SNP Weibull	4.8%	0.3%	4.2%	5.4%
Causal SNP + ancestry Weibull	4.9%	0.2%	4.3%	5.4%

***Descriptions**: Log HR: log hazard ratio; SE: standard error; CI: confidence interval*

A small reduction in power to detect association of the causal SNP with AOO with adjustment for ancestry was observed over the range of scenarios considered. This is an indication that the association of AOO is described by the causal SNP genotypes, independently of ancestry. As with AOO simulated under Cox PH model the small reduction in power was most apparent in situations of equal or near equal admixture and marked differences in RAF as illustrated in Figure 2.7 Plot 4 and Plot 7.

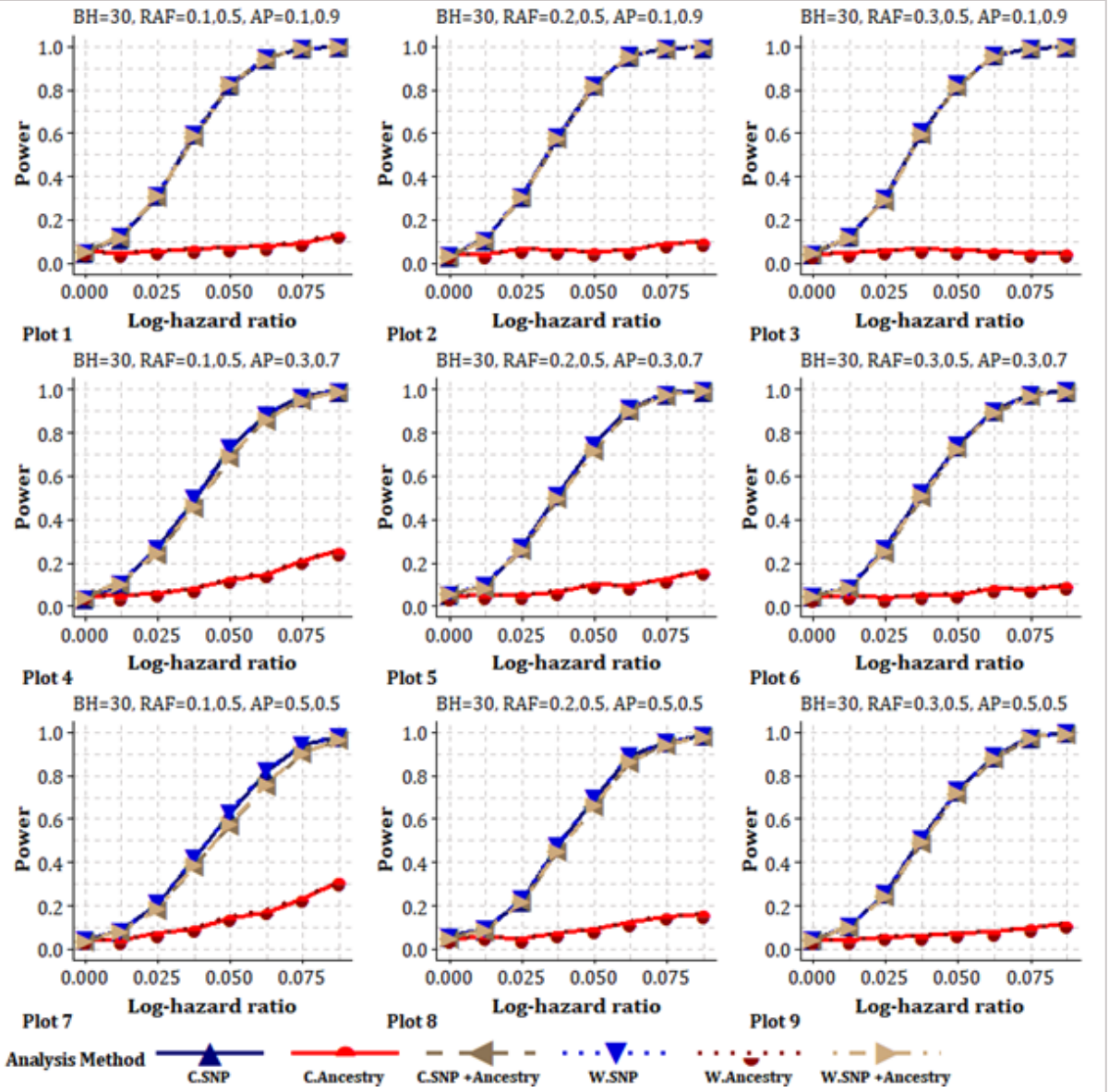


Figure 2. 7 - Power to detect association of a causal SNP with AOO of disease (simulated under a Weibull model) as a function of log HR assuming a shape parameter of 2 admixed population originating from two ancestries

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and log HR on the x axis for each TTE model analysed. Cox PH model with SNP as the single explanatory variable (navy blue); Weibull model with SNP as the single explanatory variable (blue); Cox PH model with ancestry as the single explanatory variable (red); Weibull model with ancestry as the single explanatory variable (dark red); Cox PH model with SNP as explanatory variable and ancestry as covariate (burlywood brown); Weibull model with SNP as explanatory variable and ancestry as covariate (light burlywood brown).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: ancestry proportion; C.SNP: Cox PH model with SNP variable; C.Ancestry: Cox PH model with ancestry variable; C.SNP +Ancestry: Cox PH model with SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.Ancestry: Weibull model with ancestry variable; W.SNP +Ancestry: Weibull model with SNP variable and ancestry covariate.

2.4. | Discussion

GWAS has been established as the standard approach for genetic association analysis of common complex diseases in humans. However, the threat of issues relating to the presence of population structure or admixture adds to the complexities of applying GWAS methodology. There are existing methods that can mitigate the potential impact of population structure, however, there remains many concerns particularly regarding the inflation of the false positive error rate in the presence of population structure. Simulation studies were undertaken in an effort to identify the most robust and powerful approaches for conducting AOO of disease GWAS in ancestrally diverse populations within a TTE framework.

Results from the study in relation to simulations under a Cox PH model highlighted that power under the general Weibull model is largely consistent with that of the Cox PH model, despite the additional degree of freedom required for the shape parameter. Two studies [170, 171] comparing the relative performance of the Cox PH and the Weibull regression model have indicated that there are situations where the Cox PH and the Weibull model give similar results. In these studies, data was simulated from a parametric Weibull model with an assumed shape parameter greater than 1. Both simulation studies demonstrated similar performance for testing association with a predictor under a Weibull model with unknown shape parameter and the Cox PH model. It was also noted that the shape parameter of the Weibull model does not impact the performance of the Cox PH model.

Findings from the simulation study presented here also highlighted issues relating to confounding due to ancestry. Inflated type I error rates in testing for association of AOO of the disease with a tag SNP because of confounding with ancestry was observed. The effects of confounding due to ancestry are most apparent when the RAF is markedly different between the ancestries and the proportions of the ancestries are the same (equal admixture).

To address confounding due to ancestry, ancestry was included as a covariate in the regression models. This removed the inflation in type I error rates, however, in practice ancestry might not be known, or difficult to define in the presence of population admixture [172]. However, in these circumstances, multivariate analyses of GWAS data can be used to infer “axes of genetic

variation" that can be included as covariates to account for confounding due to population structure [124]. In the case of admixed populations, local ancestry inference methods based on genotype data can be applied where the focus is on locus-specific ancestry [108, 173, 174].

Variation in the level of LD between ancestries has an impact on the power of a tag SNP. The presence of strong or high LD levels between a causal SNP and tag SNP among all ancestral populations positively impact the power of the tag SNP to detect an association with AOO of disease. The power of the tag SNP to detect an association with AOO disease may still be possible in the presence of low levels of LD within ancestral populations provided that high LD is present in at least one of the ancestral populations. However, the extent of the variation in power is influenced by the relative ancestral proportions within the population. Equal or near equal admixture positively influences power.

The main limiting factors pertaining to these simulations are that scenarios were limited primarily to two ancestral populations where an admixed population were considered independently. In reality, populations are more likely to be hierarchical, consisting of both discrete ancestry groups and admixed individuals. It was also assumed that the RAF of both the causal SNP and tag SNP was the same but in reality, this is not always the case.

In conclusion, the results of these simulations provide a resource for the development or improvement to guidelines for implementing the most powerful approaches, within admixed populations, to detect association with AOO of disease in a GWAS TTE framework, given a combination of population and genetic characteristics. These results highlight the importance of accounting for ancestry when assessing the association of AOO of disease with a tag SNP. Also provided is a more definitive understanding of the likely impact on the power of a tag SNP within admixed populations originating from ancestral populations with varying levels of LD.

Chapter 3: Investigating the utility of genetic risk scores to detect an association with age-of -onset of disease in European ancestry populations

Chapter Outline

This chapter focuses on investigating the utility of genetic risk scores (GRS) to detect an association with age-of-onset (AOO) of disease in ancestrally homogenous populations, which includes a component of both simulation and real data application. The first element entails the application of GRS to investigate the association of AOO of type 2 diabetes (T2D) in two independent genome-wide association studies (GWAS) originating from the Northwestern University Gene (NUgene) Banking Project and the Wellcome Trust Case Control Consortium (WTCCC). As part of the assessment the results of these two independent GWAS was also combined in a summary statistics meta-analysis. Additionally, a GRS AOO of disease simulation study to further assess the relative performance of the Cox proportional hazards (PH), proportional odds and binary logistic models was also undertaken. In the simulations, analyses concentrated primarily on evaluating the impact of censoring on the relative power between the three models. Data analysis in the time-to-event (TTE) framework consisting of both cases and controls, AOO is assessed at the end of the study period, where controls are censored at their current age. However, in the proportional odds model framework, AOO is viewed as an ordinal outcome which distinguishes between controls, late-age-onset (LAO) and early-age-onset (EAO), while within the binary logistic regression framework contrast is made between cases (irrespective of AOO) and controls.

3.1. | Introduction

GWAS have facilitated numerous discoveries in common disease biomedical research including diseases like cardiovascular diseases, psychiatric diseases, multiple sclerosis, various types of cancer and diabetes. However, the magnitude of common disease genetic effects has been characterized on many occasions as small to moderate with odds ratio (OR) ranging between 1.1 to 1.5 [175-179] and generally only explains a small proportion of the variance in disease risk. Heritability as described in section 1.6.1 is an indication as to the level of genetic contribution attributable to disease occurrence. Questions regarding unexplained heritability led to the application of GRS as it presented an avenue to simultaneously assess the overall

genomic risk of an individual associated with a disease or trait. It was further noted (section 1.6.3) that GRS have the potential to identify individuals at risk of early age-onset disease because they are expected to have a greater genetic burden. This is because the expectation is that individuals who are affected by the disease earlier in life will not have had as much exposure to lifestyle risk factors, and therefore would be expected to have greater genetic burden of disease risk variants than those that develop the disease later in life. To this end, the utility of GRS to detect an association with AOO of T2D, as an exemplar, in ancestrally homogenous populations is explored.

Further assessment of the relative performance of the Cox PH, proportional odds and binary logistic models was also undertaken via a GRS AOO of disease simulation study. The focus of the simulation study was geared primarily towards evaluation of the impact of censoring on relative power between different statistical approaches. As part of the simulation study, evaluation of key determinants of statistical power associated with the underlying genetic architecture of common diseases were considered. This encompassed the risk allele frequency (RAF), number of susceptibility genetic variants and the magnitude of SNP or GRS effect.

3.1.1. | Current burden of T2D

Diabetes mellitus (DM) was a recognized medical condition as far back as 400 (before Christ) BC [180]. Within the last three decades, DM has emerged to become the fastest growing disease epidemic globally, and a major health burden to health systems at every level. According to the world health organization (WHO) most recent estimates [181], globally diabetes was the seventh leading cause of death in 2016. In 2017, the International Diabetes Federation (IDF) estimated that 4 million deaths globally were attributable to diabetes [182].

Despite the remarkable progress in some aspects of diabetes care, overall, the absolute burden of the disease is rising. Currently T2D is estimated to affect 10% of the world population, accounting for 90% of all diabetes cases [183]. IDF estimated that 425 million adults aged 20 - 79 years had diabetes in 2017, however, in 1985 this figure stood at around 30 million. This dramatic increase in cases has been attributed in part to modifiable factors relating to lifestyle. However, both genetics and environmental factors or known to contribute to the development of diabetes within human populations. The main lifestyle factors attributed to the onset of diabetes include being overweight or obese, physical inactivity, and unhealthy diets. The T2D

epidemic to a large extent has been ascribed to the worldwide increase in obesity during the last 30 years, for instance, more than 60% of individuals older than 15 in the UK and US are overweight (BMI > 25) [184].

3.1.2. | AOO of T2D

From a diabetes perspective, the twentieth century marked a period of remarkable advancement in terms of understanding the mechanisms leading to hyperglycaemia which led to the formal classification of type I and type 2 diabetes in 1979 [185]. Currently, diabetes is defined as a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both (T2D formerly called non-insulin dependent or adult-onset) [184]. More recently, however, it has been argued that diabetes is not a single disease, but a composite of many diseases with a common feature of hyperglycaemia [186]. Historically, AOO and disease severity were among the distinguishing characteristics used to classify the subtypes of diabetes. Furthermore, the precise degree of hyperglycaemia defining diabetes has evolved over the years. In a clinical setting, the age at first diagnosis of T2D is used as a proxy for AOO, however, given the physiological nature of T2D the true AOO in many instances is largely unknown. AOO of T2D may be difficult to discern in some circumstances given the slow asymptomatic nature of the disease, as a result, the pre-detection period may extend over many years [184]. This limitation is further complicated by diagnosis criteria issues [185] stemming from lack of clear clinical guidelines pertaining to the classification and diagnosis of diabetes. Consequently, many cases of T2D remain undiagnosed and are only diagnosed at very advanced stages of the disease course.

3.1.3. | Genetics of T2D

Genetically, T2D is characterized as a heterogenous disease. Findings from twin and family studies have reported heritability estimates as low as 25% [187], while others have reported estimates as high as 80% [184, 188]. It has been noted that this wide range in heritability of T2D is due in part to the AOO of T2D, as studies based on cases with a lower AOO resulted in higher estimates of heritability for T2D [189-191]. Conventionally T2D was considered to be a disease that primarily affected adults, and in particular later in life, hence the initial classification of adult-onset diabetes. Today, however T2D is occurring in children and within the younger adult population, likely due to increasing rates of obesity.

GWAS have proven to be the most important contributor in relation to the identification of the genetic determinants of T2D [192]. The first T2D GWAS in 2007 identified 2 novel loci (SLC30A8 and HHEX) and confirmed the TCF7L2 locus [193] originally identified by linkage analysis or candidate gene association studies (PPARG and KCNJ11 previously confirmed). Around 2011 this TCF7L2 locus was known to have the largest effect on T2D risk (OR ~1.4), as most identified loci have small effect sizes (OR~1.1-1.3) [194]. In 2012 the number of identified T2D loci stood at 50 [195], by 2015 this number had risen to more than 120.

However, only a minority of observed T2D heritability is explained based on variation at known loci. The more than 120 loci only explain 10% of T2D heritability (based on data prior to December 2017) [139]. Previous work on the application of GRS to T2D indicates that GRS could potentially represent an improvement over existing risk assessment tools used in the diagnosis of T2D as they have demonstrated good predictive ability [196]. A 2018 study comparing the lowest T2D quintile to the highest quintile reported OR of 2.34 (95% CI: 1.59 – 3.46) [197].

3.1.4. | Association analysis of T2D

The analysis of common disease GWAS, which consists of data originating for the case-control or cohort study design, is usually undertaken in a logistic regression framework. This is often the case even in situations where the outcome though dichotomous has a time related event such as AOO of the disease. As highlighted in Chapter 1, the tendency to apply logistic regression was influenced by the fact that the model is considered less computationally expensive than modelling TTE analysis. The lack of inclusion of TTE models in GWAS software; and the need to undertake meta-analysis, which benefits from having data from both case-control and cohort studies analysed in the same way, are also contributing factors [198].

From the perspective of TTE outcomes, the logistic model, which assumes a binary outcome, in the context of disease genomics affected or unaffected by disease, is often considered to be less powerful then the TTE analytical approach particularly for diseases that occur later in life [148, 149, 199, 200]. However, there remains some scepticism regarding the relative power between the logistic and Cox PH model as previous investigations comparing these two models suggests that the Cox PH model has more statistical power to detect risk factors than logistic

models [198, 201, 202]. However, there are also some studies that favour the logistic over the Cox PH model in situations of short follow-up periods for cohorts and low incidence of event occurrences (high censoring rates) [203, 204]. However, these differences in study findings have been attributed to the rate of censoring [150]. Within a TTE framework an individual who remain unaffected by T2D at the end of the study period is considered a control and is therefore censored at their current age.

3.2. | Methods for construction of T2D GRS

This section provides details pertaining to the "base" GWAS SNPs (DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) study: published T2D GWAS used as inputs for the construction of the GRS) used in the construction of the GRS along with their associated summary statistics. Information regarding the two "target" GWAS (NUgene and WTCCC) genotyped samples used to test the performance of the GRS is also described. The methods applied to combine the results of the two datasets in a meta-analysis is also described in section 3.2.4.

3.2.1. | Identification of disease-associated SNPs

A European ancestry meta-analysis of T2D GWAS published at the end of 2017 formed the basis of the GRS constructed to test the association of GRS with AOO of T2D. This study, from the DIAGRAM Consortium [205] combined GWAS from 18 European ancestry studies totalling 26,676 T2D cases and 132,532 controls. All studies included in the meta-analysis were imputed against the March 2012 multi-ethnic 1000 Genomes Project (1000G) reference panel. To improve coverage and statistical power, several reference panels have been developed. The most commonly applied reference panels for the imputation of genotyped microarray data includes the 1000G reference panel first introduced in 2010 and more recently the Haplotype Reference Consortium (HRC) in 2016 [87]. For European ancestry populations the HRC is currently the most optimum reference panel for genotype imputation [206]. The 2017 DIAGRAM study reported a total of 128 distinct signals at 113 loci that were independently associated with T2D with OR ranging from 1.03 to 2.02. These loci incorporated lead SNPs that were identified at genome-wide significance in this 2017 DIAGRAM study or in earlier T2D GWAS [207-211]. However, it was noted that some previously reported SNPs did not attain

genome-wide significance in this 2017 DIAGRAM study but met a less stringent nominal threshold. Therefore, the GRS was constructed using two different subsets of SNPs: 36 independent SNPs attaining genome-wide significance ($p<5x10^{-8}$) in the 2017 DIAGRAM study; and 90 independent SNPs attaining genome-wide significance in the 2017 DIAGARM study or other previously reported independent SNPs attaining nominal significance ($p<0.05$) in the 2017 DIAGRAM study.

Using the summary statistics from this DIAGRAM study, it was possible to extract a list containing the SNPs know to be associated with T2D along with their corresponding p-value and effect size as measured by their OR values. In addition to SNP identifier details, other associated information collected included, details pertaining to the effect allele (EA), alternative allele (NEA) and effect allele frequency (EAF) or risk allele frequency (RAF) (Appendix B Table B.1.1). Several SNPs (9 SNPs) previously identified in non-European populations failed to attain nominal significance in this European ancestry DIAGRAM meta-analysis and therefore were not included as part of the GRS. In the event where multiple SNPs were reported for a locus, only the lead SNP, i.e the SNP with the strongest association signal (smallest p-value) was kept in the GRS list of SNPs as the GRS assumes SNPs to be independent. After removing non-lead SNPs, removal of the X chromosome SNPs (removed to eliminate confounding by sex); removal of SNPs located in the extended strong LD region; at the nominal significances p-value threshold, there were a total of 90 SNPs available for the construction of the GRS (Appendix B Table B.1.1). Based on the genome-wide threshold criteria (p-value $< 5 \times 10^{-8}$ in the meta-analysis) a total of 38 SNPs was available.

Two European ancestry genotyped GWAS datasets originating from the NUgene Project and WTCCC were used to evaluate the utility of GRS in detecting an association with AOO of T2D. The NUgene Banking Project sponsored by the Centre for Genetic Medicine at Northwestern University situated in Chicago, United States of America (USA) is a genomic biobank. Data for both cases and controls are routinely collected from patients over the age of 18 at Northwestern Medicine – affiliated hospitals and clinics for many common diseases including T2D. The data collected include DNA samples, and associated health information, which are usually sourced from patient health records. The health information provided for this study included T2D status, sex, enrolment age, year of birth, decade of birth, BMI and in relation to individuals with T2D, AOO of the disease [212]. In the second dataset, data were obtained from a case control

study carried out by the WTCCC. The WTCCC was established in 2005 and is an organization that comprise several research groups across the United Kingdom (UK). The consortium has gathered genotype data relating to seven common diseases which includes T2D. For each disease 2,000 samples are collected with the study design incorporating controls ascertained from the 1958 Birth Cohort and blood donors sourced from three national UK blood services. Provided health information included T2D status, sex, enrolment age in relation to controls, and in relation to individuals with T2D, AOO of T2D. A general overview of the characteristics of the two datasets are given in Table 3.3.

For the NUgene study, cases were ascertained on the bases of already having a diagnosis of T2D, evidenced by an International Classification of Diseases (ICD) 9 code for T2D, while previously unknown cases were ascertained on the basis of laboratory evidence of hyperglycemia and being prescribed T2D medication. Among previously known T2D cases, exclusions were made on the basis of have an ICD 9 code for ketoacidosis; being treated only with insulin and having never been on a T2D medication [212]. Cases in the WTCCC study were ascertained on the basis of current prescribed medication used to treat T2D and as defined by the World Health Organization (WHO), historical or contemporary laboratory evidence of hyperglycemia, in relation to individuals treated with diet alone; while other forms of diabetes were excluded on the basis of a standard clinical criteria based on personal and family history [213].

After the main quality control checks and genotype imputation were completed the NUgene and WTCCC, data were made available by Professor Andrew Morris (Supervisor – University of Liverpool). The genotype samples from NUgene study were genotyped using Illumina-Human660W-Quad_v1_A and Illumina-Human1M-Duov3_B microarray [212] and imputed against the Haplotype Reference Consortium (HRC) reference panel ,release 1.1 [214]. For the WTCCC study, genotyping was undertaken with Affymetrix-Affymetrix 500K chi [215] and imputed against the HRC reference panel release 1.1. Using the Michigan Imputation Server [216] the genotyped data were first phased using SHAPEIT, and then the pre-phased data were imputed using minimac3. To account for ancestry, principal components derived from a genetic relationship matrix were adjusted for in each dataset.

3.2.2. | Development and construction of GRS

To facilitate the construction of the GRS for individuals in the target GWAS, the process started with the extraction of SNPs in the target datasets that matched the selected SNPs from the base GWAS. As a result of this process 8 SNPs were excluded from the calculation of GRS because they were not available in the target datasets (Appendix B Table B.1.2). An additional SNP was removed from both datasets because of having an info score less than 0.4 in the NUgene dataset (Appendix B Table B.1.3). In the next step the effect allele (EA) in the target GWAS was then aligned with the base GWAS by flipping the EA in target GWAS to be the same as in the base GWAS in instances where there was a difference. Subsequently, the dosage associated with each EA were adjusted using the formula (2 – current dosage value) if the EA in the target GWAS were flipped. The formula used to calculate the original dosage values is as described in Equation 3.1.

$$\boldsymbol{G_{ij}} = [(\mathbf{P\_0_{ij}} \times \mathbf{0}) + (\mathbf{P\_1_{ij}} \times \mathbf{1}) + (\mathbf{P\_2_{ij}} \times \mathbf{2})]$$

Equation (3.1)

Where G_{ij} is the genotype dosage of the ith individual at the jth SNP; $\boldsymbol{P\_0_{ij}}$ refers to the probability of homozygous genotype associated with the NEA; $\boldsymbol{P\_1_{ij}}$ refers to the probability of heterozygous genotype; $\boldsymbol{P\_2_{ij}}$ refers to the probability of homozygous genotype associated with the EA and 0,1 or 2 refers to the number of EA present.

To calculate the GRS for each individual in the sample the formula used is as describe in Equation 1.3 in section 1.6.3.1 for the weighted GRS and Equation 1.4 also described in section 1.6.3.1 for the unweighted GRS. The corresponding R code used to calculate the GRS is outlined in Appendix G. Additionally, to evaluate the performance of GRS different versions of weighted and unweighted GRS determined at different p-value thresholds were developed. Four different versions of GRS were constructed based on: (1) SNPs determined at genome-wide significances p-value threshold of 5×10^{-8} with base GWAS effect size weighting; (2) SNPs determined at genome-wide significance p-value without weighting; (3) SNPs determined at nominal

significances p-value threshold of 0.05 with base GWAS effect size weighting; and (4) SNPs determined at nominal significances without weighting.

3.2.3. | Statistical analysis of individual T2D GWAS datasets

The main statistical methods and statistical software tools applied in the data analysis of the individual T2D genotyped GWAS datasets is described in this section. The procedures undertaken in the analysis of the datasets based on three different outcome measures applied to the Cox PH, proportional odds, and logistic model is described. The pseudo R^2 measure applied to assess the relative performance of the various statistical models within each statistical approach is also described.

3.2.3.1. |Statistical methods to individual T2D GWAS datasets

The T2D status of individuals in the samples which primarily distinguished between individuals affected by T2D (cases) and individuals who remained unaffected by T2D at the end of the study period (controls) were assessed using three different outcome measures. The first measure considered the AOO of T2D, where AOO was modelled in a TTE framework by means of a Cox PH model. In the TTE framework, which considered both cases and controls, controls were censored at their current age at the end of the study period. As part of the modelling process, the hazard ratio (HR) of the GRS associated with the AOO of T2D was estimated based on Equation 1.9 described in section 1.7.2. In Equation 1.9, the baseline hazard rate is given by λ (i.e. the hazard rate when all covariates are zero), which is scaled by the function of predictors or covariates, X, and corresponding regression coefficients, β, via $\exp^{(\beta X)}$. In this context, the predictor of interest was the GRS, however, to address confounding due to population structure principal component analysis (PCA) was applied to the genotyped data to form a relatedness matrix where eigen decomposition was performed to generate a smaller set of variables through a few linear combinations of the original variables. This smaller set of variables termed "principal components (PCs)" quantifies the patterns of population structure within the sample. These PC were thus included as covariates in the model. Other covariates included in the model were sex, and for the NUgene sample, BMI. The HR of the GRS for AOO of T2D was estimated using the function (coxph) of the R package (survival).

The second outcome measure entailed three ordered response categories or outcomes where an AOO of 55 was used to distinguish between early AOO T2D (age <=55) and late AOO T2D cases; while controls referred to individuals unaffected by T2D at the end of the study period. The proportional odds model was used to model the association of GRS and the ordered T2D status. The cumulative OR of the GRS for T2D were estimated using the function (polr) of the R package (MASS) based on Equation 1.11 in section 1.7.3. While the T2D GRS was included in the model as the predictor variable additional covariates were included to address potential confounding. These confounding variables are as described above in relation to the Cox PH model.

The third outcome measure considered the binary response categories cases versus controls where the association of GRS and T2D status was modelled via the binary logistic model. The OR of the GRS for T2D was estimated using the function (glm) of the R package (stats) based on Equation 1.12 in section 1.7.4. Confounding variables included in the model to address confounding is as outlined above in relation to the Cox PH model. Furthermore, age was not included as a covariate (see also Appendix H.1 – H.3).

Furthermore, due to the limited data available regarding age, age at enrolment was not included as a covariate in the logistic analysis in both datasets. The NUgene data was ascertained via routine electronic medical records with no fixed entry or end points which is likely to induced bias. Stratification by birth cohorts have been recommended as a solution to avoid bias due to age [217]. In relation to the WTCCC dataset age at enrolment was only available for controls, while AOO were available for cases as data was collected via case-control study design.

3.2.3.2. |Evaluating performance of T2D GRS models

To quantify the amount of variation attributable to the GRS the Nagelkerke pseudo R^2 measure (described in Chapter 1) was applied for all three analytical approaches. This pseudo R^2 measure is commonly applied in genetic research due to its maximum value of 1 property which is lacking in most pseudo R^2 measures. To determine the proportion of variance in AOO explained by the T2D GRS after adjustment for confounding variables, the R^2 values between nested models were compared. The proportion of variance explained represents the difference in R^2 after adjustment for confounding variables where the full model (model with confounding

variables and GRS) was compared to a reduced model (model with confounding variables only, GRS was excluded in this model). These confounding variables included sex, BMI (NUgene) and PCs described in section 3.2.3.2. The difference in R^2 between these two models (full – reduced) was used to determine the R^2 that is likely to be due to the GRS (Table 3.1).

Table 3. 1 – Description of models used in the analysis of T2D GRS

Model	**Terms included in model**
(1) GRS reduced models	**Covariate(s):** (X_s) - Sex: male=0; female =1 (X_d) - BMI: continuous covariate measured in kg/m^2 (X_{c1}) - X_{c3}) - PC1 - PC3: principal components used to account for population structure)
(2) adjusted (full) model	**Variable of interest:** (X_g) - GRSwN: weighted nominal significant GRS **Covariate(s):** **(**X_s) - Sex: male=0; female =1 (X_d) - BMI: continuous covariate measured in kg/m^2 (X_{c1}) - X_{c3}) - PC1 - PC3: principal components used to account for population structure)
Description	
Versions of GRS (X_{g1}) - GRSwN: weighted nominal significant GRS (X_{g2}) - GRSwG: weighted genome-wide significant GRS (X_{g3}) - GRSuN: unweighted nominal significant GRS (X_{g4}) - GRSuG: unweighted genome-wide significant GRS	

As indicated in section 3.1.1 obesity has been recognised as a potent independent and modifiable risk factor for T2D. Obesity is estimated to account for 80-85% of the risk of developing T2D [218, 219]. Additionally, the relationship between obesity and T2D may differ according to external factors including age and sex. In relation to obesity, BMI is the marker most commonly applied to identify the risk of T2D [220]. Therefore, in the NUgene sample the extent to which BMI explains the variance of AOO of T2D was also considered. In these models the full model consisting of the confounding variables which included the GRS and BMI was compared to a reduced model consisting of the confounding variables which included the GRS, while BMI was excluded in these models (Table 3.2).

Table 3. 2 – Description of models used in the analysis of T2D GRS and BMI

Model	Terms included in model
(1) BMI reduced models	**Covariate(s):** **Model 1:** (X_s) - Sex: male=0; female =1 (X_{c1}) - X_{c3}) - PC1 - PC3: principal components used to account for population structure) (X_g) - GRSwN: weighted nominal significant GRS
(2) adjusted (full) model	**Model 1:** **Variable of interest:** (X_d) - BMI: continuous covariate measured in kg/m^2 **Covariate(s):** **(**X_s) - Sex: male=0; female =1 (X_{c1}) - X_{c3}) - PC1 - PC3: principal components used to account for population structure) (X_g) - GRSwN: weighted nominal significant GRS
Description	
Versions of GRS (X_{g1}) - GRSwN: weighted nominal significant GRS (X_{g2}) - GRSwG: weighted genome-wide significant GRS (X_{g3}) - GRSuN: unweighted nominal significant GRS (X_{g4}) - GRSuG: unweighted genome-wide significant GRS	

3.2.4. | Statistical analysis of combined T2D GWAS datasets

The main statistical methods and statistical software tools applied to combine the results of the two datasets after independent analysis is outlined in this section. The procedures carried out in the summary statistics meta-analysis of the two T2D GWAS datasets NUgene and WTCCC which entailed two primary approaches; procedures for combining effect sizes, which in this instance are the OR and cumulative OR originating from the logistic and proportional odds analysis respectively; and (due to issues relating to the age timescale applied in individual studies) procedures for combing p-values originating from the Cox PH analysis are outlined. Methods applied to assess consistency and heterogeneity of model estimates are also outlined.

3.2.4.1. |Data extracted for inclusion in meta-analysis

The general characteristics of the NUgene and WTCCC T2D GWAS datasets have be outlined in Table 3.3. In addition to the OR outputted from the logistic model, cumulative OR outputted from the proportional odds model, and the HRs outputted from the Cox PH model, their corresponding standard error (SE) and p-value were also extracted for inclusion in the meta-analysis.

3.2.4.2. |Statistical methods applied in meta-analysis

The fixed effect model was applied to facilitate the combining of the log OR where inverse-variance weighting was applied. The fixed effect model is formed on the premise that there is one true effect size to be estimated across all studies [221]. As a result, the pooled estimate is considered to be the true common effect size, which here is denoted $\boldsymbol{\beta_{pooled}}$ and is calculated using Equation 3.2 where $\boldsymbol{w_k}$ refers to weights assigned to each dataset and $\boldsymbol{\beta_k}$ estimate based on individual datasets. ($\boldsymbol{k}$ refers to individual datasets and $\boldsymbol{K}$ overall number of datasets included in the meta-analysis).

$$\boldsymbol{\beta_{pooled}} = \sum_{k=1}^{K} \beta_k w_k \,/\, \sum_{k=1}^{K} w_k$$

Equation (3.2)

In a meta-analysis setting, the inverse variance weighting has been shown to be optimal [222], therefore the weighting applied to the pooled estimate is based on the inverse variance rather than the sample size of the individual datasets ($\boldsymbol{w_k = 1 \,/\, v_k}$). If the SE for each dataset is denoted SE ($\boldsymbol{\beta_k}$), then $\boldsymbol{v_k}$ is SE($\boldsymbol{\beta_k}$)^2 [223]. It is further assumed that any deviation in individual estimates from the pooled estimate is solely due to the play of chance. The pooled SE is calculated based on Equation 3.3.

$$\boldsymbol{SE_{pooled}} = \sqrt{1 \,/\, \sum_{k=1}^{K} w_k}$$

Equation (3.3)

It was deemed more appropriate to combine the p-values rather than the effect size of the HR from the Cox PH analysis. As the entry of controls into both the NUgene and WTCCC studies were not fixed to a reference date and consisted of different ages, this can pose a challenge when chronological age is used as the timescale within a TTE framework [217]. This is particularly so when measures implemented to account for different entry points of controls in relation to cases is unclear. In recent years the use of chronological age as a timescale has gained moderate acceptance for the analysis of TTE data [217]. However, the issue of calendar time when individuals enter the study at different times remains. As a result, to facilitate the combining of the p-values of the HR from the individual Cox PH model analysis, Equation 3.8 for the Stouffers method was applied. The Stouffer's z-score method is one of the most commonly applied methods in combined p-value meta-analysis [224, 225]. Compared to meta-analysis based on effect sizes, Stouffer's method has been found to be more robust when there was a difference in analytical approaches between studies. However, there is the possibility of a small loss of efficiency [226]. Unlike earlier approaches that have been developed, the Stouffer's z-score approach considers the direction of the effect size [1]. In the Stouffers method, it is assumed that the z-score of individual studies is given by Equation 3.4 [1, 227] , where $\boldsymbol{z_k}$ refers to the directed z-score from individual datasets and $\boldsymbol{p_k}$ refers to the p- value for the individual datasets where $\boldsymbol{p_k/2}$ is applied in the case of two-sided p-values. Additionally, ∅ refers to the standard normal cumulative distribution function and $\boldsymbol{z}$ is assumed to follow a standard normal distribution.

$$\mathbf{z_k = \emptyset^{-1}(1 - p_k/2) * sign(\beta_k)}$$

Equation (3.4)

The direction of the z-score is aligned to the same effect allele across studies. In Equation 3.4 the effect direction for an individual dataset k is denoted sign ($\boldsymbol{\beta_k}$). The overall meta-analysis z-score, denoted $\boldsymbol{Z}$ is formulated on the basis of Equation 3.5. To combine the z-score using the Stouffer's method, the inverse normal transformed p-values is summed taking into account applied weighting. To improve power when combining studies of varying sample sizes weighting is usually incorporated [1, 225, 228]. The optimal weighting applied in the Stouffer's method is given by the squared root of the sample sizes [229]. Under the null hypothesis of no association the standard cumulative normal distribution c.c.f (Φ) is expected to follow a standard normal distribution.

$$\mathbf{Z= \sum_{k=1}^{K} z_k v_k \ / \sqrt{\sum_{k=1}^{K} v_k}}$$

Equation (3.5)

The components of the z-score equation used to calculate the overall meta-analysis z-score value includes $\boldsymbol{z_k}$ which refers to the z-score value from individual datasets; v_k refers to the squared root of the sample size of the kth dataset; K represents the number of datasets included in the meta-analysis; and k which refers to the individual datasets.

3.2.4.3. |Evaluating heterogeneity in meta-analysis

The Cochran Q statistic, in conjunction with the I^2 index, was used to assess heterogeneity in the effect size in the meta-analysis [230]. Due to the low power of the Cochran Q test to detect heterogeneity, a less stringent P-value threshold is often recommended. Here we use the recommended threshold of a P-value < 0.10 instead of the conventional P < 0.05 [231-233]. The Q statistic is given by:

$$\mathbf{Q = \sum_{k=1}^{K} w_k \left(\beta_k - \beta_{pooled}\right)^2}$$

Equation (3.10)

where β_{pooled} denotes the overall common effect size assuming a fixed effect model and β_k represents the estimate based on individual datasets. Q is distributed as a chi-square statistic with K - 1 degrees of freedom (df), where K corresponds to the number of datasets included in meta-analysis.

$$\mathbf{I^2 = \left((Q - df) \ / Q\right) \times 100\%}$$

Equation (3.11)

The I^2 statistic or index which quantifies heterogeneity is the percentage of the total variability in the set effect sizes due to true heterogeneity (Equation 3.11). On the basis of the I^2 index, the extent of heterogeneity is assessed by comparing the Q value to its expected value, i.e. its df (df=K-1), with K representing the number of datasets included in the meta-analysis. A

I^2 value > 50% was considered an indicator of substantial heterogeneity between studies [231, 234, 235] while an I^2 < 25% signifies low or no heterogeneity [236, 237].

Combined p-value meta-analysis methods relative to other statistical methods are limited in respect to their ability to quantify or characterise heterogeneity [238]. As estimation of pooled effect sizes does not form part of the p-value meta-analysis, it is not possible to directly assess heterogeneity. However, in the case of Stouffer's z-score method, heterogeneity can be assessed visually by ranking the estimated z-score from individual studies, where values are plotted and investigated to see if they lay on a straight line [239]. Nevertheless, application of methods that directly combine p-values, which are relatively flexible, require minimal information and assumptions regarding individual studies [240]. Methods that directly combine p-values have the advantage of simplicity as well as being extensible to different kinds of outcome measures as p-values are used as a common metric to examine the evidence of association [241, 242].

3.2.4.4. |Statistical software used in meta-analysis

The meta-analysis of the T2D GWAS datasets was also conducted in the R software version 3.3.1 within the Linux redhat environment. The pooled OR were estimated using the fixed effects model along with its corresponding 95% confidence interval (CI) in base R. Pooling of the estimates were weighted according to the inverse variance method. Forest plots were generated to summarize information for effect size and the corresponding 95% CI of each dataset and the pooled effect using the 'forestplot' function of the "forestplot" package. Additionally, pooled p-values for the HR were estimated using the sumz function of the "metap" package (Appendix H.4).

3.3. | Association of T2D GRS with AOO of the disease

Findings from the primary analysis of the T2D genotype data from the NUgene and WTCCC are discussed in this section along with key findings from the GRS AOO T2D summary statistics meta-analysis of the two datasets. The utility of GRS to detect an association with AOO of T2D is assessed, although the association of BMI, a major risk factor of T2D, with AOO of T2D is also considered. At the outset, a general overview of the underlying characteristics and GRS profiles within each genotype dataset is provided in section 3.3.1. The single SNP association analysis is presented in section 3.3.2. Results of the GRS association analysis based on the cases and control Cox PH, proportional odds and binary logistic models is presented in section in 3.3.3. Assessment of the models is formulated in terms of the size of estimated effect and strength of association resulting from the GRS. The variation in AOO of T2D explained by the GRS were assessed on the basis of the Nagelkerke pseudo R^2 measure (section 3.3.5). Analysis focused on assessing BMI are presented in section 3.3.4; assessing model assumptions 3.3.6; and the combining of the two samples in a meta-analysis is presented in section 3.3.7.

3.3.1. | Profile of GWAS datasets

Presented in Table 3.3 is a general summary of the characteristics of the NUgene and WTCCC genotyped datasets. The general characteristics considered include sex, age (AOO for cases and age at the end of the study period for controls) and BMI (NUgene only). It is observed that the two datasets differed mainly in terms of the ratio of cases and controls and distribution of age among the cases and controls.

Table 3. 3 – Descriptive characteristics of T2D cases and controls

Characteristics	NUgene SAMPLE		WTCCC SAMPLE	
	Cases	Control	Cases	Control
Total (N)	(517)	(598)	(921)	(2,889)
Sex (n, % female)	226 (43.7%)	333 (55.7%)	409 (44.4%)	1,467 (50.7%)
Age (years)				
Mean (SD)	57.75 (11.30)	49.90(12.90)	49.41 (10.77)	50.83 (11.34)
Median	58	51	50	58
Range (Min-Max)	20 - 90	18 - 90	25 - 75	17 - 69
BMI (Body Mass Index)			-	-
Mean (SD)	33.36 (8.05)	26.92 (5.35)	-	-
Median	32.02	25.98	-	-
Range (Min-Max)	17.6 - 67.03	15.58 - 50.59	-	-

Descriptions*: N: overall sample size; n: subgroup sample size;* Age: for cases *age refers to AOO of T2D and controls age at enrolment; SD: standard deviation; Min: minimum; Max: maximum; BMI: body mass index measured in kg/m².*

The NUgene dataset consists of 1,115 individual samples of which 46.4% are cases of T2D, while the WTCCC dataset consists of 3,810 individuals with 24.1% being cases of T2D. It was also noted that on average individuals affected by T2D in the NUgene dataset are older when compared to individuals in the WTCCC dataset, mean age of onset of T2D are 57.8 years (SD 11.30) and 49.4 years (SD 10.77) for NUgene and WTCCC respectively. It was further noted that on average the controls in the WTCCC are older when compared to the controls used in the NUgene dataset (median age of controls, NUgene 51 years and WTCCC 58 years).

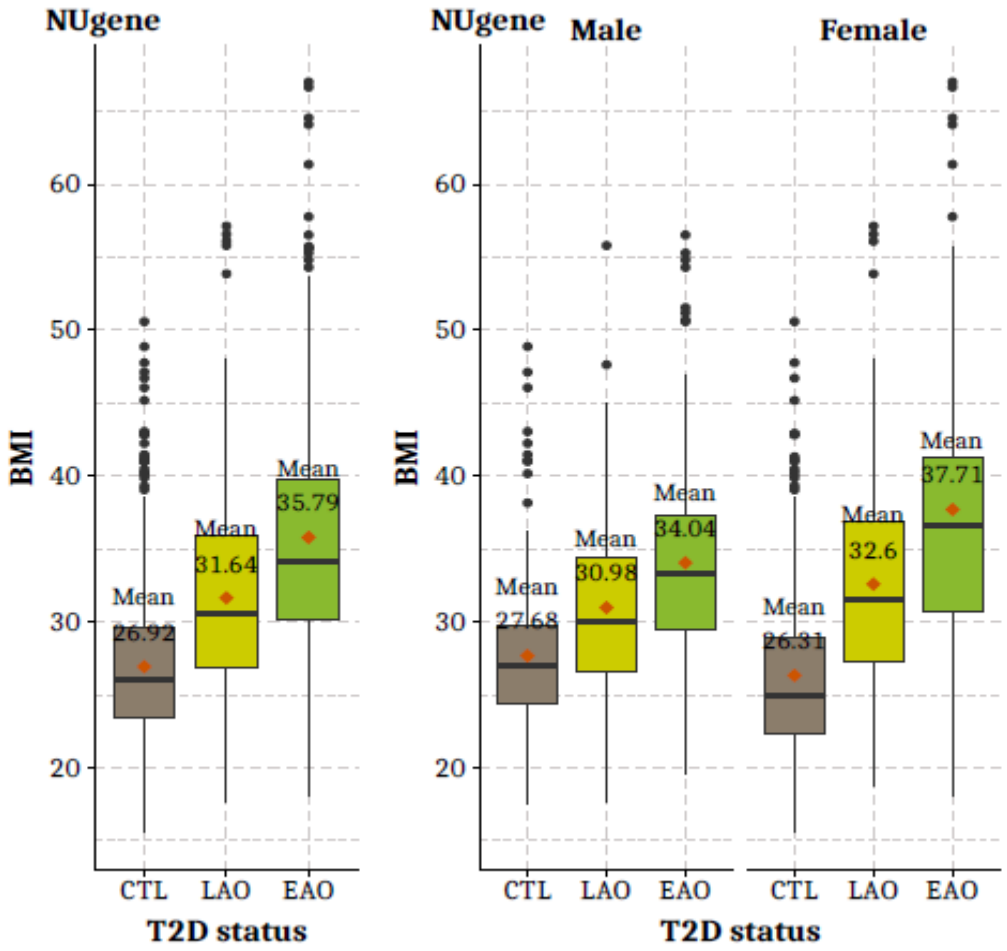


Figure 3. 1 - Distribution of BMI and T2D status in the NUgene
The boxes represent the distribution of BMI for EAO (green), LAO (yellow green) as well as CTL (dark grey). The horizonal line inside the box refers to the median. The lower and upper hinges of the box correspond to the 25th and 75th quartiles and the black dots denoted outliers.
Abbreviations: CTL: controls; LAO: late age onset; EAO: early age onset.

Using the NUgene dataset it was possible to consider the impact of BMI in relation to the onset of T2D. The distribution of BMI among cases and controls is illustrated in Figure 3.1. T2D EAO cases appear on average to have a higher BMI compared to unaffected controls and LAO cases (BMI 35.8, 31.6, and 26.9 respectively for EAO, LAO and unaffected controls). The difference in BMI is particularly noticeable among females where the mean BMI is 37.7, 32.6, and 26.3 for EAO, LAO and unaffected controls, respectively.

Table 3. 4 – Comparison of mean GRS using unpaired two sample t-test of T2D cases and controls in NUgene and WTCCC samples

Characteristics	P-value	Mean of Cases	Mean of Controls	Standard Error of difference in means
Cases versus controls				
NUgene				
Genetic Risk Scores (GRS)				
Weighted GRS (P<0.05)	$\mathbf{1.1 \times 10^{-13}}$	7.36	7.14	0.03
Weighted GRS (P<5*10^{-8})	$\mathbf{1.7 \times 10^{-14}}$	4.85	4.65	0.03
Unweighted GRS (P<0.05)	$\mathbf{2.4 \times 10^{-09}}$	81.65	79.76	0.31
Unweighted GRS (P<5*10^{-8})	$\mathbf{1.9 \times 10^{-12}}$	44.00	42.37	0.23
WTCCC				
Genetic Risk Scores (GRS)				
Weighted GRS (P<0.05)	$\mathbf{2.2 \times 10^{-53}}$	7.45	7.17	0.02
Weighted GRS (P<5*10^{-8})	$\mathbf{6.2 \times 10^{-43}}$	4.88	4.66	0.02
Unweighted GRS (P<0.05)	$\mathbf{9.3 \times 10^{-42}}$	82.69	79.96	0.20
Unweighted GRS (P<5*10^{-8})	$\mathbf{3.2 \times 10^{-37}}$	44.16	42.39	0.14

Descriptions: *GRS: genetic risk score; EAO early-age-onset; LAO late-age-onset*

Table 3.4 presents the findings from t-tests undertaken to compare the means of the T2D GRS between cases and controls in the NUgene and WTCCC datasets. The GRS was defined by the nominal and genome-wide significance thresholds. For the weighted GRS, the weighting applied was based on the OR reported in the base GWAS used to construct the GRS, while for the unweighted GRS each SNP contributing to the GRS was assumed to contribute equally to the risk of T2D. In both samples (NUgene and WTCCC) the tests indicate that the mean GRS of the cases was significantly higher than controls for both weighted and unweighted GRS. Furthermore, a general summary of the characteristics of weighted and unweighted GRS among the T2D cases and controls is presented in Appendix B B.3.1.

Table 3. 5 - Comparison of mean GRS using unpaired two sample t-test of T2D EAO cases and LAO cases in NUgene and WTCCC samples

Characteristics	P-value	Mean of EAO cases	Mean of LAO cases	Standard Error of difference in means
EAO cases versus LAO cases				
NUgene				
Genetic Risk Scores (GRS)				
Weighted GRS (P<0.05)	$\mathbf{1.7 \times 10^{-01}}$	7.32	7.38	0.04
Weighted GRS (P<5*10^{-8})	$\mathbf{2.1 \times 10^{-01}}$	4.82	4.87	0.04
Unweighted GRS (P<0.05)	$\mathbf{4.2 \times 10^{-01}}$	81.42	81.81	0.49
Unweighted GRS (P<5*10^{-8})	$\mathbf{1.6 \times 10^{-01}}$	43.71	44.21	0.35
WTCCC				
Genetic Risk Scores (GRS)				
Weighted GRS (P<0.05)	$\mathbf{4.4 \times 10^{-01}}$	7.46	7.43	0.03
Weighted GRS (P<5*10^{-8})	$\mathbf{3.9 \times 10^{-01}}$	4.89	4.86	0.03
Unweighted GRS (P<0.05)	$\mathbf{4.3 \times 10^{-01}}$	82.78	82.49	0.37
Unweighted GRS (P<5*10^{-8})	$\mathbf{3.6 \times 10^{-01}}$	44.24	44.01	0.24

Descriptions: *GRS: genetic risk score; EAO early-age-onset; LAO late-age-onset*

Table 3.5 presents the findings from t-tests undertaken to compare the mean GRS of EAO cases and LAO cases in NUgene and WTCCC datasets. Unlike the comparison between the cases and the controls, among all four T2D GRS considered in the two datasets it was found that they were not significantly different between EAO and LAO cases.

Table 3. 6 - Comparison of mean GRS in the NUgene and WTCCC samples using unpaired two sample t-test of both T2D cases and controls

Characteristics	P-value	Mean in NUgene	Mean in WTCCC	Standard Error of difference in means
NUgene controls versus WTCCC controls				
Genetic Risk Scores (GRS)				
Weighted GRS (P<0.05)	$\mathbf{2.0 \times 10^{-01}}$	7.14	7.17	0.02
Unweighted GRS (P<0.05)	$\mathbf{4.0 \times 10^{-01}}$	79.76	79.96	0.24
Weighted GRS (P <5*10^{-8})	$\mathbf{5.1 \times 10^{-01}}$	4.65	4.66	0.02
Unweighted GRS (P <5*10^{-8})	$\mathbf{9.0 \times 10^{-01}}$	42.37	42.39	0.17
NUgene cases versus WTCCC cases				
Genetic Risk Scores (GRS)				
Weighted GRS (P<0.05)	$\mathbf{6.2 \times 10^{-04}}$	7.36	7.45	0.03
Unweighted GRS (P<0.05)	$\mathbf{3.8 \times 10^{-04}}$	81.65	82.69	0.29
Weighted GRS (P <5*10^{-8})	$\mathbf{2.5 \times 10^{-01}}$	4.85	4.88	0.02
Unweighted GRS (P <5*10^{-8})	$\mathbf{4.2 \times 10^{-01}}$	44.00	44.16	0.20

Descriptions: *GRS: genetic risk score*

Table 3.6 presents the findings from t-tests undertaken to compare the mean GRS of cases in the NUgene and WTCCC studies and the mean GRS of controls in the NUgene and WTCCC studies. The GRS of the controls in the NUgene and WTCCC studies were not significantly different for all four versions of the T2D GRS. Additionally, the mean genome-wide weighted and unweighted GRS of cases were also not significantly different between the two studies. However, the mean GRS of both the weighted and unweighted nominally significant GRS of cases were significantly different.

3.3.2. | Single-SNP association with T2D status

The single SNP component of the analysis were undertaken to assess the association of each SNP independently within each sample. This was designed to gauge their potential for predicting the risk of the onset of T2D as part of the T2D GRS. In the NUgene sample, the impact of BMI on the perform of individual SNPs was also considered.

A summary of the significant results of the single SNP tests, which originally included the 81 SNP used to construct the GRS in presented in Appendix B Table B.2.1 and Table B.2.2. The analysis of the single SNP association with T2D status adjusted for age, sex and population structure, indicated that of the 81 SNPs tested in the NUgene dataset, 8 were nominally significantly associated with T2D status, while 32 SNPs were found to be significant within the WTCCC dataset (Appendix B Table B.2.2). In the NUgene sample the 8 significant SNPs were significant with and without adjustment for BMI. This suggests that these 8 SNPs are associated with T2D independently of BMI.

3.3.3. | Association of GRS with AOO of T2D

In this section, the performance of both the weighted and unweighted GRS, which consists of the genome-wide and nominal significance threshold for SNPs included in the GRS was assessed. Potential confounding in respect to sex, and population structure (using PCs), as well as BMI in the case of the NUgene sample, have been taken into consideration. An overall assessment of the three analytical approaches (cases and controls Cox PH, proportional odds, and binary logistic analysis) was also given.

3.3.3.1. |Association of weighted GRS with AOO of T2D

Illustrated in Figure 3.2 are model estimates based on Cox PH, proportional odds, and logistic models. It compares the estimated HR of AOO of T2D (from Cox PH model), OR of T2D status (from logistic model) and cumulative OR of AOO of T2D (from the proportional odds model) associated with the weighted T2D GRS produced by the NUgene and WTCCC samples. Generally, the results indicated that the weighted genome-wide GRS produced greater effect sizes in relation to AOO of T2D and T2D status compared to the weighted nominal significance

GRS models. In both samples, for the analysis based on the Cox PH model, the weighted nominally significant GRS was found to be the most highly significant of the two versions of GRS evaluated. The difference in the strength of the association between weighted GRS was less noticeable for the NUgene sample. The estimated HR pertaining to the weighted nominally significant GRS was 1.5 (95% CI 1.2 - 1.8: $p = 3.7 \times 10^{-05}$) and 2.4 (95% CI 2.1 - 2.8: $p = 2.9 \times 10^{-38}$) respectively for the NUgene and WTCCC samples. For the proportional odds model, the weighted nominally significant GRS was also found to be most highly significantly associated with AOO of T2D in the WTCCC sample. For the NUgene sample though, a similar strength of association was observed for both versions of the weighted GRS in relation to AOO of T2D.

Regarding T2D status, of the two versions of GRS evaluated in both samples, the weighted nominally significant GRS was found to be most highly associated with T2D status. The estimated OR associated with the weighted nominally significant GRS was 2.8 (95% CI 2.1 – 3.7: $p = 2.3 \times 10^{-12}$) and 3.4 (95% CI 2.9 – 4.0: $p = 5.3 \times 10^{-48}$) respectively for the NUgene and WTCCC samples. In relation to T2D status, it was further noted as with the other two methods (Cox PH and proportional odds), in the NUgene sample, the observed strength of association was found to be similar for both versions of the weighted GRS. However, for the WTCCC sample, a more marked difference in the strength of association was observed between the two versions of weighted GRS. The P-value associated with the weighted nominally significant GRS and weighted genome-wide significant GRS was $p = 5.3 \times 10^{-48}$ and $p = 1.6 \times 10^{-39}$ respectively. Furthermore, in general, these results seem to suggest that while constructing the GRS based on nominally significant SNPs (greater number of SNPs included in GRS) increases the evidence of association, constructing the GRS based on genome-wide significant SNPs (smaller number of SNPs included in GRS but more strongly associated with T2D) illustrates a greater effect on risk of T2D or AOO of T2D.

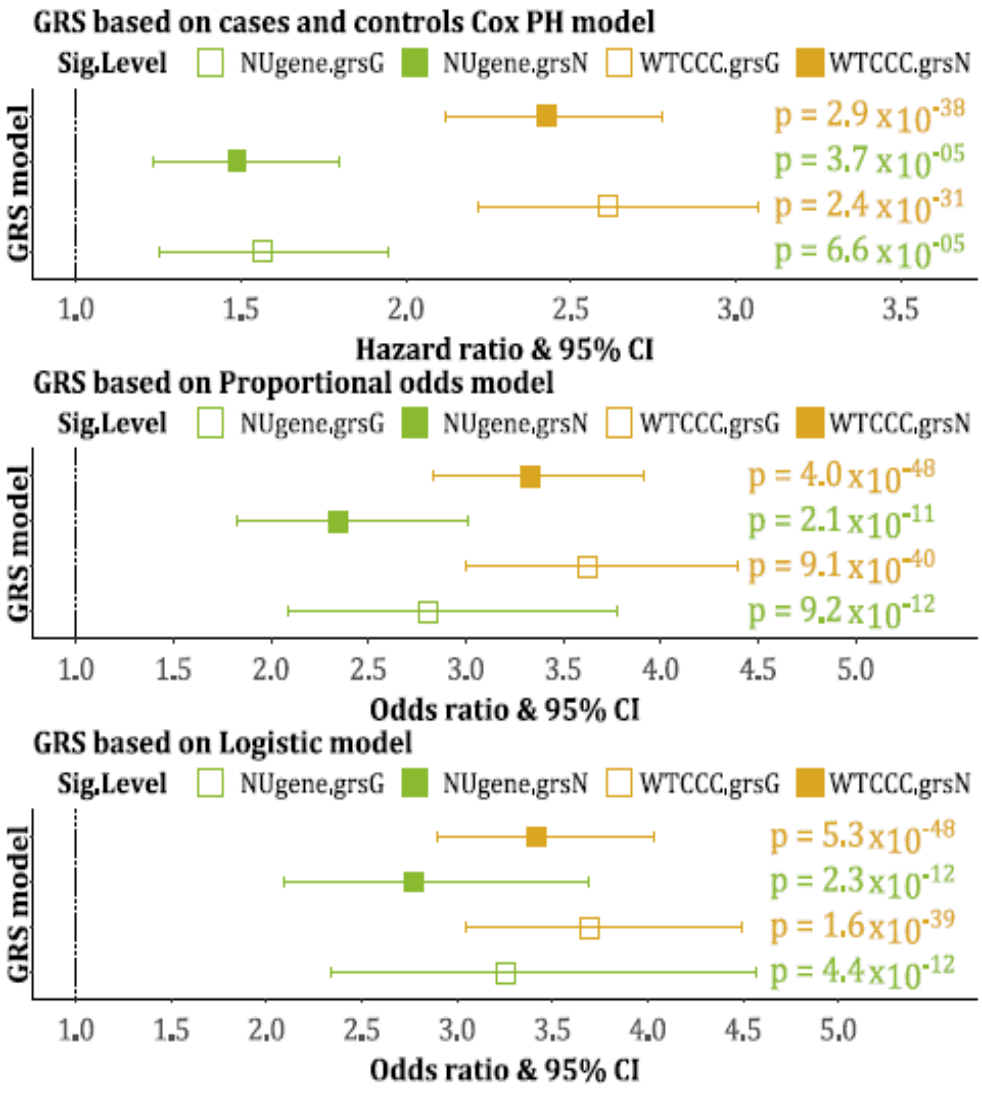


Figure 3. 2 - Comparison of estimated ES of AOO of T2D associated with the weighted GRS based on three analytical methods for NUgene and WTCCC samples.
The x-axis indicates the HR for the Cox PH model and the OR for the logistic and proportional odds model and 95% CI for each GRS model shown on the y-axis. The models have been adjusted for sex, ancestry principal components to account for population structure, GRS and for NUgene BMI. The two models considered include the adjusted model with weighted nominal significant GRS denoted (grsN) and adjusted weighted genome-wide significant GRS denoted (grsG). Models for the NUgene samples are represented by green and WTCCC gold.

Table 3.7 presents a summary of the estimates resulting from each of the three statistical methods evaluating the weighted GRS in the NUgene and WTCCC samples. Overall, the strength of association with T2D status and AOO of T2D for both versions of the weighted GRS were found to be similar based on the logistic and proportional models. However, the observed association between both versions of the weighted GRS and T2D status and AOO of T2D based on the logistic and proportional odds models were found to be substantially stronger when compared the weighted GRS based on the Cox PH model. The P-value associated with the weighted nominally significant GRS based on the proportional odds, logistic and Cox PH models was $p = 4.0 \times 10^{-48}$, $p = 5.3 \times 10^{-48}$ and $p = 2.9 \times 10^{-38}$ respectively for the WTCCC sample.

Table 3. 7 - Estimated effect of association of weighted GRS and AOO of T2D in NUgene and WTCCC samples

Analysis Method	Weighted GRS (P-value threshold P < 0.05)				Weighted GRS (P-value threshold P < 5×10^{-8})			
	ES	Lower 95% CI	Upper 95% CI	P-value	ES	Lower 95% CI	Upper 95% CI	P-value
NUGENE								
Cox PH model (cases and controls)								
Adjusted *(GRS+ BMI+ Covariates)*	1.488	1.232	1.798	3.7×10^{-05}	1.565	1.256	1.949	6.6×10^{-05}
Proportional odds model								
Adjusted *(GRS+ BMI+ Covariates)*	2.347	1.829	3.013	2.1×10^{-11}	2.804	2.085	3.772	9.2×10^{-12}
Binary logistic regression model								
Adjusted *(GRS+ BMI+ Covariates)*	2.768	2.090	3.693	2.3×10^{-12}	3.255	2.340	4.566	4.4×10^{-12}
WTCCC								
Cox PH model (cases and controls)								
Adjusted *(GRS+ Covariates)*	2.428	2.123	2.778	2.9×10^{-38}	2.611	2.222	3.069	2.4×10^{-31}
Proportional odds model								
Adjusted *(GRS+ Covariates)*	3.330	2.833	3.915	4.0×10^{-48}	3.627	2.996	4.392	9.1×10^{-40}
Binary logistic regression model								
Adjusted *(GRS+ Covariates)*	3.417	2.900	4.037	5.3×10^{-48}	3.692	3.043	4.491	1.6×10^{-39}

Descriptions: *ES: effect size which refers to the HR for Cox PH model and OR for the logistics and proportional odds models; GRS: genetic risk score; BMI: Body Mass Index; PC1-PC3: Principal Components; CI: confidence interval*

3.3.3.2. |Association of unweighted GRS with AOO of T2D

Represented in Figure 3.3 are again model estimates based on the three analytical approaches. In this instance however, the analysis compares the estimated OR of AOO of T2D associated with the unweighted GRS. In general, the results indicated that the unweighted genome-wide GRS produced greater effect sizes in relation to AOO of T2D and T2D status compared to the unweighted nominal significance GRS models. As with the weighted GRS, in the Cox PH analysis the unweighted nominally significant GRS was found to be the most highly significant of the two versions of GRS evaluated in both samples. It was further noted however, in relation to the NUgene sample, that the observed strength of association was similar for both versions of the unweighted GRS. In the NUgene and WTCCC samples the unweighted nominally significant GRS HR was estimated to be 1.03 (95% CI 1.01 - 1.05: $p = 5.0 \times 10^{-04}$) and 1.07 (95% CI 1.06 – 1.09: $p = 1.5 \times 10^{-29}$) respectively. Additionally, in both the proportional odds and logistic analysis the unweighted nominally significant GRS was found to be the most highly significant of the two versions of GRS evaluated in the WTCCC. However, for the NUgene sample, the difference in the strength of association observed for both versions of the unweighted GRS in relation to AOO of T2D was marginal. For the WTCCC sample, the estimated OR associated with the nominally significant GRS based on the logistic model was 1.1 (95% CI 1.09 – 1.12: $p = 2.4 \times 10^{-38}$) and the estimated cumulative OR based on the proportional odds model was 1.1 (95% CI 1.09 – 1.12: $p = 3.5 \times 10^{-38}$).

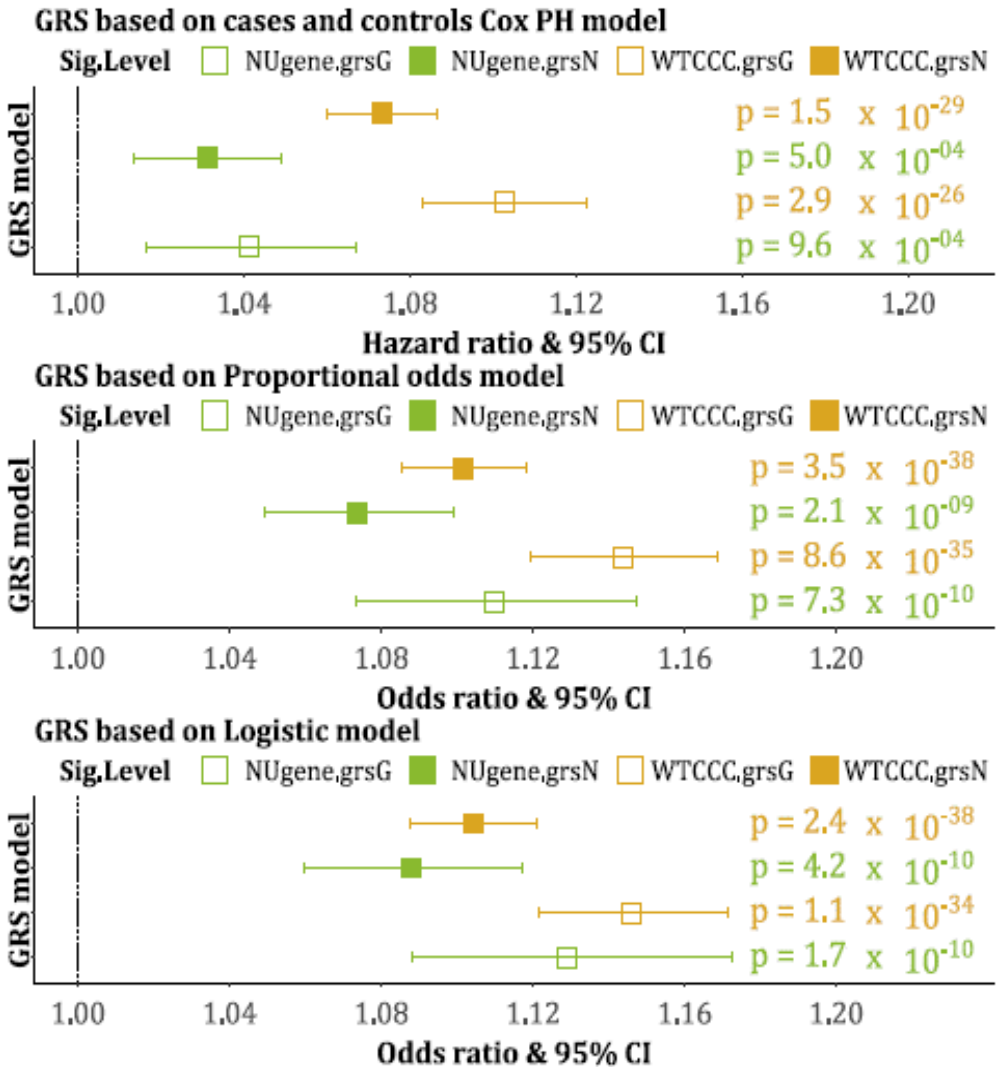


Figure 3.3 - Comparison of estimated ES of AOO of T2D associated with the unweighted GRS based on three analytical methods for NUgene and WTCCC samples.

The x-axis indicates the HR for the Cox PH model and the OR for the logistic and proportional odds model and 95% CI for each GRS model shown on the y-axis. The models have been adjusted for sex, ancestry principal components to account for population structure, GRS and for NUgene BMI. The two models considered include the adjusted model with unweighted nominal significant GRS denoted (grsN) and adjusted unweighted genome-wide significant GRS denoted (grsG). Models for the NUgene samples are represented by green and WTCCC gold.

Table 3.8 presents a summary of the estimates resulting from each of the three statistical methods, evaluating in this instance the unweighted GRS in the NUgene and WTCCC samples. Overall, as in the weighted GRS, the difference in the observed association between both versions of the unweighted GRS and T2D status and AOO of T2D based on the logistic and proportional odds models were found to be substantially stronger when compared the unweighted GRS based on the Cox PH model. However, the difference in performance between the logistic and proportional odds models were found to be marginal. The P-value associated with the unweighted nominally significant GRS based on the proportional odds, logistic and Cox PH models was $p = 3.5 \times 10^{-38}$, $p = 2.4 \times 10^{-38}$ and $p = 1.5 \times 10^{-29}$ respectively for the WTCCC sample.

Table 3. 8 - Estimated effect of association of unweighted GRS and AOO of T2D in NUgene and WTCCC samples

Analysis Method	Unweighted GRS (P-value threshold P < 0.05)				Unweighted GRS (P-value threshold P < 5×10^{-8})			
	ES	Lower 95% CI	Upper 95% CI	P-value	ES	Lower 95% CI	Upper 95% CI	P-value
NUGENE								
Cox PH model (cases and controls)								
Adjusted *(GRS+ BMI+ Covariates)*	1.031	1.013	1.049	5.0×10^{-04}	1.041	1.017	1.067	9.6×10^{-04}
Proportional odds model								
Adjusted *(GRS+ BMI+ Covariates)*	1.074	1.049	1.099	2.1×10^{-09}	1.110	1.074	1.147	7.3×10^{-10}
Binary logistic regression model								
Adjusted *(GRS+ BMI+ Covariates)*	1.088	1.060	1.117	4.2×10^{-10}	1.129	1.088	1.172	1.7×10^{-10}
WTCCC								
Cox PH model (cases and controls)								
Adjusted *(GRS+ Covariates)*	1.073	1.060	1.086	1.5×10^{-29}	1.103	1.083	1.123	2.9×10^{-26}
Proportional odds model								
Adjusted *(GRS+ Covariates)*	1.102	1.086	1.118	3.5×10^{-38}	1.144	1.120	1.169	8.6×10^{-35}
Binary logistic regression model								
Adjusted *(GRS+ Covariates)*	1.104	1.088	1.121	2.4×10^{-38}	1.146	1.121	1.171	1.1×10^{-34}

Descriptions: *ES: effect size which refers to the HR for Cox PH model and OR for the logistics and proportional odds models; GRS: genetic risk score; BMI: Body Mass Index; PC1-PC3: Principal Components; CI: confidence interval*

3.3.4. | Association of BMI with AOO of T2D

Given that overweight and obesity, measured by BMI, is an important contributing factor to the onset of T2D (discussed in section 3.1.1 and 3.2.3.2), the focus of this section was to assess the impact of BMI after adjustment for one of the four versions of the T2D GRS. Model adjustment includes confounding factors sex, ancestry principal components to account for population structure and one of the weighted or unweighted GRS.

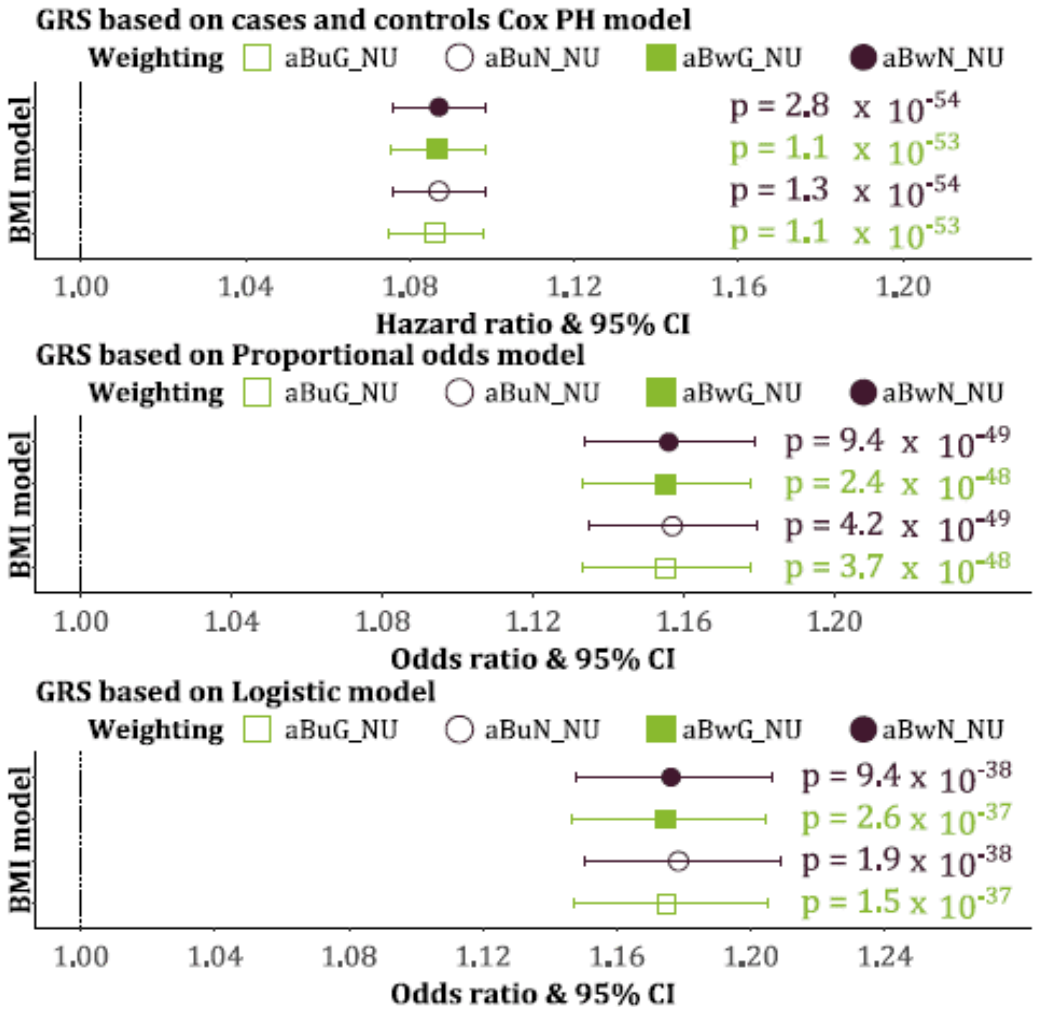


Figure 3. 4 - Comparison of estimated ES of AOO of T2D associated with the BMI based on three analytical methods for NUgene sample.

The x-axis indicates the HR for the Cox PH model and the OR for the logistic and proportional odds model and 95% CI for each BMI model shown on the y-axis. The models have been adjusted for sex, ancestry principal components to account for population structure, and GRS. The four models considered include the adjusted BMI model with weighted nominal significant GRS denoted (aBwN_NU); adjusted BMI weighted genome-wide significant GRS denoted (aBwG_NU); adjusted BMI model with unweighted nominal significant GRS denoted (aBuN_NU) and adjusted BMI unweighted genome-wide significant GRS denoted (aBuG_NU). GRS at the genome-wide level denoted (green) and GRS at the nominal level denoted (purple).

Represented in Figure 3.4 are model estimates based on the three analytical approaches. It compared the estimated HR and OR of AOO of T2D or T2D status associated with BMI while considering the impact of one version of the weighted or unweighted GRS. The models were adjusted for sex and population structure (via PCs). In all three approaches BMI was found to be highly significantly associated with AOO of T2D or T2D status. In general, for all four versions of the BMI models, the difference in performance was marginal. However, the unweighted nominally significant GRS was found to be the most strongly associated with AOO of T2D and T2D status. Additionally, for all four versions of the BMI models, the Cox PH model was found to be most strongly associated, followed by the proportional odds model among the three analytical approaches considered. The P-value associated with the unweighted nominally significant BMI based on the Cox PH, proportional odds, and logistic models was $p = 1.3 \times 10^{-54}$ with HR 1.09 (CI 1.08 – 1.10); $p = 4.2 \times 10^{-49}$ with cumulative OR 1.16 (95% CI 1.14 – 1.18) and $p = 1.9 \times 10^{-38}$ with OR 1.18 (95% CI 1.15 – 1.21) respectively (Appendix B Table B.3.2).

3.3.5. | Variance in AOO of T2D explained

This section considered the extent to which the variance in the AOO of T2D or T2D status that can be attributed to the T2D GRS and similarily for BMI. Given the maximun value of 1 property, explained variance was evaluated on the basis of the Nagelkerke pseudo R^2, a commonly applied pseudo R^2 measure in genetic research. The Nagelkerke pseudo R^2 measure was used to facilitate comparison between the four versions of T2D GRS distinguished by weighting and significance threshold for included SNPs.

3.3.5.1. | Variance in T2D Status and AOO of T2D explained by GRS

Figure 3.5 depicts the proportion of variance in AOO of T2D and T2D status that is explained by GRS, as measured by the Nagelkerke pseudo R^2, based on the three analytical approaches. In general, it seems that of the four versions of the T2D GRS considered, a weighted T2D GRS based on the nominally signifcant SNPs explained the highest proportion of variance in the onset of T2D due the T2D GRS. Furthermore, both weighted T2D GRS explains a higher proportion of the variance in T2D onset or AOO of T2D when compared to the unweighted T2D GRS. Based on the logistic model in the WTCCC dataset, the observed proportion of variance in T2D status

due to the weighted nominally signifcant GRS was 8.8%, which compares to 7.1%, 6.8% and 6.1%, respectively, for the weighted genome-wide significance GRS, unweighted nominal significance GRS, and unweighted genome-wide significance GRS adjusted models. In the NUgene sample, the proprtion of variance in T2D status in the logistic model was estimated to be 4.9% and 4.8 %, respectively, for the weighted nominal significance and genome-wide GRS models, and for the unweighted nominal significance and genome-wide GRS models was 3.8% and 4.0%, respectively.

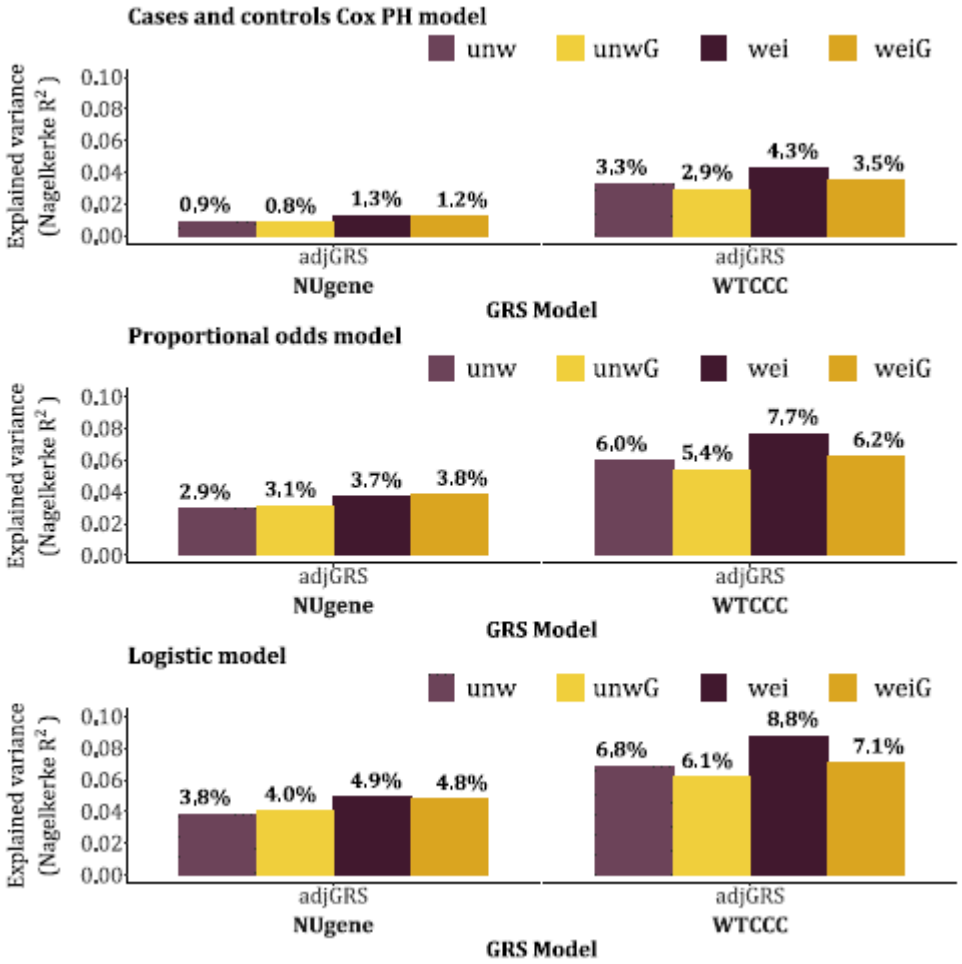


Figure 3. 5 - Proportion of variance in AOO of T2D explained by GRS in NUgene a and WTCCC samples based on Nagelkerke R^2

The proportion of variance measured by the Nagelkerke pseudo R^2 measure is shown on the y-axis while the effects of the nominal significance weighted (wei: dark purple) and unweighted (unw: light purple) GRS models as well as the genome-wide significance weighted (weiG: gold) and unweighted (unwG: yellow) GRS models are shown on the x-axis. The models which have been adjusted for potential confounding (adjGRS), the proportion of variance explained represents the difference in R^2 after adjustment for confounding variables where the full model (model with confounding variables and GRS) was compared to a reduced model (model with confounding variables only, GRS is excluded in this model). Models adjusted for sex, ancestry principal component to account for population structure and BMI.

3.3.5.2. |Variance in AOO of T2D explained by BMI

Figure 3.6 depicts the proportion of variance in AOO of T2D that was explained by BMI in the NUgene sample. It was observed that the difference in performance between the BMI models based on the four versions of GRS was marginal for all three approaches. Therefore, there is no evidence that the proportion of variance in AOO of T2D attributable to BMI is substantially impacted by GRS weighting. Based on the logistic model the observed proportion of variance in AOO of T2D due to the BMI model adjusted with the unweighted nominal significance GRS was 25.3%, which compares to 24.3%, 24.4% and 23.9% respectively for the unweighted genome-wide significance GRS; weighted nominal significance GRS; and weighted genome-wide significance GRS adjusted models.

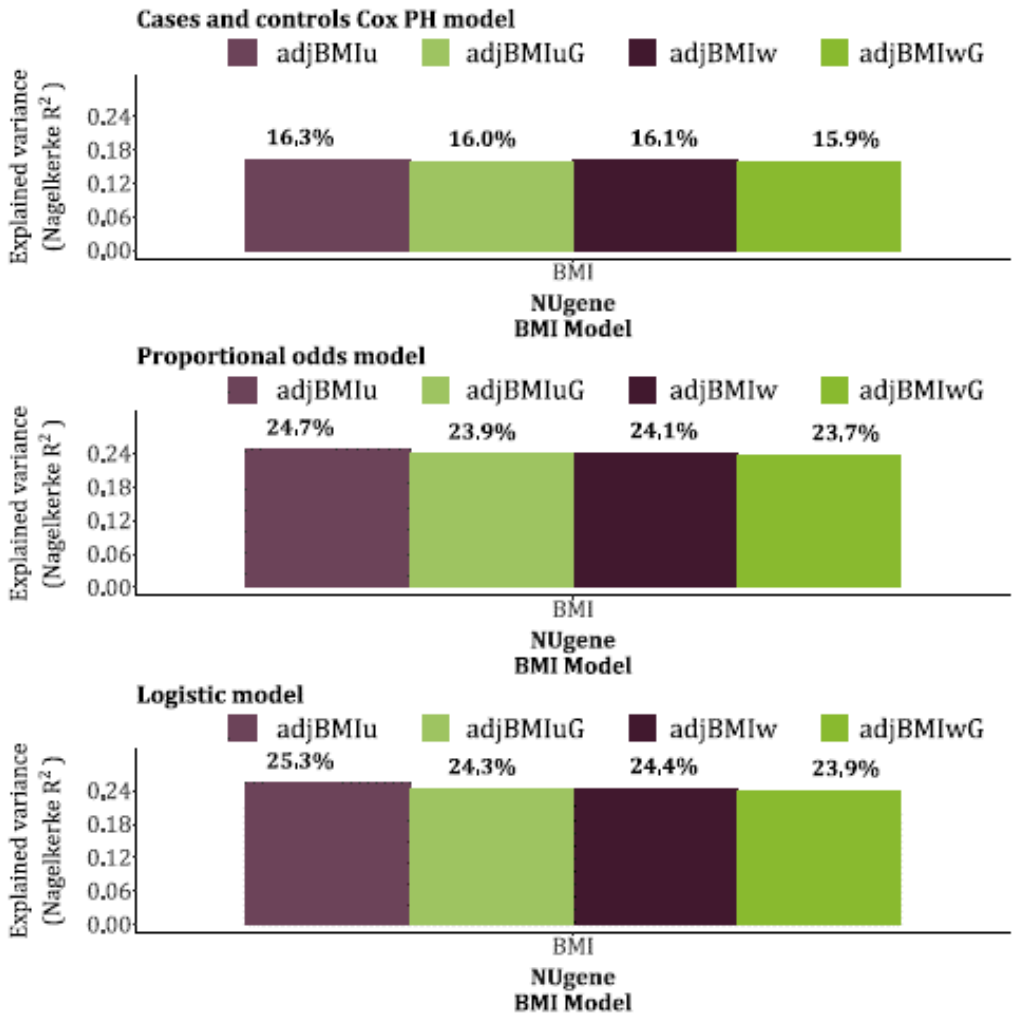


Figure 3. 6 - Proportion of variance in AOO of T2D explained by BMI in NUgene based on Nagelkerke R^2

The proportion of variance measured by the Nagelkerke pseudo R^2 measure is shown on the y-axis while the effects of adjusted BMI with nominal significance unweighted GRS model (light purple); adjusted BMI with genome-wide significance unweighted GRS model (light green); adjusted BMI with nominal significance weighted GRS model (dark purple); and adjusted BMI with genome-wide significance weighted GRS model (medium green) is shown on the x-axis. The proportion of variance explained represents the difference in R^2 after adjustment for confounding variables where the full model (model with confounding variables and GRS) was compared to a reduced model (model with confounding variables only, BMI is excluded in this model). Models adjusted for sex, ancestry principal component to account for population structure and weighted or unweighted GRS.

3.3.6. | Assessing model assumptions

This section focuses of the assessment of the proportional odds assumption key for the application of the proportional odds models. In earlier analysis, it was noted that the distribution of T2D GRS across LAO and EAO cases was not as expected, i.e lowest mean GRS in LAO cases and highest mean GRS in EAO cases, particularly in the NUgene sample. Therefore, the multinomial logistic regression model as an alternative to the proportional odds model was fitted to check the validity of the proportional odds assumption, and the residual deviances compared as a formal test of the deviation from the proportional odds model assumption.

Presented in Table 3.9 are the results of a likelihood ratio test undertaken to assess the validity of the proportional odds assumption. The residual deviance is lower in the multinomial models which suggests that the multinomial model is a better fit when compared to the proportional odds model. The likelihood ratio test indicate that the p-value associated with each GRS model assessed is significant, and therefore the proportional odds assumptions is not valid for these models.

Table 3. 9 - Likelihood ratio test between the multinomial and proportional odds models

Characteristics	Residual Deviance		Likelihood Ratio Test		
	Deviance from Multinomial Model	Deviance from Proportional Odds Model	Degrees of Freedom	Chisq	P_value
NUgene					
Genetic Risk Scores (GRS)					
Weighted GRS (P <0.05)	1874.14	1910.13	5	35.99	9.6×10^{-07}
Unweighted GRS (P <0.05)	1886.11	1919.83	5	33.71	2.7×10^{-06}
Weighted GRS (P <5*10^{-8})	1875.64	1908.49	5	32.84	4.0×10^{-06}
Unweighted GRS (P <5*10^{-8})	1883.70	1917.57	5	33.87	2.5×10^{-06}
WTCCC					
Genetic Risk Scores (GRS)					
Weighted GRS (P <0.05)	5076.35	5094.90	5	18.55	2.3×10^{-03}
Unweighted GRS (P <0.05)	5128.77	5147.05	5	18.28	2.6×10^{-03}
Weighted GRS (P <5*10^{-8})	5122.96	5139.46	5	16.49	5.6×10^{-03}
Unweighted GRS (P <5*10^{-8})	5147.75	5164.38	5	16.63	5.3×10^{-03}

Descriptions: *Chisq: chi-square distribution*

3.3.7. | Combining estimates from individual T2D GWAS

This section discusses the finding from the meta-analysis conducted to combine estimates from the NUgene and WTCCC datasets. The combined estimate of the P-value associated with the log HR produced by the Cox PH model is first presented. This is followed by the combined estimated log OR originating from the logistic models. Given that the proportional odds assumptions were found to be invalid, the GRS models produced by the proportional odds models was not included in the meta-analysis.

3.3.7.1. |Combing P-value estimated from Cox PH model

Presented in Table 3.10 are the combined z-score of the NUgene and WTCCC datasets originating from the Cox PH analysis associated with the HR of the four versions of the GRS. The results indicated that all four versions of the GRS were highly significant overall. The weighted nominally significant GRS was shown to have the largest combined z-score value.

Table 3. 10 – Stouffer meta-analysis of Cox PH model HR P-value

Study.ID	Sample size	Z-value	P-value	Weights
NUgene	1115	3.40	6.7×10^{-04}	35%
WTCCC	3810	12.93	3.0×10^{-38}	65%
All W(P<0.05)		12.85	8.6×10^{-38}	**100%**
NUgene	1115	2.39	1.7×10^{-02}	35%
WTCCC	3810	11.29	1.5×10^{-29}	65%
All uW(P<0.05)		10.89	1.3×10^{-27}	**100%**
NUgene	1115	3.55	3.9×10^{-04}	35%
WTCCC	3810	11.65	2.3×10^{-31}	65%
All W(P<5 x 10^{-8})		11.79	4.4×10^{-32}	**100%**
NUgene	1115	2.79	5.3×10^{-03}	35%
WTCCC	3810	10.60	3.0×10^{-26}	65%
All uW(P<5 x10^{-8})		10.48	1.1×10^{-25}	**100%**

Descriptions: *W: Weighted GRS; Uw: Unweighted GRS*

3.3.7.2. |Combing log OR estimated from logistic model

Figure 3.7 presents the combined estimate of the log OR based on the logistic model. Among the four versions of T2D GRS the CI of the individual studies overlap, suggesting a common effect. The magnitude of the log OR was found to be greatest for the weighted GRS based on genome-wide significant SNPs. The overall log OR for the genome wide weighted GRS was estimated to be 1.24 (which corresponds to an OR of 3.46). Additionally, as in the analysis of the individual datasets, the nominally significant weighted GRS were founded to be most strongly associated with T2D status. The pooled P-value associated with the nominally significant weighted GRS and genome-wide significant weighted GRS was $p = 1.6 \times 10^{-57}$ and $p = 3.0 \times 10^{-50}$, respectively.

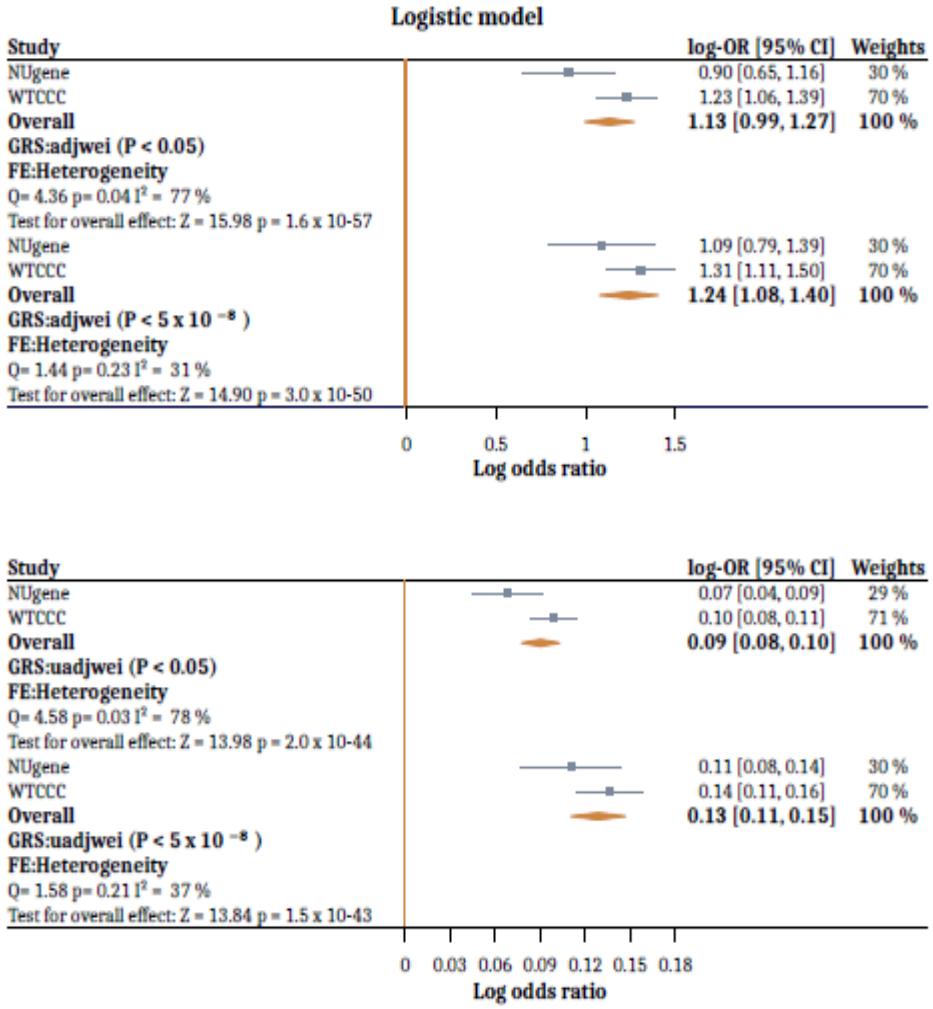


Figure 3. 7 - Fixed effect meta-analysis of estimated OR of onset of T2D associated with T2D GRS based on the logistic model(adjusted)

The x-axis indicates the log OR and 95% CI for each individual dataset(green) and combined (orange) shown on the y-axis for the adjusted model

Abbreviations: adjwei: adjusted weighted GRS models; uadjwei: adjusted unweighted GRS model; FE: Fixed effect; Q: Cochran Q statistic; p: p-value for Cochran Q statistics; I^2: I-squared statistic.

The level of heterogeneity as measured by the Cochran Q and I^2 statistics for the logistic model indicated that there was no evidence of heterogeneity for the genome-wide weighted and unweighted T2D GRS (Q p-value 0.23 and 0.21 and I^2 31% and 37% respectively). However, the level of heterogeneity reported for the nominal significance T2D GRS was found to be nominally significant (Q p-value 0.04 and 0.03 and I^2 77% and 78% respectively for weighted and unweighted nominal significance T2D GRS.

3.4. | GRS simulation methods

The methods applied to further assess the performance of the GRS via a simulation study is described in this section. Included is an outline of the scenarios explored, the Cox PH model used to generate the simulated data, along with a general description of the overall simulation process.

3.4.1. | Description of simulation study of GRS

This simulation study was undertaken as a result of persistent scepticism regarding the relative performance of the Cox PH and logistic models, coupled with preliminary findings from the NUgene and WTCCC datasets which indicated that the logistic model may have greater predictive power when compared to the Cox PH model. Therefore, in this simulation study a disease of interest in a sample of individuals ascertained from a homogeneous single ancestry population was considered along with the impact of GRS on AOO of disease. It is assumed that individuals in the study are followed for a specified period, with a record made of the age at which a disease occurs. Those that are unaffected at the end of the study period are considered as censored observations. It is also assumed that the causal SNPs have been correctly identified.

A range of scenarios were considered comprising of genetic and TTE parameters (see Appendix B.5 Figure B.5.1). The genetic component described the log HR of the GRS derived from genotypes of individual unweighted causal SNPs, under an additive model in the number of risk alleles, where each SNP is assumed to contribute equally to the risk of the onset of disease. It was further assumed that the RAF was the same among all SNPs. Primarily the number of SNPs included in the GRS, which ranged between 1 and 25 and the RAF are varied in the scenarios considered.

Finally, the TTE component described: (i) the TTE model (Cox PH); and (ii) the baseline hazard (see Appendix B.5 Figure B.5.1). Here, the Cox PH, which assumes a constant hazard over time resulting from, in this instance, the GRS forms the basis of the simulations. An important element of the simulation study was to evaluate the impact of censoring on the relative power between the Cox PH, logistic and proportional odds model, therefore, the censoring rates were varied. Initially, the baseline hazard rate and time (t), which refers to the study period was set

to achieve an equal number of cases and controls at the end of the study, i.e. a censoring rate of 50%. However, to assess the impact of censoring the t component was altered to achieve varying degrees of censoring.

3.4.1.1. |GRS simulation model

The Cox PH model outlined in Equation 1.9 of section 1.7.2 was applied to simulate AOO of disease. In this context, the component βX in the equation refers to the variable of interest, the GRS, where X represents the GRS and β the corresponding regression coefficient which is the log HR of the GRS. For the Cox PH, which is a special case of the general Weibull model, the HR is constant over time (i.e. shape parameter (ν) =1). As part of the simulation process, the baseline hazard rate and time (t), which refers to the study period or end of study time, was varied to achieve different rates of censoring. The scenarios considered were: (1) moderate censoring (50%); (2) low censoring (under 30%); and (3) censoring ranging from high (90%) to low (10%). To achieve the required censoring rate, the end of study time (t), was altered. For instance, varying the end of study time from 10 to 170 corresponded to changing from a high censoring rate of 90% to a low censoring rate of 10%.

In the range of simulation scenarios considered, the simulated GRS (independent variable) includes a range of values pertaining to the number of SNPs used to construct the GRS, SNP RAF, and log HR of the associated GRS. The values considered for the number of SNPs used to construct the GRS ranged from 1 to 25 SNPs. The values considered in relation to the log HR of the associated GRS included β values ranging from 0 to 0.30. For the associated RAF values considered included RAF of 0.05, 0.1, 0.25 and 0.5.

3.4.1.2. |GRS simulation process

The simulation encompassed three main steps, incorporating the two components of the simulation scenario where genetic and TTE data were generated, for each replicate. Each replicate of data comprised 1,000 individuals (detailed description of the three steps are outlined in Figure 3.8). In the first step, the genotype of individual causal SNPs was simulated under a binomial distribution, given a specified RAF. In the second step, the GRS, which is unweighted, is constructed for each individual. In the third and last step, conditional on the

GRS, the AOO of the disease was simulated under the Cox PH model, given the specified log HR of the GRS.

Generating genotype of individual SNPs: To begin the simulation process 1,000 datasets was generated based on the specified RAF for each causal SNP in a population of 1,000 individuals assumed to originate from a single homogenous population (see Appendix F for R code used to generate genotype of individual SNPs for each individual). The process involves: (1) specification of the RAF (with the risk and protective alleles denoted a and A, respectively); (2) generating simulated genotype for each simulated SNP based on binomial distribution (facilitated by the R rbinom() function).

Generating GRS of individuals: After the genotype for each SNP is determined, the GRS for each individual in the dataset is then constructed by summing the genotype values of each SNP. The process entails: (1) specification of the number of SNPs to be included in GRS; (2) summing the genotype values of each SNP based on the number of risk alleles present (0,1 or 2); and (3) GRS values were rescaled to have a mean of zero and standard deviation of 1 across individuals.

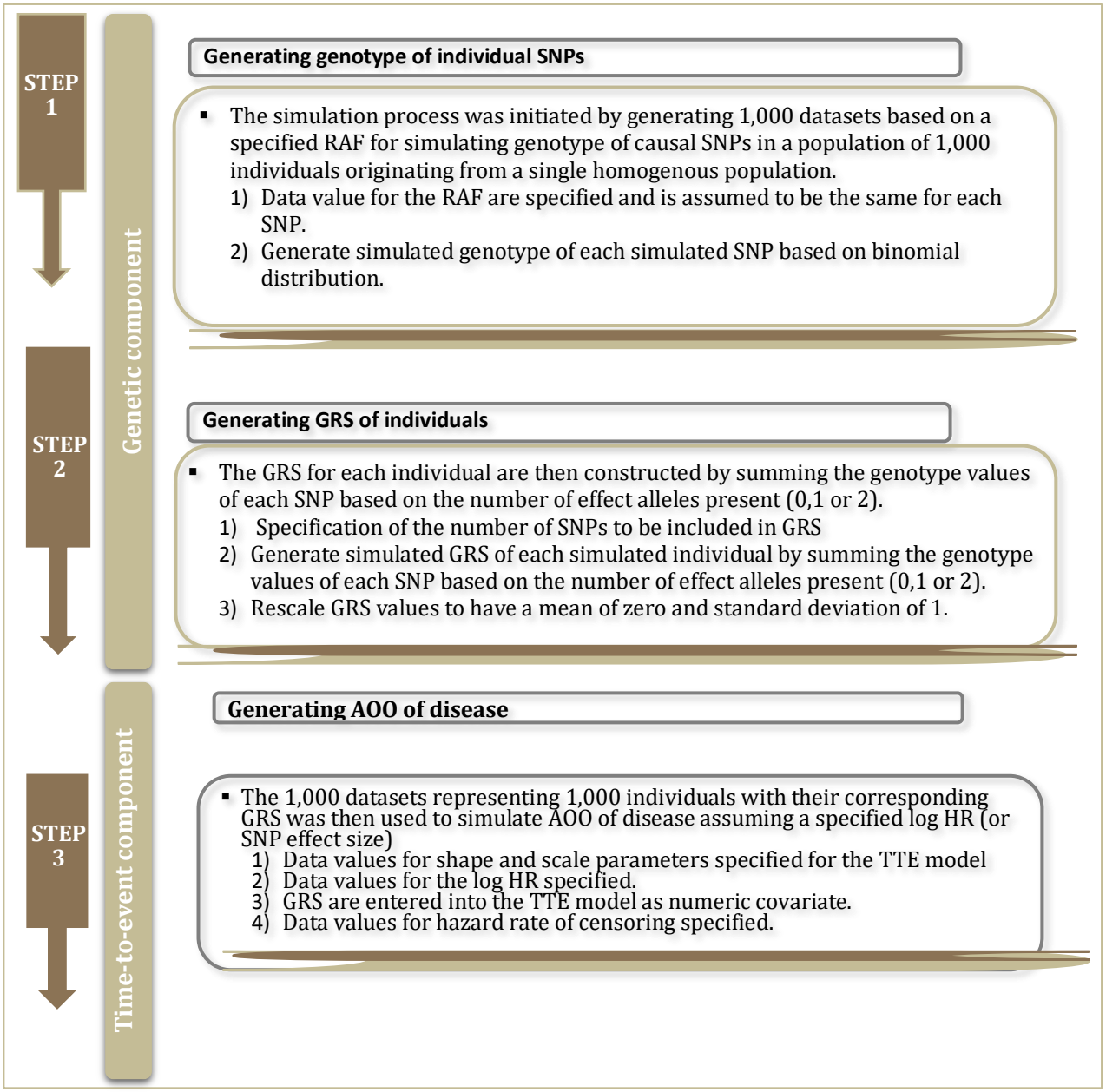


Figure 3. 8 - Description of data generating process of GRS

Generating AOO of disease: Following the simulation of the genotypes of each individual in the datasets associated SNPs, included in the construction of the GRS, the AOO of disease conditional on the GRS was simulated assuming a range of log hazard ratios (HRs) (ranging from 0.05 to 0.30). The process entailed; (1) based on the Weibull distribution, specification of the shape parameter value, which was set to l to achieve a constant HR for the Cox PH; (2) specification of the log HRs associated with the GRS; (3) GRS entered into the TTE model as a

continuous explanatory variable; (4) and specification of the baseline hazard rate which was based on the value BH= [-50/log(0.5)] to intially achieve a censoring rate of 50%.

3.4.2. | Statistical analysis of simulated GRS data

The simulated GRS data was also assessed using the three different outcome measures, TTE, ordinal and binary outcomes which have been described in detail in section 3.2.5.1. A similar process of data analysis was undertaken as in the analysis of the T2D GWAS datasets. However, the scope of the simulation study did not incorporate adjustment of the GRS for potential confounding factors. As a result, the analysis considered the GRS as a single covariate in the regression models. The disease status of individuals in the simulated datasets was based on their simulated disease status value where a value of one represented case (affected by the disease) and zero control (unaffected by the disease at end of study period). In the case of the proportional odds model, the cut-off age used to define EAO in each GRS simulation scenario were based on 50% of the study period. For example, if time t=30, the EAO cut-off age would be (< 15).

As in the case of the T2D GWAS datasets, data analysis was also conducted using the function coxph (R package survival), polr (R package MASS), and glm (R package stats) for the Cox PH, proportional odds and logistic model respectively. The relative power between the Cox PH, logistic regression, and proportional odds models to detect an association with GRS was assessed at a nominal level of significance ($p<0.05$). Outputted P-values associated with the estimated HR or OR were used to evaluate relative power between the models. A description of the models fitted are outlined in Appendix B.

3.5. | GRS simulation results

The GRS simulation study findings are discussed in this section. The GRS simulation study focused for the most part on evaluating the relative power of the Cox PH, logistic and proportional odds models considered to detect an association between GRS and AOO of disease. As differences in performance of the Cox PH and logistic model have been attributed in part to the rate of censoring, the scope of the study has focused mainly on the impact of censoring on power. The first section (3.5.1) evaluates power in a setting where there is a balance of cases and controls and where the logistic model is expected to be most powerful (i.e. 50% cases and 50% controls). The second section (3.5.2) examines the impact of both the censoring rate and GRS effect size. Here the impact of low censoring rates was evaluated. In the third section (3.5.3) the impact of censoring was again considered along with variation in the number of SNPs contributing to the GRS. Here the impact of high, moderate, and low censoring was evaluated. The scope of simulations incorporated specifically three primary characteristics pertaining to the GRS; (1) RAF; (2) Effect size (ES); and (3) number of SNPs from which the GRS was constructed.

3.5.1. | Impact of SNPs in GRS on power in presence of moderate censoring

Simulating under a Cox PH model, the power to detect association of a GRS with AOO of disease as a function of the number of SNPs included in the GRS is presented in Figure 3.9. A log HR of 0.05 (HR=1.05) attributable to the GRS is assumed throughout the simulations. A perfect balance in terms of the number of cases and controls assessed is assumed (i.e. a censoring rate of 50%), where the logistic model is expected to be most powerful. Different parameter settings pertaining to the RAF, which is assumed to be the same for all SNPs included in the GRS, are presented across the three plots. In general, it was observed that power, which was evaluated at the 5% significance level, increases in relation to the number of SNPs used to generate the GRS which is as anticipated. It was also noted that larger RAF among SNPs within the GRS resulted in greater power to detect an association with AOO of disease. Therefore, the combination of more SNPs with larger RAF within a GRS results in greater power to detect an association with AOO. In this setting where there is a balance regarding cases and controls little

difference in power is observed between the three models (Appendix B.5 Figure B.5.2 -B.5.3), however, the Cox PH model is marginally better, particularly for small effect sizes.

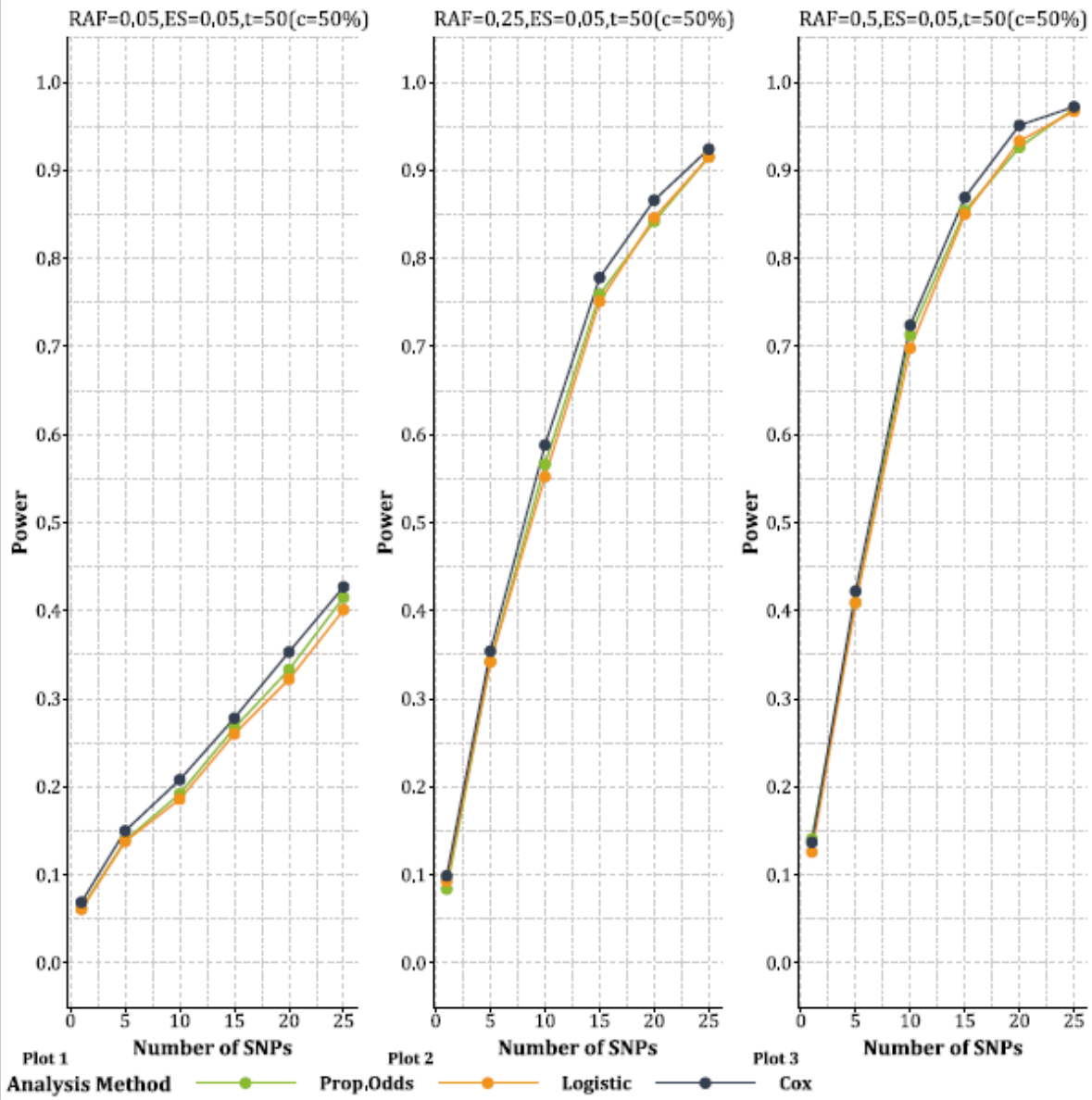


Figure 3. 9 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming an ES of 0.05

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox PH model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; ES: GRS effect size.

3.5.2. | Impact of effect size of GRS on power in the presence of low censoring

Figure 3.10 presents the power to detect association of a GRS with AOO of disease as a function of effect size. Different parameter setting for the number of SNPs included in the construction of the GRS is presented across the four plots. The simulations assumed a RAF of 0.05 which was also assumed to be the same for all SNPs included in the GRS, as well as a low censoring rate of 10%. A difference in performance in favour of the Cox PH model can be distinguished for smaller effect sizes and fewer SNPs contributing to the GRS. A similar trend is also observed in relation to low censoring rates of 20% and 30% (Appendix B.5.4 and B.5.5).

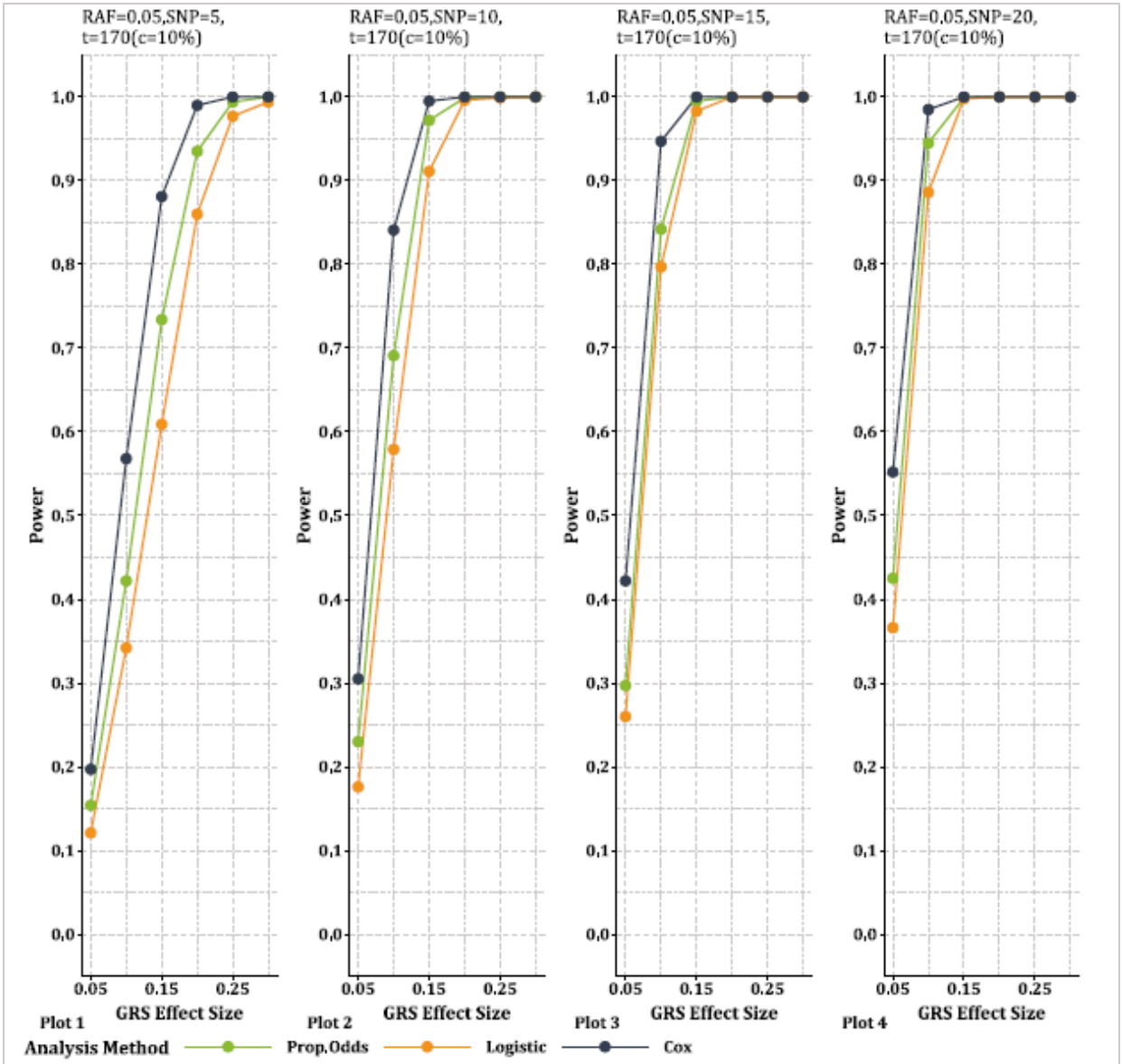


Figure 3. 10 – Power to detect association of GRS with AOO of disease as a function of the GRS effect size assuming a RAF of 0.05

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox PH model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; SNP: number of SNPs included in the GRS calculation; t: study period (follow-up time) c: censoring rate.

3.5.3. | Impact of SNPs in GRS on power in the presence of high to low censoring

Presented in Figure 3.11 the power to detect association of a GRS with AOO of disease as a function of the number of SNPs included in the GRS. A log HR of 0.05 (HR=1.05) attributable to the GRS was assumed throughout the simulation. Additionally, a RAF of 0.05 was also assumed for all SNPs included in the GRS. Different parameter settings pertaining to the censoring rate are presented across the four plots. The plots illustrate that for very high levels of censoring the performance of the three models was relatively similar, however, when there was a perfect balance of cases and controls the Cox PH was marginally better than both the logistic and proportional odds models. On the other hand, when censoring levels were low, a marked difference in favour of the Cox PH model can be observed (See also Appendix B.5.6 – B.5.9).

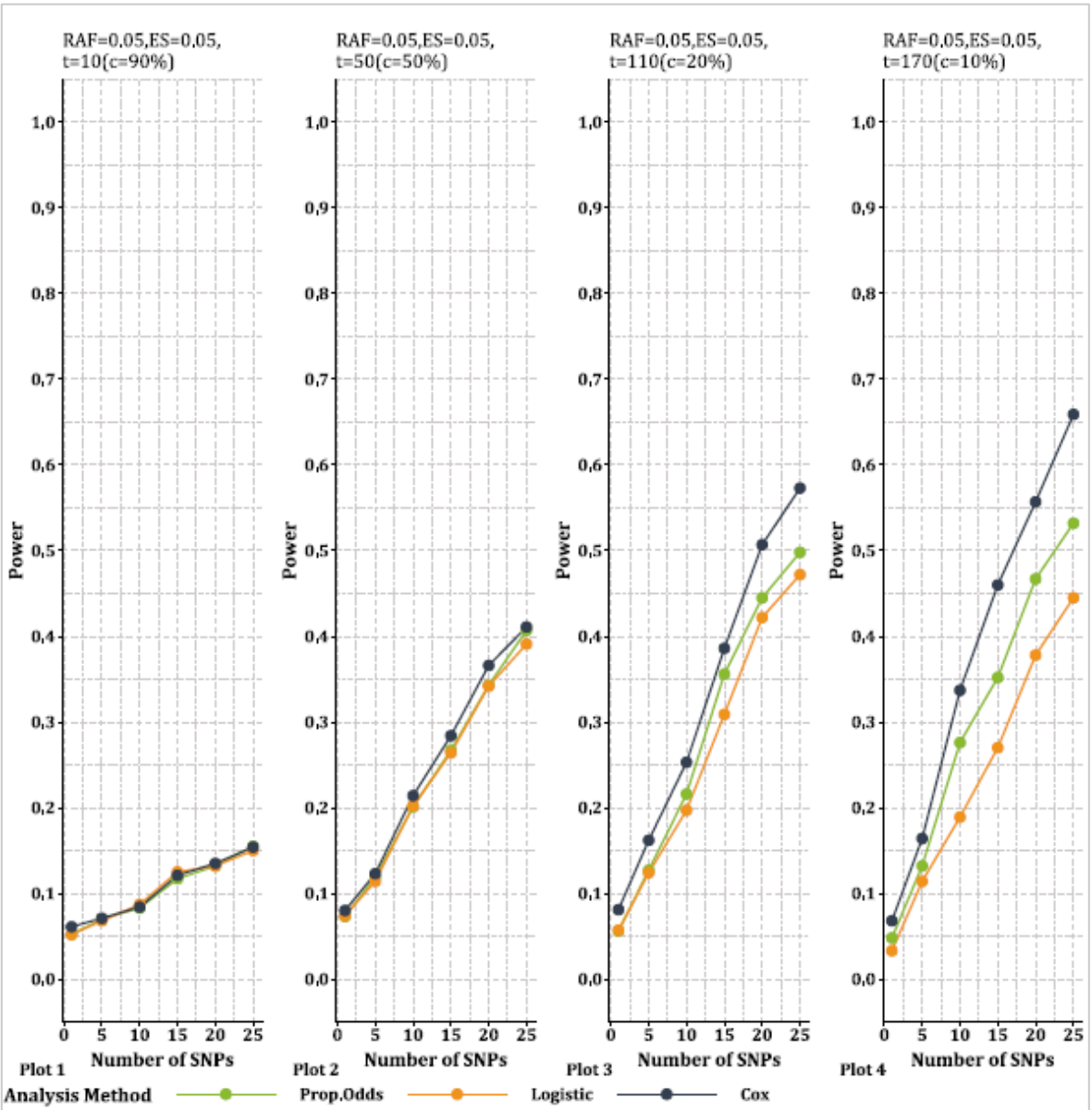


Figure 3. 11 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming a RAF of 0.05 and ES of 0.05

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox PH model with GRS as covariate (navy blue).
Abbreviations: RAF: risk allele frequency; ES: GRS effect size; t: study period (follow-up time) c: censoring rate.

3.6. | Discussion

GWAS have facilitated numerous discoveries associated with many common complex diseases which subsequently lead to the application of GRS. GRS simultaneously assesses overall genomic risk at the individual level and therefore have the potential to predict disease risk in individuals. The GRS approach has been employed to investigate the association of T2D GRS with AOO of T2D with the view to identifying the most powerful of the three statistical approaches considered, binary, ordinal and TTE outcomes framework. This was coupled with a GRS simulation study to further assess the relative statistical power between the three outcome measures. A summary statistics meta-analysis to combine the results of the two datasets was also undertaken.

In general, the logistic model performed better than the Cox PH model, which was somewhat surprising giving the expectation that EAO cases would exhibit a greater genetic burden than LAO cases. Results from the simulation study seem to indicate that high rates of censoring results in relatively similar performance between the methods. However, the Cox PH seemed to have the advantage, in terms of power, in a setting where there are very low rates of censoring. In the analysis of the NUgene and WTCCC datasets, which reflects high rates of censoring 54% and 76% respectively, it was found that the models based on the logistic method performed better in both circumstances. A likely contributing factor is the fact that the SNPs and weights applied in the construction of the GRS originated from models based on the logistic regression approach. As a result, the analysis may be biased in favour of the logistic model. Moreover, as the data used in case control studies are collected retrospectively, such data are likely to be subjected to recall bias, particularly when it comes to AOO or age at first diagnosis (particularly for patients that have relocated, or instances where medical records are unavailable). Furthermore, there are likely to be inconsistencies in the measurement of a characteristic like age particularly among cases and controls. Consequently, these limiting factors may have negatively impacted the TTE analysis.

A general finding from the three modelling approaches indicates that the utility of the T2D GRS to detect an association with T2D status under a logistic regression model was substantially better when compared to the Cox PH model, which assessed the utility of the T2D GRS to detect an association with AOO of T2D. Consideration was also given to the proportional odds

modelling framework which encompassed LAO and EAO, however, the proportional odds assumption was found not to be valid in both the NUgene and WTCCC datasets. Of the four versions of the T2D GRS considered, the weighted nominally significant GRS was found to be the best predictor of the onset of T2D based on strength of association, as measured by the p-value, with AOO of T2D and proportion of variance explained by the model. However, it was also observed that greater effect sizes in relation to AOO of T2D or T2D status were attributed to the weighted genome-wide GRS models relative to the nominally significant GRS models. This translates to a smaller number of SNPs included in the GRS but which are more strongly associated with T2D and generating a greater effect on risk of T2D or AOO of T2D.

As anticipated, BMI significantly impacts T2D status and AOO of T2D. Controlling for the effects of the GRS also increases the effect size associated with T2D attributable to BMI but not substantially. Furthermore, it was noted that controlling for the effects of GRS using weighted or unweighted GRS produced similar results, however, the nominally significant unweighted GRS were shown to have the greatest impact of the four versions of the T2D GRS considered.

The nominally significant weighted GRS was found to be most strongly associated with T2D status based on the logistic regression model in both the meta-analysis and individual analysis of the NUgene and WTCCC studies. Furthermore, assessment of heterogeneity in the meta-analysis indicated that there was evidence of heterogeneity for the nominally significant weighted and unweighted T2D GRS. However, there was no evidence of heterogeneity for the genome-wide significant weighted and unweighted T2D GRS. These results are consistent with findings from the statistical tests of differences in the mean GRS as the mean GRS of both the weighted and unweighted nominally significant GRS of cases in the NUgene study were significantly different from the cases in WTCCC study. A key limiting factor particularly for a clinically heterogeneous disease, such as T2D, is potential differences in the clinical definition of disease that may be applied in different studies or in different countries. An International Classification of Diseases (ICD) -9 codes for T2D or laboratory evidence of hyperglycemia and prescribed T2D medication formed the basis of selection of cases in the NUgene study. However, current prescribed medication for T2D and historical or contemporary laboratory evidence of hyperglycemia formed the basis of selection of cases in the WTCCC study.

In conclusion, the results of this analysis have provided further insight regarding the utility of T2D GRS to detect an association with AOO of T2D in European ancestry populations. However, the main challenge affecting the clinical implementation of GRS is their application to diverse global populations. This is particularly problematic for T2D as T2D has global impact. As a result, Chapter 4 builds on the work of this chapter as it examines the utility of the T2D GRS in diverse populations.

Chapter 4: Investigating the utility of genetic risk scores to detect an association with age-of-onset of disease in European, Asian, and African descended populations

Chapter Outline

Currently, the main challenge surrounding the clinical implementation of genetic risk scores (GRS) is that they are of far greater predictive value in European descended populations when compared to other populations. This in part is because European derived GRS are optimized to capture common variants with higher minor allele frequencies (MAF) on average in Europeans compared to non-Europeans. Also contributing to differences in performance is the fact that different ancestral populations tend to differ in respect to risk allele frequencies (RAF) and patterns of linkage disequilibrium (LD) structure. To construct non-European GRS, European derived discovery single nucleotide polymorphisms (SNPs) is often used which given the present methodology is not as accurate at predicting disease risk in non-Europeans when compared to Europeans. However, as non-European GWAS are often not large enough to produce powerful GRS, European derived SNPs are often used to construct non-European GRS. This chapter extends the work of the previous chapter, focusing on investigating the utility of GRS to detect an association with age-of-onset (AOO) of disease in ancestrally diverse populations. The first aspect entails the application of GRS to investigate the association of AOO of type 2 diabetes (T2D) using data from the UK Biobank, where application of a European ancestry derived GRS in a European ancestry population was again explored. Application of the European derived GRS was also extended to non-European populations, which comprised individuals of Asian and African descent from the UK Biobank where the relative performance of the GRS in each ancestry was compared. A second aspect of the work involved further assessment of the utility of GRS in diverse populations where the focus was on identifying the conditions within each ancestry that are likely to influence the degradation in performance of a European ancestry derived GRS in non-European populations.

4.1. | Introduction

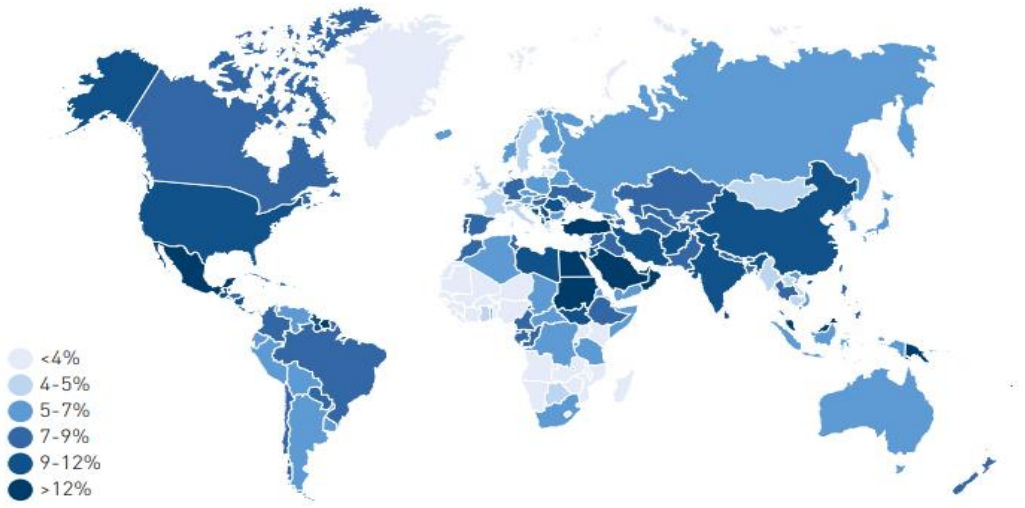
"Delivering the right treatments, at the right time, every time to the right person" were words spoken by President Barack Obama in 2015 in his State of the Union Address in reference to precision medicine. Precision medicine is broadly defined as the use of diagnostic tools and treatments targeted to the needs of the individual patient on the basis of genetic, biomarker, or psychosocial characteristics [243]. By 2018, echoes of *"A treatment plan like this — tailored to an individual's genetic risk — is one of the great promises of precision medicine" [244].* Gene discoveries resulting from common disease GWAS have aided the emergence of GRS as a potential biomarker for predicting risk for many common diseases. Currently, however, the main challenge surrounding the clinical implementation of GRS is that they are of far greater predictive value in European descended populations [190, 245]. The performance of GRS which are primarily based on SNPs derived from GWAS undertaken in European ancestry populations are affected by RAF and LD that differ among different ancestral populations. These differences, RAF, and LD structures, as well as differences in effect sizes across populations which is impacted by LD structure, overall have implications for statistical power. Therefore, to pinpoint the causes of diseases more effectively or quickly, broadening the scope of scientific inquiry to acquiring a better understanding of genomics in all populations on a global scale may be the key to improving disease risk prediction for people of all ancestries. To this end, the utility of a European ancestry derived GRS to detect an association with AOO of T2D in European, Asian, and African populations was explored.

4.1.1. | Global Impact of T2D

T2D has now attained the status of a global pandemic, affecting all regions around the world [246]. The prevalence of diabetes on a global level has been increasing over recent decades. Estimates for 2017 based on the International Diabetes Federation (IDF) diabetes atlas, standardized for the age group 20-79 years, estimated global diabetes prevalence to be 8.8% (95% confidence interval 7.2-11.3%) [247]. However, estimates from 1980, based on adults over the age of 18, reported the global prevalence of diabetes at 4.7%, which increased to 8.5% in 2014 [181].

The IDF diabetes atlas classifies the world population into 7 regions that comprise: Africa (AFR); Europe (EUR); Middle East and North Africa (MENA); North America and the Caribbean

(NAC); South and Central America (SACA); South East Asia (SEA) and Western Pacific (WP). It ranked NAC number 1 based on prevalence of diabetes for persons aged 20 to 79 (Appendix C Table C.1.1) while MENA was ranked 2. Furthermore, the region of AFR, though ranked lowest in terms of prevalence, experienced the largest proportion of all deaths due to diabetes occurring before age 60 (Appendix C, Table C.1.2). A global view of prevalence is illustrated in Figure 4.1.



Source: International diabetes federation (IDF) diabetes atlas, eighth edition 2017

Figure 4. 1- Estimated age-adjusted prevalence of diabetes in adults (20-79 years), 2017

The countries that contained the largest number of adults living with diabetes worldwide include China and India (ranked 1 & 2 respectively), both of which form part of the SEA region. Other countries that made up the top ten countries containing the largest number of adults with diabetes include: the United States of America (USA, part of NAC region); Brazil and Mexico (part of SACA region); Indonesia (part of the SEA); the Russian Federation and Germany (part of EUR region); and Egypt and Pakistan (part of MENA region). A global view of the estimated

total number of adults (20-79 years) living with diabetes is illustrated in Appendix C Figure C.2.1.

4.1.2. | Ancestral and geographic composition of individuals in GWAS

At the inception of GWAS, the practice of conducting studies in ancestrally homogeneous populations was established. This course of action was taken primarily because of concerns about the validity of GWAS findings due to unrecognized population structure within heterogeneous populations. LD the mechanism that underpins GWAS methodology, tends to vary between ancestral groups, which adds to the complexity of undertaking GWAS in diverse geographical populations. In association testing, population structure has the potential to cause inflated type I error rates (false positives) due to geographical confounding between the disease and SNPs. Simultaneously inflation in the type II error (false negatives) may also occur resulting from LD differences among ancestries which can potentially bias estimates of effect sizes [248]. Due to these concerns regarding population structure, the application of GWAS was primarily based in homogenous populations. In 2009, 96% of participants in GWAS studies were of European ancestry and, more recently, this figure was estimated to be 88% (2017) [249]. In the long run, having GWAS discoveries based primarily on European ancestry populations (people of European ancestry are estimated to represent only 16% of the world population) may present problems particularly for diseases like T2D that have global impact. The lack of GWAS in geographically diverse ancestral populations has the potential to negatively impact the implementation of precision medicine as segments of the world population may be unable to benefit from the clinical or therapeutic advances stemming from such research [250]. This is the case owing to the likelihood of a biased picture emerging regarding which variants are important in relation to disease risk. Moreover, increasing diversity of geographic ancestry is critical to prevent genomics from further contributing to healthcare inequalities [251]. Treatment of asthma and cardiovascular disease are two examples of commonly available medications that were found not to work as effectively in non-European ancestry populations [252].

4.1.3. | Recognising the benefits of ancestral diversity in GWAS

In recent years, there has been increasing recognition that, although ancestral diversity presents many challenges, it also provides many opportunities in the context of gene discoveries. *"Its not that people of different ethnic backgrounds have wildly different biology. Its more subtle, and fascinating that, we need to explore the vast range of human genetic variation: It could end up saving us all"* [244]. Furthermore, in 2013, a life science research associate at Stanford University, Andres Moreno-Estrada, MD, PhD was also quoted as saying, *"If we don't understand the origin of our genetic variants, we won't be able to design personalized, or even population-level medicine"* [253].

Ancestral diversity in GWAS is now regarded as key to providing more accurately targeted therapeutic treatments to more of the world population [254]. The additional benefits noted include extended insight underlying the genetic architecture of diseases; greater capacity to uncover rare variants with significant effect sizes (as rare variants in European populations may exist at higher rates in other ancestries), particularly in isolated populations [249]. There is also the possibly that variants that are monomorphic (a SNP is defined as monomorphic in a population if only one allele occurs at a site or locus [255]) in European populations may be present in non-Europeans.

4.1.4. | Application of GRS in ancestrally diverse populations

Current research has illustrated the potential of GRS to improve risk prediction for common diseases in the long term. However, a major shortcoming of GRS is that they are often derived of European ancestry populations and as such are therefore optimized for use within this population. Consequently, owing to the bias of GWAS to European ancestry populations, GRS tend to perform sub-optimally in other non-European ancestry populations [256]. This diminished predictive power of GWAS is due in part to differences in the pattern of LD among ancestry groups, which in turn drives differences in effect size estimates across ancestry groups [257]. Adding to this is the tendency for the risk allele associated with most significant SNPs to be more common in the population in which it was discovered. It has been noted that GWAS catalogue variants are on average more common in European descent populations when compared to Asian and African descent populations [257]. While discovered variants may be

common in European populations, they may be rare in non-European populations, and thus have less predictive power.

Although the scope of common disease GWAS has extended to include both low frequency (MAF 1 – 5%) and rare variants (MAF < 1%), the fundamental framework which forms the basis of GWAS, "common disease" " common variant" has implications on the performance of the GRS across different ancestral populations. This is due to the fact that most GWAS discoveries have MAF > 5% in the discovery population. Furthermore, in some instances, a SNP may be monomorphic (defined in section 4.1.5) in a population. Moreover, some risk variants may be ancestry-specific, and therefore monomorphic in other ancestral populations [258]. For example, the T2D *AGMO* gene locus (rs73284431) identified in an African ancestry population is monomorphic in non-African ancestry populations [258]. This is likely to substantially impact the performance of the GRS in different populations.

Approaches to the construction of GRS in relation to LD is an importannt consideration. In GWAS, the identified associated SNP is most often not the causal variant but rather a tag SNP in LD with the causal variant. Therefore, variation in LD structure between ancestry groups may result in ancestry-specific tagging of the same causal SNP. Moreover, the process of LD pruning in European ancestry populations potentially may result in the removal of non-European ancestry-specific tag SNPs. Consequently, the applied LD pruning strategy has implications for the performance of the GRS, as the causal SNP may no longer be well tagged in non-European ancestry populations after the process of LD pruning.

Given the global impact of T2D, a more in-depth understanding of population specific characteristics, particularly as it relates to the variation in patterns of LD structure across ancestral populations and allele frequency is essential. In this chapter, the application of European-ancestry derived GRS to investigate the association of AOO of T2D in European and non-European ancestry populations is explored. The relative performance of the GRS in each ancestry is compared, where the non-European populations comprised those of Asian and African descent. To identify the conditions within each ancestry that are likely to influence the degradation in performance of a European ancestry derived GRS in non-European populations, further assessment of the utility of GRS in diverse populations was also undertaken.

4.2. | T2D GRS methods

In this section, aspects of the process undertaken to construct the T2D GRS used to assess the association of AOO of T2D are described. Included are details pertaining to the base GWAS (European ancestry T2D GWAS meta-analysis published at the end of 2018) from which the selected SNPs and their associated summary statistics were sourced [259]. The target GWAS (UK Biobank dataset) genotyped samples used to test the performance of the GRS is also described.

4.2.1. | Identification of disease-associated SNPs

Base GWAS: A study published at the end of 2018, which undertook locus discovery and fine-mapping in European ancestry T2D GWAS, formed the basis of the GRS that were constructed to investigate their association with AOO of T2D [259]. This study combined GWAS association summary statistics via meta-analysis from 32 European ancestry studies imputed to two different reference panels, 1 from a population-specific reference panel, while the others were imputed from the Haplotype Reference Consortium (HRC) reference panel. The study identified 243 genome-wide significant loci with odds ratios (OR) ranging from 1.04 to 8.05 and MAF ranging from 0.02% to 50%. The analysis also incorporated assessment of the impact of obesity, where models unadjusted and adjusted for body mass index (BMI) were evaluated [259].

Using the BMI unadjusted summary statistics from this published study, it was possible to acquire information for SNPs known to be associated with T2D. It was possible to extract the SNP IDs along with their corresponding p-value and effect size as measured by their OR values. In addition to SNP identifier details, other associated information collected included details pertaining to the effect allele (EA), alternative allele (NEA) and effect allele frequency (EAF) (Appendix C Table C.3.1).

Target GWAS: The utility of the T2D GRS in detecting an association with AOO of T2D was evaluated in European, Asian, and African descended populations using data originating from the UK Biobank. The UK Biobank is a prospective cohort study consisting of approximately 500,000 individuals recruited between 2006 and 2010 from across the UK. The data collected

from these individuals, who were aged between 40 and 69, included their DNA sample, along with demographic and health information. Provided health information relevant to this investigation included T2D status; sex; ethnicity, BMI; genotyping microarray; enrolment age in relation to controls, and in relation to individuals with T2D, AOO of the disease [260]. Genotyping of DNA samples was carried out using the Affymetrix UK BiLEVE Axiom Array and Affymetrix UK Biobank Axiom Array. Phasing and imputation were based on the HRC, 1000 Genomes phase 3 and UK10K reference panels [261]. A general overview of the characteristics of the UK Biobank dataset are given in Table 4.3. Before the dataset was used in data analysis, the samples were checked for the presence of related individuals or duplicate samples. As indicated in Chapter 1 section 1.4.3.3, the identity by descent (IBD) is a measure of how strongly pairs of individuals may be related. It is often used to remove related individuals or duplicate samples based on an IBD metric (pi-hat) threshold value > 0.1875 [25].

4.2.2. | Identifying individuals of European, Asian, and African ancestry

Individuals in the UK Biobank dataset who self-reported as being of European, Asian, or African descent formed part of the initial samples. However, to confirm ancestry, the first two principal components (PCs) along with self-reported ancestry were used to identify individuals deemed to be ancestry outliers. This approach is often used as at the continental level most of the variation in ancestry is explained by the first two PCs. The procedure generally used to construct PCs is described in detail in Chapter 1 section 1.5.2.

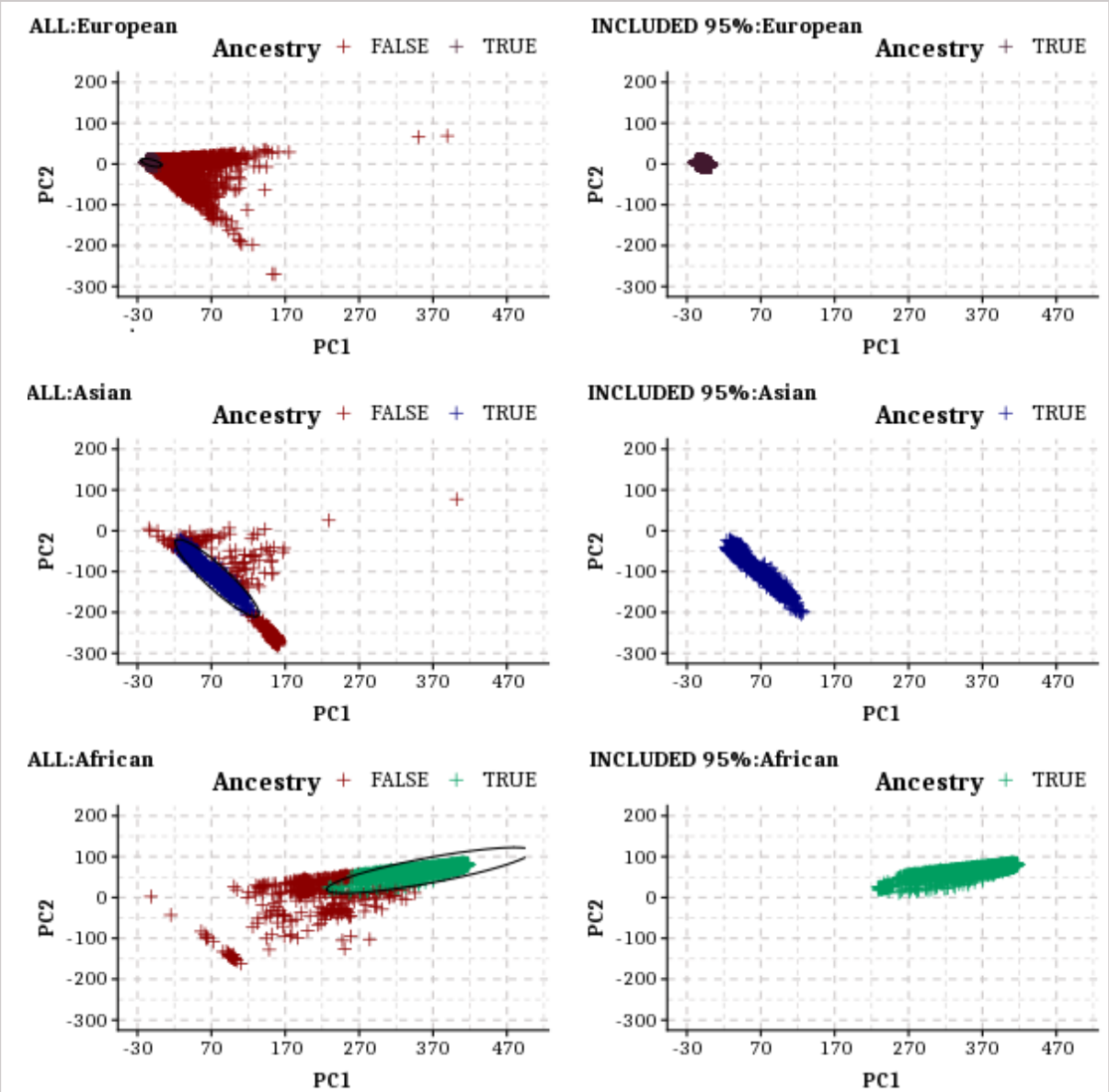


Figure 4. 2 - Identification of ancestry outliers based on first two principal components using 95% confidence levels

The x axis indicates the principal component 1, while the y axis indicates the principal component 2. Each individual is represented by a dot with the colours representing the selected and excluded (outliers) individuals in each population. Red (Outliers) selected (purple European; blue Asian; green African).

For each ancestry, the R function stat_ellipse from the R package "ggplot2" [262] was used to generate an ellipse of the data that was assumed to be normally distributed. The calculation was based on the mean of both PC1 and PC2 and the covariance matrix. The resulting ellipse contained a defined percentage, in this instance 95%, of the original data. From the ellipse, it was possible to identify the individuals that fall inside the ellipse and individuals that fall outside the ellipse (ancestry outliers). Individuals deemed to be ancestry outliers were removed from the sample (Figure 4.2).

4.2.3. | Development and construction of GRS

The 243 lead SNPs identified in the base GWAS were extracted from the UK Biobank dataset. At the start of the process, a check was made to ensure that the SNPs included in the target sample (UK Biobank data) match the selected SNPs from the base GWAS. The alignment of the EA of each SNPs was then checked. If the EA of the SNPs were not in alignment, the EA in the target GWAS was aligned with the base GWAS by flipping the EA in target GWAS to be the same as in the base GWAS. If the EA in the target GWAS was flipped, the corresponding genotype dosage for each individual was adjusted using the formula (2 – current dosage value). The formula used to calculate the GRS for each individual in the sample is outlined in section 1.6.3.1. Two different versions of the GRS were constructed, the first based on weightings derived from the effect size (log OR) in the base GWAS, and the second unweighted, which assumes that each SNP contributes equally to the risk of T2D.

4.2.4. | Statistical analysis

This section outlines the procedures undertaken in the analysis of the European, Asian, and African ancestry T2D genotyped GWAS datasets originating from the UK Biobank. This includes the main statistical methods and statistical software tools applied in the data analysis. As in Chapter 3, AOO of T2D was again analysed with the Cox PH, and logistic models, where within each statistical approach the pseudo R^2 measure was applied to assess the relative performance of nested models.

4.2.4.1. |Statistical analysis of individual T2D GWAS datasets

A similar process, as described in section 3.2.3.1, was undertaken in relation to the application of each analytical approach. The analysis for each ancestry group was conducted independently. The association of T2D GRS and AOO of T2D was again assessed in a TTE, and logistic regression framework where the relative performance of the T2D GRS in each of the ancestry groups was compared. In the case of the TTE analysis, based on the Cox PH model (described in Equation 1.9 section 1.7.2), two different analyses were carried out. As there are situations where the use of controls may not be appropriate, the first analysis entailed a case only analysis, where only individuals diagnosed with T2D were included in the model. The second analysis comprised both cases of T2D and controls, where controls were censored at their current age at the end of the study period. The hazard ratio (HR) of the GRS (predictor of interest in the model) for AOO of T2D was estimated using the function (coxph) of the R package (survival). Sex, BMI, and ancestry principal components (used to adjust for the effects of confounding due to population structure), and in the case of the European ancestry group, type of genotyping microarray, were included in the model as covariates. Adjustment for genotyping microarray was not necessary for the non-European ancestry samples as the single Affymetrix UK Biobank Axiom Array was used.

The second outcome measure relates to the binary logistic model described in section 1.7.4. Equation 1.12 illustrates the relationship of the logit-transformed probabilities associated with disease outcome which was modelled as a linear relationship with the predictor variable which in this instance was the GRS. To address confounding the variables listed above (Sex, BMI ancestry PCs, and type of genotyping microarray) were included as covariates in the model. Additionally, age was included as a covariate. The OR of the GRS for T2D was estimated using the function (glm) of the R package (stats).

Further analysis, where single SNP association analysis involving each of the 81 SNPs included in the T2D GRS with T2D were undertaken. The association of T2D status with each SNP independently were assessed using SNPtest in R and the association of AOO of T2D with each SNP independently using the function (coxph) of the R package (survival). All tests were adjusted for confounding (Sex, BMI ancestry PCs, and type of genotyping microarray included as covariates).

4.2.4.2. |Evaluating performance of T2D GRS models

As in Chapter 3, the Nagelkerke pseudo R^2 measure (discussed in Chapter 1) was applied for all three analytical approaches to quantify the amount of variation attributable to the GRS. To determine the proportion of variance in AOO explained by the T2D GRS after adjustment for confounding variables (Sex, BMI ancestry PCs, and type of genotyping microarray), the R^2 values between nested models were compared. Therefore, the proportion of variance explained represents the difference in R^2 after adjustment for confounding variables where the full model (confounders and GRS) was compared to a reduced model (confounders only), described further in Table 4.1.

Table 4. 1 - Description of models used in the analysis of T2D GRS

Model	**Terms included in model**
(1) GRS reduced models	**Covariate(s):** (X_s) - Sex: male=0; female =1 (X m) - genotyping microarray: UK BILEVE=0; UK Biobank =1 (X_d) - BMI: continuous covariate measured in kg/m^2 (X_{c1}) - X_{c3}) - PC1 – PC10: principal components used to account for population structure)
(2) adjusted (full) model	**Variable of interest:** (X_{g1}) - GRSw: weighted GRS **Covariate(s):** (X_s) - Sex: male=0; female =1 (X m) - genotyping microarray: UK BILEVE=0; UK Biobank =1 (X_d) - BMI: continuous covariate measured in kg/m^2 (X_{c1}) - X_{c3}) - PC1 - PC3: principal components used to account for population structure)
Description	
Versions of GRS (X_{g1}) - GRSw: weighted GRS (X_{g2}) - GRSu: unweighted GRS	

As indicated in section 3.1.1 and 3.2.3.2 obesity is an important modifiable risk factor for T2D. It is also known that the relationship between obesity and T2D varies with geographical areas and ancestry. To explore the association of BMI (the most commonly applied marker used to assess risk for T2D) and AOO of T2D while taking into account the different verisons of the T2D

GRS, BMI models were also assessed. Therefore, the extent to which BMI explains the variation in AOO of T2D was also considered. In these models the full model consisting of the confounding variables which included the GRS and BMI was compared to a reduced model consisting of the confounding variables which included the GRS, while BMI was excluded in these models (Table 4.2).

Table 4. 2 - Description of models used in the analysis of T2D GRS and BMI

Model	**Terms included in model**
(1) BMI reduced models	**Covariate(s):** **Model 1:** (X_s) - Sex: male=0; female =1 (X m) - genotyping microarray: UK BILEVE=0; UK Biobank =1 (X_{c1}) - X_{c3}) - PC1 - PC3: principal components used to account for population structure) (X_{g1}) - GRSw: weighted GRS
(2) adjusted (full) model	**Model 1:** **Variable of interest:** (X_d) - BMI: continuous covariate measured in kg/m^2 **Covariate(s):** **(**X_s) - Sex: male=0; female =1 (X m) - genotyping microarray: UK BILEVE=0; UK Biobank =1 (X_{c1}) - X_{c3}) - PC1 - PC3: principal components used to account for population structure) (X_{g1}) - GRSw: weighted GRS
Description	
Versions of GRS (X_{g1}) - GRSw: weighted GRS (X_{g2}) - GRSu: unweighted GRS	

4.3. | T2D GRS results

This section presents the results of the investigation of the utility of a European ancestry derived T2D GRS to detect an association with AOO of T2D in European, Asian, and African ancestry populations in the UK Biobank. Section 4.3.1 provides an overview of the underlying general characteristics and GRS profile within each ancestry group included in the ancestry-specific analysis of T2D GRS. Presented in section 4.3.2 are the results of the single SNP association analysis in relation to T2D status and Cox PH analysis for AOO of T2D. This is followed by an account of the findings of the association analysis based on cases only Cox PH, cases and controls Cox PH and binary logistic models for T2D GRS. The models are first assessed in terms of the size of estimated effect and strength of association resulting from the GRS (section 4.3.3) and second in terms of the variance in AOO of T2D explained by the GRS based on the Nagelkerke pseudo R^2 measure (section 4.3.5). Analysis focused on assessing the impact of BMI are presented in section 4.3.4 and 4.3.6.

4.3.1. | Profile of GWAS datasets

Table 4.3 presents a summary of the general characteristics of each ancestry group included. The analysis of T2D GRS consisted of 366,422 European; 7,937 Asian; and 6, 387 African ancestry individuals. Within these three main ancestry groups, European, Asian, and African, there were 15,028 (4.1%); 1,252 (15.8%); and 628 (9.8%) cases of T2D, respectively.

Table 4. 3 - General descriptive characteristics of T2D cases and controls in European, Asian, and African descended populations

Characteristics	T2D Status	
	Cases	Control
European ancestry population		
Total (N, %)	15,028 (4.1%)	351,394 (95.9%)
Sex (n, % female)	5,252 (34.9%)	192,146 (54.7%)
Age (years)		
Mean (SD)	53.77 (9.55)	56.57 (8.02)
Median	55	58
Range (Min-Max)	1 - 70	39 - 73
BMI (Body Mass Index)		
Mean (SD)	31.92 (5.81)	27.17 (4.61)
Median	31.14	26.54
Range (Min-Max)	16.47 - 74.68	12.12 - 68.41
Asian ancestry population		
Total (N, %)	1,252 (15.8%)	6,685 (84.2%)
Sex (n, % female)	407 (32.5%)	3,195 (47.8%)
Age (years)		
Mean (SD)	48.06 (11.37)	52.48 (8.42)
Median	50	52
Range (Min-Max)	1 - 69	40 - 70
BMI (Body Mass Index)		
Mean (SD)	28.69 (5.05)	26.89 (4.2)
Median	27.81	26.37
Range (Min-Max)	17.24 - 60	14.87 - 58.91
African ancestry population		
Total (N, %)	628 (9.8%)	5,759 (90.2%)
Sex (n, % female)	305 (48.6%)	3,292 (57.2%)
Age (years)		
Mean (SD)	48.86 (10.62)	51.22 (7.79)
Median	49	50
Range (Min-Max)	1 - 69	39 - 70
BMI (Body Mass Index)		
Mean (SD)	31.53 (5.93)	29.3 (5.21)
Median	30.46	28.59
Range (Min-Max)	17.69 - 68.13	19.28 - 59.37

***Descriptions**: N: overall sample size; n: subgroup sample size;* Age: for cases *age refers to AOO of T2D and controls age at enrolment; SD: standard deviation; Min: minimum; Max: maximum; BMI: body mass index measured in kg/m².*

The general characteristics considered include sex, age (AOO for cases) and BMI. It was noted that more than 50% of controls in European and African population were female (54.7% and 57.2% respectively), compared to 47.8% in the Asian population. Among cases, however, the proportion of females was lower than males, with the highest proportion of females occurring in the African population (48.6%) compared to 34.9% and 32.5% respectively in the European and Asian population. Furthermore, it was noted that both cases and controls, on average, were younger in Asian and African populations when compared to the European population. Among cases, the mean AOO was 48.06, 48.86, and 53.77, respectively, in Asian, African, and European ancestry populations. It was further noted for T2D, which typically occurs in adulthood, that inspection of the distribution of AOO revealed a small number of outliers with earlier AOO than expected. However, the percentage of T2D cases in the population with AOO 20 years or less was just 0.06% (228), 0.57% (45) and 0.22% (14) in the European, Asian, and African ancestry populations, respectively, and thus would not be expected to impact on results. It was also observed that, in general, cases had on average a higher BMI when compared to controls in all populations. Additionally, cases in the European and African populations had a higher BMI on average when compared to the Asian population.

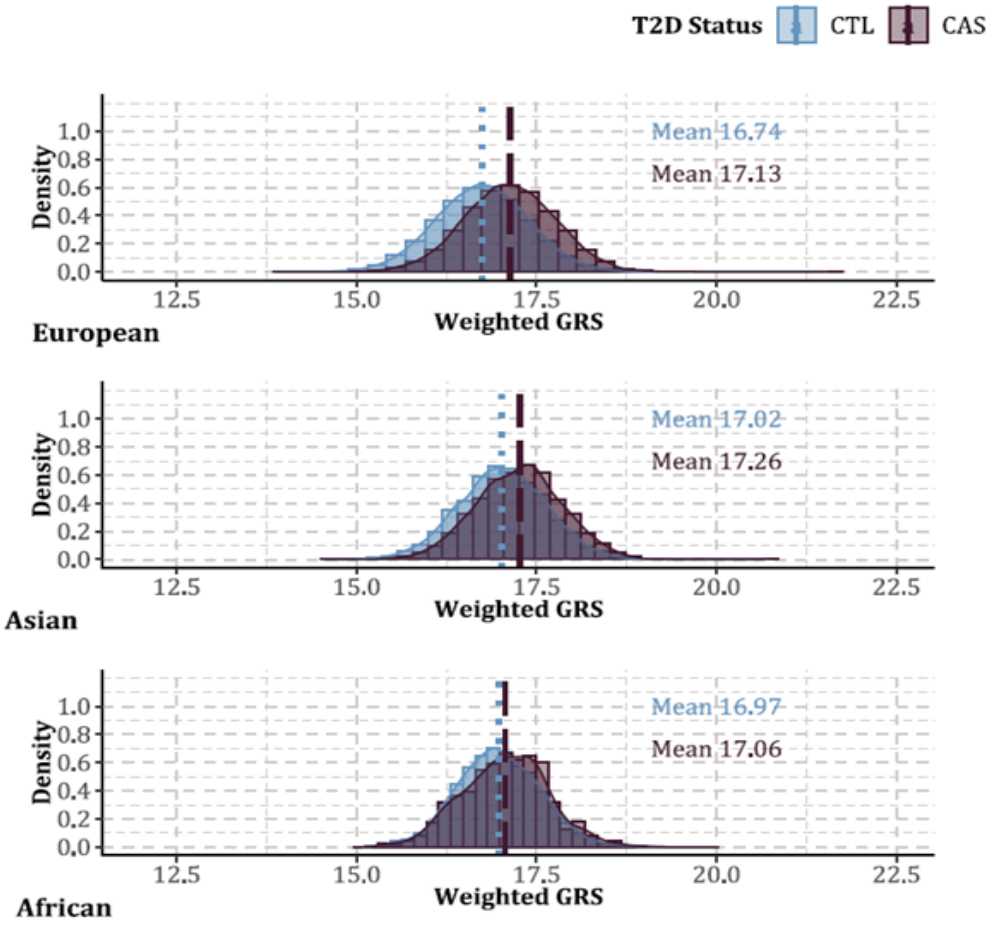


Figure 4. 3 - Distribution of weighted GRS and T2D status in European, Asian, and African descended populations

The distribution of GRS for cases T2D (purple) and controls (light blue) is shown. The horizonal dotted line represent the means.

Abbreviations: CTL: controls; CAS: cases of T2D.

Figure 4.3 presents the profile of the distribution of the weighted GRS within each ancestry group according to T2D status for the European, Asian, and African ancestry populations. The GRS distribution of cases of T2D (CAS) and controls (CTL) within the three ancestry groups is presented. The mean weighted GRS across the three ancestry groups in relation to T2D cases was 17.13, 17.26 and 17.06 respectively for European, Asian, and African ancestry populations. As it relates to controls within the European, Asian, and African ancestry groups the mean weighted GRS was 16.74, 17.02 and 16.97, respectively.

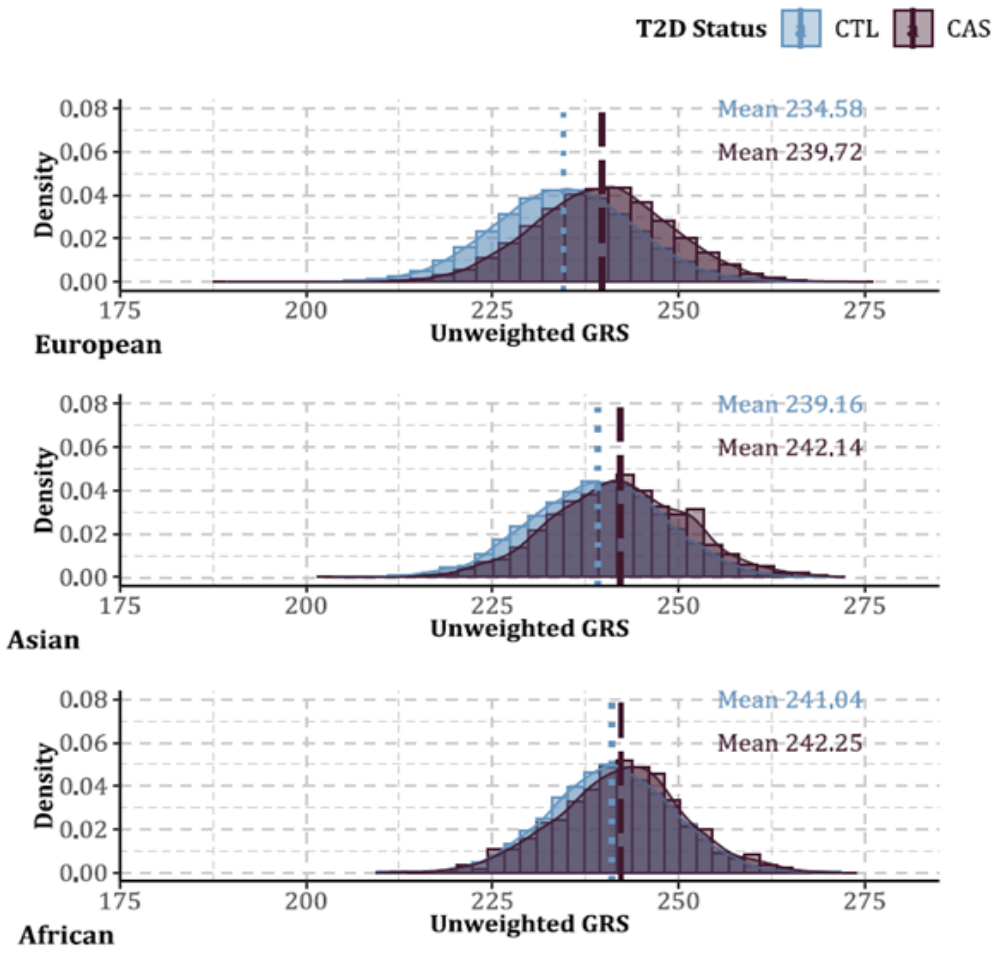


Figure 4. 4 - Distribution of unweighted GRS and T2D status in European, Asian, and African descended populations

The distribution of GRS for cases of T2D (purple) and controls (light blue) is shown. The horizonal dotted line represent the means.

Abbreviations: CTL: controls; CAS: cases of T2D.

Figure 4.4 presents the profile of the distribution of the unweighted GRS within each ancestry group according to T2D status. Illustrated is the GRS distribution in relation to cases of T2D (CAS) and controls (CTL) within the three ancestry groups. Among T2D cases the mean unweighted GRS across the three ancestry groups were 239.72, 242.14 and 242.25 respectively for European, Asian, and African ancestry groups. For controls the mean unweighted GRS for European, Asian, and African ancestry groups were 234.58, 239.16, and 241.01, respectively.

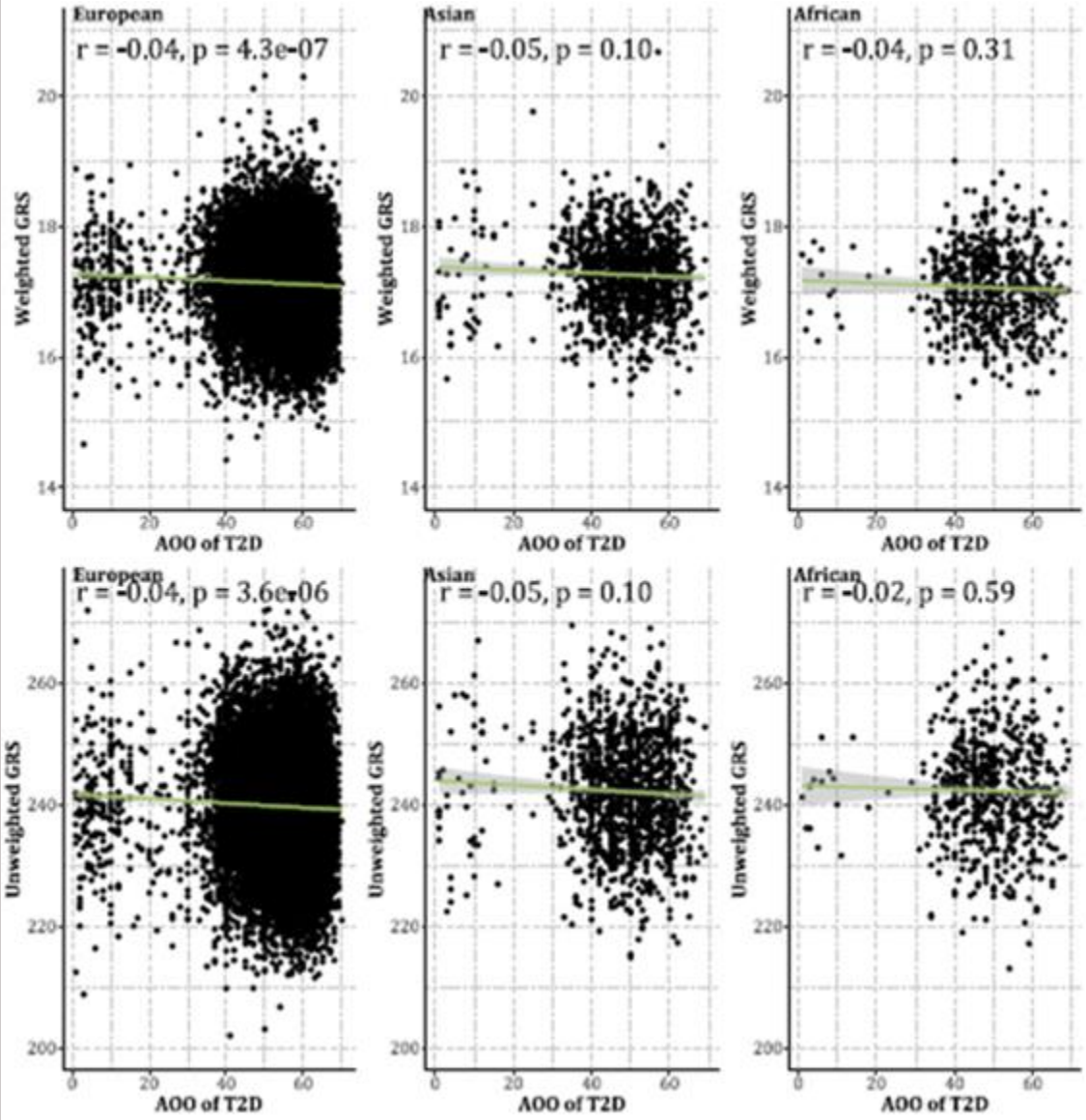


Figure 4. 5 - Relationship of GRS and AOO of T2D in European, Asian, and African populations

The x-axis indicates the AOO of T2D and the corresponding GRS is shown on the y-axis for each individual in the European, Asian, and African populations. Each point represents the individuals included in the samples.

Figure 4.5 depicts the relationship between GRS and AOO of T2D in the European, Asian, and African populations. A small, but consistent, negative correlation between GRS and AOO of T2D was observed across all three ancestries. The correlations were not significant in the Asian and

African ancestry populations because of the smaller sample size, but the trend of earlier AOO for higher GRS was consistent.

4.3.2. | Single-SNP association with T2D status and AOO of T2D

Single SNP analysis was undertaken to assess the association of each European derived SNP independently within the European, Asian, and African ancestry populations from the UK Biobank in respect to both risk of T2D and AOO of T2D. This assessment was essential to gauge how well the European derived SNPs that form the basis of the T2D GRS were likely to perform in non-European ancestry populations. Given the much smaller sample size in the Asian and African ancestry populations, relative to the European ancestry population, consideration was given to both nominal and genome-wide significance across the three ancestry groups.

Figure 4.6 summarises the results of the test of the single SNP association with T2D status adjusted for age, sex, BMI, population structure via 10 PCs and, in the case of the European ancestry population, type of genotyping microarray used (further details are also summarised in Appendix C.4 Table C.4.1 -Table C.4.5). Given that the SNPs used to construct the T2D GRS were primarily European ancestry derived, coupled with a much larger sample size, it was unsurprising that the majority of SNPs achieved at least nominal significance (234 SNPs) in the European ancestry population. In contrast, the number of SNPs achieving nominal significance was far less in the Asian and African ancestry populations, where 33 and 18 SNPs respectively, were nominally significant. Furthermore, it was noted that among the SNPs attaining nominal significance, six were common to all three ancestral populations; rs2972144 (near gene *IRS1*), rs11708067 (near gene *ADCY5*), rs3774723 (near gene PSMD6), rs7903146 (near gene *TCF7L2*), rs141521721 (near gene *PDE3B*), and rs10406431 (near gene *GIPR*).

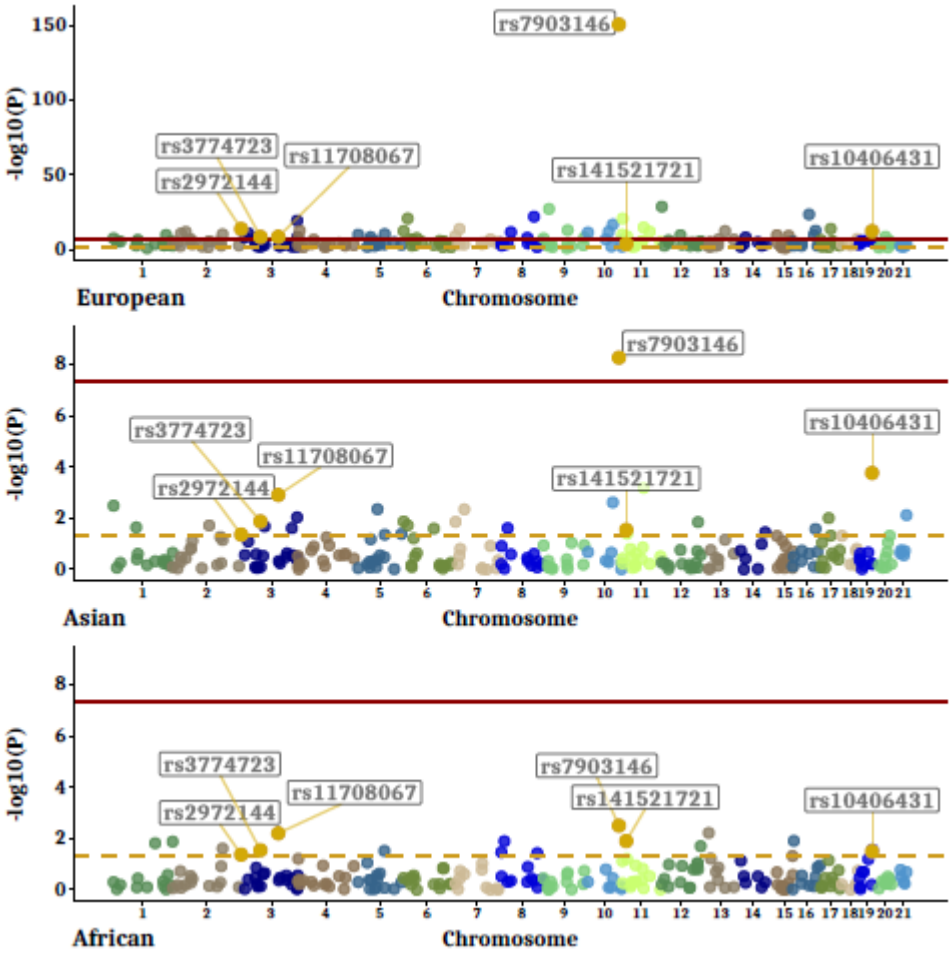


Figure 4. 6 - Single SNP association with T2D using 243 genotyped SNPs in European, Asian, and African descended populations.

The x axis indicates the chromosomal position of each SNP, while the y axis indicates the statistical significance of the association of each SNP with T2D represented by -log10 (p-values). Each SNP is represented by a dot with the colours representing the different chromosomes (SNPs attaining nominal significance in all three ancestries are highlighted by gold dots). The horizontal red line indicates the genome-wide significance threshold, while the gold indicates the nominal significance threshold.

At the genome-wide significance level, overall, 51 SNPs were found to be significant with T2D status in the European population and one in the Asian population. None of the SNPs evaluated achieved genome-wide significances in the African ancestry population although this could partly be due to the much smaller sample size for this population. It was further noted that the SNP rs7903146, mapping near the *TCF7L2* gene, was the most highly significant SNP across all three populations, achieving genome-wide significance in both the European and Asian ancestry populations.

The results of the test for a single SNP association with AOO of T2D based on the Cox PH model is given in Appendix C.5 Table C.5.1 – Table C.5.3. At the genome-wide significance level, 57 SNPs were found to be associated with AOO of T2D in the European ancestry group. As with T2D risk, rs7903146, mapping near the *TCF7L2* gene, was also found to be associated with AOO of T2D in the Asian ancestry population. Considering nominal significance for the Asian and African ancestry populations (as none of the SNPs achieved genome-wide significance), 34 and 16 SNPs respectively were found to be associated with AOO of T2D.

4.3.3. | Association of GRS with AOO of T2D

In this section, the relative performance of the T2D GRS in the European, Asian, and African ancestry populations was evaluated in each of the three analytical approaches (cases only Cox PH, cases and controls Cox PH, and binary logistic analysis). Additionally, as part of the evaluation, the performance of the weighted and unweighted T2D GRS was considered.

4.3.3.1. |Association of weighted GRS with AOO of T2D

Illustrated in Figure 4.7 are model estimates relating to the European, Asian, and African ancestry populations. For each of the three analytical approaches that featured the Cox PH and logistic models, the estimated HR of AOO of T2D or OR of T2D status associated with the weighted T2D GRS was compared across the three ancestries considered. Potential confounding in respect to sex, population structure (using 10 PCs), type of genotyping microarray, BMI, and age at enrolment in the case of the logistic model have been taken into consideration.

In the analysis based on the cases only Cox PH model the estimated HR in the adjusted GRS model in the European population was found to be the most highly significantly associated with AOO of T2D of the three ancestral populations evaluated. Additionally, it was observed that the CI of all three populations overlapped. However, although the estimated HR in the European ancestry population was the smallest of the three groups, the CI in the European was much narrower compared to the Asian and African ancestry populations. The estimated HR pertaining to the European, Asian, and African ancestry populations was 1.12 (95% CI 1.09 - 1.15: $p = 4.4 \times 10^{-18}$) and 1.13 (95% CI 1.03 – 1.24: $p = 1.1 \times 10^{-02}$) and 1.15 (95% CI 1.01 – 1.32: $p = 4.0 \times 10^{-02}$) respectively. Based on the cases and controls Cox PH model, the European population was also found to be the most highly significantly associated with AOO of T2D among the three ancestries evaluated. Additionally, it was observed that the estimated HR in the European population was both larger and CI much tighter when compared to the Asian and African ancestry populations. The estimated HR in the adjusted GRS model in the European population, the largest of the three ancestries considered, was 2.5 (CI: 2.5 - 2.6: $p = 4.7 \times 10^{-1214}$); Asian 1.9 (CI: 1.8 - 2.1: $p = 3.6 \times 10^{-44}$); and African 1.3 (CI: 1.2 - 1.5: $p = 1.8 \times 10^{-05}$).

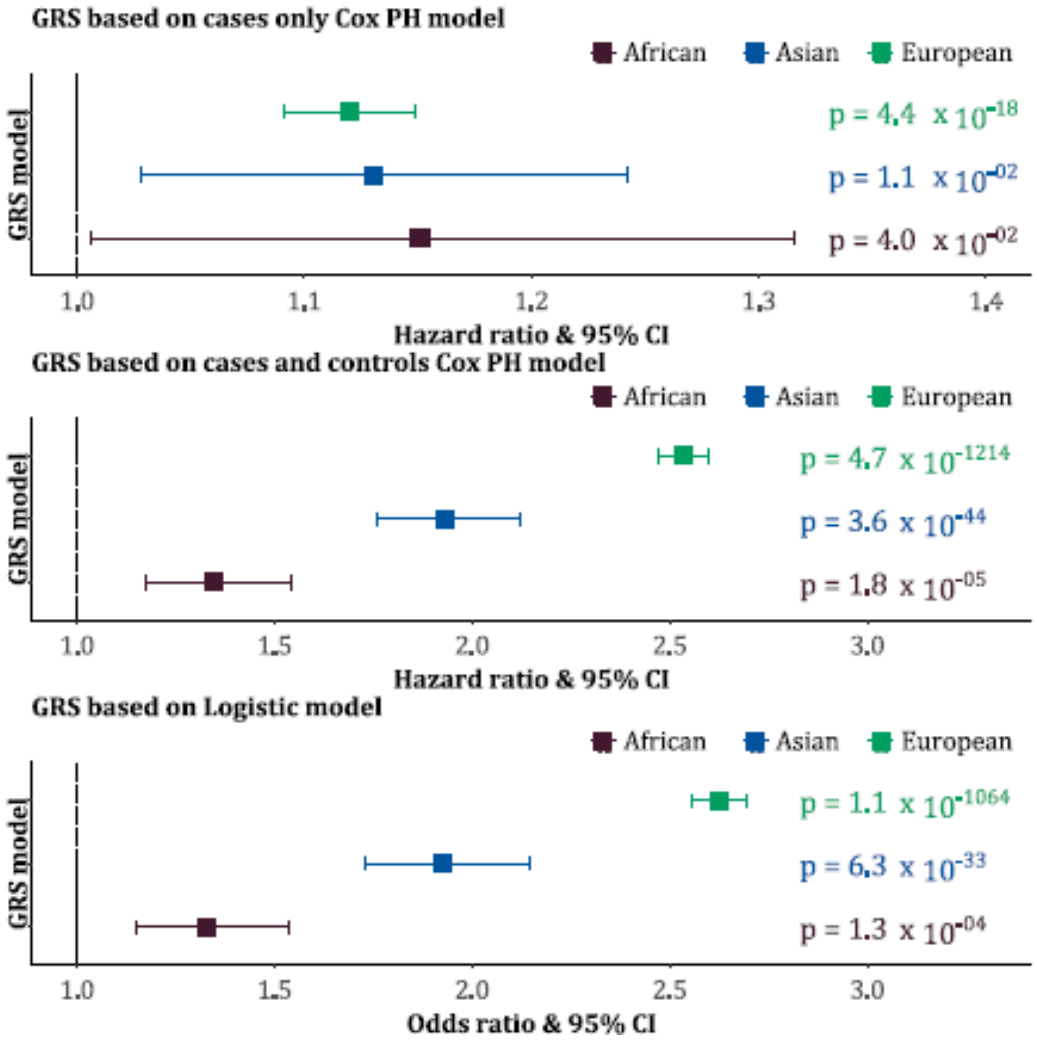


Figure 4. 7 - Comparison of estimated ES of AOO of T2D associated with the weighted GRS for European, Asian, and African descended populations

The x-axis indicates the HR or OR and 95% CI for each GRS model shown on the y-axis. The analysis methods included are the cases only Cox PH model; cases and controls Cox PH model; and logistic model.. The weighted (wei) GRS models which have been adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and BMI are denoted (purple), (blue) and (green) which refers to the African, Asian and European ancestry populations.

Regarding T2D status, the T2D GRS was found to be most highly associated with T2D status, in the European ancestry population followed by the Asian ancestry population. Additionally, for the European ancestry population, it was observed that the estimated OR was both larger and CI much tighter when compared to the Asian and African ancestry populations. In the logistic model the estimated OR in the adjusted GRS model in the European population was 2.6 (CI: 2.6 - 2.7: $p = 1.1 \times 10^{-1064}$); Asian 1.9 (CI: 1.7 - 2.1: $p = 6.3 \times 10^{-33}$); and African 1.3 (CI: 1.2 - 1.5: $p = 1.3 \times 10^{-04}$).

In general, among the three analytical approaches considered, the weighted T2D GRS models were highly significantly associated with AOO of T2D across the three ancestries for the cases and controls Cox PH, and logistic models. The cases only Cox PH model, however, was highly significant only in the European ancestry population. Within the European and Asian ancestry populations, the Cox PH model consisting of both cases and controls were found to be the most strongly associated of the three approaches. The performance of the weighted GRS model in the African ancestry population was similar for both the cases and controls Cox PH model and logistic model.

4.3.3.2. |Association of unweighted GRS with AOO of T2D

Represented in Figure 4.8 are model estimates pertaining to the European, Asian, and African ancestry populations. In this instance the estimated HR of AOO of T2D or OR of T2D status associated with the unweighted T2D GRS were considered, which were based on three analytical approaches incorporating the Cox PH, and logistic models. The European population was found to be the most highly significantly associated with AOO of T2D in the case only Cox PH analysis (European: $p = 2.1 \times 10^{-12}$). This was followed by the Asian ancestry population. However, the T2D GRS was found not to be significantly associated with AOO of T2D in the African ancestry population. It was noted that although the estimated HR in the European ancestry population was the smallest of the three groups, the CI in the European was narrower compared to the Asian and African ancestry populations. The GRS in the European population was also found to be the most highly significantly associated with AOO of T2D among the three ancestries evaluated in the cases and control analysis. The estimated HR in the European population, which was larger relative to the Asian and African ancestry were also characterized

by a narrower CI. The HR estimated for each ancestry based on the cases and controls Cox PH model was European 1.06 (CI: 1.058 – 1.062: p = 5.8 x 10^{-955}); Asian 1.04 (CI: 1.03 – 1.04: p = 3.9 x 10^{-29}); and African 1.02 (CI: 1.01 - 1.03: p = 6.7 x 10^{-05}).

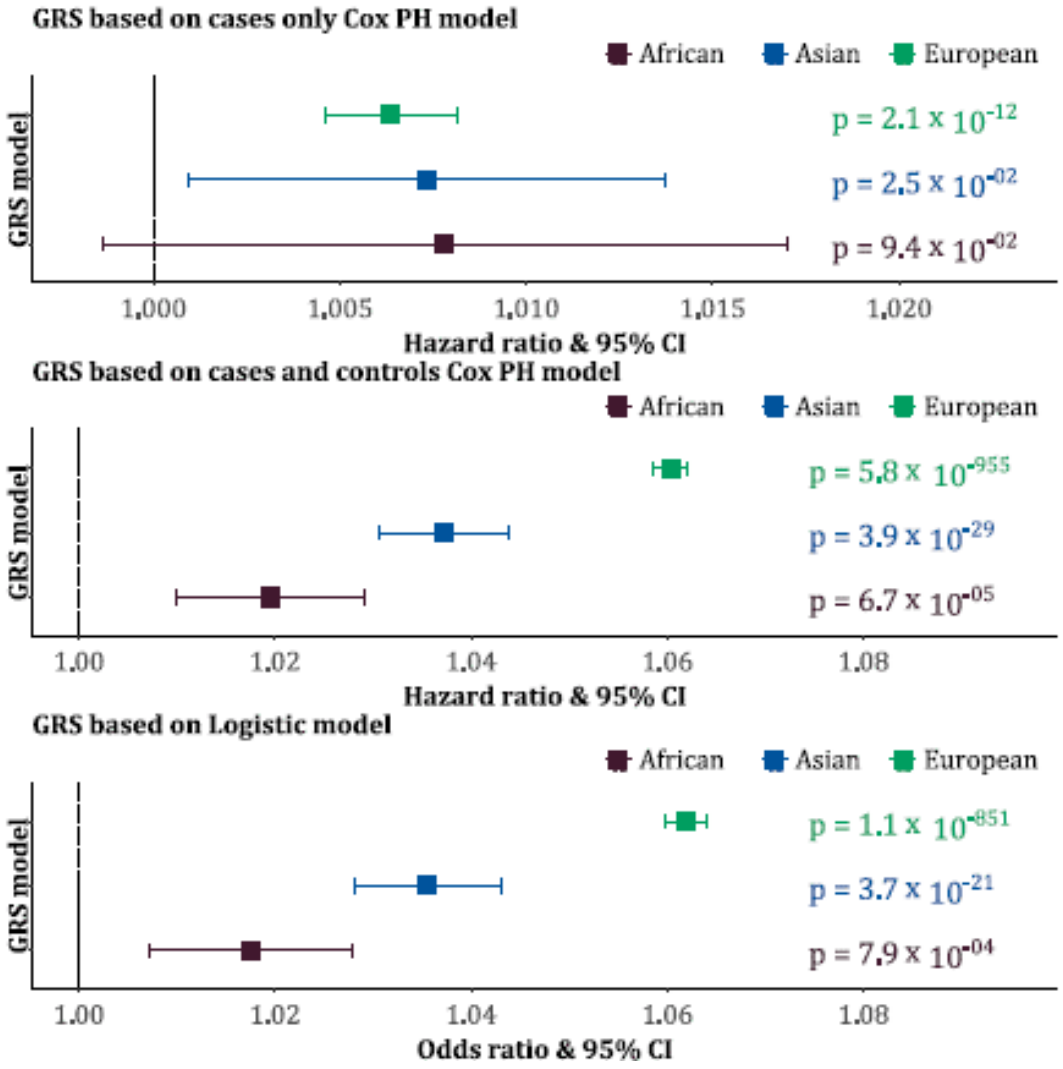


Figure 4. 8 - Comparison of estimated ES of AOO of T2D associated with the unweighted GRS for European, Asian, and African descended populations.

The x-axis indicates the HR or OR and 95% CI for each GRS model shown on the y-axis. The analysis methods included are the cases only Cox PH model; cases and controls Cox PH model; and logistic model. The unweighted (unw) GRS models which have been adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and BMI are denoted (purple), (blue) and (green) which refers to the African, Asian and European ancestry populations.

In the logistic analysis, where the association between T2D status and T2D GRS was evaluated, the most strongly associated T2D GRS was once again observed in the European ancestry population along with the largest estimated OR. It was further noted that the African ancestry population which produced the smallest OR of the three ancestries considered, larger SE were observed relative to the Asian population (SE was estimated to be 0.005 and 0.004 respectively in the African and Asian population).

Overall, as in the weighed GRS models, unweighted GRS models were significantly associated with AOO across the three populations assessed for the cases and controls Cox PH, and logistic models. Additionally, the Cox PH model based on cases and controls was again most highly associated in the three ancestry populations. As for the weighted T2D GRS, the on average younger ages of cases and controls seen in the Asian and African population likely contributed to differences in model performance among the three ancestries in the analysis based on the cases only Cox PH model.

A summary of the estimated HR of AOO of T2D or OR of T2D status associated with both the weighted and unweighted T2D GRS in the European, Asian, and African ancestry populations are presented in Table 4.4. In general, the GRS models in the European were found to be more strongly associated in the Cox PH model relative to the logistic model.

Table 4. 4 - Estimated effect of association of GRS and AOO of T2D in European, Asian, and African ancestry populations

Analysis Method	Weighted GRS				Unweighted GRS			
	ES	Lower 95% CI	Upper 95% CI	P-value	ES	Lower 95% CI	Upper 95% CI	P-value
European ancestry population								
Cox PH model (cases only)								
Adjusted *(GRS+ BMI+Covariates)*	1.120	1.092	1.149	4.4×10^{-18}	1.006	1.005	1.008	2.1×10^{-12}
Cox PH model (cases and controls)								
Adjusted *(GRS+ BMI+Covariates)*	2.533	2.472	2.595	4.7×10^{-1214}	1.060	1.058	1.062	5.8×10^{-955}
Binary logistic regression model								
Adjusted *(GRS+ BMI+Covariates)*	2.623	2.553	2.695	1.1×10^{-1064}	1.062	1.060	1.064	1.1×10^{-851}
Asian ancestry population								
Cox PH model (cases only)								
Adjusted *(GRS+ BMI+Covariates)*	1.130	1.028	1.243	1.1×10^{-02}	1.007	1.001	1.014	2.5×10^{-02}
Cox PH model (cases and controls)								
Adjusted *(GRS+ BMI+Covariates)*	1.933	1.762	2.121	3.6×10^{-44}	1.037	1.031	1.044	3.9×10^{-29}
Binary logistic regression model								
Adjusted *(GRS+ BMI+Covariates)*	1.927	1.731	2.146	6.3×10^{-33}	1.036	1.028	1.043	3.7×10^{-21}
African ancestry population								
Cox PH model (cases only)								
Adjusted *(GRS+ BMI+Covariates)*	1.151	1.006	1.316	4.0×10^{-02}	1.008	0.999	1.017	9.4×10^{-02}
Cox PH model (cases and controls)								
Adjusted *(GRS+ BMI+Covariates)*	1.348	1.176	1.545	1.8×10^{-05}	1.019	1.010	1.029	6.7×10^{-05}
Binary logistic regression model								
Adjusted *(GRS+ BMI+Covariates)*	1.328	1.148	1.536	1.3×10^{-04}	1.018	1.007	1.028	7.9×10^{-04}

Descriptions: *GRS: genetic risk score; ES: Effect Size (hazard ratio or odds ratio); CI: confidence interval; Covariates: Models adjusted for Sex; BMI: Body Mass Index; array: genotype microarray; ancestry via PC1-PC10: Principal components.*

4.3.4. | Association of BMI with AOO of T2D

Since the relationship between obesity and T2D is known to vary according to geographical areas and ancestry, models incorporating BMI and T2D GRS were evaluated. Thus, the focus of this section was to assess the impact of being overweight or obese after adjustment for one of the two versions of the T2D GRS (weighted and unweighted GRS).

Depicted in Figure 4.9 are model estimates pertaining to the European, Asian, and African ancestry populations. It compares the estimated HR and OR of AOO of T2D associated with BMI across the three populations produced by each of the three analytical approaches considered. The weighted and unweighted BMI models both account for sex, population structure (via 10 PCs), type of genotype microarray and age at enrolment in the case of the logistic model, but differ in whether adjustment for GRS were based on the weighted T2D GRS or the unweighted T2D GRS.

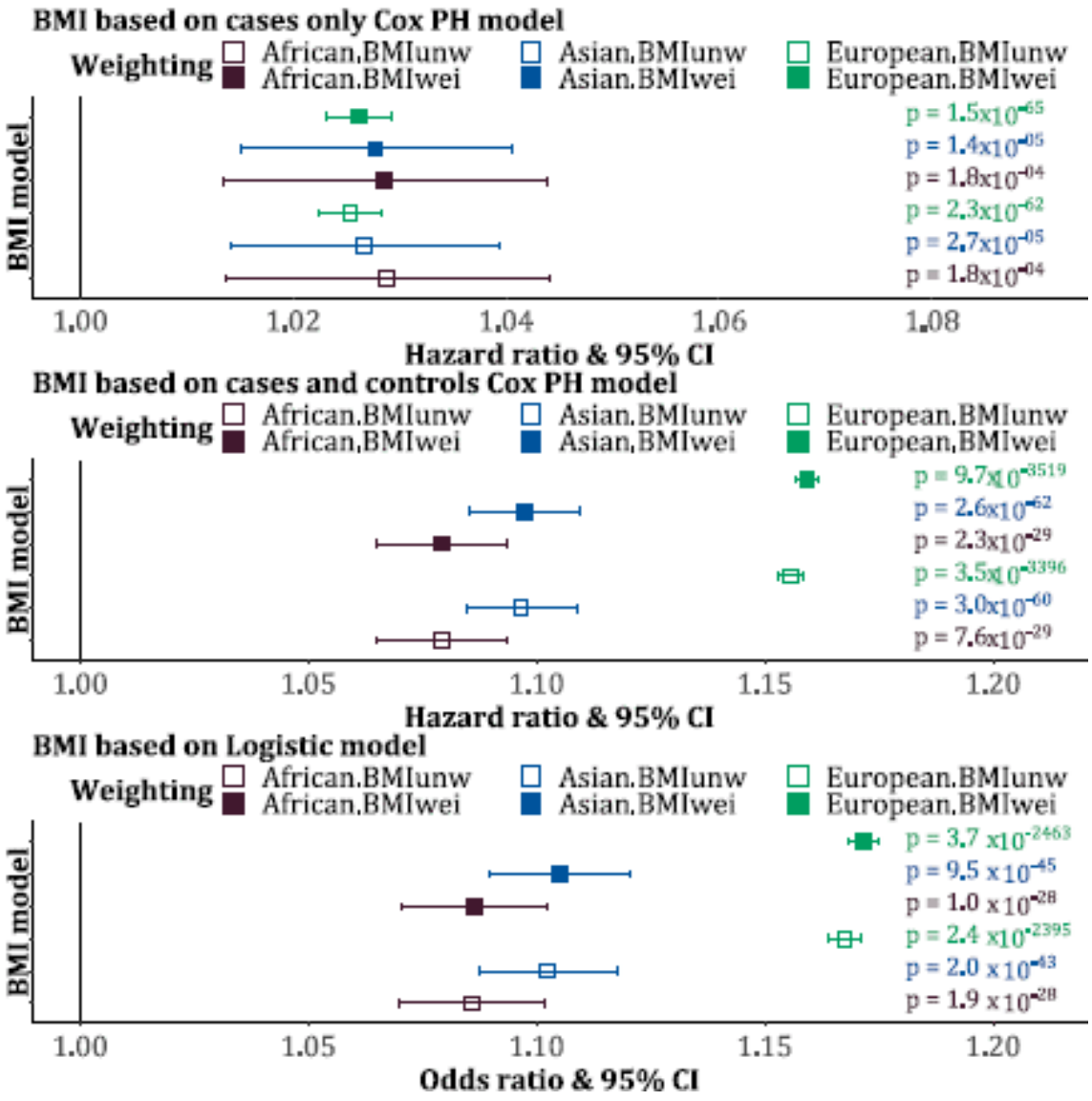


Figure 4. 9 - Comparison of estimated ES of AOO of T2D associated with BMI for European, Asian, and African descended populations

The x-axis indicates the HR or OR and 95% CI for each BMI model shown on the y-axis. The analysis methods included are the cases only Cox PH model; cases and controls Cox PH model; and logistic model. The models have been adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and GRS. The two models considered include the adjusted model with weighted GRS denoted (BMIwei) and adjusted model with unweighted GRS denoted (BMIunw), where(purple), (blue) and (green) refers to African, Asian, and European ancestry population.

BMI was found to be significantly associated with AOO of T2D or T2D status in all three ancestries and in all three analytical approaches. Additionally, it was observed that the BMI models that incorporated the weighted GRS were more strongly associated with AOO of T2D compared to the unweighted GRS. However, the estimated effect sizes were similar between the weighted and unweighted BMI models. It was also noted that the European derived BMI models were more strongly associated with AOO of T2D when compared to the Asian and African ancestry populations. The European derived BMI models were also characterized by the largest effect sizes and narrower CI relative to the Asian and African ancestry populations. Furthermore, overlap in the CI were observed between the Asian and African ancestry populations in the analysis based on the cases and controls Cox PH, and logistic models, while overlap in CI was observed between all three ancestries in the cases only Cox PH model. A summary table of the estimated HR of AOO of T2D or OR of T2D status associated with BMI in the European, Asian, and African ancestry populations are also presented in Appendix C.7.1

4.3.5. | Variance in AOO of T2D explained by GRS

Depicted in Figure 4.10 is the proportion of variance in AOO of T2D that is explained by the T2D GRS based on the three analytical approaches. Explained variance was measured by Nagelkerke pseudo R^2 where comparison was made across the three ancestry groups for both the weighted and unweighted GRS. For the models that have been adjusted for confounding, the proportion of variance explained represents the difference in R^2 after adjustment for confounding variables.

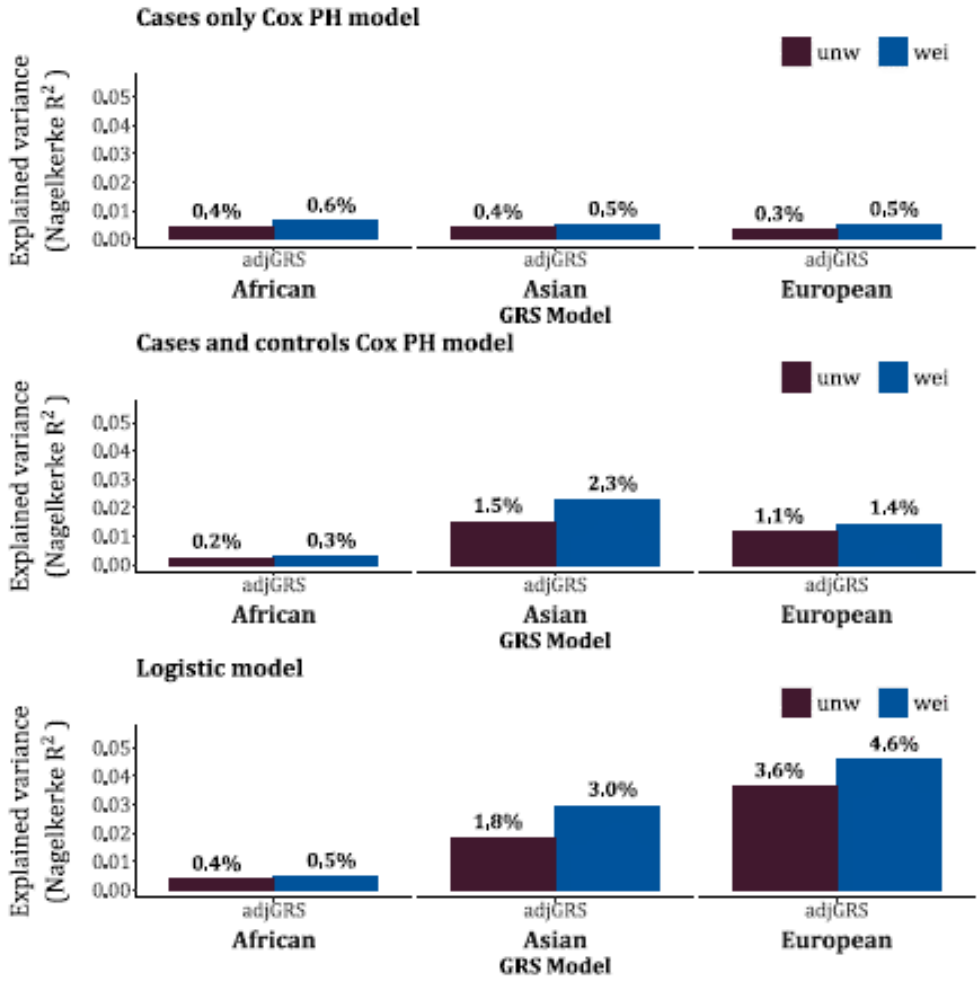


Figure 4. 10 - Proportion of variance in AOO of T2D explained by GRS in European, Asian, and African ancestry populations based on Nagelkerke R^2

The proportion of variance measured by the Nagelkerke pseudo R^2 measure is shown on the y-axis while the effects of both the weighted(wei) and unweighted (unw) GRS models is shown on the x-axis. The models which have been adjusted for potential confounding (adjGRS), the proportion of variance explained represents the difference in R^2 after adjustment for confounding variables where the full model (model with confounding variables and GRS) was compared to a reduced model (model with confounding variables only, GRS is excluded in this model). Models adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and BMI.

Overall, the weighted GRS models explained a higher proportion of variance relative to the unweighted GRS models. Looking at the weighted GRS based on the logistic model, across the three poplations, a greater proportion of variance in T2D status was explained in the European ancestry population, where 4.6% of variance was explained which compares to 3% in Asian and 0.5% in the African population. However, for the cases and controls Cox PH models, a higher proportion of variance in AOO of T2D due to the weighted GRS were explained in the Asian population (2.3%) when compared to the European and African populations (1.4% and 0.3% respectively).

4.3.6. | Variance in AOO of T2D explained by BMI

Figure 4.11 depicts the proportion of variance in AOO of T2D that is explained by BMI, as measured by the Nagelkerke pseudo R^2 , based on the three analytical approaches. The proportion of variance, for the models which have been adjusted for confounding, represents the difference in R^2 after adjstment for confounding variables. Model comparison was based on two models; the adjusted with weighted GRS; and the adjusted with unweighted GRS.

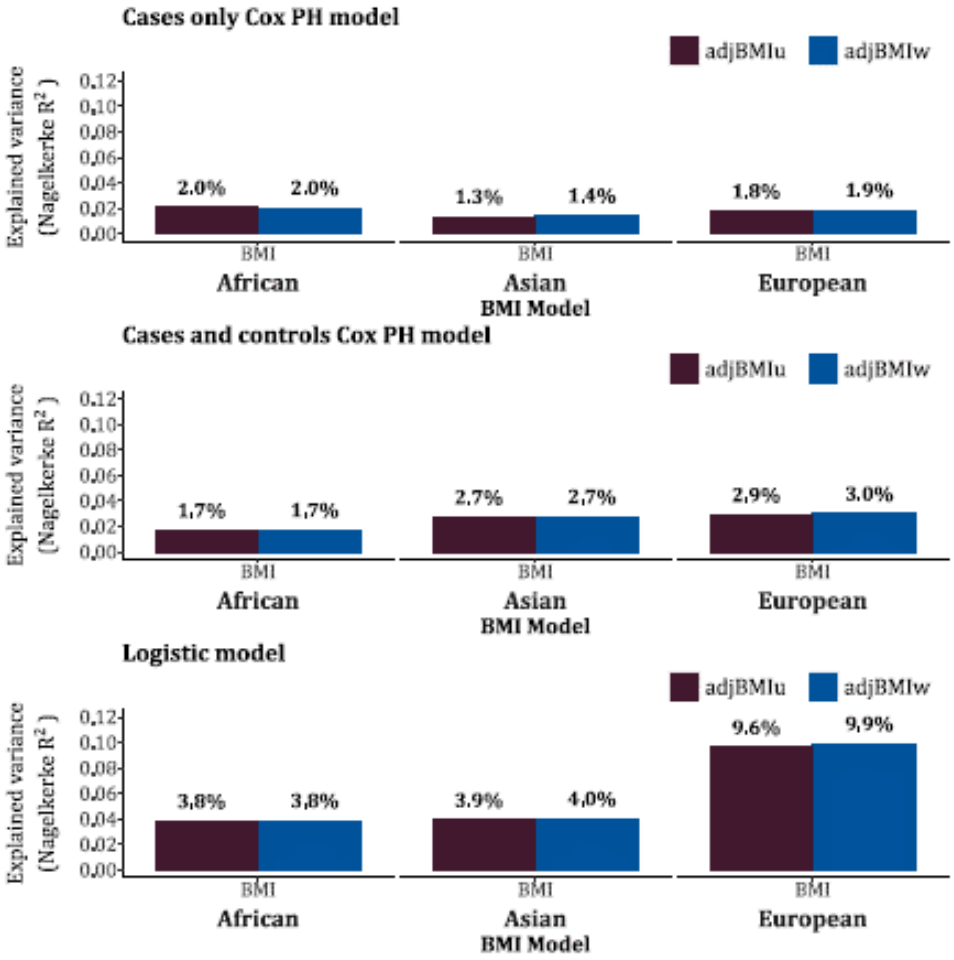


Figure 4. 11 - Proportion of variance in AOO of T2D explained by BMI in European, Asian, and African descended populations based on Nagelkerke R^2

The proportion of variance measured by the Nagelkerke pseudo R^2 measure is shown on the y-axis while the effects of adjusted BMI with unweighted GRS model (purple: adjBMIu); and adjusted BMI with weighted GRS model (blue: adjBMIw); models is shown on the x-axis. The proportion of variance explained represents the difference in R^2 after adjustment for confounding variables where the full model (model with confounding variables and GRS) was compared to a reduced model (model with confounding variables only, BMI is excluded in this model). Models adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and weighted or unweighted GRS.

Across the three ancestry groups, the BMI weighted, and BMI unweighted models performed similarly, however, overall, the BMI weighted models explained a slightly higher proportion of the variance in AOO due to BMI. Furthermore, based on the logistic model, a much greater proportion of variance in T2D status due to BMI was explained in the European population. The observed proportion of variance in T2D status due to the weighted BMI were 9.9% in the European population which compares to 4% and 3.8% respectively in the Asian and African populations based on the logistic model.

4.4. | Dissecting the ancestry specific T2D GRS

Generally, assessing the clinical utility of individual risk loci in different populations is challenging given the differences in RAF and LD patterns across populations. This is compounded by the limited availabilty of GWAS data in non-European populations, as sample sizes may vary considerably among different ancestry groups and in some instances sample size is insufficent for good SNP selection for the construction of GRS. Therefore to gain a better insight as to why the GRS performs differently in the different populations, the role of sample size, ancestry specific RAF and impact of LD was assessed in greater detail. The methods employed in carrying out this assessment is described in section 4.4.1 and the results presented in section 4.4.2.

4.4.1. | Methods employed to assess the T2D GRS

The sampling process undertaken to obtaining samples of the same size for each ancestry is described in section 4.4.1.1, while section 4.4.1.2 describes the process of calculating the RAF for each ancestry based on the ancestry specific association analysis and to identify the risk allele, the summary statistics from published GWAS used to construct the GRS. Section 4.4.1.3 describes the process of extracting information pertaining to LD differences for each GRS SNP from the LDProxy website operated by the National Cancer Institute of the United States Department of Health and Human Serves.

4.4.1.1. |Selecting ancestry specific subsamples

To assess the impact of sample size on the performance of the T2D GRS, independent analyses using the same sample size in each ancestry were conducted. The size of the sample was based on the African ancestry group as it was the smallest of the three ancestry groups. Therefore a random sample of cases and also a random sample of controls reflecting the African ancestry group were obtained for analysis of the European and Asian ancestry groups.

4.4.1.2. |Calculating ancestry specific RAF

To assess the impact of variation in the RAF among the three ancestry groups on the predictive power of the GRS, ancestry specific RAFs were calculated and comparisons made between the ancestry groups. Information provided in summary statistics of the published GWAS used to construct the GRS was used to identify the risk allele associated with each SNP included in the GRS. In conjunction with information relating to allele frequencies provided in the ancestry-specific association analyses of the UK Biobank data, RAF for each ancestry group were calculated.

4.4.1.3. |Determining ancestry specific tag SNPs

Genomic coverage based on the number of SNPs that were in pairwise LD with the SNPs included in the GRS was assessed. Using the data available on the LDProxy website [263] the number of SNPs in pairwise LD with the SNPs included in the GRS were extracted for the European, Asian, and African ancestry populations based on the Phase 3 (Version 5) of the 1000 Genomes Project [264]. SNPs with an LD of r^2 0.8 or above were extracted. The data extracted included the SNP ID, chromosome, position, alleles, MAF and r^2 value. The number of SNPs in LD with the GRS SNPs in European ancestry population was compared graphically with the Asian and African ancestry population. However, there are T2D GRS SNPs that have been excluded from the LDproxy database and/or have been identified as monoallelic in at least one ancestral population. These SNPs are listed in Appendix C.9.1.

4.4.2. | Results of T2D GRS assessment

This section presents the results of the assessment of T2D GRS performance in the different ancestries. The impact of sample size on the performance of the T2D GRS is presented in section 4.4.2.1. In section 4.4.2.2 are the results of the assessment of the impact of RAF among ancestries. Assessment of the impact of LD among ancestries is presented in section 4.4.2.3.

4.4.2.1. |Assessing the impact of sample size

An analysis based on a subset of the original samples was undertaken to assess the potential impact of sample size on the T2D GRS analysis in the European, Asian, and African ancestry populations. In general, the findings were consistent with the original samples. The magnitude of the estimated effect of the T2D GRS was still greatest in the European ancestry population for the cases and controls Cox PH, and logistic models in the weighted T2D GRS (Figure 4.12). The fact that estimated effect, in the Cox PH cases only model, was not the greatest in the European ancestry population could potentially be influence by the mean age of cases in the three ancestries. Individuals with an earlier AOO of disease are expected to have a higher genetic loading of risk variants. The mean AOO of cases in the European, Asian, and African ancestry population were 54, 48, and 49, respectively.

Additionally, with a smaller sample size, the difference in performance of the unweighted T2D GRS in the European ancestry population was less distinguishable among the three analytical approaches (Appendix C.10.1). The P-value pertaining to the logistic, and Cox PH cases and controls models were found to be $p = 1.7 \times 10^{-36}$; and $p = 3.9 \times 10^{-34}$ respectively in the European ancestry population. Furthermore, findings pertaining to the BMI models were similar to the original overall sample (Appendix C.10.2).

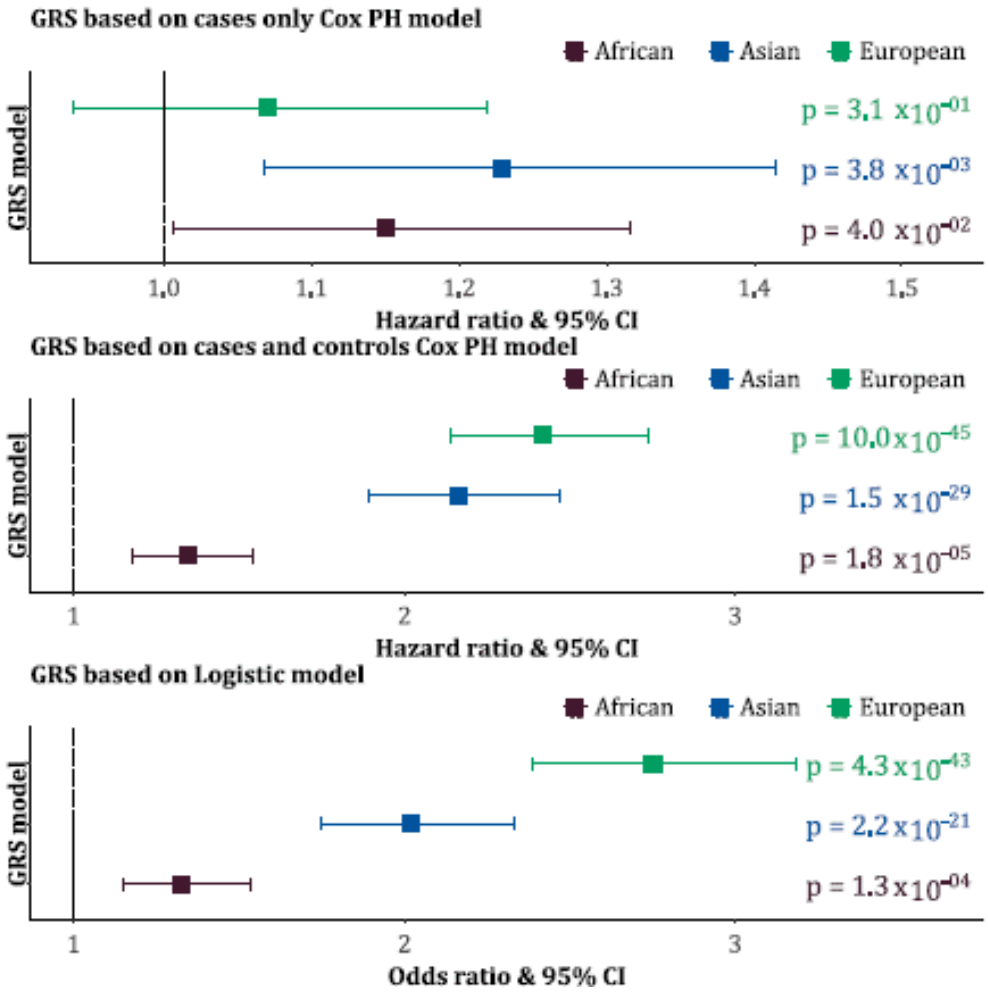


Figure 4. 12 - Subsample comparison of estimated ES of AOO of T2D associated with the weighted GRS for European, Asian and African descended populations

The x-axis indicates the HR or OR and 95% CI for each GRS model shown on the y-axis. The analysis methods included are the cases only Cox PH model; cases and controls Cox PH model; and logistic model. The weighted (wei) GRS models which have been adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and BMI are denoted (purple), (blue) and (green) which refers to the African, Asian and European ancestry populations.

4.4.2.2. |Assessing impact of RAF among ancestries

Allele frequency differences between ancestries to a large extent results in variation in the relative power to detect association between SNPs and a disease of interest. Genetic variants that are common in the study population, which is primarily European, are more likely to be discovered, but might be of a lower frequency in other population groups [265]. Considering the major role RAF plays in driving the level of power in GWAS, the RAF of SNPs in the T2D GRS was examined to determine their overall impact on the performance of the GRS in the three populations.

Figure 4.13 compares RAF of each SNP in the T2D GRS between European and Asian ancestry populations. A consistent pattern of the RAF being more common in the European ancestry population was not evident as there were instances where the RAF was higher in the Asian ancestry population. It was also observed that a low frequency or rare variant SNP in the European ancestry population used in the construction of the T2D GRS had the largest base GWAS log OR (log OR > 1), while the SNP with the second largest base GWAS log OR was a high frequency T2D SNP (log OR > 0.40) in the European ancestry population. SNPs with small base GWAS log OR were observed mainly in the intermediate RAF range (RAF 0.10 to 0.90).

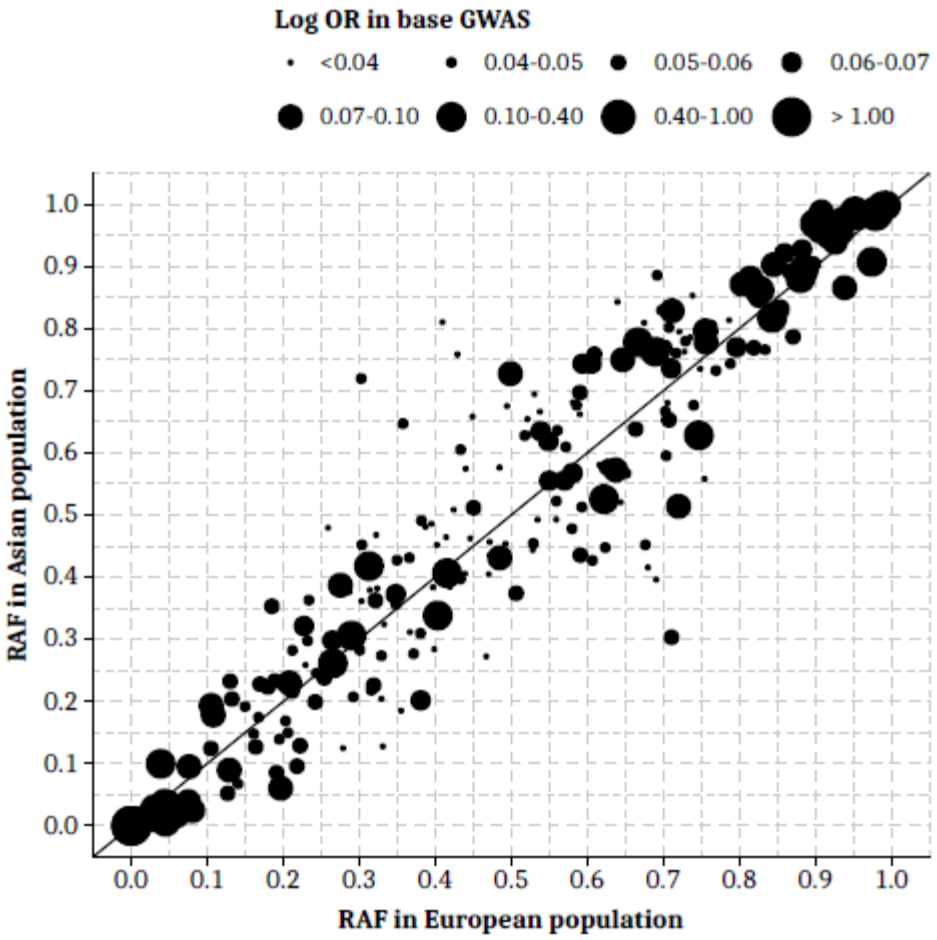


Figure 4. 13 - Relationship between RAF in European population compared to RAF in an Asian population.
The x-axis indicates the RAF in the European population for each SNP included in the GRS and the corresponding RAF for each SNP in the Asian population is shown on the y-axis. Each point which is weighted by the logOR in the base GWAS used to construct the GRS represents each SNP included in the GRS.

Figure 4.14 compares RAF of each SNP in the T2D GRS between European and African ancestry populations. The extent of the difference in the RAF between the European and African population was substantially more than that observed between the European and Asian population. The SNPs with small base GWAS log OR were also observed mainly in the intermediate RAF range (RAF 0.10 to 0.90). Additionally, it was observed that several SNPs in the European ancestry population with RAF that were low frequency or rare had base GWAS log OR greater than 0.07, however, these SNPs were rarer or absent in the African ancestry population. This is consistent with past research that has indicated that many disease-

associated alleles segregate at intermediate frequencies in non-Africans but are found at extremely low or high frequencies in Africans. This is the result of statistical power in European GWAS being maximized at intermediate allele frequencies [266].

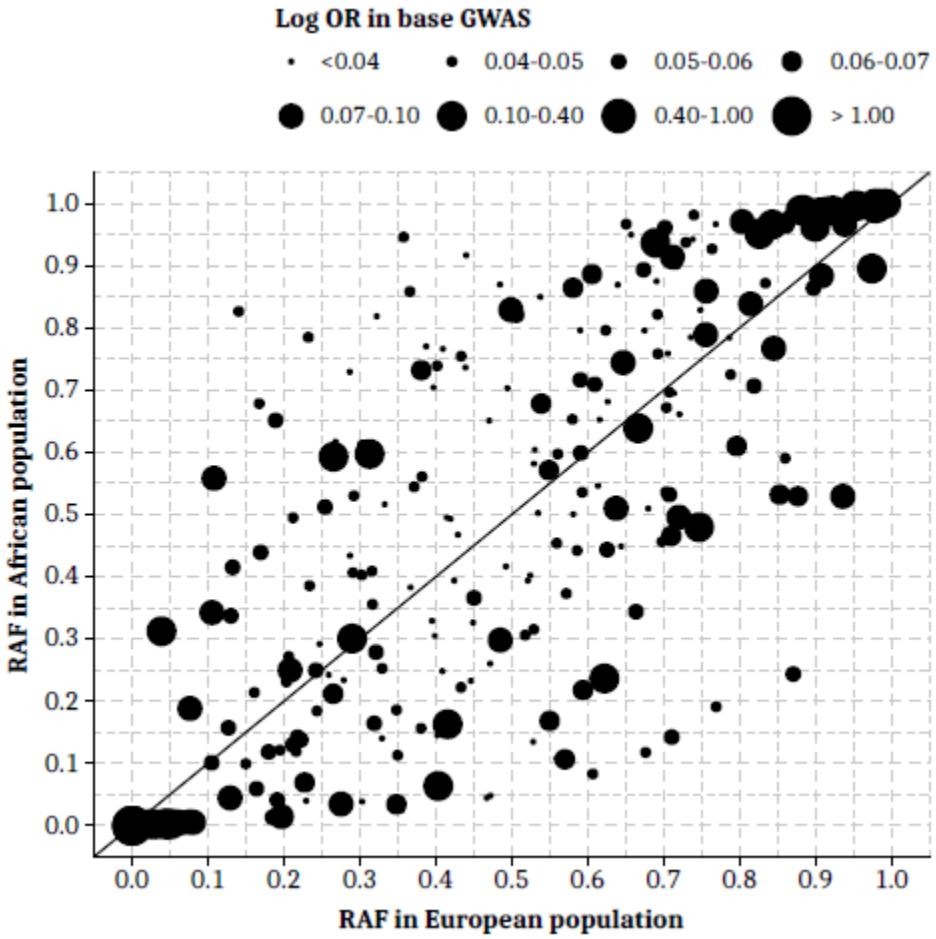


Figure 4. 14 - Relationship between RAF in European population compared to RAF in an African population.
The x-axis indicates the RAF in the European population for each SNP included in the GRS and the corresponding RAF for each SNP in the African population is shown on the y-axis. Each point which is weighted by the logOR in the base GWAS used to construct the GRS represents each SNP included in the GRS.

4.4.2.3. |Assessing impact of LD among ancestries

As the level of pairwise LD between a potential causal SNP and a tag SNP is an important feature within GWAS that drives power, the number of SNPs in LD with the SNPs that form the GRS was assessed at of $r^2 \geq 0.8$. Figure 4.15 compares the number of SNPs in LD with each GRS SNP in the European and Asian ancestry populations. Figure 4.15 illustrates that there was some variation in the number of SNPs tagging the GRS SNPs between the two populations. At the $r^2 \geq 0.8$ threshold level in the European ancestry population, the median number of SNPs tagged by the GRS was 14 compared to 9 in the Asian ancestry population. The drop in the number of tag SNPs in the Asian ancestry population relative to the European ancestry population reduces the likelihood of the causal SNP being tagged by the GRS SNPs and therefore lower predictive power in the Asian ancestry population. At the LD threshold $r^2 = 1$ the number of SNPs in pairwise LD with the SNPs included in the GRS are more in the European ancestry population relative to the Asian ancestry population overall.

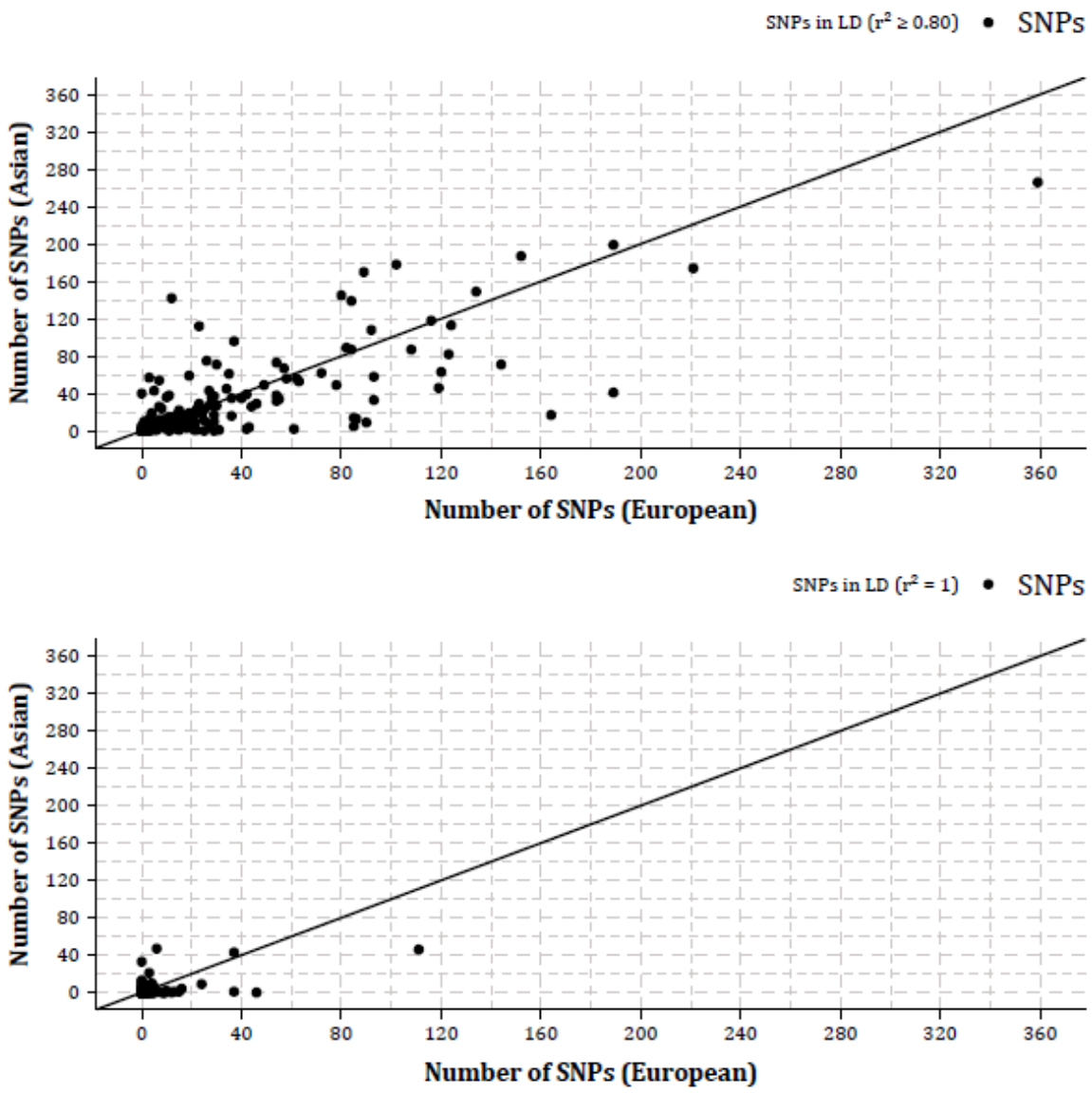


Figure 4. 15 - Relationship between the number of SNPs in LD with the GRS SNPs in European population compared to the number of SNPs in LD with the GRS SNPs in Asian population.

The x-axis indicates the number of SNPs in LD in the European population for each SNP included in the GRS and the corresponding number of SNPs in LD for each SNP in the Asian population is shown on the y-axis. Each point represents each SNP included in the GRS.

Figure 4.16 compares the number of SNPs in LD with the GRS SNPs in the European and African ancestry populations. Figure 4.16 illustrates that for most GRS SNPs, the number of tag SNPs was greater in the European ancestry population when compared to the African ancestry population at the at the $r^2 \geq 0.8$ threshold level. There were only a handful of GRS SNPs in the African ancestry population that had more SNPs tagging it at an LD level of r^2 0.8 or more when compared to the European ancestry population. At the $r^2 \geq 0.8$ threshold level, the median number of SNPs tagged by the GRS in the African ancestry population was 2 which compares to 14 in the European ancestry population. At the LD threshold $r^2 = 1$ the number of SNPs in LD with the SNPs included in the GRS was less than the number at the $r^2 \geq 0.8$ threshold level in both the European and African ancestry population. But overall, the number of SNPs in pairwise LD were more in the European ancestry population relative to the African ancestry population. It was noted also, that SNPs rs117001013 and rs117483894 included in the GRS were in pairwise LD with more SNPs in the African ancestry population compared to the European ancestry population.

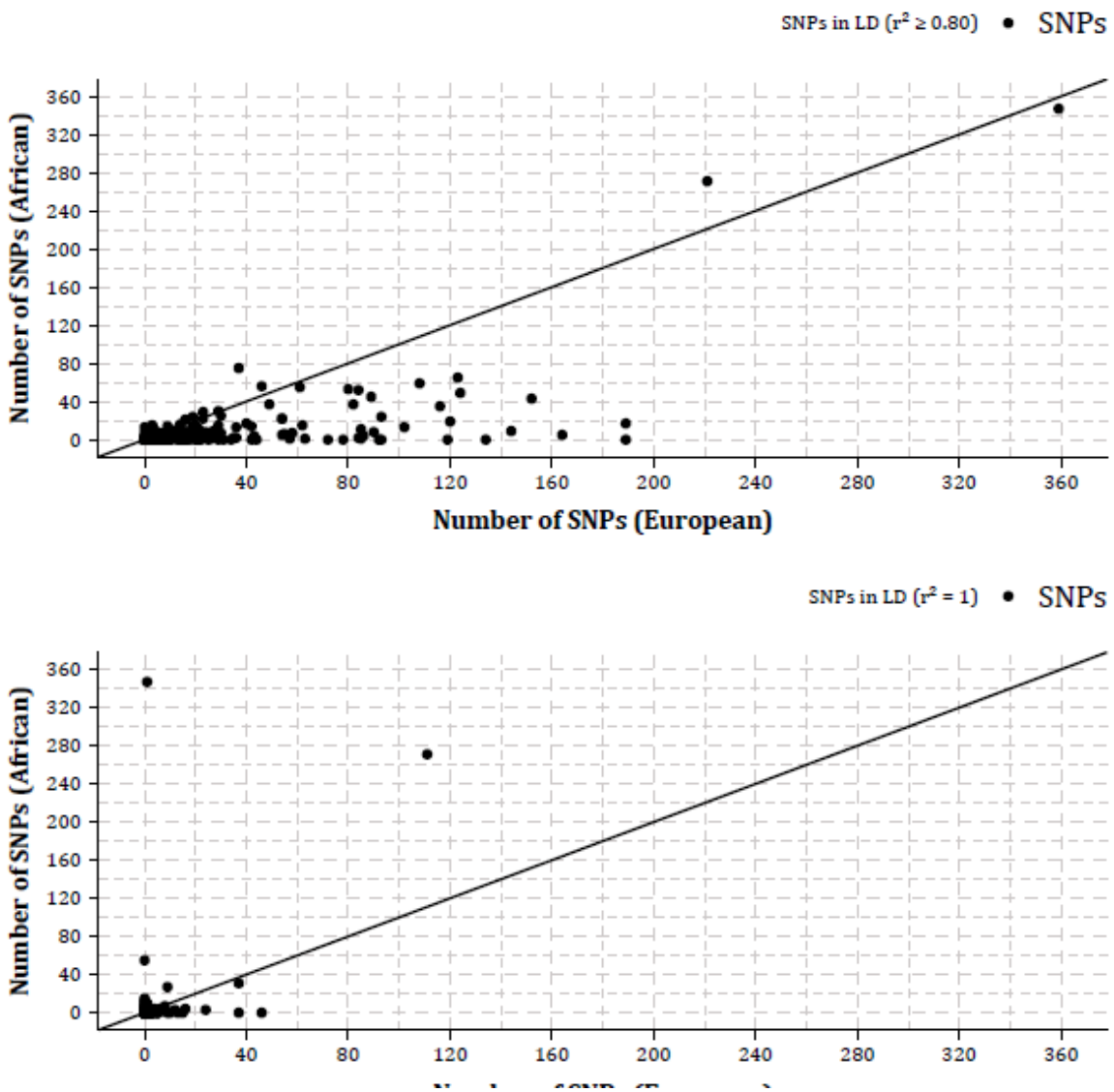


Figure 4. 16 - Relationship between the number of SNPs in LD with the GRS SNPs in European population compared to the number of SNPs in LD with the GRS SNPs in African population.

The x-axis indicates the number of SNPs in LD in the European population for each SNP included in the GRS and the corresponding number of SNPs in LD for each SNP in the African population is shown on the y-axis. Each point represents each SNP included in the GRS.

4.5. | Discussion

Discovery GWAS have aided the emergence of GRS for predicting risk for many common diseases. Owing to increasing concerns regarding the application of GRS in non-European ancestry populations, the utility of a European ancestry derived T2D GRS to detect an association with AOO of T2D in European, Asian, and African ancestry populations was assessed. The scope of the assessment was also extended to include a detailed examination of the role of sample size, ancestry specific RAF and impact of LD in relation to the performance of the T2D GRS.

Overall, assessment of the three analytical approaches, based on strength of association, the best approach was the Cox PH model with both cases and controls. Apart from the cases only Cox PH model, all models produced highly significant associations of the GRS with AOO of T2D or T2D status in all populations. As expected, given its much larger sample size, the strongest associations were observed in the European ancestry population and therefore more precise estimates, resulting from smaller standard errors, are produced within this population relative to the Asian and African ancestry populations. In relation to the most potent independent modifiable risk factor for T2D, obesity, it was noted that BMI (both weighted and unweighted models) was found to be significantly associated with AOO of T2D or T2D status in all three ancestries.

In terms of the proportion of variance in AOO of T2D or T2D status attributable to the GRS, a higher proportion based on the logistic model was explained in the European ancestry population relative to the Asian and African populations. In the cases and controls Cox PH analysis, however, the highest proportion of variance in AOO of T2D attributable to the GRS was explained in the Asian ancestry population. It is noted that cases in the Asian ancestry population on average had a lower AOO of T2D when compared to the African and European ancestry populations. Individuals with an earlier AOO of disease are expected to have a higher genetic loading of risk variants.

The impact of allele frequency and LD on the performance of the GRS was assessed by comparing the RAF in the European ancestry population with that of the Asian and African ancestry populations and similarly the number of SNPs in pairwise LD with GRS SNPs.

Differences in RAF were observed between the European and Asian ancestry populations and between the European and African ancestry populations. However, the extent of the variance in RAF was far greater between the European and African ancestry populations. Additionally, the number of SNPs in pairwise LD with those contributing to the T2D GRS in most instances were less in the African ancestry population relative to the European ancestry population. A drop in the number of SNPs tagging the GRS SNPs were also observed between the Asian and European, however, not to the extent of that observed between the African and European ancestry population. Given these differences, a single GRS that is optimal in all populations may not be possible.

In the absence of African ancestry discovery SNPs and other non-European ancestry discovery SNPs that can be used to build GRS, several alternative approaches have been explored. These methods differ primarily from the standard approach in terms of: (1) how allele frequency differences between ancestries are addressed; (2) how differing LD structures among ancestries is accounted for; and (3) how ancestry is accounted for. GRS are constructed for different ancestries by employing the corresponding human genome reference panel (usually from the 1000 Genomes Project) to determine allele frequency distribution and LD structure within each ancestry group. However, the most basic alternative approach to constructing GRS in different ancestries involves only retaining SNPs that attain at least nominal significance in target samples and combining with ancestry specific SNPs [267, 268].

In relation to allele frequency determination, options include constructing standardized GRS distributions for each ancestry centred on an overall global mean, which gives consideration as to whether a risk allele is derived (risk alleles resulting from new mutations) or ancestral (risk alleles that are shared across ancestries) [266]. This approach appeared to be potentially beneficial for metabolic diseases, as corrected GRS have shown that the African ancestry population GRS distribution overlapped heavily with the other ancestries after correction [266]. Other approaches utilise Bayesian or other statistical procedures to infer the allele frequency distribution in each ancestral population [269].

Alternative avenues pertaining to LD include explicitly modelling or accounting for LD where both linked and unlinked SNPs are included in the calculation of the GRS [270] or excluding linked SNPs based on a LD threshold. Lowering LD thresholds used for clumping to lower the

risk of removing non-European ancestry tags has been explored [271]. Research in this area has shown that lower r^2 thresholds for clumping tend to make worldwide ancestral population polygenic risk score distributions more similar [271].

In conclusion, some of the key areas that need further consideration with regards to the application of GRS in diverse global populations have been highlighted. Although, there is shared genetic contribution to T2D at established loci across different populations, a single GRS that is optimal in different populations may not be possible. Considering that different ancestral populations tend to differ in respect to RAF and patterns of LD structure which drives differences in effect size, different approaches to constructing GRS need to be explored further to ensure the clinical useful of the GRS across global populations.

Chapter 5: Discussion and future work

Chapter Outline

This final chapter provides a synopsis of the research that has been carried out in this thesis. It encompasses a discussion of the key findings and limitations, and implications for epidemiological and clinical research. Additionally, recommendations for future work are also considered.

5.1. | Introduction

Genome-wide association study (GWAS) methodology has evolved considerably since its inception in 2005. This evolution was achieved even in the face of many challenges which has be ascribed primarily in the realm of statistical, computational, and methodological. At the outset of GWAS, concerns about the validity of findings due to unrecognized population structure within diverse populations resulted in the practice of conducting GWAS in ancestrally homogeneous populations, which was most often populations of European ancestry. More recently however, there has been increasing recognition that, although ancestral diversity presents many challenges, it also provides many opportunities for gene discoveries. Moreover, broadening the scope of scientific inquiry to acquiring a better understanding of genomics in all populations on a global scale may be the key to improving disease risk prediction for people of all ancestries. As a result, methods to effectively address population structure, the most prominent confounding issue to consider in the design and analysis of GWAS, is therefore essential.

GWAS discoveries have contributed tremendously to the current knowledge of genetics and its implication for human health. Analysis of contributing genomic variants associated with many common complex diseases have enabled the application of genetic risk scores (GRS). However, currently, the main challenge for the clinical implementation of GRS is that they are of far greater predictive value in European ancestry populations when compared to other populations, owing to bias emanating from European derived GRS. Therefore, a better understanding of the mechanisms influencing the degradation in performance of a European ancestry derived GRS in non-European populations is crucial to ensure accurate prediction of disease risk in all populations given the importance of genomics to the future of healthcare. The

focus of the research in this thesis was concentrated on evaluating methods for detecting associations of age-of-onset (AOO) of disease with a single nucleotide polymorphism (SNP) and GRS in the presence of population structure due to both substructure and admixture.

5.2. | Summary of main findings of the thesis

Chapter 2 focused on investigating methods to account for population structure and admixture in GWAS of time-to-event (TTE) outcomes via simulation. The simulation study based on an admixed population comprised of two ancestral populations, evaluated the power to detect association between a single SNP and AOO of disease under an additive genetic model in a TTE framework. Investigations compared the performance of the Cox proportional hazards (PH) model and the general Weibull model. Based on comparison of these two models it was demonstrated that the power of the general Weibull model was largely consistent with that of the Cox PH model.

From the results of the simulation study which consisted of an admixed population, it was observed, in general, that the association of AOO of disease was described by the causal SNP genotype independently of ancestry. Furthermore, the type I error rates overall also appeared consistent with the nominal significance level (5%) under the null hypothesis (i.e. log HR of 0). However, a small reduction in power to detect an association with the causal SNP with adjustment for ancestry was observed. Given the relationship between the type II error rate (β) and power ($1 - \beta$), this signifies the potential for inflation in the type II error rate if ancestry is not appropriately accounted for in the analysis.

Additionally, testing for association with a tag SNP, which is the more common occurrence in GWAS, was impacted considerably by population admixture as inflation in the type I error rate was observed if ancestry was not accounted for in the analysis. In assessing the impact of the level of LD between the causal SNP and tag SNP, assuming that LD levels were the same in both ancestries, it was observed that equal admixture combined with RAF that are markedly different between ancestries reduces power. Results also indicates that higher levels of LD may be required between the causal SNP and tag SNP when there was equal admixture in order to facilitate detection of an association with AOO of disease based on the tag SNP. This suggests

that greater levels of admixture could potentially cause a reduction in the power to detect associations. Additionally, assuming differing levels of LD between ancestries, the greater the level of LD between the causal SNP and tag SNP in the ancestry where LD was assumed to be fixed at a specified level, the greater the observed lost in power due to ancestry as LD was varied in the second ancestral population. Based on the local ancestry of each admixed individual, which was defined by the genetic ancestry of an individual at the causal SNP, the power to detect association with AOO is greatest when the admixed population is comprised of an equal mixture of the two ancestral populations with very different RAF. In such situations the greater the relative difference in RAF between the two ancestries, the greater the power to detect an association with AOO of disease.

Chapter 3 focused on exploring the association of type 2 diabetes (T2D) GRS and AOO of T2D through the application of the Cox PH, proportional odds and logistic regression models in two independent European ancestry GWAS originating from the Northwestern University Gene (NUgene) Banking Project and Wellcome Trust Case Control Consortium (WTCCC). In the Cox PH framework, in which both cases and controls were considered, the AOO of T2D cases was considered, while controls were censored at their current age. For the proportional odds model, AOO of T2D was viewed as an ordinal outcome that distinguished between controls, late-age-onset (LAO) cases and early-age-onset (EAO) cases. Within the binary logistic regression framework, contrast was made between cases (irrespective of AOO) and controls. The results of these two studies were also combined via meta-analysis. Additionally, a simulation study to further assess the relative performance of the three analytical approaches on power to detect association of a GRS with AOO of disease was undertaken, concentrating primarily on the impact of censoring. As part of the T2D analysis, four versions of the T2D GRS were considered: (1) SNPs determined at the genome-wide significance p-value threshold of 5×10^{-8} with base GWAS effect-size weighting; (2) SNPs determined at the genome-wide significance p-value threshold without weighting; (3) SNPs determined at a nominal significance p-value threshold of 0.05 with base GWAS effect size weighting; and (4) SNPs determined at a nominal significance p-value threshold without weighting.

Generally, based on strength of association measured by the P-value, the utility of the T2D GRS to detect an association with T2D status under a logistic regression model was substantially better when compared to the time-to-event (TTE) modelling framework (Cox PH model), which

assessed the utility of the T2D GRS to detect an association with AOO of T2D. The utility of the T2D GRS to detect an association with AOO of T2D within a proportional odds modelling framework encompassing LAO and EAO T2D was also considered, however, the proportional odds assumption was not valid in the datasets evaluated. Additionally, of the four versions of the T2D GRS considered, the weighted GRS with nominally significant SNPs was found to be the best predictor of the onset of T2D based on strength of association, as measured by the p-value and proportion of variance explained by the model. Results from the simulation study indicated that high rates of censoring did not impact on the relative performance of the methods. However, the Cox PH model seemed to have the advantage, in terms of power, in a setting where there were very low rates of censoring.

Chapter 4 extended the work of Chapter 3 as it focused on investigating the utility of GRS to detect an association with AOO of T2D in ancestrally diverse populations. Here, the utility of a European ancestry derived T2D GRS in detecting an association with AOO the disease in European, Asian, and African ancestry populations using data originating from the UK Biobank was evaluated. The results indicated that the T2D GRS was found to be most strongly associated with AOO of T2D in the Cox PH, comprised of cases and controls, where controls were censored at their current age. However, the utility of the T2D GRS to detect an association with T2D status within a logistic regression framework was stronger than the association with AOO of T2D in the Cox PH model based on cases only. Considering the proportion of variance explained by the models, it was observed that the logistic model explained a greater proportion of the variance in T2D status attributable to T2D GRS in the European ancestry population. However, in the Cox PH model, a greater proportion of variance in AOO of T2D attributable to T2D GRS was explained in the Asian ancestry population, relative to the European and African ancestry populations.

An assessment of the RAF and number of SNPs having an r^2 of ≥ 0.8 with SNPs used to construct the T2D GRS showed differences between the three ancestry groups. The assessment showed that the number of SNPs in LD with the T2D SNPs used to construct the GRS was much less in the African ancestry population when compared with the European ancestry population. This suggests that if the GRS SNPs are not causal, they are less likely to tag the causal SNP in African ancestry populations. It was also observed in comparing the African ancestry population with the European population that the RAFs between the two populations were far more variable. In a situation where the RAF is at a moderate to low level, the RAF of GRS SNPs were typically

lower in the African ancestry population than the European ancestry population, meaning that the GRS will be more powerful in European ancestry populations.

5.3. | Implications for epidemiological and clinical research

The primary implications of the findings in this thesis in relation to AOO of common disease GWAS are highlighted in this section. Implications of undertaking GWAS in diverse and admixed global populations and the implementation of GRS in a clinical setting is considered. Additionally, implications in terms of the statistical analysis of AOO of disease in GWAS and approaches to constructing a GRS that is optimum for different ancestry populations are also considered.

The findings in this thesis have important implications for common disease GWAS in relation to detecting an association with AOO of disease with GRS. With the application of GRS to T2D in three different ancestry populations, it was observed that European ancestry derived T2D GRS were not transferrable to two non-European ancestry populations when compared to the European ancestry group. Therefore, implementation of a T2D GRS in a clinical setting for diverse global populations based on the current methodology would prove to be in general less effective in non-Europeans. Deviation in RAF and the number of SNPs in LD with the SNPs used to construct the GRS was observed in both the Asian and African ancestry populations when compared to the European ancestry population. The extent of the deviation was greatest in the African ancestry population. As a result, it may be necessary to develop ancestry-specific GRS. However, due to discovery GWAS in non-European being comparatively based on smaller sample sizes which are not as well powered as European ancestry GWAS, the number discovery SNPs available for constructing GRS are often limited in non-European populations. Therefore, different approaches to selecting the SNPs or constructing the GRS may be necessary to improve the utility of GRS in non-European ancestry populations [272, 273]. The weighted T2D GRS was found to have exhibited the strongest association with AOO of T2D in all ancestry populations when compared to the unweighted T2D GRS. Hence, the weighted T2D GRS would be the preferred model for clinical implementation.

The findings pertaining to the application of the T2D GRS in European, Asian, and African ancestry population have also demonstrated that the Cox PH model were effective in detecting

an association between AOO of disease and T2D GRS. The stronger performance of the logistic model, which demonstrated the strongest association between AOO and T2D GRS when applied in the NUgene and WTCCC may in part be attributed to the fact that the discovery SNPs and weights applied in the construction of the GRS originated from models based on the logistic regression approach, which were derived from a T2D case-control GWAS. Furthermore, there are other limiting factors in relation to AOO being ascertained retrospectively in a case-control study framework which increasingly is becoming more common [274-276]. Age at diagnosis was used as a proxy measure of AOO of disease, which is limited by several factors including longitudinal biases resulting from changes in the clinical diagnostic criteria overtime as well as advances in technology that can be used to aid the clinical diagnosis of diseases. Additionally, in conducting GWAS globally, or comparing studies across populations, potential differences in clinical definitions of phenotypes or adherence to international disease classification standards across countries is important to consider, particularly for a disease such as T2D, which is clinically heterogeneous [277]. A recent review of the classification of diabetes has highlighted the challenges associated with defining T2D and the existing overlap between the different diabetes subtypes [278].

5.4. | Recommendations for future work

Areas where further research can be initiated have been highlighted in this thesis. Extensions to the simulation work undertaken in this thesis are considered as well as work relating to GRS for the prediction of AOO of common diseases in global worldwide populations.

5.4.1. | Single SNP association with AOO of disease

In this thesis to facilitate the development or improvement to existing methods designed to correct for the potential effects of population structure, a simulation study was undertaken to evaluate statistical power in relation to detecting an association between a single SNP and AOO of disease in a TTE framework. The simulations, which assumed a study period that spanned 50 years, considered the scenarios where it is assumed that the population comprised of an admixed population. The main limiting factors pertaining to these simulations are that scenarios primarily considered only two ancestral populations which formed an admixed

population were considered independently. In reality, populations are more likely to be hierarchical, consisting of both discrete ancestry groups and admixed individuals. It was also assumed that the RAF of both the causal SNP and tag SNP was the same, but in reality, this is not always the case. The simulation scenarios could be extended by increasing the level of complexity by: (1) incorporating three way or four way admixture which could represent for example admixture in South American populations (three way admixture: African, European, and Native American; and four way admixture: African, European, Native American and Asian); (2) incorporating ancestry inference based on genotype data rather than the more simplistic assumption of known ancestry that was applied in the simulations.

5.4.2. | GRS association with AOO of disease

Application of T2D GRS in three independent genotype T2D datasets has demonstrated that T2D GRS has the potential to detect an association with the AOO of disease. Several areas for further research have been highlighted in this thesis particularly as it relates to the application of T2D GRS in non-European ancestry populations. These include identifying the most appropriate statistical method for assessing AOO of disease and GRS, alternative approaches for the application of T2D GRS in non-European ancestry populations and establishment of standard age-related attributes to aid research that involves AOO of disease.

The GRS simulation results indicated that, in the presence of high censoring rates, the performance of the Cox PH, and logistic models was similar. Furthermore, the results of the analysis based on the NUgene and WTCCC datasets as well as the much larger UK Biobank dataset suggests the T2D GRS are likely to have the greatest ability to detection an association with AOO of T2D within a Cox PH modelling framework. Therefore, the Cox PH model should be assessed further using other common complex disease phenotypes such as Alzheimer's disease, arthritis and other metabolic diseases.

Further assessment of application of European ancestry derived T2D GRS in non-European populations should be considered. This could include different criteria for selection of SNPs for inclusion in the construction of GRS. As the number discovery SNPs for most non-European ancestry population are often limited due to smaller sample sizes, which are less well powered

than European ancestry GWAS, development of ancestry-specific GRS based on a combination of European derived discovery SNPs and inclusion of ancestry-specific SNPs should also be explored. Given the limited number of ancestry-specific GWAS discovery SNPs for non-European populations, it may be beneficial to also explore other approaches to constructing GRS that would allow modelling of ancestry specific LD. Such an approach would potentially reduce the loss of information in relation to ancestry-specific tagging due to LD differences between ancestries. Other approaches including mixed modelling approaches, which often comprise fixed and random effects models, could be explored for this purpose, as well as for accounting for differences in effect sizes and population structure [279, 280]. In such settings population structure is accounted for by means of genetic relationship matrix (GRM). The GRM which can account for both population structure and relatedness (described in section 1.5.2) is constructed from individual sampled genotyped data comprising genome-wide SNPs. The relationship of this GRM with the phenotype of interest is evaluated in a random effects model framework, which can also incorporate a fixed effects component to accounts for other potential confounding effects by including them as covariates in the [123]. The application of GRS to AOO of disease to other common complex disease phenotypes should also be considered.

With the decrease in prospective cohort studies and increasing use of electronic medical and health data systems, data for research is increasingly being collected retrospectively. Therefore, the data used in research is vulnerable to different data quality issues. In respect to age, the impact of errors in AOO need to be given further consideration. In the area of age or AOO of disease, efforts to improve or standardize the recording of age and age-related attributes globally is essential. As cases and controls are often reused in different studies and at different time points, establishing a mandatory component to the recording of age and age-related attributes may be beneficial in terms of improving the quality of AOO captured in electronic medical and health data systems. Attributes mandatory for cases could include age at diagnosis, year of diagnosis, year of birth, age at DNA extraction and year of DNA sample extraction. In relation to controls, mandatory elements could include year of birth, age at DNA extraction and year of DNA sample extraction. Another important consideration relates to the selection of controls. Careful consideration should be given to the fact that the use of controls younger than the cases reduces the power of the analysis as controls given time may develop the disease.

5.5. | Concluding remarks

In conclusion, the research in this thesis has demonstrated that GRS have the potential to advance common disease genetic research in relation to AOO. GWAS discoveries to date have provided valuable insight in terms of disease risk, but increasing research dedicated to understanding the disease biology of complex diseases in relation to AOO is an important next step. Implementation of better targeted screening strategies informed by knowledge of the AOO of disease is key to reducing treatment costs as well as facilitating improvements to survival rates. Additionally, improvements to methods developed to detect and account for population structure is paramount for GWAS discoveries as sample sizes continue to grow and for the clinical implementation of risk prediction models based on GRS. The application of GRS in ancestrally diverse or admixed populations is key to the realization of the vision of personalized medicine or personalized healthcare for all.

References

[1] Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Research 2012;40(9):3777-84.

[2] Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. American journal of human genetics 2010;86(1):6-22.

[3] Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Computational Biology 2012;8(12):e1002822.

[4] Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. Science 2005;308(5720):419.

[5] Jansen PR, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag AR, et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nature genetics 2019;51(3):394-+.

[6] Eun PH, Ji WP. Sample Size and Statistical Power Calculation in Genetic Association Studies. Genomics & Informatics 2012;10(2):117-22.

[7] Price AL, Spencer CCA, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. Proceedings of the Royal Society of London - Series B: Biological Sciences 2015;282(1821):20151684.

[8] Zaitlen N, Kraft P. Heritability in the genome-wide association era. Human genetics 2012;131(10):1655-64.

[9] Riancho JA. Genome-wide association studies (GWAS) in complex diseases: Advantages and limitations. Reumatologia Clinica 2012;8(2):56-7.

[10] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics 2017;101(1):5-22.

[11] Carolina M-G, Janine Frédérique F, Estrada K, Marjoline JP, Herrera L, Claudia JK, et al. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. European journal of epidemiology 2015;30(4):317.

[12] Cook JP, Morris AP. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. European Journal of Human Genetics 2016;24(8):1175.

[13] Morris AP. Transethnic meta-analysis of genomewide association studies. Genetic epidemiology 2011;35(8):809-22.

[14] Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. Current protocols in human genetics 2017;95(1):1.22. 1-1.. 3.

[15] Teo YY. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Current opinion in lipidology* 2008;19(2):133-43.

[16] Marigorta UM, Rodríguez JA, Gibson G, Navarro A. Replicability and prediction: lessons and challenges from GWAS. Trends in Genetics 2018;34(7):504-17.

[17] de los Campos G, Vazquez AI, Hsu S, Lello L. Complex-trait prediction in the era of big data. Trends in Genetics 2018;34(10):746-54.

[18] Joiret M, John JMM, Gusareva ES, Van Steen K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. BioData Mining 2019;12(1):11.
[19] King RC, Mulligan PK, Stansfield WD. A dictionary of genetics. Eighth edition. ed. New York: Oxford University Press; 2013.
[20] Porta MS, Last JM. A dictionary of public health. 2nd ed. ed. New YorK: Oxford University Press; 2018.
[21] Dorak TM. Common Terms in Genetics; 2015. Available from: http://www.dorak.info/genetics/glosgen.html. [Accessed 11/17 2019].
[22] Liu Y. Genetic Diversity and Disease Susceptibility. InTech; 2018.
[23] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature 2015;526(7571):68-74.
[24] Elmas A, Yang T-HO, Wang X, Anastassiou D. Discovering Genome-Wide tag SNPs based on the mutual information of the variants. PloS one 2016;11(12):e0167994.
[25] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols 2010;5(9):1564.
[26] Charles BA, Shriner D, Rotimi CN. Accounting for linkage disequilibrium in association analysis of diverse populations. Genetic epidemiology 2014;38(3):265-73.
[27] Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 2008;9(6):477-85.
[28] Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annual Review of Genomics & Human Genetics 2008;9:403-33.
[29] Rice JP. Human Linkage and Association Analysis. Psychiatric Genetics: A Primer for Clinical and Basic Scientists 2018.
[30] Waksmunski AR, Main LR, Haines JL. Segregation, linkage, GWAS, and sequencing. Genetics and Genomics of Eye Disease. Elsevier; 2020, p. 7-23.
[31] Bailey-Wilson JE, Wilson AF. Linkage analysis in the next-generation sequencing era. Human heredity 2011;72(4):228-36.
[32] Stein CM. Identifying genes underlying human inherited disease. e LS 2010:1-7.
[33] Rodriguez-Murillo L, Greenberg DA. Genetic association analysis: a primer on how it works, its strengths and its weaknesses. International journal of andrology 2008;31(6):546-56.
[34] Ellinghaus E, Ellinghaus D. Genetic Association Analysis / GWAS; Available from: https://www.ikmb.uni-kiel.de/research/genetics-bioinformatics/genetic-association-analysis-gwas. [Accessed 11/17 2019].
[35] Modena BD, Doroudchi A, Patel P, Sathish V. Leveraging genomics to uncover the genetic, environmental and age-related factors leading to asthma. Genomic and Precision Medicine : Infectious and Inflammatory Disease. Elsevier; 2019, p. 331-81.
[36] Yuan J, Tickner J, Mullin B, Zhao J, Zeng Z, Morahan G, et al. Advanced genetic approaches in discovery and characterisation of genes involved with osteoporosis in mouse and human. Frontiers in genetics 2019;10:288.
[37] Norris ET, Wang L, Conley AB, Rishishwar L, Mariño-Ramírez L, Valderrama-Aguirre A, et al. Genetic ancestry, admixture and health determinants in Latin America. BMC genomics 2018;19(8):75-87.
[38] Shriner D. Overview of admixture mapping. Current protocols in human genetics 2013;76(1):1.23. 1-1.. 8.
[39] Smith MW, O'Brien SJ. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nature Reviews Genetics 2005;6(8):623.

[40] Du Y, Xie J, Chang W, Han Y, Cao G. Genome-wide association studies: inherent limitations and future challenges. Fronteras en Medicina 2012;6(4):444-50.
[41] Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661.
[42] Huang TD. An Overview of Genome-Wide Association Studies. New York, NY: Humana Press; 2018.
[43] (NHGRI) NHGRI. Genome-Wide Association Studies Fact Sheet; 2019. Available from: https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet. [Accessed 08/24 2019].
[44] Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273(5281):1516-7.
[45] Fadista J, Manning AK, Florez JC, Groop L. The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. European Journal of Human Genetics 2016;24(8):1202-5.
[46] Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Current protocols in human genetics 2011;68(1):1.19. 1-1.. 8.
[47] Diabetologia, Journal of the European Association for the Study of D. Guidelines for Genetic Association Studies; 2020. Available from: https://diabetologia-journal.org/for-authors-and-reviewers/guidelines-for-genetic-association-studies/. [Accessed 02/01 2020].
[48] Evangelou E. Genetic epidemiology : methods and protocols. New York (NY): Humana Press; 2018.
[49] Huffman JE. Examining the current standards for genetic discovery and replication in the era of mega-biobanks. Nature communications 2018;9(1):5054.
[50] Hsu Y-H, Kiel DP. Genome-wide association studies of skeletal phenotypes: what we have learned and where we are headed. The Journal of Clinical Endocrinology & Metabolism 2012;97(10):E1958-E77.
[51] Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26(17):2190-1.
[52] Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2010;34(1):60-6.
[53] Zintzaras E, Lau J. Trends in meta-analysis of genetic association studies. Journal of human genetics 2008;53(1):1.
[54] Bai W-Y, Zhu X-W, Cong P-K, Zhang X-J, Richards JB, Zheng H-F. Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. Briefings in Bioinformatics 2019.
[55] Berlanga-Taylor AJ. From Identification to Function: Current Strategies to Prioritise and Follow-Up GWAS Results. Genetic Epidemiology. Springer; 2018, p. 259-75.
[56] Staples J, Maxwell EK, Gosalia N, Gonzaga-Jauregui C, Snyder C, Hawes A, et al. Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. The American Journal of Human Genetics 2018;102(5):874-89.
[57] Sul JH, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models. PLoS genetics 2018;14(12):e1007309.
[58] Helgason A, Yngvadóttir B, Hrafnkelsson B, Gulcher J, Stefánsson K. An Icelandic example of the impact of population structure on association studies. Nature genetics 2005;37(1):90-5.

[59] Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, et al. The genome-wide patterns of variation expose significant substructure in a founder population. The American Journal of Human Genetics 2008;83(6):787-94.
[60] Barton N, Hermisson J, Nordborg M. Population Genetics: Why structure matters. Elife 2019;8:e45380.
[61] Bouaziz M, Ambroise C, Guedj M. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. PloS one 2011;6(12):e28845.
[62] Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Statistical Science 2009;24(4):451-71.
[63] Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of african americans, latinos, and european Americans across the United States. The American Journal of Human Genetics 2015;96(1):37-53.
[64] Murray T, Beaty TH, Mathias RA, Rafaels N, Grant AV, Faruque MU, et al. African and non-African admixture components in African Americans and an African Caribbean population. Genetic epidemiology 2010;34(6):561-8.
[65] Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, et al. Genomic insights into the ancestry and demographic history of South America. PLoS genetics 2015;11(12):e1005602.
[66] Seldin MF, Pasaniuc B, Price AL. New approaches to disease mapping in admixed populations. Nature Reviews Genetics 2011;12(8):523.
[67] Feng Q, Lu Y, Ni X, Yuan K, Yang Y, Yang X, et al. Genetic history of Xinjiang's Uyghurs suggests Bronze Age multiple-way contacts in Eurasia. Molecular biology and evolution 2017;34(10):2572-82.
[68] Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics & informatics 2012;10(2):117-22.
[69] Yang Y, Remmers EF, Ogunwole CB, Kastner DL, Gregersen PK, Li W. Effective sample size: Quick estimation of the effect of related samples in genetic case–control association analyses. Computational biology and chemistry 2011;35(1):40-9.
[70] Pirinen M, University of H. GWAS 3: Statistical power; 2019. Available from: https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/2019/material/GWAS3.pdf. [Accessed 11/19 2019].
[71] Vukcevic D, Hechter E, Spencer C, Donnelly P. Disease model distortion in association studies. Genetic epidemiology 2011;35(4):278-90.
[72] Molster CM, Bowman FL, Bilkey G, Cho AS, Burns BL, Nowak KJ, et al. The evolution of public health genomics: exploring its past, present and future. Frontiers in public health 2018;6:247.
[73] Spector SA, Brummel SS, Nievergelt CM, Maihofer AX, Singh KK, Purswani MU, et al. Genetically determined ancestry is more informative than self-reported race in HIV-infected and -exposed children. Medicine 2016;95(36).
[74] Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M. The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. PloS one 2013.
[75] Schrodi SJ. The Impact of Diagnostic Code Misclassification on Optimizing the Experimental Design of Genetic Association Studies. Journal of healthcare engineering 2017;2017.
[76] Ryan J, Fransquet P, Wrigglesworth J, Lacaze P. Phenotypic heterogeneity in dementia: a challenge for epidemiology and biomarker studies. Frontiers in public health 2018;6:181.

[77] Llinares-López F, Grimm DG, Bodenham DA, Gieraths U, Sugiyama M, Rowan B, et al. Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. Bioinformatics 2015;31(12):i240-i9.
[78] Orlova E, Carlson JC, Lee M-K, Feingold E, McNeil DW, Crout RJ, et al. Pilot GWAS of caries in African-Americans shows genetic heterogeneity. BMC oral health 2019;19(1):215.
[79] Wei C, Elston RC, Lu Q. A weighted u statistic for association analyses considering genetic heterogeneity. Statistics in medicine 2016;35(16):2802-14.
[80] Keith BP, Robertson DL, Hentges KE. Locus heterogeneity disease genes encode proteins with high interconnectivity in the human protein interaction network. Frontiers in genetics 2014;5:434.
[81] Leiserson MDM, Eldridge JV, Ramachandran S, Raphael BJ. Network analysis of GWAS data. Current opinion in genetics & development 2013;23(6):602-10.
[82] McClellan J, King M-C. Genetic heterogeneity in human disease. Cell 2010;141(2):210-7.
[83] Rees SD, Hydrie MZI, Shera AS, Kumar S, O'Hare JP, Barnett AH, et al. Replication of 13 genome-wide association (GWA)-validated risk variants for type 2 diabetes in Pakistani populations. Diabetologia 2011;54(6):1368-74.
[84] Alpert JE, Fava M. Handbook of chronic depression: Diagnosis and therapeutic management. New York: CRC Press; 2014.
[85] Hong H, Xu L, Su Z, Liu J, Ge W, Shen J, et al. Pitfall of genome-wide association studies: Sources of inconsistency in genotypes and their effects. Journal of Biomedical Science and Engineering 2012;5:557-73.
[86] Yuan M, Fang H, Zhang H. Correcting for differential genotyping error in genetic association analysis. Journal of human genetics 2013;58(10):657-66.
[87] Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. Annual review of genomics and human genetics 2018;19:73-96.
[88] Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics 2010;11(7):499-511.
[89] Shi S, Yuan N, Yang M, Du Z, Wang J, Sheng X, et al. Comprehensive Assessment of Genotype Imputation Performance. Human heredity 2018;83(3):107-16.
[90] Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International Journal Of Methods In Psychiatric Research 2018;27(2):e1608-e.
[91] Coleman JRI, Euesden J, Patel H, Folarin AA, Newhouse S, Breen G. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. Briefings in functional genomics 2016;15(4):298-304.
[92] Hine RE. Hardy-Weinberg equilibrium. Oxford University Press; 2019.
[93] Mills MC, Barban N, Tropf FC. An Introduction to Statistical Genetic Data Analysis. Cambridge, Massachusetts USA: MIT Press; 2020.
[94] Zeggini E, Southam L, Panoutsopoulou K, Rayner NW, Chapman K, Durrant C, et al. The effect of genome-wide association scan quality control on imputation outcome for common variants. European Journal of Human Genetics 2011.
[95] Wang MH, Cordell HJ, Van Steen K. Statistical methods for genome-wide association studies. *Seminars in cancer biology.* 55. Elsevier:53-60.
[96] Van Leeuwen EM, Kanterakis A, Deelen P, Kattenberg MV, Abdellaoui A, Hofman A, et al. Population-specific genotype imputations using minimac or IMPUTE2. Nature protocols 2015;10(9):1285.

[97] Hancock DB, Levy JL, Gaddis NC, Bierut LJ, Saccone NL, Page GP, et al. Assessment of genotype imputation performance using 1000 Genomes in African American studies. PloS one 2012;7(11):e50610.
[98] Zheng H-F, Rong J-J, Liu M, Han F, Zhang X-W, Richards JB, et al. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PLoS One 2015;10(1):1-10.
[99] Gilly A, Kuchenbaecker K, Southam L, Suveges D, Moore R, Melloni G, et al. Very low depth whole genome sequencing in complex trait association studies. bioRxiv 2017.
[100] Pulit SL, Leusink M, Menelaou A, Paul I W de B. Association Claims in the Sequencing Era. Genes 2014;5(1):196-213.
[101] Lin Y. The multiple comparison problem in GWAS: Bonferroni correction, FDR control, and permutation testing; 2015. Available from: http://lybird300.github.io/2015/10/19/multiple-test-correction.html. [Accessed 02/01 2020].
[102] Padhukasahasram B. Inferring ancestry from population genomic data and its applications. Front Genet 2014; 5: 204. *Frontiers in genetics* 2014;5:204.
[103] Yuan K, Zhou Y, Ni X, Wang Y, Liu C, Xu S. Models, methods and tools for ancestry inference and admixture analysis. Quantitative Biology 2017;5(3):236-50.
[104] Thornton TA, Bermejo JL. Local and global ancestry inference and applications to genetic association analysis for admixed populations. Genetic epidemiology 2014;38(S1):S5-S12.
[105] Alhusain L, Hafez AM. Nonparametric approaches for population structure analysis. Human genomics 2018;12(1):25.
[106] Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo Á, Lareu M. An overview of STRUCTURE: applications, parameter settings, and supporting software. Frontiers in genetics 2013;4:98.
[107] Lin DY, Zeng D. Correcting for population stratification in genomewide association studies. Journal of the American Statistical Association 2011;106(495):997-1008.
[108] Yushi L, Toru N, Shuguang L, Belinsky Steven A, Yohannes T, Shannon B. Softwares and methods for estimating genetic ancestry in human populations. *Human genomics* 2013;7(1):1-7.
[109] Winkler CA, Nelson GW, Smith MW. Admixture mapping comes of age. Annual review of genomics and human genetics 2010;11:65-89.
[110] Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Science International-Genetics 2007;1(3-4):273-80.
[111] Chen G, Shriner D, Zhou J, Doumatey A, Huang H, Gerry NP, et al. Development of admixture mapping panels for African Americans from commercial high-density SNP arrays. BMC genomics 2010;11(1):417.
[112] Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2002;11(6):505-12.
[113] Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theoretical Population Biology 2001;60(3):155-66.
[114] Dorak MT. Genetic association studies: background, conduct, analysis, interpretation. United Kingdom, Europe: Garland Science, Taylor & Francis Group; 2017.
[115] Wu C, DeWan A, Hoh J, Wang Z. A comparison of association methods correcting for population stratification in case–control studies. Annals of Human Genetics 2011;75(3):418-27.

[116] Lee JJ, McGue M, Iacono WG, Chow CC. The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. Genetic epidemiology 2018;42(8):783-95.
[117] Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics 2015;47(3):291.
[118] Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. The American Journal of Human Genetics 2016;98(4):653-66.
[119] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 2006;38(8):904.
[120] Umit S. Hands-on tutorial to Genome-wide Association Studies (GWAS); 2015. Available from: http://www.transplantdb.eu/sites/transplantdb.eu/files/HandsOnTutorialtoGWAS_Seren-030715.pdf. [Accessed 08/23 2019].
[121] Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 2010;11(7):459.
[122] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature genetics 2014;46(2):100.
[123] Cook JP, Mahajan A, Morris AP. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. European Journal of Human Genetics 2017;25(2):240-5.
[124] Wollstein A, Lao O. Detecting individual ancestry in the human genome. Investigative genetics 2015;6(1):7.
[125] Shawky RM. Reduced penetrance in human inherited disease. Egyptian Journal of Medical Human Genetics 2014;15(2):103-11.
[126] Hughes DJ. Use of association studies to define genetic modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers. Familial cancer 2008;7(3):233-44.
[127] Frayling TM. Genome-wide association studies: the good, the bad and the ugly. Clinical medicine (London, England) 2014;14(4):428-31.
[128] Rieger R, Michaelis A, Green MM. Glossary of genetics: classical and molecular. Fifth Edition ed. Berlin: Springer Science & Business Media; 2012.
[129] Giral H, Landmesser U, Kratzer A. Into the Wild: GWAS Exploration of Non-coding RNAs. Frontiers in cardiovascular medicine 2018;5:181.
[130] Gallagher MD, Chen-Plotkin AS. The post-GWAS era: from association to function. The American Journal of Human Genetics 2018;102(5):717-30.
[131] Parkes M, Investigators IBDB. IBD BioResource: an open-access platform of 25 000 patients to accelerate research in Crohn's and Colitis.. Gut. 2019;68(9):1538-40.
[132] Hill-Burns EM, Ross OA, Wissemann WT, Soto-Ortolaza AI, Zareparsi S, Siuda J, et al. Identification of genetic modifiers of age-at-onset for familial Parkinson's disease. Human molecular genetics 2016;25(17):3849-62.
[133] Guerreiro R, Bras J. The age factor in Alzheimer's disease. Genome medicine 2015;7(1):106.
[134] Oliynyk RT. Age-related late-onset disease heritability patterns and implications for genome-wide association studies. PeerJ 2019;7:e7168.
[135] Kamboh MI, Barmada MM, Demirci FY, Minster RL, Carrasquillo MM, Pankratz VS, et al. Genome-wide association analysis of age-at-onset in Alzheimer's disease. Molecular psychiatry 2012;17(12):1340.

[136] Lill CM, Hansen J, Olsen JH, Binder H, Ritz B, Bertram L. Impact of Parkinson's disease risk loci on age at onset. Movement Disorders 2015;30(6):847-50.
[137] Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nature Reviews Genetics 2016;17(7):392.
[138] Evangelou E. Methods for Polygenic Traits. New York (NY): Humana Press; 2018.
[139] Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nature communications 2018;9(1):1-14.
[140] Statistics S. Coefficient of Determination; 2019. Available from: https://www.statisticssolutions.com/coefficient-of-determination/. [Accessed 12/04 2019].
[141] Flandre P, Deutsch R, O'Quigley J. Accuracy of predictive ability measures for survival models. Statistics in medicine 2017;36(20):3171-80.
[142] McFadden D. Conditional logit analysis of qualitative choice behavior. In: Frontiers in Economics. P Zarembka, eds New York: Academic Press 1974.
[143] Cox DR, Snell EJ. The Analysis of Binary Data. 2nd ed ed. London: Chapman and Hall; 1989.
[144] Nagelkerke N, J D. A note on the general definition of the coefficient of determination. Biometrika 1991;78(3):691-2.
[145] Hansen TF, Møller RS. The first step towards personalized risk prediction for common epilepsies. Brain 2019;142(11):3316-8.
[146] Florez JC. The genetics of type 2 diabetes and related traits. Switzerland: Springer International Publishing 2016.
[147] Chegg S. Chegg Study Textbook Solutions Expert Q and A: Coefficient of Partial Determination; 2019. Available from: https://www.chegg.com/homework-help/definitions/coefficient-of-partial-determination-31. [Accessed 12/04 2019].
[148] George B, Seals S, Aban I. Survival analysis and regression models. Journal of Nuclear Cardiology 2014;21(4):686-94.
[149] He L, Kulminski AM. Genome-wide association analysis of age-at-onset traits using Cox mixed-effects models. bioRxiv 2019:729285.
[150] Syed H, Jorgensen AL, Morris AP. Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. Pharmacogenomics 2016;17(8):907-15.
[151] Kleinbaum DG, Klein M. Survival analysis. a self-learning text. 3rd ed. ed. New York: Springer; 2012.
[152] Emmert-Streib F, Dehmer M. Introduction to Survival Analysis in Practice. Machine Learning and Knowledge Extraction 2019;1(3):1013-38.
[153] Schober P, Vetter TR. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesthesia and Analgesia 2018;127(3):792-8.
[154] Hassanzadeh J, Moradzadeh R, Rajaee Fard A, Tahmasebi S, Golmohammadi P. A comparison of case-control and case-only designs to investigate gene-environment interactions using breast cancer data. Iranian journal of medical sciences 2012;37(2):112-8.
[155] Harrell FE. Cox proportional hazards regression model. Regression modeling strategies. Springer; 2015, p. 475-519.
[156] Moolgavkar SH, Chang ET, Watson HN, Lau EC. An assessment of the Cox proportional hazards regression model for epidemiologic studies. Risk Analysis 2018;38(4):777-94.

[157] Khan SA, Khosa SK. Generalized log-logistic proportional hazard model with applications in survival analysis. Journal of Statistical Distributions and Applications 2016;3(1):1-18.
[158] McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological) 1980;42(2):109-27.
[159] McCullagh P. Proportional-odds model. Encyclopedia of Biostatistics 2005;6.
[160] Cramer JS. The origins of logistic regression. Tinbergen Institute Working Paper; 2002.
[161] Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al. Assessing the impact of population stratification on genetic association studies. Nature genetics 2004;36(4):388.
[162] Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nature genetics 2004;36(5):512.
[163] Campbell MC, Hirbo JB, Townsend JP, Tishkoff SA. The peopling of the African continent and the diaspora into the new world. Current opinion in genetics & development 2014;29:120-32.
[164] Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nature Reviews Genetics 2010;11(5):356.
[165] Mersha TB. Mapping asthma-associated variants in admixed populations. Frontiers in genetics 2015;6:292.
[166] Balding DJ. A tutorial on statistical methods for population association studies. Nature reviews genetics 2006;7(10):781-91.
[167] Matloff NS. The art of R programming: a tour of statistical software design. San Francisco, USA: No Starch Press; 2011.
[168] Moore DF. Applied survival analysis using R. 1 ed. Switzerland: Springer; 2016.
[169] Therneau TM, Lumley T. Package 'survival'. R Top Doc 2015;128(10):28-33.
[170] Crumer AM. Comparison between Weibull and Cox proportional hazards models. Manhattan, Kan. : Kansas State University, 2011.; 2011.
[171] Adejumo AO, Ahmadu AO. A Study of The Slope of Cox Proportional Hazard and Weibull Models: Simulated and Real Life Data Approach. Science World Journal 2016;11(3):31-5.
[172] Lee YL, Teitelbaum S, Wolff MS, Chen J, Wetmur JG. Comparing genetic ancestry and self-reported race/ethnicity in a multiethnic population in New York City. Journal of Genetics 2010;89(4):417-23.
[173] Martin ER, Tunc I, Liu Z, Slifer SH, Beecham AH, Beecham GW. Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. Genetic epidemiology 2018;42(2):214-29.
[174] Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating Local Ancestry in Admixed Populations. 2008.
[175] Lee YH. Meta-analysis of genetic association studies. Annals of laboratory medicine 2015;35(3):283-7.
[176] Suzuki A, Yamamoto K. From genetics to functional insights into rheumatoid arthritis. Clinical and experimental rheumatology 2015;33(4 Suppl 92):40-3.
[177] Persico AM, Verdecchia M, Pinzone V, Guidetti V. Migraine genetics: current findings and future lines of research. Neurogenetics 2015;16(2):77-95.
[178] Struck TJ, Mannakee BK, Gutenkunst RN. The impact of genome-wide association studies on biomedical research publications. Human Genomics 2018;12(1):1-9.
[179] Sanghera DK, Blackett PR. Type 2 Diabetes Genetics: Beyond GWAS. Journal of diabetes & metabolism 2012;3(198):1-23.
[180] Brian C, Leutholtz, Ignacio R. Chapter 2 Diabetes. Exercise and Disease Management. USA: Taylor & Francis Group, LLC; 2011, p. 25-47.

[181] Organisation WH. Diabetes: fact sheets; 2018. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed 08/23 2019].
[182] Karuranga S. IDF Diabetes Atlas. Eighth edition ed. Belgium: International Diabetes Federation; 2017.
[183] UK D. Diabetes: facts and stats. Diabetes UK 2014;3:1-21.
[184] Prasad RB, Groop L. Genetics of Type 2 Diabetes—Pitfalls and Possibilities. Genes 2015;6(1):87-123.
[185] Zaccardi F, Webb DR, Yates T, Davies MJ. Pathophysiology of type 1 and type 2 diabetes mellitus: a 90-year perspective. Postgraduate medical journal 2016;92(1084):63-9.
[186] Leslie RD, Palmer J, Schloot NC, Lernmark A. Diabetes at the crossroads: relevance of disease classification to pathophysiology and treatment. Diabetologia 2016;59(1):13-20.
[187] Florez JC, Udler MS, Hanson RL. Genetics of type 2 diabetes. In: Cowie CC CS, Menke A, et al., editor Diabetes in America. Bethesda: National Institutes of Health; 2016.
[188] Ali O. Genetics of type 2 diabetes. World journal of diabetes 2013;4(4):114-23.
[189] Almgren P LM, Isomaa BO, Sarelin LE, Taskinen MR, Lyssenko V, Tuomi T, Groop L. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 2011;54(11):2811-9.
[190] Willemsen G, Ward, K.J., Bell, C.G., Christensen, K., Bowden, J., Dalgård, C., Harris, J.R., Kaprio, J., Lyle, R., Magnusson, P.K. and Mather, K.A.,. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers:The Discordant Twin (DISCOTWIN) Consortium. *Twin Research and Human Genetics* 2015;18(6):762-71.
[191] Läll K, Mägi R, Morris A, Metspalu A, Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics in Medicine* 2017;19(3):322-9.
[192] Langenberg C, Lotta LA. Genomic insights into the causes of type 2 diabetes. The Lancet 2018;391(10138):2463-74.
[193] Shahvazian E, Yazd EF, Sheikhha MH, Rahmanian M. Genetics of Type 2 Diabetes- A Review Article. Iranian Journal of Diabetes & Obesity (IJDO) 2015;7(4):187-95.
[194] Thanabalasingham G, Gloyn AL, Owen KR. Genome-wide association studies of Type 2 diabetes: are these ready to make an impact in the clinic? Diabetes Management - Future Medicine 2011(4):379-87.
[195] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. The American Journal of Human Genetics 2012;90(1):7-24.
[196] Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nature Reviews Genetics 2018(3):175.
[197] Goto A, Noda M, Goto M, Yasuda K, Mizoue T, Yamaji T, et al. Predictive performance of a genetic risk score using 11 susceptibility alleles for the incidence of Type 2 diabetes in a general Japanese population: a nested case-control study. Diabetic Medicine 2018(5):602.
[198] Staley JR, Jones E, Kaptoge S, Butterworth AS, Sweeting MJ, Wood AM, et al. A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. European Journal of Human Genetics 2017;25(7):854-62.
[199] Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of chronic diseases* 1983;36(10):715-23.

[200] Callas PW, Pastides H, Hosmer DW. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. American Journal of Industrial Medicine 1998;33(1):33-47.
[201] van der Net JB, Janssens AC, Eijkemans MJC, Kastelein JJP, Sijbrands EJG, Steyerberg EW. Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. European Journal Of Human Genetics: EJHG 2008;16(9):1111-6.
[202] Leonenko G, Sims R, Shoai M, Frizzati A, Bossù P, Spalletta G, et al. Polygenic risk and hazard scores for Alzheimer's disease prediction. Annals of Clinical & Translational Neurology 2019;6(3):456-65.
[203] Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Statistics in medicine 1989;8(12):1515-21.
[204] Nunez E, Steyerberg EW, Nunez J. Regression Modeling Strategies. *Revista Española de Cardiología (English Edition)* 2011;*64*(6):501-7.
[205] Scott RA, Scott LJ, Magi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes 2017;66(11):2888-902.
[206] Iglesias AI, Van Der Lee SJ, Bonnemaijer PWM, Höhn R, Nag A, Gharahkhani P, et al. Haplotype reference consortium panel: practical implications of imputations with large reference panels. Human mutation 2017;38(8):1025-32.
[207] Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics* 2014;46(3):234-44.
[208] Voight BF, Scott LJ, McCulloch LJ, Ferreira T, Grallert H, Amin N, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature genetics 2010;42(7):579-89.
[209] Morris AP, Voight BF, Prokopenko I, Hyun Min K, Dina C, Eskoe T, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics 2012;44(9):981-90.
[210] Zeggini E, Scott LJ, De Bakker PIW, Abecasis GR, Almgren P, Andersen G, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature genetics 2008;40(5):638-45.
[211] Dupuis J, Langenberg C, Lindgren CM, MÄGi R, Morris AP, Randall J, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature genetics 2010;42(2):105-16.
[212] National Center for Biotechnology Information USNLoM. Northwestern NUgene Project: Type 2 Diabetes; 2019. Available from: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/dataset.cgi?study_id=phs000237.v1.p1&phv=158824&phd=3361&pha=&pht=2162&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1. [Accessed 11/11 2019].
[213] The Wellcome Trust Case Control C, Zeggini E, Weedon Michael N, Lindgren Cecilia M, Frayling Timothy M, Elliott Katherine S, et al. Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. Science 2007;316(5829):1336-41.
[214] Haplotype Reference C. Haplotype Reference Consortium (HRC); 2019. Available from: http://www.haplotype-reference-consortium.org/. [Accessed 11/18 2019].
[215] Institute WS. Wellcome Trust Case Control Consortium (WTCCC); 2019. Available from: https://www.wtccc.org.uk/ccc1/overview.html. [Accessed 11/11 2019].

[216] Health USNIo. Michigan Imputation Server; 2019. Available from: https://imputationserver.sph.umich.edu/index.html#!pages/home. [Accessed 11/18 2019].
[217] Hurley MA. A reference relative time-scale as an alternative to chronological age for cohorts with long follow-up. Emerging themes in epidemiology 2015;12(1):18.
[218] England PH. Adult obesity and type 2 diabetes. England: Public Health England London; 2014.
[219] Rupal S. Assessing the risk of diabetes. *BMJ: British Medical Journal* 2015;351:1-3.
[220] Wang S, Ma W, Yuan Z, Wang SM, Yi X, Jia H, et al. Association between obesity indices and type 2 diabetes mellitus among middle-aged and elderly people in Jinan, China: a cross-sectional study. BMJ open 2016;6(11):e012742.
[221] EunJin A, Hyun K. Introduction to systematic review and meta-analysis. Korean Journal of Anesthesiology 2018;71(2):103-12.
[222] Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics 2013.
[223] Lee CH, Cook S, Lee JS, Han B. Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. Genomics & informatics 2016;14(4):173-80.
[224] Heard NA, Rubin-Delanchy P. Choosing between methods of combining p -values. Biometrika 2018;105(1):239-46.
[225] Whitlock MC. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. Journal of Evolutionary Biology 2005;18(5):1368-73.
[226] Center for Statistical Genetics SoPH, University of Michigan. Stouffer Method for Meta-Analysis; 2010. Available from: https://genome.sph.umich.edu/wiki/Stouffer_Method_for_Meta-Analysis. [Accessed 07/08 2020].
[227] Toro-Domínguez D, Villatoro-García JA, Martorell-Marugán J, Román-Montoya Y, Alarcón-Riquelme ME, Carmona-Sáez P. A survey of gene expression meta-analysis: methods and applications. Briefings in bioinformatics 2020.
[228] Zaykin DV. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. Journal of Evolutionary Biology 2011;24(8):1836-41.
[229] Chen Z, Saralees N. On the optimally weighted z-test for combining probabilities from independent studies. Computational statistics & data analysis 2014;70:387-94.
[230] Pathak M, Dwivedi SN, Deo SVS, Sreenivas V, Thakur B. Which is the preferred measure of heterogeneity in meta-analysis and why? A revisit. Biostat Biometrics Open Acc 2017;1:1-7.
[231] Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane handbook for systematic reviews of interventions. Second ed.: John Wiley & Sons; 2019.
[232] West SL. Comparative Effectiveness Review Methods: Clinical Heterogeneity. Rockville (MD)- USA: Agency for Healthcare Research and Quality; 2010.
[233] Neupane B, Loeb M, Anand SS, Beyene J. Meta-analysis of genetic association studies under heterogeneity. European journal of human genetics 2012;20(11):1174-81.
[234] Rao G, Lopez-Jimenez F, Boyd J, D'Amico F, Durant NH, Hlatky MA, et al. Methodological Standards for Meta- Analyses and Qualitative Systematic Reviews of Cardiac Prevention and Treatment Studies. Circulation 2017;136(10):e172-e94.
[235] Spineli LM, Pandis N. Statistical heterogeneity: Notion and estimation in meta-analysis. American Journal of Orthodontics and Dentofacial Orthopedics 2020;157(6):856-9.e2.
[236] Pei Y-F, Tian Q, Zhang L, Deng H-W. Exploring the Major Sources and Extent of Heterogeneity in a Genome-Wide Association Meta-Analysis. Annals of Human Genetics 2016;80(2):113-22.

[237] Wang A, Tan Y, Zhang Y, Xu D, Fang Y, Chen X, et al. The prognostic role of angiolymphatic invasion in N0 esophageal carcinoma: a meta-analysis and systematic review. Journal of Thoracic Disease 2019;11(8):3276.
[238] Umaporn S, Archer KJ. Estimation of random effects and identifying heterogeneous genes in meta-analysis of gene expression studies. Briefings in Bioinformatics 2017;18(4):602-18.
[239] Nawrot TS, Thijs L, Den Hond EM, Roels HA, Staessen JA. An epidemiological re-appraisal of the association between blood pressure and blood lead: a meta-analysis. Journal of Human Hypertension 2002;16(2):123.
[240] Yihan L, Ghosh D. Meta-analysis based on weighted ordered P-values for genomic data with heterogeneity. BMC Bioinformatics 2014;15(1):1-22.
[241] Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic acids research 2012;40(9):3785-99.
[242] Schimmack U. Visual Inspection of Strength of Evidence: P-Curve vs. Z-Curve; 2018. Available from: https://replicationindex.com/2018/04/05/visual-inspection-of-strength-of-evidence-p-curve-vs-z-curve/. [Accessed 07/08 2020].
[243] Ramaswami R, Bayer R, Galea S. Precision Medicine from a Public Health Perspective. Annual Review of Public Health 2018;39:153-68.
[244] Resnick B. The overwhelming whiteness of genetics research is holding back medicine; 2018. Available from: https://www.vox.com/science-and-health/2018/10/22/17983568/dna-tests-precision-medicine-genetics-gwas-diversity-all-of-us. [Accessed 10/04 2019].
[245] Reisberg S, Iljasenko T, Läll K, Fischer K, Vilo J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. PLoS ONE 2017;12(7):1-9.
[246] Unnikrishnan R, Pradeepa R, Joshi SR, Mohan V. Type 2 Diabetes: Demystifying the Global Epidemic. Diabetes 2017;66(6):1432-42.
[247] International Diabetes F. International Diabetes Federation Diabetes Atlas. 8th ed.; 2017.
[248] Lacour A, Schüller V, Drichel D, Herold C, Jessen F, Leber M, et al. Novel genetic matching methods for handling population stratification in genome-wide association studies. BMC bioinformatics 2015;16(1):84.
[249] Mills MC, Rahal C. A scientometric review of genome-wide association studies. *Communications biology* 2019;2(1):1-11.
[250] Wolford B. The Nascent Transcript – November 2018 - ASHG 2018 Recap: Polygenic Risk Scores. 2019. 2019.
[251] Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome biology 2016;17(1):157.
[252] Jamuar S, Picker J, Jamuar S, Turpaz Y. From human genome to global genome–why precision medicine must address genetic diversity. Drug Discovery 2019:51.
[253] Conger K. Researchers home in on roots of Caribbean populations using new DNA analysis method; 2013. Available from: https://med.stanford.edu/news/all-news/2013/11/researchers-home-in-on-roots-of-caribbean-populations-using-new-dna-analysis-method.html. [Accessed 08/23 2019].
[254] Peprah E, Xu H, Tekola-Ayele F, Royal CD. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. Public Health Genomics 2015;18(1):40-51.

[255] Elston RC, Satagopan JM, Sun S. Statistical Human Genetics (section: Genetic terminology). New York (NY), USA: Humana Press; 2012.
[256] Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocrine reviews 2019;40(6):1500-20.
[257] Alicia R Martin MKYKYOBMNMJD. Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities. *bioRxiv* 2018:1-26.
[258] Doumatey AP, Ekoru K, Adeyemo A, Rotimi CN. Genetic Basis of Obesity and Type 2 Diabetes in Africans: Impact on Precision Medicine. Current diabetes reports 2019;19(10):105.
[259] Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature genetics 2018;50(11):1505.
[260] Biobank U. Main demographic fields likely to be of interest to researchers; 2019. Available from: http://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=1001. [Accessed 12/06 2019].
[261] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on~ 500,000 UK Biobank participants. BioRxiv 2017:166298.
[262] Wickham H, Chang W, Wickham MH. Package 'ggplot2'. Create Elegant Data Visualisations Using the Grammar of GraphicsVersion 2016;2(1):1-189.
[263] National Cancer Institute USDoHaHS. LDproxy Tool; 2019. Available from: https://ldlink.nci.nih.gov/?tab=ldproxy. [Accessed 12/06 2019].
[264] Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31(21):3555-7.
[265] Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nature genetics 2019;51(4):584-91.
[266] Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. Genetic disease risks can be misestimated across global populations. Genome biology 2018;19(1):1-14.
[267] Smith JA, Ware EB, Middha P, Beacher L, Kardia SLR. Current applications of genetic risk scores to cardiovascular outcomes and subclinical phenotypes. Current epidemiology reports 2015;2(3):180-90.
[268] Chikowore T, van Zyl T, Feskens EJM, Conradie KR. Predictive utility of a genetic risk score of common variants associated with type 2 diabetes in a black South African population. Diabetes research and clinical practice 2016;122:1-8.
[269] Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature communications 2019;10(1):1-10.
[270] Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Human molecular genetics 2019;28(R2):R133-R42.
[271] Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nature communications 2019;10(1):1-9.
[272] Sebastiani P, Solovieff N, Sun J. Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all! Frontiers in genetics 2012;3:26.
[273] So H-C, Sham PC. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Scientific reports 2017;7:41262.
[274] Perlis RH, Ellen B. Dennehy, David J. Miklowitz, Melissa P. DelBello, Michael Ostacher, Joseph R. Calabrese, Rebecca M. Retrospective age at onset of bipolar disorder and

outcome during two-year follow-up: results from the STEP-BD study. Bipolar disorders 2009;11(4):391-400.
[275] Rhebergen D, van der Steenstraten IM, van Balkom AJLM, van Oppen P, Stek ML, Comijs HC, et al. Admixture analysis of age of onset in generalized anxiety disorder. 50. Elsevier Ltd; 2017:47-51.
[276] Tutuncu M, Tang J, Zeid NA, Kale N, Crusan DJ, Atkinson EJ, et al. Onset of progressive phase is an age-dependent clinical milestone in multiple sclerosis. Multiple Sclerosis Journal 2013;19(2):188-98.
[277] Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Current clinical use of polygenic scores will risk exacerbating health disparities. BioRxiv 2019:441261.
[278] Organization WH. Classification of diabetes mellitus. Geneva: World Health Organization; 2019.
[279] Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. Current opinion in genetics & development 2015;33:10-6.
[280] Golan D, Rosset S. Effective genetic-risk prediction using mixed models. The American Journal of Human Genetics 2014;95(4):383-93.

Appendices

Appendix A: Supporting information relating to simulation studies based on admixed populations

List of supporting tables

List of supporting figures

Table of Contents

Supporting information relating to simulation studies based on admixed populations

A.1: Supporting tables relating to simulation process

Table A.1. 1 - Description of causal SNP and tag SNP

Allele		Allele frequency	
Name	**Description**	**Name**	**Description**
A1	Causal SNP alternative allele	**p1**	Frequency of causal SNP alternative allele
a2	Causal SNP risk allele	**p2**	Frequency of causal SNP risk allele
B1	Tag SNP alternative allele	**q1**	Frequency of tag SNP alternative allele
b2	Tag SNP risk allele	**q2**	Frequency of tag SNP risk allele

Table A.1. 2 - Description of allele frequencies in relation to haplotype frequencies and LD

Allele	Allele Frequency assuming LD	Allele frequency based on haplotype frequency
A1	p1= (p1q1 + D) + (p1q2 - D)	p1 = p11 + p12
a2	p2 = (p2q1 - D) + (p2q2 + D)	p2 = p21 + p22
B1	q1 = (p1q1 + D) + (p2q1 - D)	q1 = p11 + p21
b2	q2 = (p1q2 - D) +(p2q2 + D)	q2 = p12 + p22

Table A.1. 3 - Description of haplotype

Haplotype		Haplotype frequency	
Name	**Description**	**Name**	**Description**
A1B1	Haplotype derived from causal and tag SNP alternative allele	**p11**	Frequency of causal and tag SNP alternative allele haplotype
A1b2	Haplotype derived from causal SNP alternative and tag SNP risk allele	**p12**	Frequency of causal SNP alternative and tag SNP risk allele haplotype
a2B1	Haplotype derived from causal SNP risk and tag SNP alternative allele	**p21**	Frequency of causal SNP risk and tag SNP alternative allele haplotype
a2b2	Haplotype derived from causal and tag SNP risk allele	**p22**	Frequency of causal and tag SNP risk allele haplotype

Table A.1. 4 - Description of haplotype frequency based on allele frequency assuming LD

Haplotype	Measured with deviation (D)	Measured with squared correlation coefficient(r^2)	Frequency
A1B1	p11 = (p1q1 + D)	p(BA)= p(B)*p(A) + (r * SQRT(p(B)*p(b)+p(A)*p(a)))	p11 = p(BA)
A1b2	p12 = (p1q2 - D)	p(bA)= p(A) - p(BA)	p12 = p(bA)
a2B1	p21 = (p2q1 - D)	p(Ba)= p(B) - p(BA)	p21 = p(Ba)
a2b2	p22 = (p2q2 + D)	p(ba)= 1 - (p(BA) +P(bA) + p(Ba))	p22 = p(ba)
	D = (p11*p22) - (p12*p21)	**r^2 = D/(pA*pa*pB*pb)**	

A.2: Supporting figures relating to simulation process

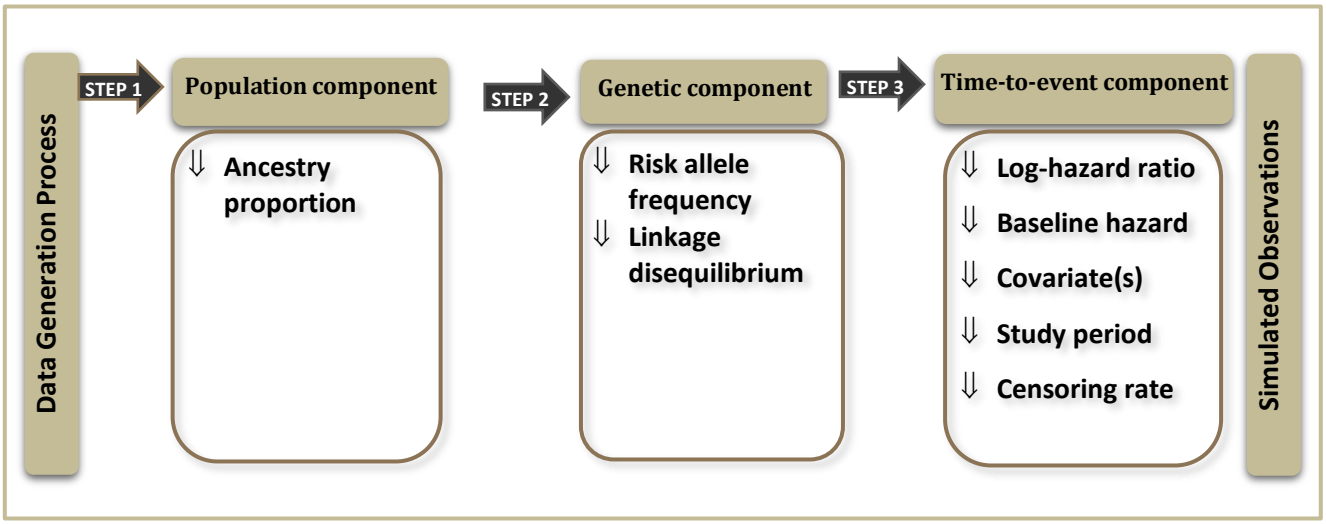


Figure A.2. 1 - General structure of simulation model

A.3: Supporting figures with further results relating to the admixed population simulations.

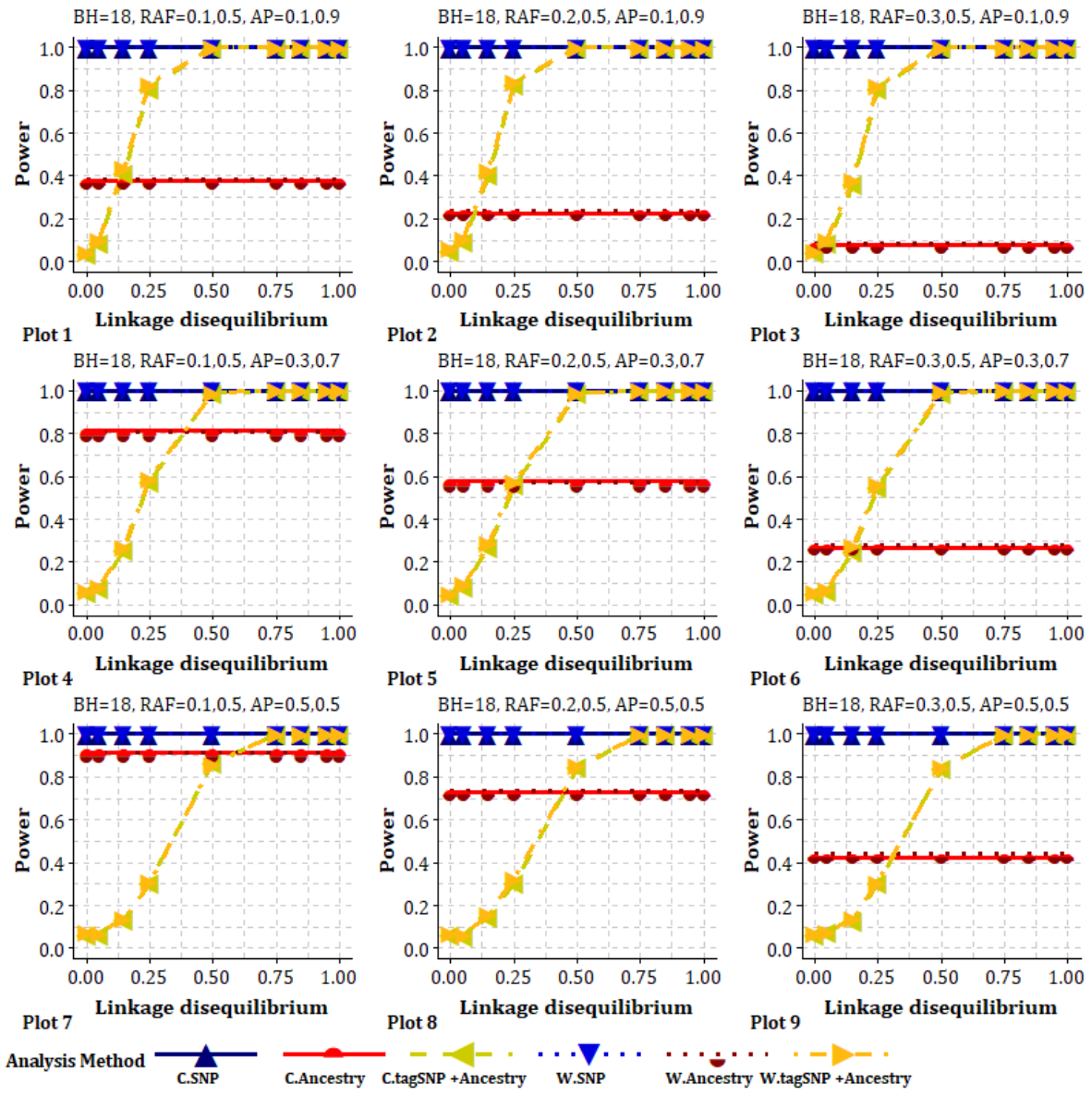


Figure A.3. 1 - Effect of LD on power to detect an association with AOO of disease assuming levels of LD between tag SNP and causal SNP are the same in the ancestral populations

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and linkage disequilibrium on the x axis, for each TTE model analysed. Cox PH model with causal SNP as the single explanatory variable (navy blue); Weibull model with causal SNP as the single explanatory variable (blue); Cox PH model with ancestry as the single explanatory variable (light red); Weibull model with ancestry as the single explanatory variable (dark red); Cox PH model with tag SNP as explanatory variable and ancestry as covariate (yellow green); Weibull model with tag SNP as explanatory variable and ancestry as covariate (gold).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: Ancestry proportion; C.SNP: Cox PH model with SNP variable; C.Ancestry: Cox PH model with ancestry variable; C.tagSNP +Ancestry: Cox PH model with tag SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.Ancestry: Weibull model with ancestry variable; W.tagSNP +Ancestry: Weibull model with tag SNP variable and ancestry covariate.

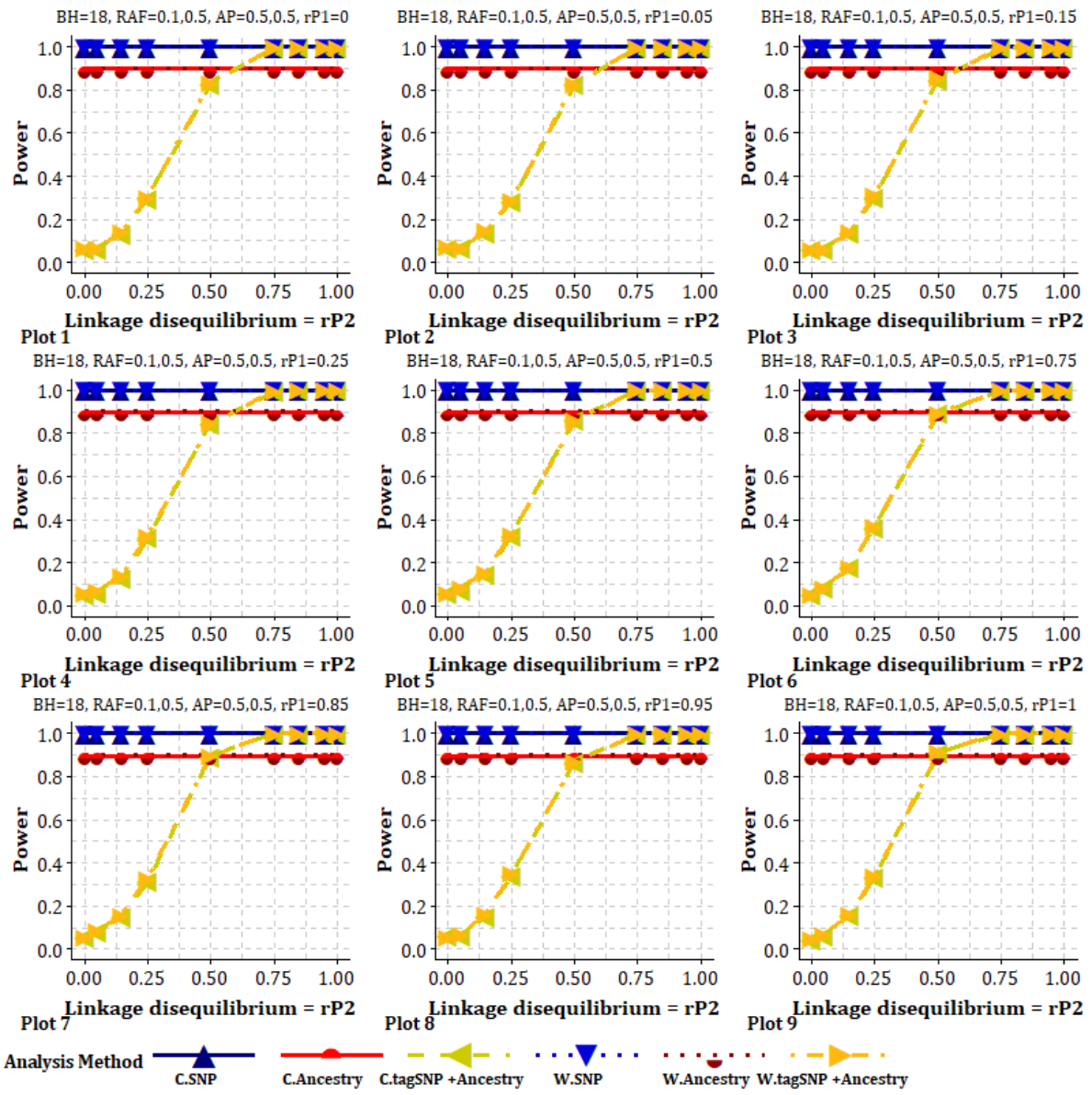


Figure A.3. 2 - Effect of LD on power to detect an association with AOO of disease assuming levels of LD between tag SNP and causal SNP are different among ancestral populations

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and linkage disequilibrium on the x axis, for each TTE model analysed. Cox PH model with causal SNP as the single explanatory variable (navy blue); Weibull model with causal SNP as the single explanatory variable (blue); Cox PH model with ancestry as the single explanatory variable (light red); Weibull model with ancestry as the single explanatory variable (dark red); Cox PH model with tag SNP as explanatory variable and ancestry as covariate (yellow green); Weibull model with tag SNP as explanatory variable and ancestry as covariate (gold).

Abbreviations: BH: baseline hazard; RAF: risk allele frequency; AP: Ancestry proportion; C.SNP: Cox PH model with SNP variable; C.Ancestry: Cox PH model with ancestry variable; C.tagSNP +Ancestry: Cox PH model with tag SNP variable and ancestry covariate; W.SNP: Weibull model with SNP variable; W.Ancestry: Weibull model with ancestry variable; W.tagSNP +Ancestry: Weibull model with tag SNP variable and ancestry covariate.

Appendix B: Supporting information relating to application of T2D GRS in ancestrally homogenous populations

List of supporting tables

List of supporting figures

Table of Contents

Supporting information relating to application of T2D GRS in ancestrally homogenous populations

B.1: Supporting tables relating to construction of T2D GRS

Table B.1. 1 – List of associated T2D SNPs selected from base GWAS

#	Nearest gene	Chrom-osome	Base Pair Position	rsid	EA	NEA	EAF	OR	P_value
1	MACF1	1	40035928	rs3768321	T	G	0.19	1.08	8.1×10^{-07}
2	FAF1	1	51109269	rs12031920	T	A	0.56	1.05	3.8×10^{-05}
3	NOTCH2	1	120554048	rs406767	C	T	0.09	1.14	7.5×10^{-07}
4	ATP8B2	1	154336716	rs67156297	A	G	0.25	1.03	2.7×10^{-02}
5	PROX1	1	214159256	rs340874	C	T	0.55	1.07	3.4×10^{-08}
6	GCKR	2	27748539	rs145819220	G	C	0.01	1.26	3.0×10^{-03}
7	THADA	2	43734847	rs6757251	C	T	0.9	1.14	1.9×10^{-10}
8	BCL11A	2	60552476	rs10193447	T	C	0.6	1.07	1.3×10^{-08}
9	RBMS1	2	161131694	rs1563575	A	G	0.74	1.07	6.7×10^{-07}
10	GRB14	2	165689720	rs28584669	T	C	0.83	1.05	2.0×10^{-03}
11	IRS1	2	227117778	rs2972156	G	C	0.61	1.08	1.2×10^{-09}
12	PPARG	3	12344730	rs11712037	C	G	0.87	1.14	8.6×10^{-13}
13	UBE2E2	3	23455582	rs35352848	T	C	0.78	1.09	1.5×10^{-08}
14	ADAMTS9	3	64710850	rs7428936	T	C	0.59	1.07	1.0×10^{-08}
15	ADCY5	3	123065778	rs11708067	A	G	0.79	1.12	8.8×10^{-13}
16	IGF2BP2	3	185511687	rs4402960	T	G	0.31	1.15	2.7×10^{-25}
17	ST6GAL1	3	186663868	rs9820223	C	T	0.38	1.06	1.5×10^{-05}
18	LPP	3	187741842	rs6777684	G	A	0.61	1.05	5.9×10^{-05}
19	MAEA	4	744972	rs1531583	T	G	0.05	1.15	3.9×10^{-06}
20	WFS1	4	6299940	rs3821943	T	C	0.54	1.1	4.2×10^{-16}
21	TMEM154	4	153397823	rs7660590	C	T	0.72	1.06	6.8×10^{-05}
22	ACSL1	4	185708807	rs60780116	T	C	0.84	1.09	7.4×10^{-08}
23	ARL15	5	53301561	rs11747901	G	C	0.19	1.07	1.2×10^{-05}
24	ANKRD55	5	55861601	rs9687833	A	G	0.19	1.1	1.6×10^{-09}
25	ZBED3	5	76453765	rs6453287	C	A	0.3	1.07	4.5×10^{-06}
26	PAM	5	102726073	rs74944275	T	C	0.04	1.16	4.4×10^{-06}
27	SSR1/RREB1	6	7258847	rs6923241	C	T	0.71	1.07	1.6×10^{-06}
28	CDKAL1	6	20673880	rs7451008	C	T	0.26	1.19	3.8×10^{-37}
29	ZFAND3	6	38228979	rs143308245	T	A	0	2.02	3.0×10^{-03}
30	KCNK16	6	39331930	rs139514607	T	C	0	1.48	8.0×10^{-03}

#	Nearest gene	Chrom-osome	Base Pair Position	rsid	EA	NEA	EAF	OR	P_value
31	CENPW	6	126792095	rs11759026	G	A	0.24	1.1	5.8×10^{-10}
32	SLC35D3	6	137287702	rs6918311	A	G	0.53	1.07	6.7×10^{-07}
33	DGKB	7	15054232	rs10238625	A	G	0.54	1.07	3.2×10^{-08}
34	JAZF1	7	28189411	rs1635852	T	C	0.5	1.1	3.0×10^{-14}
35	GCK	7	44255643	rs878521	A	G	0.24	1.05	6.1×10^{-04}
36	GCC1	7	127631181	rs73455744	A	G	1	1.93	3.0×10^{-03}
37	KLF14	7	130463758	rs10954284	T	A	0.5	1.06	1.8×10^{-05}
38	MNX1	7	157027753	rs1182436	C	T	0.8	1.08	8.3×10^{-07}
39	ANK1	8	41519248	rs516946	C	T	0.78	1.08	8.6×10^{-07}
40	TP53INP1	8	95957984	rs11786613	C	A	0.03	1.21	1.6×10^{-06}
41	SLC30A8	8	118185025	rs3802177	G	A	0.68	1.12	1.7×10^{-17}
42	GLIS3	9	4292083	rs10758593	A	G	0.41	1.05	2.9×10^{-04}
43	PTPRD	9	8288059	rs186838848	T	C	0.01	1.46	2.4×10^{-04}
44	CDKN2A/B	9	22132878	rs10965248	T	C	0.82	1.15	6.5×10^{-17}
45	TLE4	9	81900744	rs13301067	G	A	0.92	1.11	1.5×10^{-05}
46	TLE1	9	84311800	rs9410573	T	C	0.6	1.08	2.0×10^{-08}
47	ABO	9	136155000	rs635634	T	C	0.18	1.08	3.6×10^{-07}
48	GPSM1	9	139252148	rs11787792	A	G	0.67	1.04	1.3×10^{-02}
49	CDC123/CAMK1D	10	12309269	rs11257659	T	C	0.23	1.08	2.7×10^{-08}
50	VPS26A	10	70859204	rs10998572	C	A	0.93	1.09	2.8×10^{-04}
51	ZMIZ1	10	80942620	rs810517	C	T	0.51	1.09	1.3×10^{-12}
52	HHEX/IDE	10	94466910	rs11187140	G	A	0.62	1.14	4.2×10^{-26}
53	TCF7L2	10	114758349	rs7903146	T	C	0.29	1.34	9.2×10^{-108}
54	PLEKHA1	10	124186714	rs2292626	C	T	0.5	1.09	1.8×10^{-12}
55	KCNQ1	11	2858546	rs2237897	C	T	0.95	1.25	4.9×10^{-13}
56	KCNJ11	11	17409572	rs5219	T	C	0.38	1.07	4.3×10^{-08}
57	HSD17B12	11	43877934	rs1061810	A	C	0.28	1.08	5.3×10^{-09}
58	MAP3K11	11	65364385	rs111669836	A	T	0.25	1.07	7.4×10^{-07}
59	ARAP1 (CENTD2)	11	72428172	rs76550717	A	G	0.83	1.1	3.8×10^{-09}
60	MTNR1B	11	92708710	rs10830963	G	C	0.27	1.08	1.7×10^{-07}
61	CCND2	12	4376089	rs4238013	C	T	0.2	1.1	3.6×10^{-09}
62	KLHDC5	12	27962719	rs7953190	T	C	0.8	1.08	4.2×10^{-07}
63	HMGA2	12	66221060	rs2258238	T	A	0.1	1.11	1.6×10^{-07}
64	TSPAN8/LGR5	12	71656723	rs6581998	C	T	0.27	1.06	1.2×10^{-05}
65	HNF1A (TCF1)	12	121432117	rs56348580	G	C	0.68	1.08	2.5×10^{-08}
66	MPHOSPH9	12	123653592	rs2851437	A	C	0.72	1.07	2.6×10^{-06}
67	SPRY2	13	80705315	rs11616380	G	T	0.71	1.09	3.9×10^{-11}
68	NRXN3	14	79945162	rs10146997	G	A	0.21	1.07	4.6×10^{-06}
69	RASGRP1	15	38822905	rs7403531	T	C	0.21	1.04	1.4×10^{-02}

#	Nearest gene	Chrom-osome	Base Pair Position	rsid	EA	NEA	EAF	OR	P_value
70	C2CD4A	15	62117975	rs4774420	C	T	0.7	1.08	2.7×10^{-08}
71	HMG20A	15	77776498	rs952471	G	C	0.69	1.08	4.0×10^{-10}
72	ZFAND6	15	80411245	rs62006309	A	G	0.52	1.05	4.6×10^{-05}
73	AP3S2	15	90289162	rs62023387	C	A	0.18	1.07	5.7×10^{-04}
74	PRC1	15	91563513	rs12595616	C	T	0.37	1.07	5.6×10^{-07}
75	FTO	16	53803574	rs1558902	A	T	0.42	1.13	4.7×10^{-25}
76	BCAR1	16	75252327	rs8056814	G	A	0.92	1.16	3.7×10^{-11}
77	CMIP	16	81534790	rs2925979	T	C	0.3	1.08	2.7×10^{-08}
78	SRR	17	2309188	rs9911305	A	G	0.7	1.05	1.0×10^{-03}
79	ZZEF1	17	4014384	rs7224685	T	G	0.3	1.07	2.0×10^{-07}
80	GLP2R	17	9780387	rs78761021	G	A	0.34	1.07	5.5×10^{-08}
81	HNF1B (TCF2)	17	36102833	rs757209	G	A	0.58	1.09	1.1×10^{-09}
82	GIP	17	46967038	rs79349575	A	T	0.51	1.07	2.6×10^{-07}
83	LAMA1	18	7067652	rs7234111	C	T	0.36	1.06	7.7×10^{-07}
84	MC4R	18	57793209	rs1942880	T	C	0.33	1.07	2.8×10^{-07}
85	BCL2A	18	60845884	rs12454712	T	C	0.62	1.05	2.0×10^{-03}
86	CILP2	19	19456917	rs58489806	T	C	0.09	1.09	1.0×10^{-04}
87	PEPD	19	33943994	rs139990642	A	G	0.01	1.25	2.0×10^{-03}
88	APOE	19	45411941	rs429358	T	C	0.85	1.13	1.4×10^{-10}
89	HNF4A	20	43042364	rs1800961	T	C	0.04	1.17	4.4×10^{-06}
90	MTMR3/HORMAD2	22	30599562	rs2023681	G	A	0.89	1.13	3.9×10^{-09}

Descriptions: N***earest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID;* ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37;* ***rsid:*** *Cluster ID;* ***EA:*** *Discovery SNP effect allele;* ***NEA:*** *Discovery SNP alternative allele;* ***EAF:*** *Discovery SNP effect allele frequency;* ***OR:*** *Odds ratio associated with SNP effect allele;* ***P_value:*** *P_value associated with SNP effect allele.*

Table B.1. 2 - List of associated T2D SNPs not available in target GWAS

#	Nearest gene	Chrom-osome	Base Pair Position	rsid	EA	NEA	EAF	OR	P_value
1	IRS1	2	227117778	rs2972156	G	C	0.61	1.08	1.2×10^{-09}
2	ARL15	5	53301561	rs11747901	G	C	0.19	1.07	1.2×10^{-05}
3	KCNK16	6	39331930	rs139514607	T	C	0	1.48	8.0×10^{-03}
4	PTPRD	9	8288059	rs186838848	T	C	0.01	1.46	2.4×10^{-04}
5	CDC123/CAMK1D	10	12309269	rs11257659	T	C	0.23	1.08	2.7×10^{-08}
6	MAP3K11	11	65364385	rs111669836	A	T	0.25	1.07	7.4×10^{-07}
7	AP3S2	15	90289162	rs62023387	C	A	0.18	1.07	5.7×10^{-04}
8	GIP	17	46967038	rs79349575	A	T	0.51	1.07	2.6×10^{-07}

Descriptions: N***earest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID;* ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37;* ***rsid:*** *Cluster ID;* ***EA:*** *Discovery SNP effect allele;* ***NEA:*** *Discovery SNP alternative allele;* ***EAF:*** *Discovery SNP effect allele frequency;* ***OR:*** *Odds ratio associated with SNP effect allele;* ***P_value:*** *P_value associated with SNP effect allele.*

Table B.1. 3 - List of associated T2D SNPs excluded due poor imputation score

#	Nearest gene	Chrom-osome	Base Pair Position	rsid	EA	NEA	EAF	OR	P_value
1	GCC1	7	127631181	rs73455744	A	G	1	1.93	3.0×10^{-03}

Descriptions: N***earest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID;* ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37;* ***rsid:*** *Cluster ID;* ***EA:*** *Discovery SNP effect allele;* ***NEA:*** *Discovery SNP alternative allele;* ***EAF:*** *Discovery SNP effect allele frequency;* ***OR:*** *Odds ratio associated with SNP effect allele;* ***P_value:*** *P_value associated with SNP effect allele.*

B.2: Supporting tables with further results relating to single SNP association with T2D status

Table B.2. 1 - Single SNP association with T2D status in NUgene dataset

#	Nearest gene	Chrom-osome	Base Pair Position	rsid	P-value: Adjusting for age, sex and ancestry	P-value: Adjusting for age, sex, ancestry and BMI
1	PROX1	1	214159256	rs340874	3.6×10^{-02}	3.7×10^{-02}
2	MACF1	1	40035928	rs3768321	6.3×10^{-03}	6.7×10^{-03}
3	WFS1	4	6299940	rs3821943	7.2×10^{-04}	7.2×10^{-04}
4	CDKN2A/B	9	22132878	rs10965248	2.1×10^{-02}	2.0×10^{-02}
5	TCF7L2	10	114758349	rs7903146	1.1×10^{-05}	1.2×10^{-05}
6	FTO	16	53803574	rs1558902	1.3×10^{-03}	1.4×10^{-03}
7	HNF1B (TCF2)	17	36102833	rs757209	3.7×10^{-02}	3.4×10^{-02}
8	BCL2A	18	60845884	rs12454712	7.0×10^{-03}	6.3×10^{-03}

Descriptions: *N**earest gene:** refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID**;** **Base pair position:** Base pair position of the SNP on the human genome based on the human reference genome build 37**; rsid:** Cluster ID**;***
P-value: *P-value associated with each SNP adjusted by covariates age, sex, ancestry and BMI.*

Table B.2. 2 - Single SNP association with T2D status in WTCCC dataset

#	Nearest gene	Chrom-osome	Base Pair Position	rsid	P-value: Adjusting for age, sex and ancestry
1	PROX1	1	214159256	rs340874	4.5×10^{-02}
2	BCL11A	2	60552476	rs10193447	2.2×10^{-03}
3	ADCY5	3	123065778	rs11708067	2.2×10^{-02}
4	PPARG	3	12344730	rs11712037	2.8×10^{-03}
5	UBE2E2	3	23455582	rs35352848	3.3×10^{-02}
6	IGF2BP2	3	185511687	rs4402960	6.8×10^{-03}
7	ADAMTS9	3	64710850	rs7428936	2.8×10^{-02}
8	ZBED3	5	76453765	rs6453287	2.7×10^{-03}
9	ANKRD55	5	55861601	rs9687833	4.3×10^{-02}
10	CDKAL1	6	20673880	rs7451008	1.1×10^{-05}
11	DGKB	7	15054232	rs10238625	1.8×10^{-02}
12	JAZF1	7	28189411	rs1635852	4.8×10^{-03}
13	CDKN2A/B	9	22132878	rs10965248	1.3×10^{-03}
14	TLE4	9	81900744	rs13301067	1.6×10^{-02}
15	ABO	9	136155000	rs635634	3.4×10^{-02}
16	TLE1	9	84311800	rs9410573	4.1×10^{-04}
17	HHEX/IDE	10	94466910	rs11187140	7.6×10^{-04}
18	TCF7L2	10	114758349	rs7903146	6.8×10^{-10}
19	HSD17B12	11	43877934	rs1061810	8.6×10^{-03}
20	KCNQ1	11	2858546	rs2237897	3.3×10^{-02}
21	KCNJ11	11	17409572	rs5219	2.0×10^{-02}
22	MPHOSPH9	12	123653592	rs2851437	3.0×10^{-02}
23	HNF1A (TCF1)	12	121432117	rs56348580	1.8×10^{-03}
24	TSPAN8/LGR5	12	71656723	rs6581998	7.9×10^{-05}
25	KLHDC5	12	27962719	rs7953190	2.9×10^{-02}
26	PRC1	15	91563513	rs12595616	3.3×10^{-02}
27	FTO	16	53803574	rs1558902	6.0×10^{-05}
28	ZZEF1	17	4014384	rs7224685	2.7×10^{-02}
29	MC4R	18	57793209	rs1942880	2.5×10^{-02}
30	LAMA1	18	7067652	rs7234111	3.1×10^{-02}
31	CILP2	19	19456917	rs58489806	1.5×10^{-02}
32	HNF4A	20	43042364	rs1800961	2.4×10^{-02}

Descriptions: ***Nearest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID;* ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37;* ***rsid:*** *Cluster ID;*
P-value: *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

B.3: Supporting tables with further results relating to association of GRS with AOO of T2D

Table B.3. 1 - Descriptive characteristics of weighted and unweighted GRS among T2D cases and controls

Characteristics	NUGENE SAMPLE		WTCCC SAMPLE	
	Cases	Controls	Cases	Controls
Genetic Risk Scores (GRS)				
Weighted GRS (P<0.05)				
Mean (SD)	7.36 (0.50)	7.14 (0.48)	7.45 (0.48)	7.17 (0.48)
Median	7.38	7.15	7.46	7.18
Range (Min-Max)	5.72 - 9.08	5.58 - 8.53	5.79 - 8.95	5.57 - 8.78
Unweighted GRS (P<0.05)				
Mean (SD)	81.65 (5.46)	79.76 (5.02)	82.69 (5.2)	79.96 (5.28)
Median	81.90	80.04	82.45	80.06
Range (Min-Max)	66.06 - 100.93	64.19 - 94.93	64.87 - 97.63	62.93 - 97.30
Weighted GRS (P <5*10^{-8})				
Mean (SD)	4.85 (0.44)	4.65 (0.42)	4.88 (0.39)	4.66 (0.41)
Median	4.86	4.66	4.88	4.66
Range (Min-Max)	3.29 - 6.18	3.41 - 5.74	3.73 - 6.14	3.33 - 5.98
Unweighted GRS (P <5*10^{-8})				
Mean (SD)	44.00 (3.96)	42.37 (3.68)	44.16 (3.46)	42.39 (3.68)
Median	44.13	42.67	44.20	42.53
Range (Min-Max)	29.00 - 55.97	29.66 - 52.12	34.26 - 55.07	29.08 - 54.86

Descriptions: *GRS: genetic risk score; SD: standard deviation*

B.4: Supporting tables with further results relating to association of BMI with AOO of T2D

Table B.4. 1 - Estimated effect of association of BMI and AOO of T2D in NUgene samples (weighted GRS)

Analysis Method	BMI with Weighted GRS (P-value threshold P < 0.05)				BMI with Weighted GRS (P-value threshold P < 5 x 10^{-8})			
	ES	Lower 95% CI	Upper 95% CI	P-value	ES	Lower 95% CI	Upper 95% CI	P-value
NUgene								
Cox PH model (cases and controls)								
Adjusted *(BMI+ Covariates)*	1.087	1.076	1.099	2.8×10^{-54}	1.087	1.075	1.098	1.1×10^{-53}
Proportional odds model								
Adjusted *(BMI+ Covariates)*	1.156	1.134	1.179	9.4×10^{-49}	1.155	1.133	1.178	2.4×10^{-48}
Binary logistic regression model								
Adjusted *(BMI+ Covariates)*	1.176	1.148	1.206	9.4×10^{-38}	1.174	1.146	1.205	2.6×10^{-37}

Descriptions: ES: *effect size which refers to the HR for Cox PH model and OR for the logistics and proportional odds models;* ***GRS****: genetic risk score;* ***BMI****: Body Mass Index;* ***CI****: confidence interval;* ***Covariates:*** *include sex and Principal Components PC1-PC2 to account for population structure.*

Table B.4. 2 - Estimated effect of association of BMI and AOO of T2D in NUgene samples (unweighted GRS)

Analysis Method	BMI with Unweighted GRS (P-value threshold P < 0.05)				BMI with Unweighted GRS (P-value threshold P < 5 x 10 -8)			
	ES	Lower 95% CI	Upper 95% CI	P-value	ES	Lower 95% CI	Upper 95% CI	P-value
NUgene								
Cox PH model (cases and controls)								
Adjusted *(BMI+ Covariates)*	1.087	1.076	1.099	1.3×10^{-54}	1.086	1.075	1.098	1.1×10^{-53}
Proportional odds model								
Adjusted *(BMI+ Covariates)*	1.157	1.135	1.180	4.2×10^{-49}	1.155	1.133	1.178	3.7×10^{-48}
Binary logistic regression model								
Adjusted *(BMI+ Covariates)*	1.178	1.150	1.209	1.9×10^{-38}	1.175	1.147	1.205	1.5×10^{-37}

Descriptions: ES: *effect size which refers to the HR for Cox PH model and OR for the logistics and proportional odds models;* ***GRS****: genetic risk score;* ***BMI****: Body Mass Index;* ***CI****: confidence interval;* ***Covariates:*** *include sex and Principal Components PC1-PC2 to account for population structure.*

B.5: Supporting figures with further results relating to GRS simulation of AOO

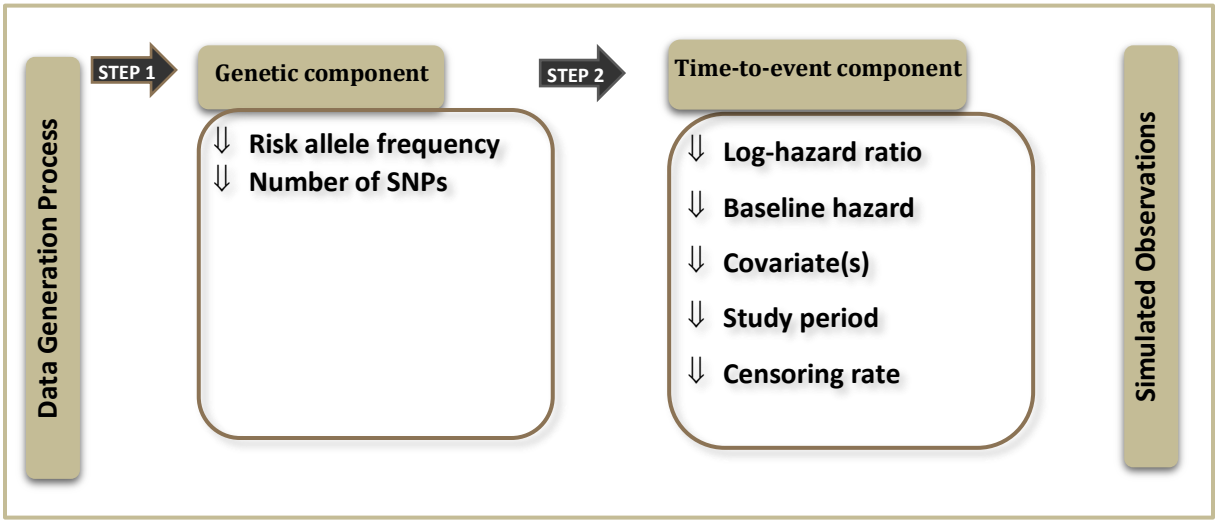


Figure B.5. 1 - General structure of GRS simulation model

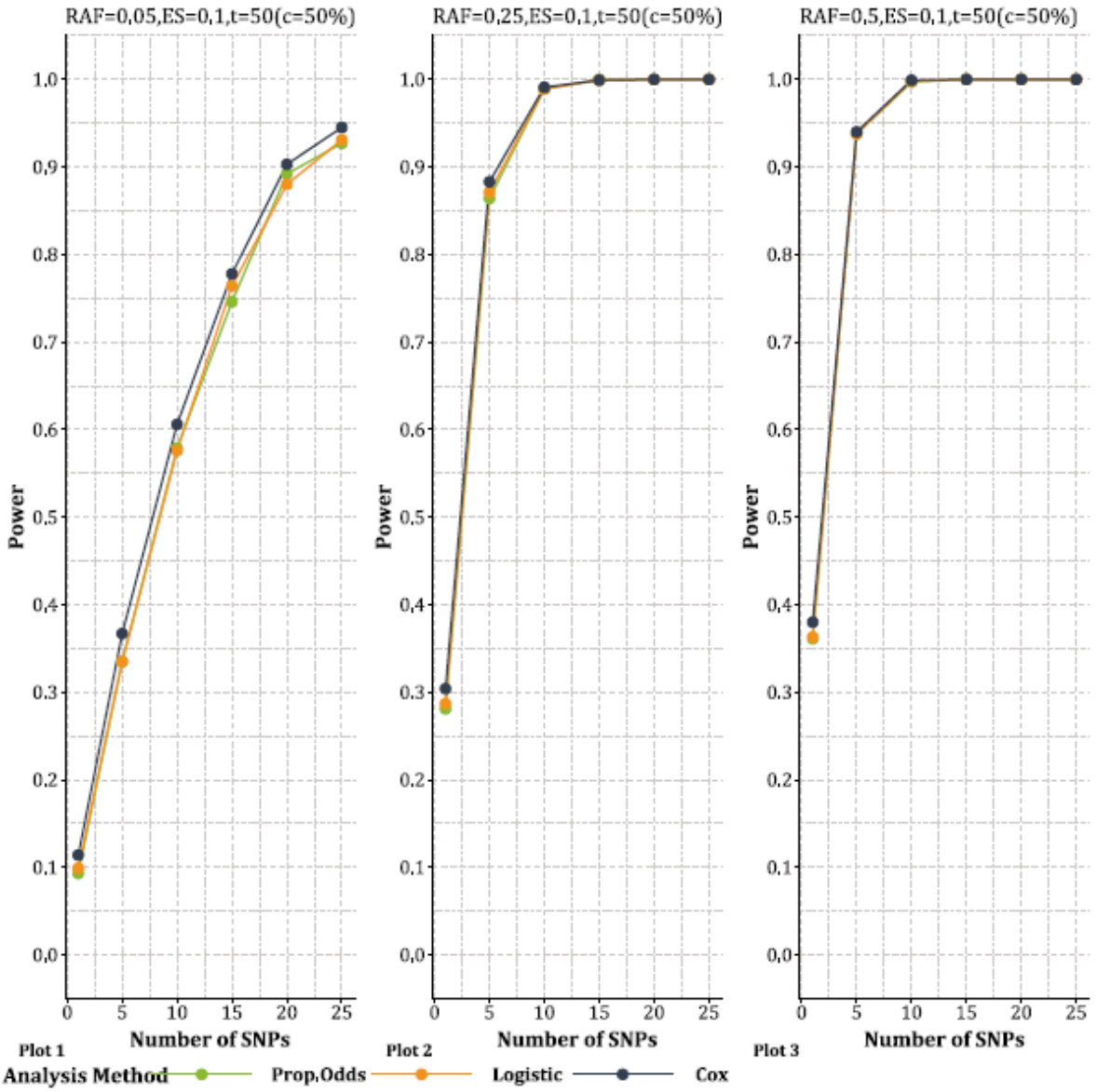


Figure B.5. 2 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming an ES of 0.10

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with PRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; ES: GRS effect size.

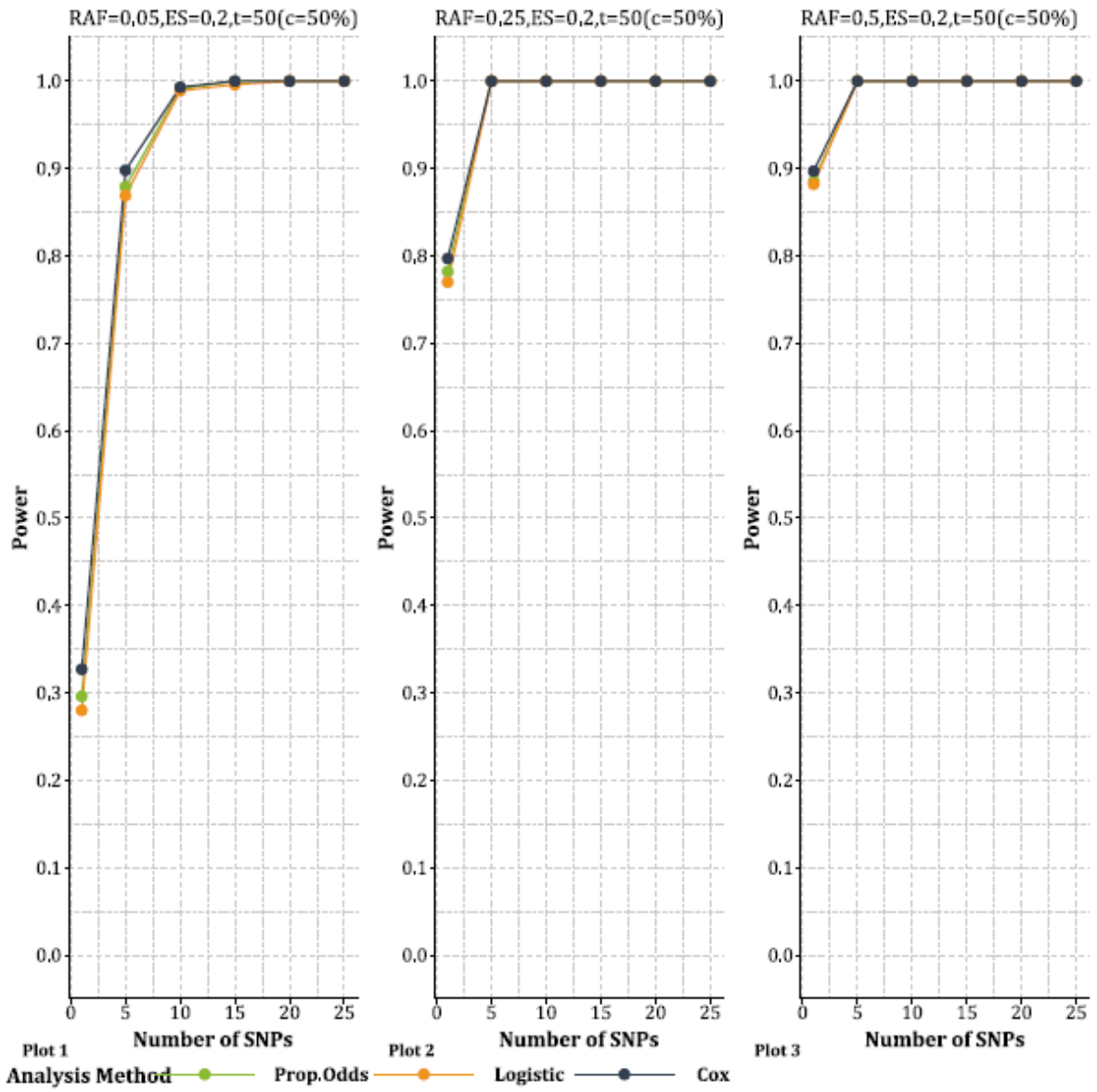


Figure B.5. 3 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming an ES of 0.20

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with PRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; ES: GRS effect size.

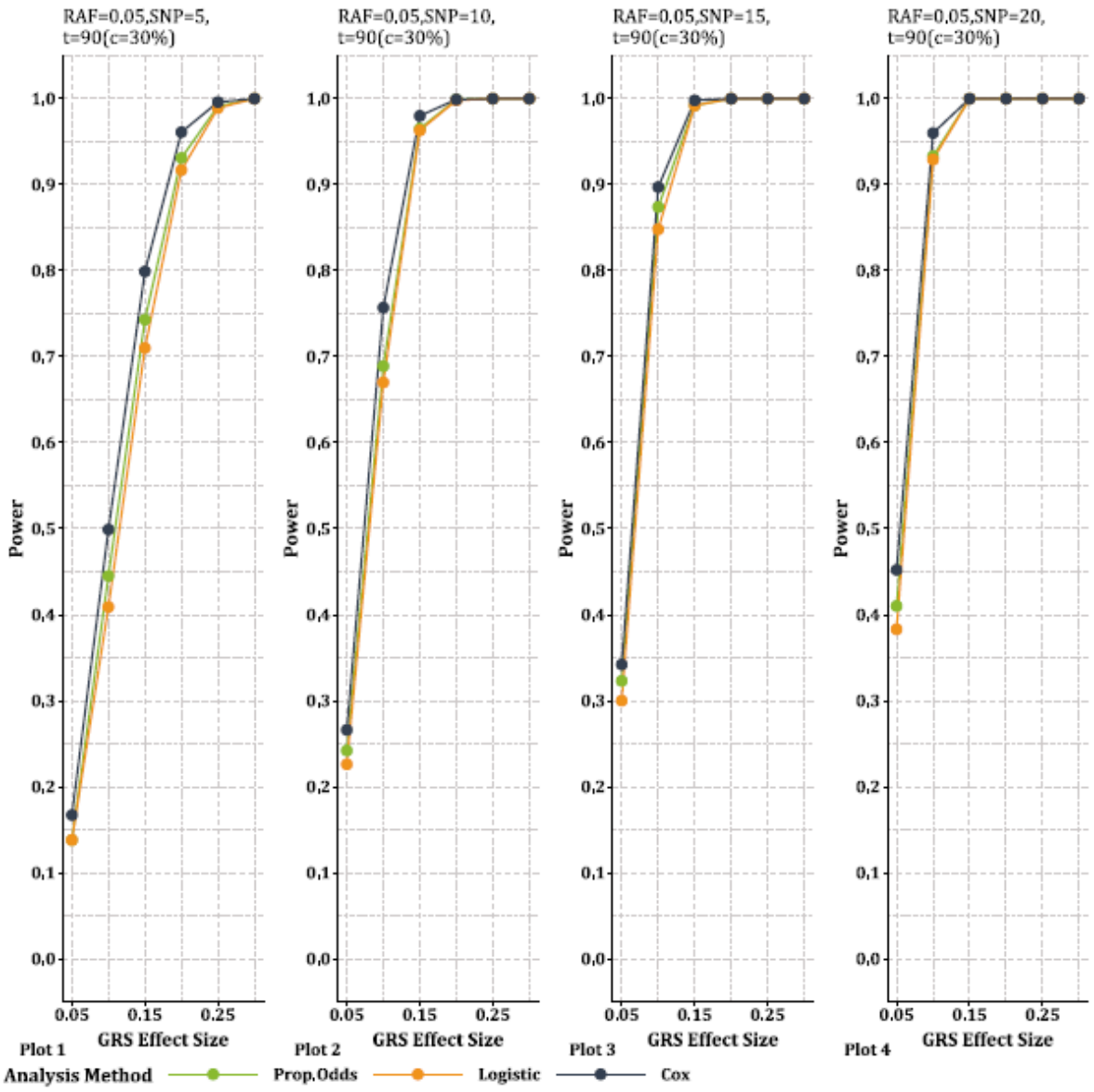


Figure B.5. 4 - Power to detect association of GRS with AOO of disease as a function of the GRS effect size assuming a RAF of 0.05 and C of 30%

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; SNP: number of SNPs included in the GRS calculation; t: study period (follow-up time) c: censoring rate.

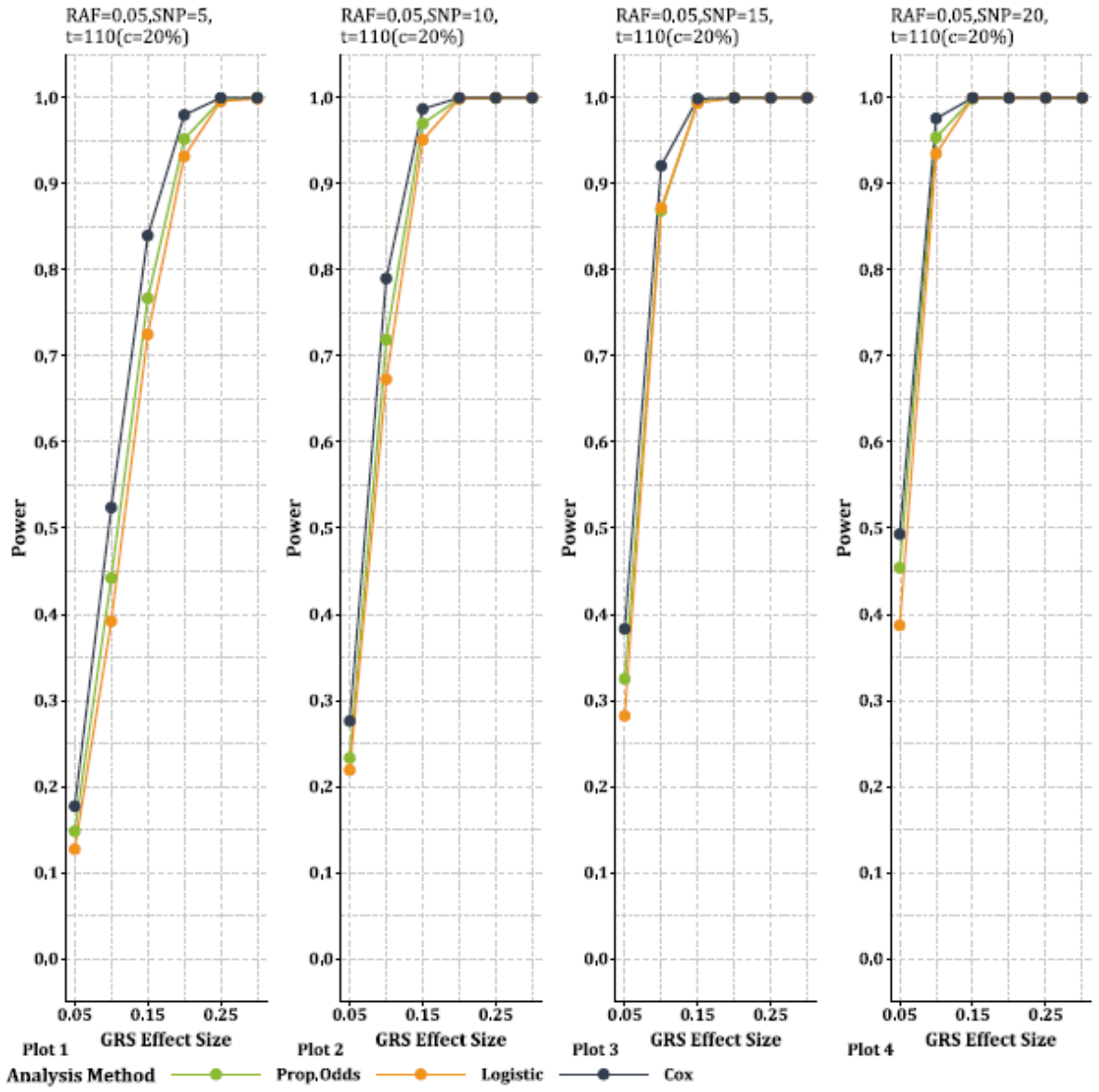


Figure B.5. 5 - Power to detect association of GRS with AOO of disease as a function of the GRS effect size assuming a RAF of 0.05 and C of 20%

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; SNP: number of SNPs included in the GRS calculation; t: study period (follow-up time); c: censoring rate.

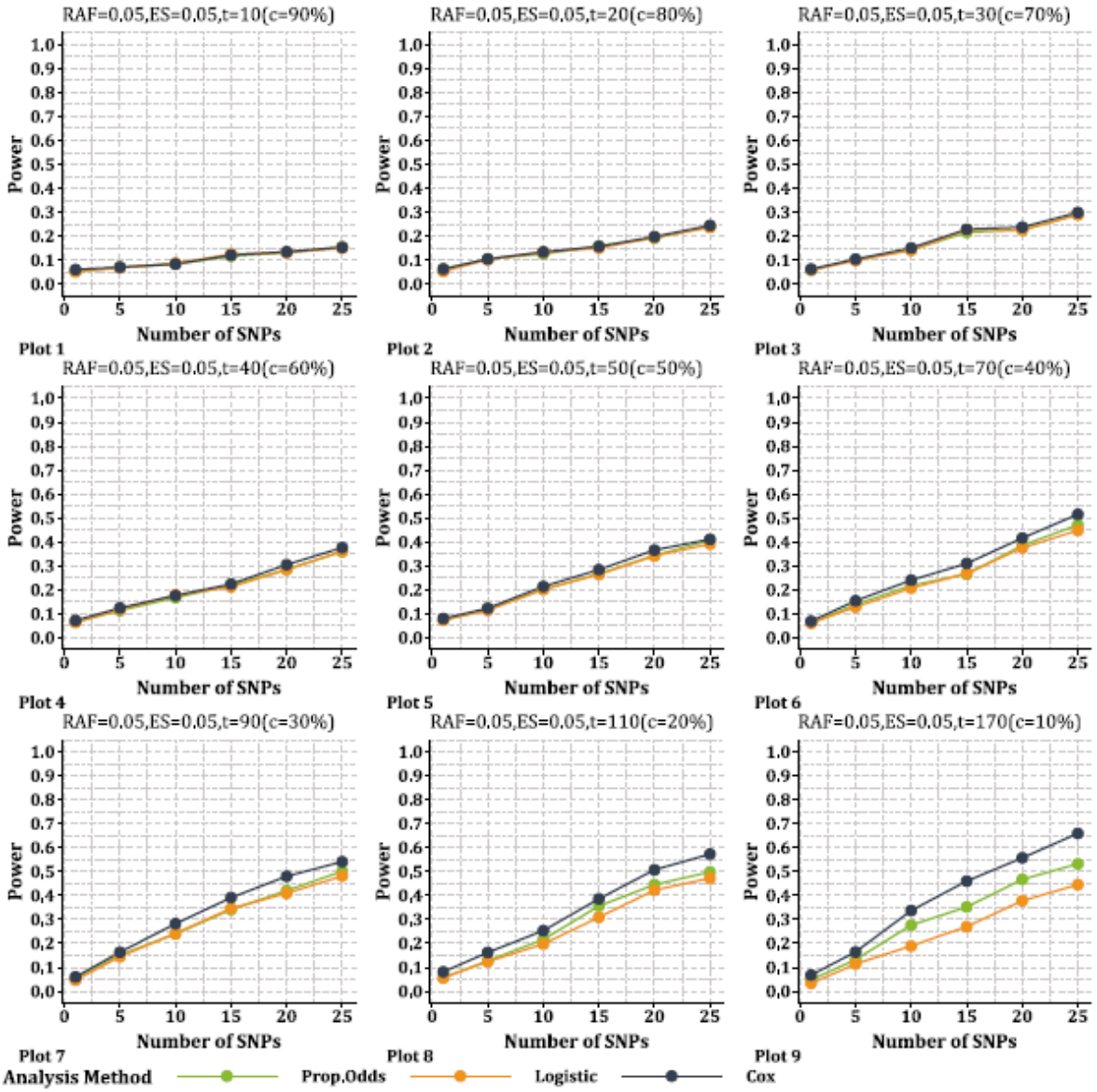


Figure B.5. 6 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming a RAF of 0.05 and ES of 0.05

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; ES: GRS effect size; t: study period (follow-up time) c: censoring rate.

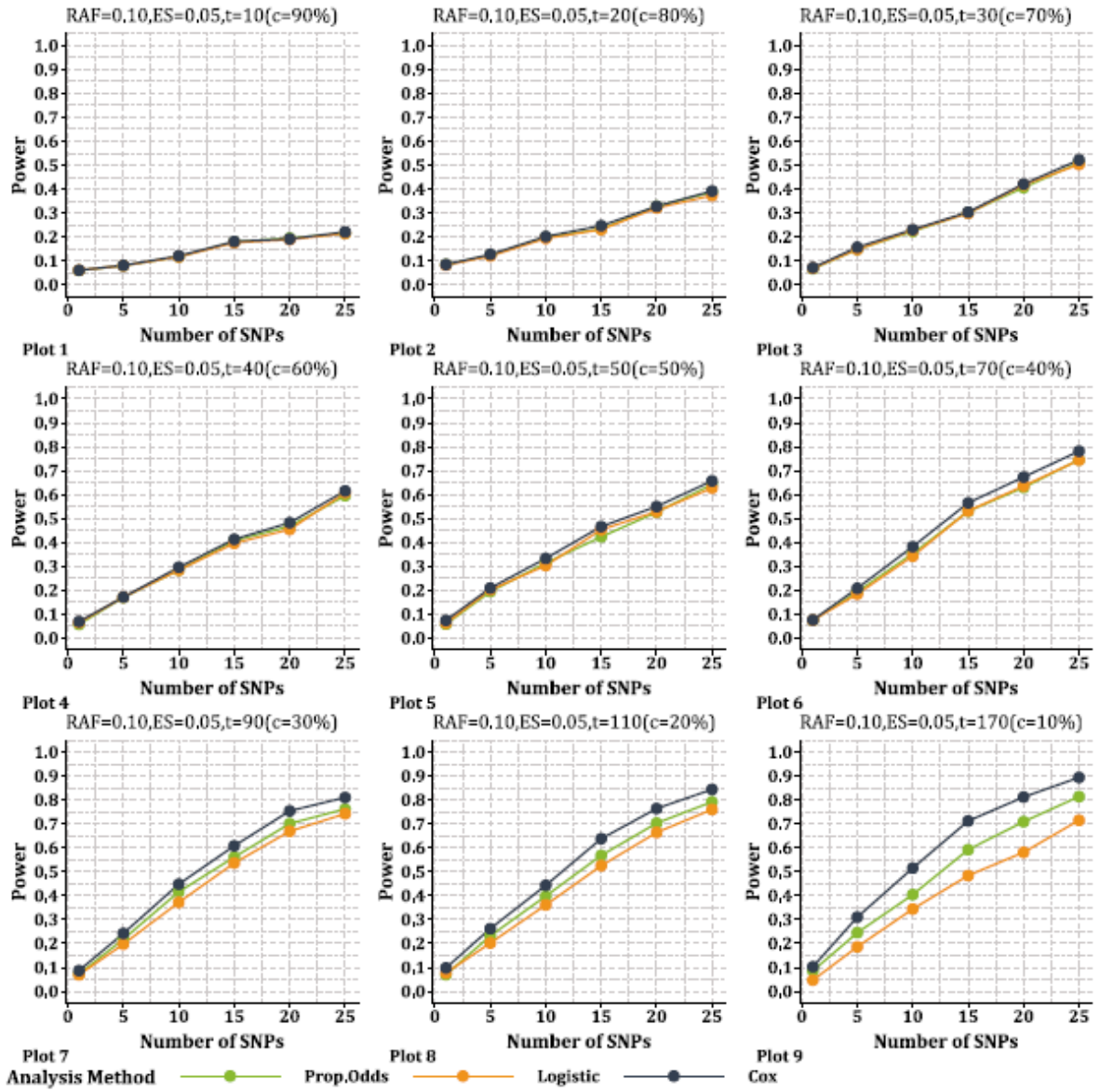


Figure B.5. 7 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming a RAF of 0.10 and ES of 0.05

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; ES: GRS effect size; t: study period (follow-up time) c: censoring rate.

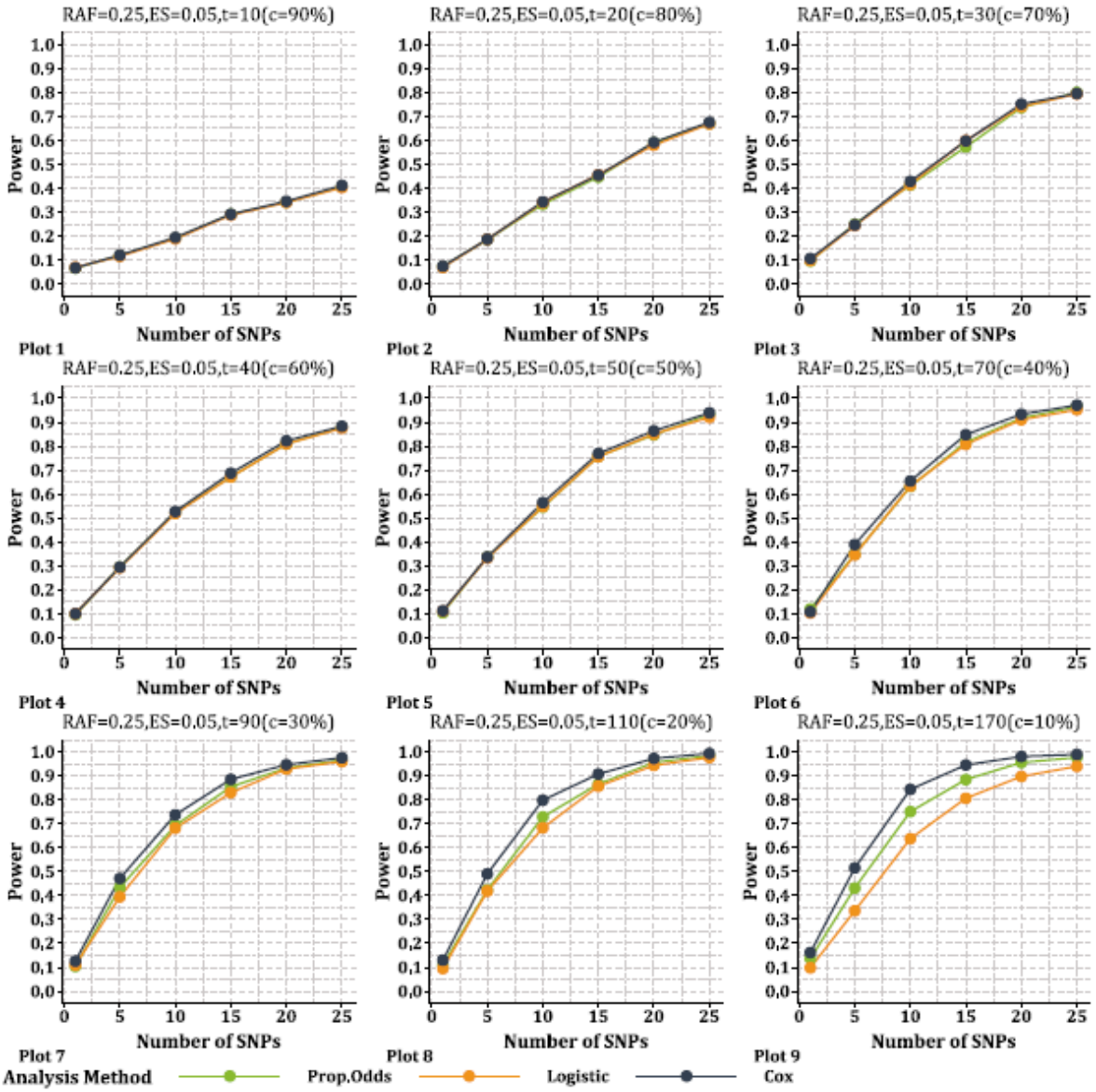


Figure B.5. 8 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming a RAF of 0.25 and ES of 0.05

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; ES: GRS effect size; t: study period (follow-up time) c: censoring rate.

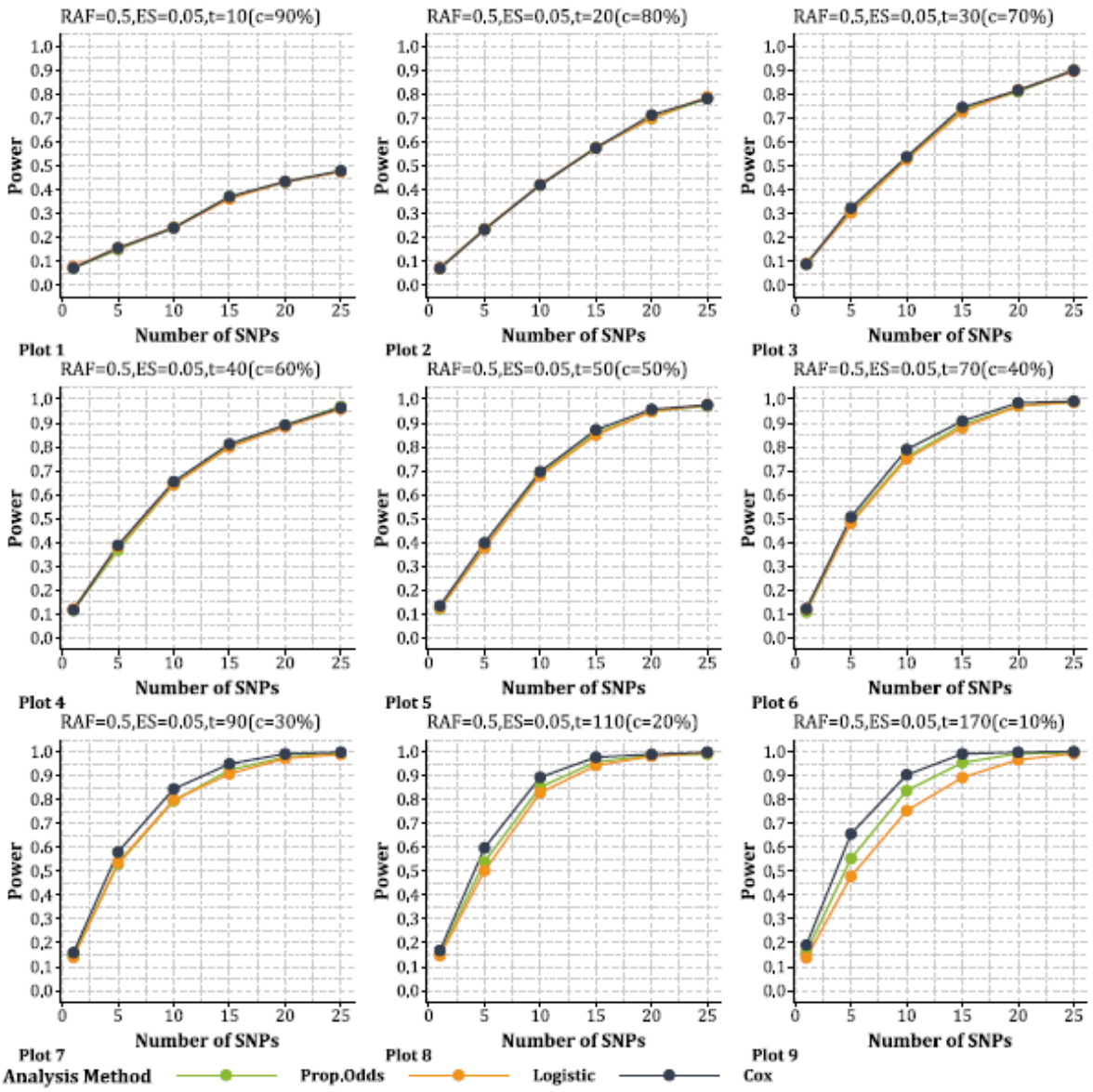


Figure B.5. 9 - Power to detect association of GRS with AOO of disease as a function of the number of SNPs in the GRS assuming a RAF of 0.5 and ES of 0.05

Statistical power which is evaluated at the 5% significance threshold is shown on the y axis and GRS effect on the x axis for each analysis model evaluated, Proportional odds model with GRS as covariate (green); logistic model with GRS as covariate (gold); Cox model with GRS as covariate (navy blue).

Abbreviations: RAF: risk allele frequency; ES: GRS effect size; t: study period (follow-up time) c: censoring rate.

Appendix C: Supporting information relating to application of T2D GRS in ancestrally diverse populations

List of supporting tables

List of supporting figures

Table of Contents

Supporting information relating to application of T2D GRS in ancestrally diverse populations

C.1: Supporting tables relating to global Impact of T2D

Table C.1. 1 - IDF regions ranked by prevalence (%) of diabetes (20-79 years) per region in 2017

Rank	IDF regions	Age-adjusted comparative diabetes prevalence		Raw diabetes prevalence	
		Estimate	CI	Estimate	CI
1	North America and Caribbean	11.0%	9.2 - 12.5%	13.0%	10.8 - 14.5%
2	Middle East and North Africa	10.8%	7.5 - 14.2%	9.6%	6.7 - 12.7%
3	South-East Asia	10.1%	7.9 - 12.8%	8.5%	6.5 - 10.7%
4	Western Pacific	8.6%	7.6 - 11.0%	9.5%	8.4 - 12.0%
5	South and central America	7.6%	6.3 - 9.5%	8.0%	6.7 - 9.8%
6	Europe	6.8%	5.4 - 9.9%	8.8%	7.0% - 12.0%
7	Africa	4.4%	2.9 - 7.8%	3.3%	2.1 - 6.0%

Source: International diabetes federation (IDF) diabetes atlas, eighth edition 2017

Table C.1. 2 - Proportion (%) of people who died from diabetes in 2017 before the age of 60 in IDF regions

IDF regions	Number of deaths due to diabetes before age 60	Proportion of all deaths due to diabetes occurring before age 60
Africa	0.23 million (0.16-0.39)	77.0%
Europe	0.16 million (0.13-0.22)	32.9%
Middle East and North Africa	0.16 million (0.12-0.21)	51.8%
North America and Caribbean	0.13 million (0.11-0.14)	45.0%
South and central America	0.09 million (0.08-0.11)	44.9%
South-East Asia	0.58 million (0.47-0.69)	51.5%
Western Pacific	0.48 million (0.43-0.60)	38.0%

Source: International diabetes federation (IDF) diabetes atlas, eighth edition 2017

C.2: Supporting figures relating to global Impact of T2D

Source: International diabetes federation (IDF) diabetes atlas, eighth edition 2017

Figure C.2. 1 - Estimated total number of adults (20-79 years) living with diabetes, 2017

C.3: Supporting tables relating to T2D base and target GWAS

Table C.3. 1 - List of associated T2D SNPs selected from base GWAS

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	EA	NEA	EAF (%)	OR	*p*-value
1	MACF1	1	40035928	rs3768321	T	G	20.04	1.09	$2.6x10^{-26}$
2	FAF1	1	51256091	rs58432198	C	T	88.11	1.07	$2.1x10^{-10}$
3	PATJ	1	62579891	rs12140153	G	T	90.51	1.07	$1.3x10^{-8}$
4	DENND2C	1	115144899	rs184660829	C	T	0.02	8.05	$2.5x10^{-8}$
5	PTGFRN	1	117532790	rs1127215	C	T	58.38	1.05	$1.6x10^{-13}$
6	NOTCH2	1	120526982	rs1493694	T	C	10.89	1.09	$2.7x10^{-16}$
7	FAM63A	1	151017991	rs145904381	T	C	98.67	1.19	$2.6x10^{-8}$
8	SEC16B	1	177889025	rs539515	C	A	19.81	1.05	$1.6x10^{-10}$
9	DSTYK	1	205114873	rs12048743	G	C	44.17	1.04	$3.5x10^{-9}$
10	SRGAP2	1	206593900	rs9430095	C	G	49.4	1.04	$1.9x10^{-8}$
11	PROX1	1	214159256	rs340874	C	T	55.55	1.07	$1.6x10^{-22}$
12	LYPLAL1	1	219748818	rs2820446	C	G	70.55	1.06	$3.3x10^{-16}$
13	ABCB10	1	229672955	rs348330	G	A	36.05	1.05	$2.7x10^{-14}$
14	GNG4	1	235690800	rs291367	G	A	63.22	1.04	$4.7x10^{-10}$
15	TMEM18	2	422144	rs62107261	T	C	95.36	1.12	$3.8x10^{-12}$
16	FAM49A	2	16574669	rs11680058	A	G	86.3	1.06	$1.4x10^{-8}$
17	DTNB	2	25643221	rs17802463	G	T	73.14	1.04	$2.9x10^{-8}$
18	GCKR	2	27730940	rs1260326	C	T	60.69	1.07	$6.5x10^{-25}$
19	THADA	2	43698028	rs80147536	A	T	90.43	1.13	$2.7x10^{-29}$
20	BNIPL	2	59307725	rs6545714	G	A	39.2	1.04	$8.9x10^{-9}$
21	BCL11A	2	60583665	rs243024	A	G	45.99	1.06	$2.5x10^{-20}$
22	CEP68	2	65287896	rs2249105	A	G	63.43	1.10	$2.2x10^{-14}$
23	TMEM127	2	96913918	rs79046683	T	G	0.48	2.34	$3.0x10^{-8}$
24	DDX18	2	118071061	rs562386202	G	A	0.06	3.20	$4.2x10^{-8}$
25	GLI2	2	121347612	rs11688682	G	C	72.78	1.05	$4.2x10^{-9}$
26	PABPC1P2	2	147861633	rs35999103	T	C	15.47	1.05	$9.7x10^{-9}$
27	CYTIP	2	158339550	rs13426680	A	G	93.73	1.09	$6.7x10^{-10}$
28	RBMS1	2	161135544	rs3772071	T	C	71.35	1.05	$1.2x10^{-11}$
29	GRB14/COBLL1	2	165513091	rs10195252	T	C	58.64	1.07	$6.0x10^{-25}$
30	CRYBA2	2	219859171	rs113414093	A	G	5.14	1.12	$6.6x10^{-9}$
31	IRS1	2	227101411	rs2972144	G	A	63.85	1.10	$2.1x10^{-46}$
32	PPARG	3	12336507	rs11709077	G	A	87.65	1.14	$1.8x10^{-36}$
33	UBE2E2	3	23455582	rs35352848	T	C	78.78	1.07	$1.3x10^{-17}$

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	EA	NEA	EAF (%)	OR	*p*-value
34	KIF9	3	46925539	rs11926707	C	T	62.62	1.27	$2.1x10^{-8}$
35	RBM6	3	49980596	rs4688760	T	C	68.42	1.04	$3.5x10^{-10}$
36	RFT1	3	53127677	rs2581787	T	G	56.34	1.04	$2.4x10^{-8}$
37	CACNA2D3	3	54828827	rs76263492	T	G	4.52	1.09	$6.3x10^{-9}$
38	PSMD6	3	63962339	rs3774723	G	A	84.42	1.07	$1.6x10^{-13}$
39	ADAMTS9	3	64701146	rs9860730	A	G	70.36	1.06	$4.9x10^{-15}$
40	SHQ1	3	72865183	rs13085136	C	T	92.83	1.08	$1.5x10^{-8}$
41	ROBO2	3	77671721	rs2272163	C	A	61.84	1.04	$9.6x10^{-9}$
42	ADCY5	3	123065778	rs11708067	A	G	77.23	1.09	$5.2x10^{-32}$
43	SLC12A8	3	124926637	rs649961	T	C	46.51	1.04	$9.9x10^{-10}$
44	TMCC1	3	129333182	rs9828772	C	G	89.83	1.06	$4.2x10^{-8}$
45	TSC22D2	3	150066540	rs62271373	A	T	5.53	1.09	$1.0x10^{-9}$
46	MBNL1	3	152086533	rs13065698	A	G	60.02	1.05	$8.1x10^{-13}$
47	EGFEM1P	3	168218841	rs7629630	A	T	85.67	1.05	$2.5x10^{-8}$
48	SLC2A2	3	170733076	rs9873618	G	A	71	1.07	$4.8x10^{-21}$
49	ABCC5	3	183738460	rs2872246	A	C	45.38	1.04	$1.5x10^{-8}$
50	IGF2BP2	3	185503456	rs6780171	A	T	31.38	1.14	$9.0x10^{-56}$
51	ST6GAL1	3	186665645	rs3887925	T	C	54.68	1.07	$3.1x10^{-22}$
52	LPP	3	187740899	rs4686471	C	T	61.04	1.06	$1.7x10^{-20}$
53	PCGF3	4	744972	rs1531583	T	G	4.58	1.13	$3.5x10^{-14}$
54	MAEA	4	1784403	rs56337234	C	T	50.26	1.06	$8.6x10^{-18}$
55	HTT	4	3241845	rs362307	T	C	7.68	1.08	$1.1x10^{-9}$
56	WFS1	4	6306763	rs10937721	C	G	58.8	1.06	$1.5x10^{-8}$
57	LCORL	4	17792869	rs12640250	C	A	71.49	1.04	$3.7x10^{-8}$
58	GNPDA2	4	45186139	rs10938398	A	G	42.89	1.05	$3.6x10^{-12}$
59	USP46	4	52818664	rs2102278	G	A	31.86	1.04	$3.7x10^{-8}$
60	SCD5	4	83578271	rs12642790	A	G	33.78	1.04	$4.4x10^{-10}$
61	FAM13A	4	89740894	rs1903002	G	C	50.05	1.04	$2.7x10^{-8}$
62	SMARCAD1	4	95091911	rs6821438	A	G	53.42	1.04	$4.0x10^{-11}$
63	SLC9B1	4	104140848	rs1580278	C	A	47.28	1.04	$2.2x10^{-10}$
64	PABPC4L	4	137083193	rs1296328	A	C	44.57	1.04	$3.5x10^{-8}$
65	TMEM154	4	153513369	rs7669833	T	A	70.45	1.06	$1.2x10^{-14}$
66	PDGFC	4	157652753	rs28819812	C	A	67.67	1.04	$2.2x10^{-8}$
67	ACSL1	4	185717759	rs58730668	T	C	85.8	1.07	$1.3x10^{-13}$
68	ANKH	5	14751305	rs146886108	C	T	99.38	1.41	$7.8x10^{-13}$
69	MRPS30	5	44682589	rs6884702	G	A	39.32	1.04	$1.5x10^{-10}$
70	ITGA1	5	52100489	rs3811978	G	A	16.68	1.06	$7.7x10^{-11}$
71	ARL15	5	53271420	rs702634	A	G	69	1.05	$7.7x10^{-14}$
72	ANKRD55	5	55808475	rs465002	T	C	74.21	1.11	$6.1x10^{-38}$

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	EA	NEA	EAF (%)	OR	*p*-value
73	PIK3R1	5	67714246	rs4976033	G	A	41.05	1.05	$1.0x10^{-9}$
74	POC5	5	75003678	rs2307111	T	C	60.53	1.05	$2.1x10^{-16}$
75	ZBED3	5	76424949	rs4457053	G	A	30.36	1.06	$8.4x10^{-18}$
76	DMGDH	5	78430607	rs1316776	C	A	64.76	1.05	$2.6x10^{-12}$
77	RASA1	5	86577352	rs7719891	G	A	25.85	1.04	$2.4x10^{-8}$
78	SLCO6A1	5	101232944	rs138337556	G	A	0.36	1.56	$4.7x10^{-9}$
79	PAM	5	102422968	rs115505614	T	C	4.99	1.19	$1.3x10^{-30}$
80	PHF15	5	133864599	rs329122	A	G	42.86	1.04	$3.6x10^{-9}$
81	EBF1	5	157928196	rs3934712	C	T	20.57	1.05	$3.2x10^{-8}$
82	RREB1	6	7231843	rs9379084	G	A	88.73	1.11	$3.3x10^{-21}$
83	CDKAL1	6	20679709	rs7756992	G	A	27.35	1.15	$2.4x10^{-88}$
84	MHC	6	32573415	rs601945	G	A	17.75	1.06	$4.7x10^{-8}$
85	HMGA1	6	34247047	rs77136196	T	C	4.2	1.11	$1.6x10^{-8}$
86	LRFN2	6	40409243	rs34298980	T	C	49.67	1.04	$9.3x10^{-10}$
87	VEGFA	6	43814190	rs6458354	C	T	28.9	1.05	$2.1x10^{-12}$
88	TFAP2B	6	50788778	rs3798519	C	A	18.44	1.06	$2.6x10^{-12}$
89	SLC25A51P1	6	67387490	rs555402748	T	C	0.04	3.67	$4.6x10^{-8}$
90	BEND3	6	107431688	rs4946812	G	A	67.43	1.04	$8.2x10^{-9}$
91	CENPW	6	126792095	rs11759026	G	A	23.21	1.07	$2.4x10^{-18}$
92	SOGA3	6	127416930	rs2800733	A	G	71.65	1.05	$6.0x10^{-11}$
93	SLC35D3	6	137300960	rs9494624	A	G	28.99	1.04	$6.1x10^{-9}$
94	MIR3668	6	139835329	rs2982521	A	T	38	1.05	$1.3x10^{-9}$
95	SLC22A3	6	160770312	rs474513	A	G	51.69	1.04	$8.1x10^{-10}$
96	QKI	6	164133001	rs4709746	C	T	86.76	1.06	$5.8x10^{-9}$
97	DGKB	7	15063569	rs10228066	T	C	53.73	1.07	$1.1x10^{-28}$
98	IGF2BP3	7	23512896	rs4279506	G	C	61.02	1.06	$4.8x10^{-8}$
99	JAZF1	7	28198677	rs1708302	C	T	51.24	1.10	$1.1x10^{-48}$
100	CRHR2	7	30728452	rs917195	C	T	77	1.05	$4.2x10^{-11}$
101	GCK	7	44255643	rs878521	A	G	24.51	1.06	$1.9x10^{-13}$
102	FBXL13	7	102486254	rs11496066	T	C	81.81	1.08	$1.1x10^{-8}$
103	RELN	7	103444978	rs39328	T	C	43.34	1.04	$3.7x10^{-8}$
104	CTTNBP2	7	117495667	rs6976111	A	C	31.27	1.04	$1.2x10^{-8}$
105	KLF14	7	130457914	rs1562396	G	A	31.86	1.06	$9.9x10^{-18}$
106	AOC1	7	150537635	rs62492368	A	G	30.81	1.05	$1.1x10^{-10}$
107	MNX1	7	156930550	rs6459733	G	C	67.29	1.06	$2.4x10^{-17}$
108	MSRA	8	9974824	rs17689007	G	A	53.29	1.04	$2.5x10^{-9}$
109	XKR6	8	10808687	rs57327348	A	T	78.2	1.04	$4.5x10^{-8}$
110	LPL	8	19830921	rs10096633	C	T	87.66	1.07	$1.1x10^{-12}$
111	PURG	8	30863938	rs10954772	T	C	31.35	1.04	$1.8x10^{-9}$

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	EA	NEA	EAF (%)	OR	*p*-value
112	ANK1	8	41508577	rs13262861	C	A	82.92	1.07	$4.0x10^{-12}$
113	TP53INP1	8	95961626	rs10097617	T	C	48.47	1.04	$3.3x10^{-11}$
114	CPQ	8	97737741	rs149364428	A	G	1.04	1.27	$1.8x10^{-12}$
115	TRHR	8	110123183	rs12680028	C	G	53.42	1.04	$2.5x10^{-8}$
116	SLC30A8	8	118185025	rs3802177	G	A	68.51	1.11	$1.1x10^{-55}$
117	CASC11	8	128711742	rs17772814	G	A	91.51	1.08	$5.4x10^{-10}$
118	PVT1	8	129568078	rs1561927	C	T	26.86	1.04	$1.5x10^{-9}$
119	BOP1	8	145507304	rs4977213	C	T	37.49	1.05	$9.1x10^{-14}$
120	GLIS3	9	4291928	rs10974438	C	A	35.67	1.05	$1.5x10^{-14}$
121	HAUS6	9	19067833	rs7022807	G	A	40.14	1.04	$2.7x10^{-10}$
122	FOCAD	9	20241069	rs7867635	C	T	41.23	1.04	$4.0x10^{-8}$
123	CDKN2A/B	9	22134068	rs10811660	G	A	82.82	1.27	$1.4x10^{-115}$
124	LINGO2	9	28410683	rs1412234	C	T	32.29	1.04	$1.9x10^{-10}$
125	UBAP2	9	34074476	rs12001437	C	T	37.22	1.04	$2.8x10^{-10}$
126	MTND2P8	9	81359113	rs11137820	C	G	57.51	1.04	$2.9x10^{-8}$
127	TLE4	9	81905590	rs17791513	A	G	93.17	1.10	$3.1x10^{-14}$
128	TLE1	9	84308948	rs2796441	G	A	59.24	1.07	$4.4x10^{-24}$
129	ZNF169	9	97001682	rs55653563	A	C	73.21	1.04	$2.2x10^{-9}$
130	ABO	9	136149229	rs505922	C	T	33.17	1.05	$3.9x10^{-12}$
131	GPSM1	9	139241030	rs28505901	G	A	75.2	1.09	$6.7x10^{-26}$
132	CDC123/CAMK1D	10	12307894	rs11257655	T	C	21.84	1.09	$1.5x10^{-32}$
133	NEUROG3	10	71466578	rs2642588	G	T	70.16	1.05	$2.2x10^{-14}$
134	ZMIZ1	10	80952826	rs703972	G	C	53.3	1.07	$1.7x10^{-29}$
135	PTEN	10	89769340	rs11202627	T	C	15.18	1.06	$4.7x10^{-8}$
136	HHEX/IDE	10	94462427	rs10882101	T	C	58.72	1.06	$1.4x10^{-8}$
137	TCF7L2	10	114758349	rs7903146	T	C	29.5	1.37	$5.8x10^{-447}$
138	WDR11	10	122915345	rs72631105	A	G	18.99	1.06	$3.7x10^{-9}$
139	PLEKHA1	10	124193181	rs2280141	T	G	51.61	1.05	$1.4x10^{-13}$
140	INS/IGF2	11	2197286	rs4929965	A	G	38.29	1.07	$4.0x10^{-26}$
141	KCNQ1	11	2857194	rs2237895	C	A	42.6	1.12	$6.0x10^{-52}$
142	PDE3B	11	14763828	rs141521721	A	C	2.36	1.13	$2.7x10^{-8}$
143	KCNJ11	11	17408404	rs5213	C	T	36.24	1.07	$3.5x10^{-27}$
144	METTL15	11	28534898	rs4923543	A	G	33.2	1.04	$4.5x10^{-8}$
145	QSER1	11	32927778	rs145678014	G	T	95.67	1.11	$2.0x10^{-10}$
146	PDHX	11	34982148	rs2767036	C	A	29.08	1.04	$3.3x10^{-8}$
147	HSD17B12	11	43877934	rs1061810	A	C	28.8	1.05	$6.0x10^{-13}$
148	CRY2	11	45912013	rs7115753	A	G	44.94	1.04	$3.8x10^{-9}$
149	CELF1	11	47529947	rs7124681	A	C	40.97	1.04	$5.1x10^{-9}$
150	MAP3K11	11	65294799	rs1783541	T	C	20.35	1.06	$2.0x10^{-14}$

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	EA	NEA	EAF (%)	OR	*p*-value
151	CCND1	11	69448758	rs11820019	T	C	97.33	1.16	$5.1x10^{-12}$
152	CENTD2/ARAP1	11	72460398	rs77464186	A	C	83.63	1.11	$4.7x10^{-33}$
153	MTNR1B	11	92708710	rs10830963	G	C	27.65	1.10	$4.8x10^{-43}$
154	ETS1	11	128398938	rs67232546	T	C	20.7	1.06	$1.3x10^{-11}$
155	CCND2	12	4384844	rs76895963	T	G	98.02	1.62	$1.4x10^{-69}$
156	CDKN1B	12	12871099	rs2066827	G	T	23.5	1.05	$4.2x10^{-8}$
157	ITPR2	12	26453283	rs718314	G	A	25.32	1.05	$8.4x10^{-11}$
158	KLHDC5	12	27965150	rs10842994	C	T	80.54	1.08	$4.1x10^{-20}$
159	HMGA2	12	66221060	rs2258238	T	A	10.42	1.10	$4.5x10^{-21}$
160	TSPAN8/LGR5	12	71522953	rs1796330	G	C	57.11	1.05	$2.2x10^{-14}$
161	USP44	12	95928560	rs2197973	T	C	53.75	1.04	$3.6x10^{-8}$
162	RMST	12	97848775	rs77864822	A	G	93.24	1.08	$1.1x10^{-8}$
163	WSCD2	12	108629780	rs1426371	G	A	73.89	1.05	$8.2x10^{-12}$
164	KSR2	12	118412373	rs34965774	A	G	14.38	1.06	$2.0x10^{-9}$
165	HNF1A	12	121432117	rs56348580	G	C	68.89	1.05	$2.3x10^{-13}$
166	MPHOSPH9	12	123450765	rs4148856	C	G	78.14	1.05	$1.7x10^{-10}$
167	ZNF664	12	124468572	rs7978610	G	C	66.55	1.27	$2.0x10^{-8}$
168	FBRSL1	12	133069698	rs12811407	A	G	33.05	1.05	$1.7x10^{-12}$
169	RNF6	13	26776999	rs34584161	A	G	75.98	1.05	$2.2x10^{-10}$
170	HMGB1	13	31042452	rs11842871	G	T	73.45	1.04	$1.2x10^{-8}$
171	KL	13	33554302	rs576674	G	A	16.94	1.05	$8.3x10^{-10}$
172	DLEU1	13	51096095	rs963740	A	T	71.28	1.04	$2.1x10^{-8}$
173	PCDH17	13	58366634	rs9537803	C	T	27.71	1.04	$4.6x10^{-8}$
174	SRGAP2D	13	59077406	rs9563615	A	T	71.01	1.05	$6.4x10^{-11}$
175	SPRY2	13	80717156	rs1359790	G	A	72.01	1.09	$2.4x10^{-31}$
176	IRS2	13	109947213	rs7987740	T	C	60.94	1.04	$4.0x10^{-8}$
177	SLC7A7	14	23288935	rs17122772	G	C	22.8	1.04	$1.6x10^{-8}$
178	AKAP6	14	33302882	rs17522122	T	G	47.42	1.04	$3.2x10^{-9}$
179	CLEC14A	14	38848419	rs8017808	G	T	74.31	1.04	$2.1x10^{-8}$
180	NRXN3	14	79932041	rs17836088	C	G	21.71	1.06	$6.7x10^{-14}$
181	SMEK1	14	91963722	rs8010382	G	A	42.14	1.04	$6.5x10^{-9}$
182	MARK3	14	103894071	rs62007683	G	T	65.32	1.04	$3.1x10^{-8}$
183	RASGRP1	15	38873115	rs34715063	C	T	12.35	1.10	$2.3x10^{-19}$
184	LTK	15	41809205	rs11070332	A	G	35.78	1.05	$1.1x10^{-13}$
185	ONECUT1	15	53091553	rs2456530	T	C	12.72	1.06	$5.4x10^{-9}$
186	WDR72	15	53747228	rs528350911	G	C	0.68	1.27	$2.1x10^{-8}$
187	TCF12	15	57456802	rs117483894	G	A	3.69	1.10	$3.9x10^{-8}$
188	C2CD4A/B	15	62394264	rs8037894	G	C	56.63	1.05	$2.6x10^{-13}$
189	USP3	15	63871292	rs7178762	C	T	45.95	1.04	$5.4x10^{-10}$

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	EA	NEA	EAF (%)	OR	*p*-value
190	MAP2K5	15	68080886	rs4776970	A	T	64.06	1.04	$5.0x10^{-9}$
191	PTPN9	15	75932129	rs13737	G	T	75.86	1.05	$5.6x10^{-10}$
192	HMG20A	15	77818128	rs1005752	A	C	71.54	1.08	$2.5x10^{-29}$
193	AP3S2	15	90423293	rs4932265	T	C	26.72	1.07	$4.2x10^{-20}$
194	PRC1	15	91511260	rs12910825	G	A	36.12	1.05	$1.6x10^{-15}$
195	ITFG3	16	295795	rs6600191	T	C	82.46	1.06	$9.3x10^{-13}$
196	CLUAP1	16	3583173	rs3751837	T	C	22	1.04	$1.4x10^{-8}$
197	ATP2A1	16	28915217	rs8046545	G	A	35.89	1.04	$1.9x10^{-8}$
198	FAM57B	16	30045789	rs11642430	G	C	39.9	1.04	$2.2x10^{-9}$
199	FTO	16	53800954	rs1421085	C	T	41.5	1.13	$3.1x10^{-84}$
200	NFAT5	16	69651866	rs862320	C	T	57.83	1.04	$3.9x10^{-11}$
201	BCAR1	16	75234872	rs72802342	C	A	92.31	1.17	$4.0x10^{-32}$
202	CMIP	16	81534790	rs2925979	T	C	29.96	1.05	$1.4x10^{-14}$
203	SPG7	16	89564055	rs12920022	A	T	15.75	1.05	$3.4x10^{-9}$
204	ZZEF1	17	4045440	rs1377807	C	G	31.18	1.05	$4.2x10^{-13}$
205	ATP1B2	17	7549681	rs1641523	C	T	42.76	1.05	$1.2x10^{-10}$
206	GLP2R	17	9785187	rs7222481	C	G	32.38	1.04	$1.4x10^{-8}$
207	RAI1	17	17661802	rs4925109	A	G	31.64	1.05	$2.8x10^{-12}$
208	NF1	17	29413019	rs71372253	C	T	6.42	1.08	$4.4x10^{-8}$
209	HNF1B	17	36099952	rs10908278	T	A	48.08	1.08	$6.4x10^{-36}$
210	MLX	17	40731411	rs34855406	C	G	27.72	1.05	$2.3x10^{-12}$
211	TTLL6	17	47060322	rs35895680	C	A	67.8	1.06	$2.5x10^{-15}$
212	KIF2B	17	52140805	rs569511541	G	A	0.02	7.63	$1.5x10^{-8}$
213	ACE	17	62203304	rs60276348	T	C	13.97	1.05	$2.6x10^{-8}$
214	BPTF	17	65892507	rs61676547	C	G	19.24	1.06	$2.9x10^{-11}$
215	LAMA1	18	7070642	rs7240767	C	T	37.62	1.04	$1.6x10^{-8}$
216	COMMD9	18	36278709	rs62080313	C	T	12.33	1.06	$1.0x10^{-8}$
217	TCF4	18	53050646	rs72926932	C	A	8.39	1.09	$1.0x10^{-14}$
218	WDR7	18	54675384	rs17684074	G	C	74.03	1.04	$2.9x10^{-8}$
219	GRP	18	56876228	rs9957145	G	A	82.9	1.05	$8.1x10^{-9}$
220	MC4R	18	57848369	rs523288	T	A	23.77	1.05	$7.6x10^{-13}$
221	BCL2A	18	60845884	rs12454712	T	C	61.42	1.05	$4.6x10^{-13}$
222	UHRF1	19	4948862	rs7249758	A	G	20.39	1.05	$3.4x10^{-9}$
223	INSR	19	7240848	rs75253922	C	T	19.09	1.05	$2.7x10^{-8}$
224	MAP2K7	19	7970635	rs4804833	A	G	39.02	1.05	$7.7x10^{-13}$
225	FARSA	19	13038415	rs3111316	A	G	58.85	1.05	$6.3x10^{-13}$
226	TM6SF2	19	19388500	rs8107974	T	A	7.69	1.10	$3.3x10^{-15}$
227	PEPD	19	33890838	rs10406327	C	G	52.26	1.04	$3.8x10^{-8}$
228	TOMM40/APOE	19	45411941	rs429358	T	C	84.58	1.08	$2.6x10^{-18}$

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	EA	NEA	EAF (%)	OR	*p*-value
229	GIPR	19	46157019	rs10406431	A	G	56.25	1.05	$9.6x10^{-14}$
230	ZC3H4	19	47569003	rs3810291	A	G	67.3	1.05	$8.9x10^{-12}$
231	NKX2.2	20	21466795	rs13041756	C	T	10.72	1.06	$1.4x10^{-8}$
232	RALY	20	32596704	rs2268078	A	G	65.72	1.04	$2.3x10^{-10}$
233	HNF4A	20	43042364	rs1800961	T	C	3.53	1.18	$2.3x10^{-22}$
234	EYA2	20	45598564	rs6063048	G	A	72.46	1.05	$2.2x10^{-11}$
235	CEBPB	20	48832135	rs11699802	C	T	53.59	1.04	$1.8x10^{-11}$
236	TSHZ2	20	51223594	rs34454109	A	T	77.09	1.04	$7.1x10^{-9}$
237	GNAS	20	57394628	rs6070625	G	C	51.74	1.05	$5.3x10^{-14}$
238	TCEA2	20	62693175	rs59944054	A	G	23.82	1.06	$1.5x10^{-8}$
239	MTMR3/ASCC2	22	30609554	rs6518681	G	A	91.36	1.09	$1.1x10^{-12}$
240	YWHAH	22	32348841	rs117001013	C	T	91.17	1.07	$1.7x10^{-8}$
241	EP300	22	41489920	rs5758223	A	G	71.67	1.04	$3.8x10^{-8}$
242	PNPLA3	22	44324730	rs738408	T	C	22.61	1.05	$1.4x10^{-10}$
243	PIM3	22	50356850	rs1801645	C	T	27.5	1.04	$1.5x10^{-8}$

Descriptions: N***earest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID;* ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37;* ***rsid:*** *Cluster ID;* ***EA:*** *Discovery SNP effect allele;* ***NEA:*** *Discovery SNP alternative allele;* ***EAF:*** *Discovery SNP effect allele frequency;* ***OR:*** *Odds ratio associated with SNP effect allele;* ***P_value:*** *P_value associated with SNP effect allele.*

C.4: Supporting tables with further results for single-SNP association with T2D status

Table C.4. 1 - Single SNP association with T2D status in European ancestry population (nominal level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
1	DENND2C	1	115144899	rs184660829	3.2×10^{-03}
2	PTGFRN	1	117532790	rs1127215	1.2×10^{-05}
3	NOTCH2	1	120526982	rs1493694	5.1×10^{-07}
4	SEC16B	1	177889025	rs539515	4.2×10^{-04}
5	DSTYK	1	205114873	rs12048743	6.5×10^{-04}
6	SRGAP2	1	206593900	rs9430095	2.5×10^{-03}
7	PROX1	1	214159256	rs340874	2.8×10^{-10}
8	LYPLAL1	1	219748818	rs2820446	1.7×10^{-03}
9	ABCB10	1	229672955	rs348330	3.9×10^{-06}
10	GNG4	1	235690800	rs291367	4.7×10^{-03}
11	MACF1	1	40035928	rs3768321	4.7×10^{-08}
12	FAF1	1	51256091	rs58432198	9.2×10^{-06}
13	PATJ	1	62579891	rs12140153	7.0×10^{-06}
14	GLI2	2	121347612	rs11688682	4.1×10^{-06}
15	PABPC1P2	2	147861633	rs35999103	1.8×10^{-05}
16	CYTIP	2	158339550	rs13426680	4.3×10^{-05}
17	RBMS1	2	161135544	rs3772071	3.3×10^{-03}
18	GRB14/COBLL1	2	165513091	rs10195252	9.2×10^{-11}
19	FAM49A	2	16574669	rs11680058	3.5×10^{-05}
20	CRYBA2	2	219859171	rs113414093	2.9×10^{-03}
21	IRS1	2	227101411	rs2972144	2.1×10^{-14}
22	DTNB	2	25643221	rs17802463	7.9×10^{-03}
23	GCKR	2	27730940	rs1260326	3.2×10^{-11}
24	TMEM18	2	422144	rs62107261	2.4×10^{-05}
25	THADA	2	43698028	rs80147536	2.3×10^{-12}
26	BNIPL	2	59307725	rs6545714	1.2×10^{-06}
27	BCL11A	2	60583665	rs243024	2.9×10^{-07}
28	CEP68	2	65287896	rs2249105	4.9×10^{-04}
29	TMEM127	2	96913918	rs79046683	1.3×10^{-02}
30	ADCY5	3	123065778	rs11708067	3.9×10^{-09}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
31	PPARG	3	12336507	rs11709077	1.3×10^{-08}
32	SLC12A8	3	124926637	rs649961	3.9×10^{-03}
33	TMCC1	3	129333182	rs9828772	8.4×10^{-03}
34	TSC22D2	3	150066540	rs62271373	2.2×10^{-04}
35	MBNL1	3	152086533	rs13065698	4.7×10^{-04}
36	EGFEM1P	3	168218841	rs7629630	1.8×10^{-02}
37	SLC2A2	3	170733076	rs9873618	2.1×10^{-04}
38	ABCC5	3	183738460	rs2872246	2.4×10^{-02}
39	IGF2BP2	3	185503456	rs6780171	6.4×10^{-20}
40	ST6GAL1	3	186665645	rs3887925	7.9×10^{-09}
41	LPP	3	187740899	rs4686471	7.0×10^{-08}
42	UBE2E2	3	23455582	rs35352848	2.0×10^{-11}
43	KIF9	3	46925539	rs11926707	1.9×10^{-04}
44	RBM6	3	49980596	rs4688760	2.4×10^{-08}
45	RFT1	3	53127677	rs2581787	9.2×10^{-03}
46	CACNA2D3	3	54828827	rs76263492	1.9×10^{-04}
47	PSMD6	3	63962339	rs3774723	4.4×10^{-09}
48	ADAMTS9	3	64701146	rs9860730	2.5×10^{-02}
49	SHQ1	3	72865183	rs13085136	6.1×10^{-05}
50	ROBO2	3	77671721	rs2272163	1.2×10^{-03}
51	SLC9B1	4	104140848	rs1580278	2.7×10^{-03}
52	PABPC4L	4	137083193	rs1296328	6.6×10^{-05}
53	TMEM154	4	153513369	rs7669833	5.0×10^{-04}
54	PDGFC	4	157652753	rs28819812	3.3×10^{-03}
55	MAEA	4	1784403	rs56337234	2.0×10^{-08}
56	ACSL1	4	185717759	rs58730668	1.9×10^{-03}
57	HTT	4	3241845	rs362307	3.9×10^{-05}
58	GNPDA2	4	45186139	rs10938398	3.4×10^{-04}
59	USP46	4	52818664	rs2102278	3.5×10^{-07}
60	WFS1	4	6306763	rs10937721	3.2×10^{-13}
61	PCGF3	4	744972	rs1531583	2.4×10^{-03}
62	SCD5	4	83578271	rs12642790	5.2×10^{-04}
63	FAM13A	4	89740894	rs1903002	4.1×10^{-02}
64	SMARCAD1	4	95091911	rs6821438	6.8×10^{-03}
65	SLCO6A1	5	101232944	rs138337556	5.6×10^{-04}
66	PAM	5	102422968	rs115505614	6.7×10^{-11}
67	PHF15	5	133864599	rs329122	2.0×10^{-05}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
68	ANKH	5	14751305	rs146886108	1.5×10^{-10}
69	EBF1	5	157928196	rs3934712	1.1×10^{-03}
70	MRPS30	5	44682589	rs6884702	2.5×10^{-02}
71	ITGA1	5	52100489	rs3811978	9.3×10^{-07}
72	ARL15	5	53271420	rs702634	2.4×10^{-04}
73	ANKRD55	5	55808475	rs465002	6.8×10^{-09}
74	PIK3R1	5	67714246	rs4976033	7.3×10^{-03}
75	POC5	5	75003678	rs2307111	8.6×10^{-06}
76	ZBED3	5	76424949	rs4457053	4.3×10^{-04}
77	DMGDH	5	78430607	rs1316776	8.4×10^{-04}
78	RASA1	5	86577352	rs7719891	2.7×10^{-04}
79	BEND3	6	107431688	rs4946812	1.6×10^{-03}
80	CENPW	6	126792095	rs11759026	2.3×10^{-06}
81	SOGA3	6	127416930	rs2800733	1.1×10^{-04}
82	SLC35D3	6	137300960	rs9494624	4.4×10^{-02}
83	MIR3668	6	139835329	rs2982521	4.6×10^{-06}
84	SLC22A3	6	160770312	rs474513	1.0×10^{-04}
85	QKI	6	164133001	rs4709746	4.6×10^{-02}
86	CDKAL1	6	20679709	rs7756992	2.1×10^{-21}
87	MHC	6	32573415	rs601945	2.0×10^{-07}
88	HMGA1	6	34247047	rs77136196	3.7×10^{-03}
89	LRFN2	6	40409243	rs34298980	2.1×10^{-06}
90	VEGFA	6	43814190	rs6458354	1.6×10^{-03}
91	TFAP2B	6	50788778	rs3798519	3.2×10^{-08}
92	SLC25A51P1	6	67387490	rs555402748	9.9×10^{-06}
93	RREB1	6	7231843	rs9379084	5.9×10^{-13}
94	FBXL13	7	102486254	rs11496066	3.5×10^{-04}
95	RELN	7	103444978	rs39328	4.4×10^{-05}
96	CTTNBP2	7	117495667	rs6976111	4.4×10^{-05}
97	KLF14	7	130457914	rs1562396	2.6×10^{-07}
98	AOC1	7	150537635	rs62492368	5.9×10^{-05}
99	DGKB	7	15063569	rs10228066	1.1×10^{-09}
100	MNX1	7	156930550	rs6459733	1.6×10^{-05}
101	IGF2BP3	7	23512896	rs4279506	3.0×10^{-04}
102	JAZF1	7	28198677	rs1708302	2.7×10^{-14}
103	CRHR2	7	30728452	rs917195	3.0×10^{-04}
104	GCK	7	44255643	rs878521	5.8×10^{-04}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
105	XKR6	8	10808687	rs57327348	2.9×10^{-03}
106	TRHR	8	110123183	rs12680028	6.4×10^{-05}
107	SLC30A8	8	118185025	rs3802177	1.3×10^{-22}
108	CASC11	8	128711742	rs17772814	3.6×10^{-02}
109	PVT1	8	129568078	rs1561927	1.0×10^{-04}
110	BOP1	8	145507304	rs4977213	2.9×10^{-05}
111	LPL	8	19830921	rs10096633	5.1×10^{-04}
112	PURG	8	30863938	rs10954772	1.4×10^{-04}
113	ANK1	8	41508577	rs13262861	3.5×10^{-12}
114	TP53INP1	8	95961626	rs10097617	2.1×10^{-08}
115	CPQ	8	97737741	rs149364428	1.3×10^{-04}
116	MSRA	8	9974824	rs17689007	7.1×10^{-05}
117	ABO	9	136149229	rs505922	1.9×10^{-03}
118	GPSM1	9	139241030	rs28505901	1.2×10^{-08}
119	HAUS6	9	19067833	rs7022807	4.3×10^{-03}
120	FOCAD	9	20241069	rs7867635	1.1×10^{-03}
121	CDKN2A/B	9	22134068	rs10811660	9.4×10^{-28}
122	LINGO2	9	28410683	rs1412234	6.3×10^{-04}
123	UBAP2	9	34074476	rs12001437	3.1×10^{-02}
124	GLIS3	9	4291928	rs10974438	4.6×10^{-08}
125	TLE4	9	81905590	rs17791513	1.9×10^{-03}
126	TLE1	9	84308948	rs2796441	1.5×10^{-13}
127	ZNF169	9	97001682	rs55653563	1.1×10^{-02}
128	TCF7L2	10	114758349	rs7903146	1.5×10^{-151}
129	WDR11	10	122915345	rs72631105	1.4×10^{-06}
130	CDC123/CAMK1D	10	12307894	rs11257655	5.9×10^{-12}
131	PLEKHA1	10	124193181	rs2280141	1.8×10^{-02}
132	NEUROG3	10	71466578	rs2642588	2.1×10^{-07}
133	ZMIZ1	10	80952826	rs703972	1.7×10^{-12}
134	PTEN	10	89769340	rs11202627	1.2×10^{-02}
135	HHEX/IDE	10	94462427	rs10882101	2.6×10^{-17}
136	ETS1	11	128398938	rs67232546	1.1×10^{-05}
137	PDE3B	11	14763828	rs141521721	4.3×10^{-04}
138	KCNJ11	11	17408404	rs5213	2.1×10^{-09}
139	INS/IGF2	11	2197286	rs4929965	1.3×10^{-10}
140	METTL15	11	28534898	rs4923543	4.7×10^{-03}
141	KCNQ1	11	2857194	rs2237895	2.5×10^{-21}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
142	QSER1	11	32927778	rs145678014	1.1×10^{-02}
143	PDHX	11	34982148	rs2767036	5.4×10^{-03}
144	HSD17B12	11	43877934	rs1061810	1.2×10^{-02}
145	CRY2	11	45912013	rs7115753	1.7×10^{-03}
146	CELF1	11	47529947	rs7124681	1.2×10^{-04}
147	MAP3K11	11	65294799	rs1783541	8.0×10^{-06}
148	CCND1	11	69448758	rs11820019	2.1×10^{-06}
149	CENTD2/ARAP1	11	72460398	rs77464186	2.3×10^{-15}
150	MTNR1B	11	92708710	rs10830963	1.6×10^{-12}
151	WSCD2	12	108629780	rs1426371	3.4×10^{-05}
152	KSR2	12	118412373	rs34965774	1.6×10^{-04}
153	HNF1A	12	121432117	rs56348580	4.0×10^{-04}
154	MPHOSPH9	12	123450765	rs4148856	5.6×10^{-03}
155	ZNF664	12	124468572	rs7978610	1.9×10^{-03}
156	CDKN1B	12	12871099	rs2066827	4.1×10^{-03}
157	FBRSL1	12	133069698	rs12811407	3.6×10^{-06}
158	ITPR2	12	26453283	rs718314	1.1×10^{-06}
159	KLHDC5	12	27965150	rs10842994	5.3×10^{-05}
160	CCND2	12	4384844	rs76895963	2.6×10^{-29}
161	HMGA2	12	66221060	rs2258238	3.3×10^{-10}
162	TSPAN8/LGR5	12	71522953	rs1796330	8.5×10^{-04}
163	USP44	12	95928560	rs2197973	4.3×10^{-05}
164	RMST	12	97848775	rs77864822	6.3×10^{-03}
165	IRS2	13	109947213	rs7987740	3.1×10^{-07}
166	RNF6	13	26776999	rs34584161	1.3×10^{-02}
167	HMGB1	13	31042452	rs11842871	1.5×10^{-05}
168	KL	13	33554302	rs576674	1.8×10^{-02}
169	DLEU1	13	51096095	rs963740	5.5×10^{-03}
170	PCDH17	13	58366634	rs9537803	1.6×10^{-03}
171	SRGAP2D	13	59077406	rs9563615	2.6×10^{-07}
172	SPRY2	13	80717156	rs1359790	6.2×10^{-13}
173	MARK3	14	103894071	rs62007683	4.0×10^{-04}
174	SLC7A7	14	23288935	rs17122772	2.0×10^{-02}
175	AKAP6	14	33302882	rs17522122	3.2×10^{-04}
176	CLEC14A	14	38848419	rs8017808	1.1×10^{-08}
177	NRXN3	14	79932041	rs17836088	1.8×10^{-05}
178	SMEK1	14	91963722	rs8010382	3.8×10^{-03}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
179	RASGRP1	15	38873115	rs34715063	3.4×10^{-10}
180	LTK	15	41809205	rs11070332	1.4×10^{-04}
181	ONECUT1	15	53091553	rs2456530	1.1×10^{-03}
182	WDR72	15	53747228	rs528350911	2.1×10^{-02}
183	C2CD4A/B	15	62394264	rs8037894	2.2×10^{-03}
184	USP3	15	63871292	rs7178762	2.9×10^{-02}
185	PTPN9	15	75932129	rs13737	1.2×10^{-05}
186	HMG20A	15	77818128	rs1005752	8.4×10^{-10}
187	AP3S2	15	90423293	rs4932265	3.7×10^{-06}
188	PRC1	15	91511260	rs12910825	5.4×10^{-05}
189	ATP2A1	16	28915217	rs8046545	5.4×10^{-03}
190	ITFG3	16	295795	rs6600191	1.6×10^{-05}
191	FAM57B	16	30045789	rs11642430	1.6×10^{-04}
192	CLUAP1	16	3583173	rs3751837	4.7×10^{-03}
193	FTO	16	53800954	rs1421085	4.2×10^{-24}
194	NFAT5	16	69651866	rs862320	3.8×10^{-10}
195	BCAR1	16	75234872	rs72802342	3.1×10^{-13}
196	CMIP	16	81534790	rs2925979	5.1×10^{-03}
197	SPG7	16	89564055	rs12920022	1.0×10^{-04}
198	RAI1	17	17661802	rs4925109	5.0×10^{-04}
199	HNF1B	17	36099952	rs10908278	2.0×10^{-14}
200	ZZEF1	17	4045440	rs1377807	6.8×10^{-07}
201	MLX	17	40731411	rs34855406	6.0×10^{-06}
202	TTLL6	17	47060322	rs35895680	2.3×10^{-05}
203	KIF2B	17	52140805	rs569511541	1.0×10^{-04}
204	ACE	17	62203304	rs60276348	1.7×10^{-03}
205	BPTF	17	65892507	rs61676547	3.3×10^{-03}
206	ATP1B2	17	7549681	rs1641523	1.8×10^{-02}
207	GLP2R	17	9785187	rs7222481	5.2×10^{-03}
208	COMMD9	18	36278709	rs62080313	2.8×10^{-03}
209	TCF4	18	53050646	rs72926932	2.6×10^{-04}
210	WDR7	18	54675384	rs17684074	2.9×10^{-04}
211	GRP	18	56876228	rs9957145	5.1×10^{-04}
212	MC4R	18	57848369	rs523288	1.2×10^{-07}
213	BCL2A	18	60845884	rs12454712	2.4×10^{-06}
214	LAMA1	18	7070642	rs7240767	5.8×10^{-04}
215	FARSA	19	13038415	rs3111316	2.6×10^{-06}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
216	TM6SF2	19	19388500	rs8107974	5.3×10^{-06}
217	PEPD	19	33890838	rs10406327	5.2×10^{-07}
218	TOMM40/APOE	19	45411941	rs429358	4.2×10^{-06}
219	GIPR	19	46157019	rs10406431	6.6×10^{-13}
220	ZC3H4	19	47569003	rs3810291	2.5×10^{-06}
221	UHRF1	19	4948862	rs7249758	7.1×10^{-04}
222	INSR	19	7240848	rs75253922	7.3×10^{-05}
223	MAP2K7	19	7970635	rs4804833	8.8×10^{-06}
224	RALY	20	32596704	rs2268078	1.1×10^{-03}
225	HNF4A	20	43042364	rs1800961	4.3×10^{-09}
226	EYA2	20	45598564	rs6063048	8.0×10^{-06}
227	CEBPB	20	48832135	rs11699802	6.1×10^{-05}
228	GNAS	20	57394628	rs6070625	4.2×10^{-06}
229	TCEA2	20	62693175	rs59944054	2.6×10^{-04}
230	MTMR3/ASCC2	22	30609554	rs6518681	1.7×10^{-03}
231	YWHAH	22	32348841	rs117001013	1.5×10^{-02}
232	EP300	22	41489920	rs5758223	1.7×10^{-02}
233	PNPLA3	22	44324730	rs738408	3.4×10^{-03}
234	PIM3	22	50356850	rs1801645	4.1×10^{-04}

Descriptions: *N**earest gene:** refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID**;*** ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37**;*** ***rsid:*** *Cluster ID**;***
P-value: *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

Table C.4. 2 - Single SNP association with T2D status in European ancestry population (genome-wide level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
1	PROX1	1	214159256	rs340874	2.8×10^{-10}
2	MACF1	1	40035928	rs3768321	4.7×10^{-08}
3	GRB14/COBLL1	2	165513091	rs10195252	9.2×10^{-11}
4	IRS1	2	227101411	rs2972144	2.1×10^{-14}
5	GCKR	2	27730940	rs1260326	3.2×10^{-11}
6	THADA	2	43698028	rs80147536	2.3×10^{-12}
7	ADCY5	3	123065778	rs11708067	3.9×10^{-09}
8	PPARG	3	12336507	rs11709077	1.3×10^{-08}
9	IGF2BP2	3	185503456	rs6780171	6.4×10^{-20}
10	ST6GAL1	3	186665645	rs3887925	7.9×10^{-09}
11	UBE2E2	3	23455582	rs35352848	2.0×10^{-11}
12	RBM6	3	49980596	rs4688760	2.4×10^{-08}
13	PSMD6	3	63962339	rs3774723	4.4×10^{-09}
14	MAEA	4	1784403	rs56337234	2.0×10^{-08}
15	WFS1	4	6306763	rs10937721	3.2×10^{-13}
16	PAM	5	102422968	rs115505614	6.7×10^{-11}
17	ANKH	5	14751305	rs146886108	1.5×10^{-10}
18	ANKRD55	5	55808475	rs465002	6.8×10^{-09}
19	CDKAL1	6	20679709	rs7756992	2.1×10^{-21}
20	TFAP2B	6	50788778	rs3798519	3.2×10^{-08}
21	RREB1	6	7231843	rs9379084	5.9×10^{-13}
22	DGKB	7	15063569	rs10228066	1.1×10^{-09}
23	JAZF1	7	28198677	rs1708302	2.7×10^{-14}
24	SLC30A8	8	118185025	rs3802177	1.3×10^{-22}
25	ANK1	8	41508577	rs13262861	3.5×10^{-12}
26	TP53INP1	8	95961626	rs10097617	2.1×10^{-08}
27	GPSM1	9	139241030	rs28505901	1.2×10^{-08}
28	CDKN2A/B	9	22134068	rs10811660	9.4×10^{-28}
29	GLIS3	9	4291928	rs10974438	4.6×10^{-08}
30	TLE1	9	84308948	rs2796441	1.5×10^{-13}
31	TCF7L2	10	114758349	rs7903146	1.5×10^{-151}
32	CDC123/CAMK1D	10	12307894	rs11257655	5.9×10^{-12}
33	ZMIZ1	10	80952826	rs703972	1.7×10^{-12}
34	HHEX/IDE	10	94462427	rs10882101	2.6×10^{-17}
35	KCNJ11	11	17408404	rs5213	2.1×10^{-09}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
36	INS/IGF2	11	2197286	rs4929965	1.3×10^{-10}
37	KCNQ1	11	2857194	rs2237895	2.5×10^{-21}
38	CENTD2/ARAP1	11	72460398	rs77464186	2.3×10^{-15}
39	MTNR1B	11	92708710	rs10830963	1.6×10^{-12}
40	CCND2	12	4384844	rs76895963	2.6×10^{-29}
41	HMGA2	12	66221060	rs2258238	3.3×10^{-10}
42	SPRY2	13	80717156	rs1359790	6.2×10^{-13}
43	CLEC14A	14	38848419	rs8017808	1.1×10^{-08}
44	RASGRP1	15	38873115	rs34715063	3.4×10^{-10}
45	HMG20A	15	77818128	rs1005752	8.4×10^{-10}
46	FTO	16	53800954	rs1421085	4.2×10^{-24}
47	NFAT5	16	69651866	rs862320	3.8×10^{-10}
48	BCAR1	16	75234872	rs72802342	3.1×10^{-13}
49	HNF1B	17	36099952	rs10908278	2.0×10^{-14}
50	GIPR	19	46157019	rs10406431	6.6×10^{-13}
51	HNF4A	20	43042364	rs1800961	4.3×10^{-09}

Descriptions: *N**earest gene:** refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID**;*** ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37**;*** ***rsid:*** *Cluster ID**;***
P-value: *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

Table C.4. 3 - Single SNP association with T2D status in Asian ancestry population (nominal level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
1	DENND2C	1	115144899	rs184660829	2.3×10^{-02}
2	MACF1	1	40035928	rs3768321	3.2×10^{-03}
3	GLI2	2	121347612	rs11688682	1.9×10^{-02}
4	IRS1	2	227101411	rs2972144	4.3×10^{-02}
5	ADCY5	3	123065778	rs11708067	1.2×10^{-03}
6	EGFEM1P	3	168218841	rs7629630	2.4×10^{-02}
7	IGF2BP2	3	185503456	rs6780171	9.2×10^{-03}
8	PSMD6	3	63962339	rs3774723	1.3×10^{-02}
9	ROBO2	3	77671721	rs2272163	2.1×10^{-02}
10	PAM	5	102422968	rs115505614	4.4×10^{-02}
11	EBF1	5	157928196	rs3934712	4.1×10^{-02}
12	DMGDH	5	78430607	rs1316776	4.5×10^{-03}
13	BEND3	6	107431688	rs4946812	2.5×10^{-02}
14	CDKAL1	6	20679709	rs7756992	1.9×10^{-02}
15	RREB1	6	7231843	rs9379084	1.3×10^{-02}
16	DGKB	7	15063569	rs10228066	1.4×10^{-02}
17	GCK	7	44255643	rs878521	4.5×10^{-03}
18	PURG	8	30863938	rs10954772	2.4×10^{-02}
19	TCF7L2	10	114758349	rs7903146	5.5×10^{-09}
20	HHEX/IDE	10	94462427	rs10882101	2.4×10^{-03}
21	PDE3B	11	14763828	rs141521721	2.8×10^{-02}
22	KCNJ11	11	17408404	rs5213	4.0×10^{-02}
23	CENTD2/ARAP1	11	72460398	rs77464186	6.5×10^{-04}
24	ZNF664	12	124468572	rs7978610	1.4×10^{-02}
25	MARK3	14	103894071	rs62007683	3.5×10^{-02}
26	RASGRP1	15	38873115	rs34715063	4.9×10^{-02}
27	BCAR1	16	75234872	rs72802342	2.6×10^{-02}
28	NF1	17	29413019	rs71372253	9.8×10^{-03}
29	HNF1B	17	36099952	rs10908278	4.8×10^{-02}
30	LAMA1	18	7070642	rs7240767	4.8×10^{-02}
31	GIPR	19	46157019	rs10406431	1.7×10^{-04}
32	GNAS	20	57394628	rs6070625	4.8×10^{-02}
33	PIM3	22	50356850	rs1801645	7.5×10^{-03}

Descriptions: *N**earest gene:** refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID**;** **Base pair position:** Base pair position of the SNP on the human genome based on the human reference genome build 37**; rsid:** Cluster ID**;***
P-value: *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

Table C.4. 4 - Single SNP association with T2D status in Asian ancestry population (genome-wide level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
1	TCF7L2	10	114758349	rs7903146	5.5×10^{-09}

Descriptions: *N**earest gene:** refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID**;** **Base pair position:** Base pair position of the SNP on the human genome based on the human reference genome build 37**; rsid:** Cluster ID**;***
P-value: *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

Table C.4. 5 - Single SNP association with T2D status in African ancestry population (nominal level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	P-value: Adjusting for age, sex, BMI, and ancestry
1	SEC16B	1	177889025	rs539515	1.5×10^{-02}
2	GNG4	1	235690800	rs291367	1.3×10^{-02}
3	GRB14/COBLL1	2	165513091	rs10195252	2.4×10^{-02}
4	IRS1	2	227101411	rs2972144	4.2×10^{-02}
5	ADCY5	3	123065778	rs11708067	6.2×10^{-03}
6	PSMD6	3	63962339	rs3774723	2.8×10^{-02}
7	SLCO6A1	5	101232944	rs138337556	3.0×10^{-02}
8	CASC11	8	128711742	rs17772814	3.7×10^{-02}
9	LPL	8	19830921	rs10096633	1.3×10^{-02}
10	MSRA	8	9974824	rs17689007	3.4×10^{-02}
11	TCF7L2	10	114758349	rs7903146	3.1×10^{-03}
12	PDE3B	11	14763828	rs141521721	1.2×10^{-02}
13	FBRSL1	12	133069698	rs12811407	2.0×10^{-02}
14	RNF6	13	26776999	rs34584161	6.0×10^{-03}
15	AP3S2	15	90423293	rs4932265	4.8×10^{-02}
16	CLUAP1	16	3583173	rs3751837	1.2×10^{-02}
17	GIPR	19	46157019	rs10406431	3.2×10^{-02}
18	ZC3H4	19	47569003	rs3810291	2.7×10^{-02}

Descriptions: ***Nearest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID****;*** ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37****;*** ***rsid:*** *Cluster ID****;***
P-value: *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

C.5: Supporting tables with further results for single-SNP association with AOO of T2D

Table C.5. 1 - Single SNP association with AOO of T2D in European ancestry population (genome-wide level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	HR	Lower 95% CI	Upper 95% CI	P-value
1	NOTCH2	01	120526982	rs1493694	1.114	1.076	1.154	1.7×10^{-09}
2	PROX1	01	214159256	rs340874	1.080	1.056	1.105	3.5×10^{-11}
3	GRB14/COBLL1	02	165513091	rs10195252	1.110	1.084	1.136	1.5×10^{-18}
4	GCKR	02	27730940	rs1260326	1.073	1.048	1.098	3.1×10^{-09}
5	BCL11A	02	60583665	rs243024	1.066	1.042	1.091	3.0×10^{-08}
6	IRS1	02	227101411	rs2972144	1.111	1.085	1.139	7.3×10^{-18}
7	THADA	02	43698028	rs80147536	1.153	1.108	1.200	2.3×10^{-12}
8	ADCY5	03	123065778	rs11708067	1.105	1.075	1.135	4.0×10^{-13}
9	PPARG	03	12336507	rs11709077	1.147	1.107	1.190	1.0×10^{-13}
10	UBE2E2	03	23455582	rs35352848	1.113	1.081	1.146	3.8×10^{-13}
11	ST6GAL1	03	186665645	rs3887925	1.079	1.055	1.104	6.3×10^{-11}
12	LPP	03	187740899	rs4686471	1.069	1.044	1.094	2.3×10^{-08}
13	IGF2BP2	03	185503456	rs6780171	1.139	1.112	1.166	3.0×10^{-26}
14	SLC2A2	03	170733076	rs9873618	1.080	1.053	1.108	2.1×10^{-09}
15	WFS1	04	6306763	rs10937721	1.093	1.067	1.118	1.1×10^{-13}
16	MAEA	04	1784403	rs56337234	1.084	1.060	1.109	3.6×10^{-12}
17	PAM	05	102422968	rs115505614	1.207	1.150	1.266	2.2×10^{-14}
18	ANKH	05	14751305	rs146886108	1.788	1.506	2.123	3.2×10^{-11}
19	PHF15	05	133864599	rs329122	1.071	1.047	1.096	4.2×10^{-09}
20	ANKRD55	05	55808475	rs465002	1.099	1.071	1.129	2.5×10^{-12}
21	MHC	06	32573415	rs601945	1.085	1.055	1.116	1.4×10^{-08}
22	CDKAL1	06	20679709	rs7756992	1.146	1.118	1.175	1.0×10^{-26}
23	RREB1	06	7231843	rs9379084	1.162	1.118	1.208	2.3×10^{-14}
24	DGKB	07	15063569	rs10228066	1.071	1.047	1.096	4.0×10^{-09}
25	KLF14	07	130457914	rs1562396	1.090	1.064	1.117	2.0×10^{-12}
26	JAZF1	07	28198677	rs1708302	1.110	1.085	1.135	1.5×10^{-19}
27	TP53INP1	08	95961626	rs10097617	1.066	1.043	1.091	2.5×10^{-08}
28	ANK1	08	41508577	rs13262861	1.122	1.088	1.158	4.4×10^{-13}
29	SLC30A8	08	118185025	rs3802177	1.149	1.121	1.179	1.8×10^{-27}
30	CDKN2A/B	09	22134068	rs10811660	1.193	1.156	1.231	6.6×10^{-28}

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	HR	Lower 95% CI	Upper 95% CI	P-value
31	GLIS3	09	4291928	rs10974438	1.074	1.049	1.100	2.5×10^{-09}
32	TLE1	09	84308948	rs2796441	1.102	1.077	1.128	1.5×10^{-16}
33	GPSM1	09	139241030	rs28505901	1.089	1.060	1.118	5.7×10^{-10}
34	HHEX/IDE	10	94462427	rs10882101	1.114	1.089	1.141	5.9×10^{-20}
35	CDC123/CAMK1D	10	12307894	rs11257655	1.101	1.071	1.131	4.9×10^{-12}
36	ZMIZ1	10	80952826	rs703972	1.087	1.062	1.112	7.2×10^{-13}
37	TCF7L2	10	114758349	rs7903146	1.432	1.398	1.466	2.9×10^{-195}
38	MTNR1B	11	92708710	rs10830963	1.106	1.079	1.134	1.8×10^{-15}
39	KCNQ1	11	2857194	rs2237895	1.123	1.098	1.149	1.2×10^{-23}
40	INS/IGF2	11	2197286	rs4929965	1.081	1.056	1.106	4.3×10^{-11}
41	KCNJ11	11	17408404	rs5213	1.085	1.059	1.110	1.4×10^{-11}
42	CENTD2/ARAP1	11	72460398	rs77464186	1.147	1.111	1.185	1.0×10^{-16}
43	HMGA2	12	66221060	rs2258238	1.133	1.094	1.174	4.2×10^{-12}
44	ITPR2	12	26453283	rs718314	1.075	1.048	1.103	3.7×10^{-08}
45	CCND2	12	4384844	rs76895963	2.076	1.854	2.325	1.1×10^{-36}
46	SPRY2	13	80717156	rs1359790	1.104	1.076	1.132	3.9×10^{-14}
47	CLEC14A	14	38848419	rs8017808	1.099	1.070	1.129	7.6×10^{-12}
48	HMG20A	15	77818128	rs1005752	1.080	1.053	1.108	4.2×10^{-09}
49	RASGRP1	15	38873115	rs34715063	1.097	1.062	1.134	3.0×10^{-08}
50	AP3S2	15	90423293	rs4932265	1.089	1.062	1.117	4.5×10^{-11}
51	BCAR1	16	75234872	rs72802342	1.199	1.145	1.255	1.3×10^{-14}
52	HNF1B	17	36099952	rs10908278	1.097	1.072	1.122	1.4×10^{-15}
53	BCL2A	18	60845884	rs12454712	1.075	1.050	1.101	1.4×10^{-09}
54	PEPD	19	33890838	rs10406327	1.072	1.048	1.097	1.6×10^{-09}
55	GIPR	19	46157019	rs10406431	1.080	1.056	1.105	3.4×10^{-11}
56	TM6SF2	19	19388500	rs8107974	1.127	1.082	1.174	1.1×10^{-08}
57	HNF4A	20	43042364	rs1800961	1.238	1.167	1.314	2.0×10^{-12}

Descriptions: ***Nearest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID****;*** ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37****;*** ***rsid:*** *Cluster ID****;*** *HR: estimated SNP HR associated with AOO of T2D;* ***95% CI:*** *Lower and upper 95% confidence interval of estimated SNP HR associated with AOO of T2D;* ***P-value:*** *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

Table C.5. 2 - Single SNP association with AOO of T2D in Asian ancestry population (nominal level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	HR	Lower 95% CI	Upper 95% CI	P-value
1	DENND2C	01	115144899	rs184660829	4.816	1.429	16.229	1.1×10^{-02}
2	MACF1	01	40035928	rs3768321	1.216	1.041	1.420	1.4×10^{-02}
3	CEP68	02	65287896	rs2249105	1.085	1.003	1.175	4.3×10^{-02}
4	IRS1	02	227101411	rs2972144	1.135	1.033	1.247	8.7×10^{-03}
5	ADCY5	03	123065778	rs11708067	1.193	1.080	1.318	5.1×10^{-04}
6	ROBO2	03	77671721	rs2272163	1.094	1.009	1.186	3.0×10^{-02}
7	UBE2E2	03	23455582	rs35352848	1.126	1.024	1.238	1.5×10^{-02}
8	IGF2BP2	03	185503456	rs6780171	1.114	1.029	1.206	7.6×10^{-03}
9	SLC2A2	03	170733076	rs9873618	1.097	1.002	1.203	4.6×10^{-02}
10	DMGDH	05	78430607	rs1316776	1.108	1.022	1.201	1.2×10^{-02}
11	EBF1	05	157928196	rs3934712	1.113	1.022	1.211	1.4×10^{-02}
12	ANKRD55	05	55808475	rs465002	1.090	1.005	1.183	3.7×10^{-02}
13	CDKAL1	06	20679709	rs7756992	1.139	1.043	1.244	3.8×10^{-03}
14	RREB1	06	7231843	rs9379084	1.210	1.049	1.396	8.9×10^{-03}
15	DGKB	07	15063569	rs10228066	1.095	1.010	1.187	2.8×10^{-02}
16	JAZF1	07	28198677	rs1708302	1.100	1.006	1.203	3.7×10^{-02}
17	GCK	07	44255643	rs878521	1.095	1.001	1.199	4.7×10^{-02}
18	PURG	08	30863938	rs10954772	0.896	0.824	0.975	1.1×10^{-02}
19	GPSM1	09	139241030	rs28505901	1.155	1.046	1.275	4.3×10^{-03}
20	HHEX/IDE	10	94462427	rs10882101	1.145	1.058	1.239	7.6×10^{-04}
21	CDC123/CAMK1D	10	12307894	rs11257655	1.111	1.014	1.217	2.4×10^{-02}
22	TCF7L2	10	114758349	rs7903146	1.336	1.231	1.449	3.6×10^{-12}
23	PDE3B	11	14763828	rs141521721	1.383	1.022	1.872	3.6×10^{-02}
24	KCNJ11	11	17408404	rs5213	1.088	1.004	1.178	3.9×10^{-02}
25	CENTD2/ARAP1	11	72460398	rs77464186	1.192	1.073	1.324	1.1×10^{-03}
26	HNF1B	17	36099952	rs10908278	1.098	1.012	1.191	2.4×10^{-02}
27	ATP1B2	17	7549681	rs1641523	1.092	1.006	1.186	3.5×10^{-02}
28	NF1	17	29413019	rs71372253	0.693	0.484	0.994	4.6×10^{-02}
29	GIPR	19	46157019	rs10406431	1.191	1.101	1.289	1.4×10^{-05}
30	CEBPB	20	48832135	rs11699802	1.095	1.009	1.188	2.9×10^{-02}
31	GNAS	20	57394628	rs6070625	0.908	0.838	0.984	1.9×10^{-02}
32	PIM3	22	50356850	rs1801645	1.093	1.012	1.181	2.3×10^{-02}
33	EP300	22	41489920	rs5758223	1.124	1.017	1.242	2.2×10^{-02}
34	PNPLA3	22	44324730	rs738408	1.126	1.028	1.233	1.1×10^{-02}

Descriptions: ***Nearest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID;* ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37;* ***rsid:*** *Cluster ID;* *HR: estimated SNP HR associated with AOO of T2D;* ***95% CI:*** *Lower and upper 95% confidence interval of estimated SNP HR associated with AOO of T2D;* ***P-value:*** *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

Table C.5. 3 - Single SNP association with AOO of T2D in African ancestry population (nominal level)

#	Nearest gene	Chrom-osome	Base Pair Position	rs Number	HR	Lower 95% CI	Upper 95% CI	P-value
1	GNG4	01	235690800	rs291367	1.208	1.070	1.364	2.3×10^{-03}
2	SEC16B	01	177889025	rs539515	1.165	1.031	1.317	1.4×10^{-02}
3	GRB14/COBLL1	02	165513091	rs10195252	1.144	1.004	1.304	4.3×10^{-02}
4	ADCY5	03	123065778	rs11708067	1.289	1.082	1.536	4.5×10^{-03}
5	PSMD6	03	63962339	rs3774723	1.170	1.046	1.308	6.0×10^{-03}
6	LPL	08	19830921	rs10096633	1.148	1.026	1.285	1.6×10^{-02}
7	MSRA	08	9974824	rs17689007	1.134	1.009	1.274	3.5×10^{-02}
8	TCF7L2	10	114758349	rs7903146	1.315	1.169	1.479	4.8×10^{-06}
9	PDE3B	11	14763828	rs141521721	2.609	1.275	5.338	8.7×10^{-03}
10	FBRSL1	12	133069698	rs12811407	1.167	1.032	1.321	1.4×10^{-02}
11	HNF1A	12	121432117	rs56348580	0.850	0.738	0.979	2.4×10^{-02}
12	RNF6	13	26776999	rs34584161	0.732	0.607	0.883	1.1×10^{-03}
13	AP3S2	15	90423293	rs4932265	1.161	1.021	1.321	2.2×10^{-02}
14	SPG7	16	89564055	rs12920022	1.160	1.015	1.326	2.9×10^{-02}
15	CLUAP1	16	3583173	rs3751837	0.835	0.729	0.957	9.4×10^{-03}
16	GIPR	19	46157019	rs10406431	1.189	1.065	1.326	2.0×10^{-03}

Descriptions: ***Nearest gene:*** *refers to the name of the nearest gene to a DNA polymorphism (SNP in this instance);* ***Chromosome:*** *chromosome number or SNP ID;* ***Base pair position:*** *Base pair position of the SNP on the human genome based on the human reference genome build 37;* ***rsid:*** *Cluster ID;* *HR: estimated SNP HR associated with AOO of T2D;* ***95% CI:*** *Lower and upper 95% confidence interval of estimated SNP HR associated with AOO of T2D;* ***P-value:*** *P-value associated with each SNP adjusted by covariates age, sex, and ancestry.*

C.6: Supporting tables with further results for BMI association with AOO of T2D and T2D status

Table C.6. 1 - Estimated effect of association of BMI and AOO of T2D in European, Asian, and African ancestry populations

Analysis Method	Weighted GRS				Unweighted GRS			
	ES	Lower 95% CI	Upper 95% CI	P-value	ES	Lower 95% CI	Upper 95% CI	P-value
European ancestry population								
Cox PH model (cases only)								
Adjusted *(BMI+ GRS+Covariates)*	1.026	1.023	1.029	1.5×10^{-65}	1.025	1.022	1.028	2.3×10^{-62}
Cox PH model (cases and controls)								
Adjusted *(BMI+ GRS+Covariates)*	1.159	1.157	1.162	9.7×10^{-3519}	1.156	1.153	1.158	3.5×10^{-3396}
Binary logistic regression model								
Adjusted *(BMI+ GRS+Covariates)*	1.172	1.168	1.175	3.7×10^{-2463}	1.167	1.164	1.171	2.4×10^{-2395}
Asian ancestry population								
Cox PH model (cases only)								
Adjusted *(BMI+ GRS+Covariates)*	1.028	1.015	1.040	1.4×10^{-05}	1.027	1.014	1.039	2.7×10^{-05}
Cox PH model (cases and controls)								
Adjusted *(BMI+ GRS+Covariates)*	1.097	1.086	1.110	2.6×10^{-62}	1.097	1.085	1.109	3.0×10^{-60}
Binary logistic regression model								
Adjusted *(BMI+ GRS+Covariates)*	1.105	1.090	1.121	6.1×10^{-01}	1.102	1.087	1.118	5.0×10^{-01}
African ancestry population								
Cox PH model (cases only)								
Adjusted *(BMI+ GRS+Covariates)*	1.029	1.014	1.044	1.8×10^{-04}	1.029	1.014	1.044	1.8×10^{-04}
Cox PH model (cases and controls)								
Adjusted *(BMI+ GRS+Covariates)*	1.079	1.065	1.094	2.3×10^{-29}	1.079	1.065	1.094	7.6×10^{-29}
Binary logistic regression model								
Adjusted *(BMI+ GRS+Covariates)*	1.086	1.071	1.102	4.0×10^{-01}	1.086	1.070	1.102	3.8×10^{-01}

Descriptions: *GRS: genetic risk score; ES: Effect Size (hazard ratio or odds ratio); CI: confidence interval; Covariates: Models adjusted for Sex; BMI: Body Mass Index; array: genotype microarray; ancestry via PC1-PC10: Principal components.*

C.7: Supporting tables with further results for dissecting the ancestry specific T2D GRS

Table C.7. 1 - GRS SNPs excluded from LDproxy database and/or monoallelic in at least one ancestral population

#	Nearest gene	chromo-some	Base Pair Position	rs Number	Status
1	DENND2C	1	115144899	rs184660829	Monoallelic in Asian and African ancestry populations
2	FAM63A	1	151017991	rs145904381	Monoallelic in African ancestry populations
3	TMEM127	2	96913918	rs79046683	Monoallelic in European and African ancestry populations
4	DDX18	2	118071061	rs562386202	Not included in 1000G reference panel
5	MBNL1	3	152086533	rs13065698	Does not match RS number at 1000G position
6	SCD5	4	83578271	rs12642790	Does not match RS number at 1000G position
7	SLCO6A1	5	101232944	rs138337556	Monoallelic in Asian ancestry populations
8	SLC25A51P1	6	67387490	rs555402748	Not included in 1000G reference panel
9	ABO	9	136149229	rs505922	Does not match RS number at 1000G position
10	KCNJ11	11	17408404	rs5213	Is not a biallelic variant
11	WDR72	15	53747228	rs528350911	Monoallelic in African ancestry populations
12	KIF2B	17	52140805	rs569511541	Monoallelic in European ancestry populations

Descriptions: ***LDproxy:*** *online database that can be used to assess the number of SNPs in pairwise LD with SNPs included in the T2D GRS;* ***Monoallelic:*** *when only one of the two gene copies (alleles) at a site or locus is actively expressed in a population;* ***Excluded from reference panel:*** *SNPs excluded from the phase 3 (version 3) 1000 genome project reference panel;* ***Biallelic variant:*** *only variant RS numbers that are bi-allelic are included in the LDproxy database.*

C.8: Supporting figures with further results for dissecting the ancestry specific T2D GRS

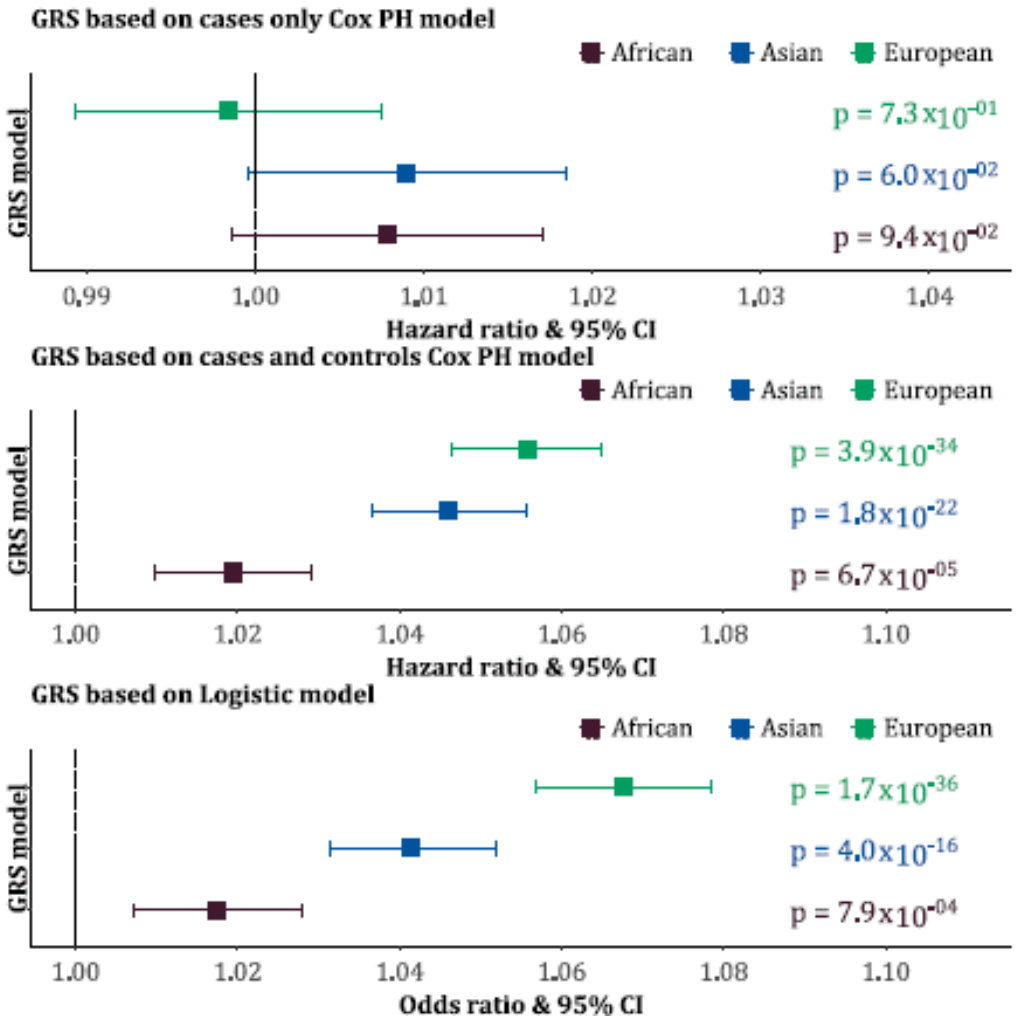


Figure C.8. 1 - Subsample comparison of estimated ES of AOO of T2D associated with the unweighted GRS for European, Asian, and African descended populations

The x-axis indicates the HR or OR and 95% CI for each GRS model shown on the y-axis. The analysis methods included are the cases only Cox PH model; cases and controls Cox PH model; and logistic model. The weighted (wei) GRS models which have been adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and BMI are denoted (purple), (blue) and (green) which refers to the African, Asian and European ancestry populations.

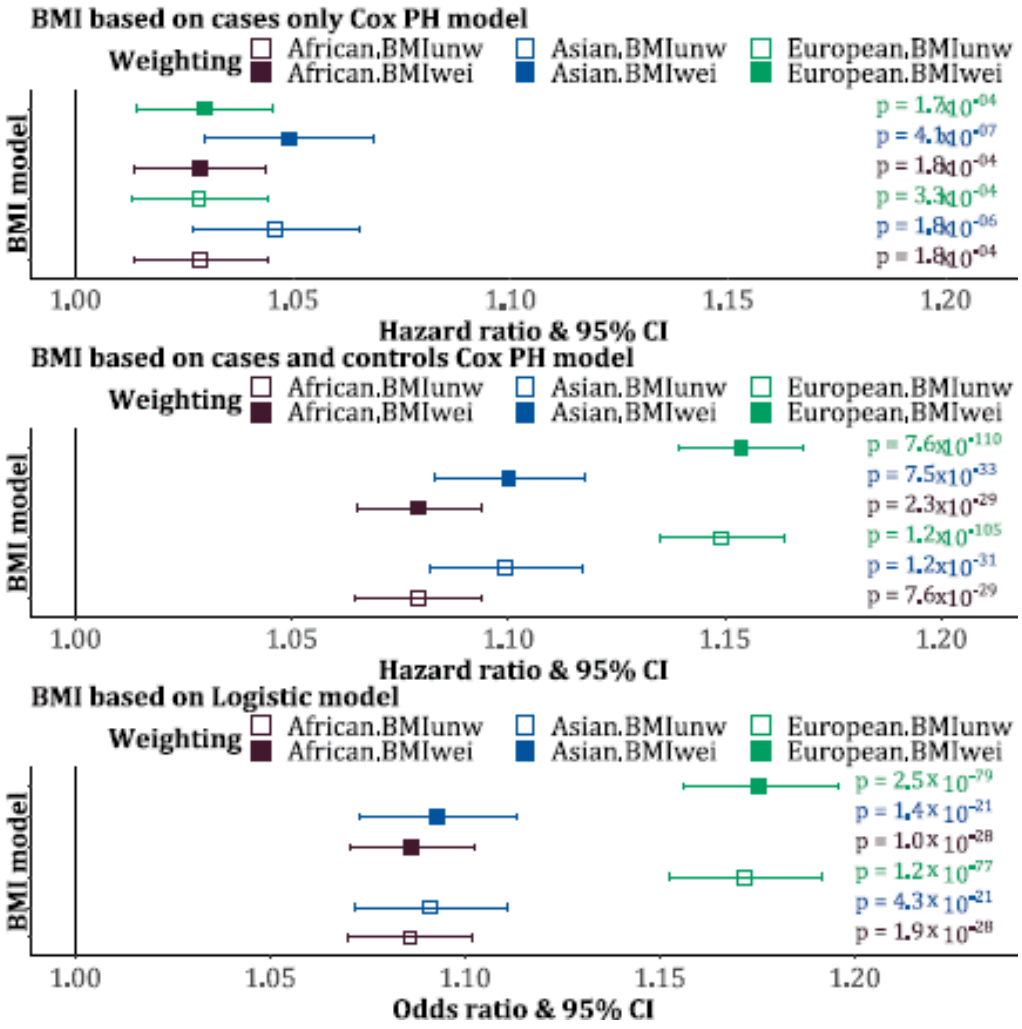


Figure C.8. 2 - Subsample comparison of estimated ES of AOO of T2D associated with BMI based on cases only Cox model for European, Asian, and African descended populations

The x-axis indicates the HR or OR and 95% CI for each BMI model shown on the y-axis. The analysis methods included are the cases only Cox PH model; cases and controls Cox PH model; and logistic model. The models have been adjusted for sex, ancestry principal component to account for population structure, type of microarray used for genotyping of SNPs and GRS. The two models considered include the adjusted model with weighted GRS denoted (BMIwei) and adjusted model with unweighted GRS denoted (BMIunw), where(purple), (blue) and (green) refers to African, Asian, and European ancestry population.

Appendix D: R syntax used for generating admixture simulation data

Table of contents

Appendix D: R syntax used for generating admixture simulation data

D.1: R syntax used to generate ancestry

```
#################################################################
###Included is the R syntax used to generate 1,000 random samples each consisting
### of 1,000 individuals assumed to originate from a population formed of discrete
###subpopulations (D.1.1).  In the case of an admixed population the syntax used
###to generate ancestry of the maternal and paternal chromosome associated with a
###tested causal SNP at a specified locus is also outlined (D.1.2 and D.1.3).
#################################################################
```

D.1.1: R syntax used to generate ancestry of maternal chromosome associated with a tested causal SNP at specified locus

```
#################################################################
### Simulating ancestry of chromosome 1 at a specified locus associated with a
### causal SNP from an admixed individual assumed to originate from two
### ancestral populations based on a given ancestry proportion
#################################################################
# Create working directories
#################################################################
#Create main directory
mainDir <- paste("C: /FOLDER ADDRESS PART 1",
          "/ FOLDER ADDRESS PART 2",
          "/Data Generation/Datasets", sep="")
#Create sub directory
subDir <- "S1Datasets"
dir.create(file.path(mainDir, subDir))
#Set working directory
setwd(file.path(mainDir, subDir))
#Get Working Directory
getwd()

#################################################################
#Specification of scenario data values - ancestry proportion
#################################################################
# Three main scenarios were considered for the ancestry proportion
# (scenario 1; o1_1 <- 0.1 (population 1)   and o2_1 <- 0.9 (population 2);
# scenario 2; o1_1 <- 0.3 (population 1)   and o2_1 <- 0.7 (population 2);
# and scenario 3; o1_1 <- 0.5 (population 1)   and o2_1 <- 0.5 (population 2);
```

```
#####################################################################
o1_1 <-  0.1  #Ancestry proportion for population 1
o2_1 <-  0.9  #Ancestry proportion for population 2
#####################################################################
# Create empty dataframes for simulated samples
#####################################################################
Population <- 0
reps = 1000 # Number of datasets to be generated in each condition
N_list = c(1000) # Population size
for (N in N_list) {
 for (i in 1:reps) {
  for (j in length(Population)) {
   Population <- matrix(data=5,nrow = N, ncol = 1)
   colnames(Population) <-  "ID"
   file=as.character(paste("Sam",N,"_",i,".csv", sep=" "))
   write.table(Population,file,row.names = FALSE)
    }
 }
}

#####################################################################
# Save path to folder that holds multiple .csv files
#####################################################################################

folder <- paste("C:/FOLDER ADDRESS PART 1",
          "/ FOLDER ADDRESS PART 2/Data Generation",
          "/Datasets/S1Datasets/",sep="")
#####################################################################
# Get list of file names in from directory folder containing files that
#will be used to create list
#####################################################################
```

```
library("tools")
myAncCHMNames <- list.files(path=folder, pattern="^Sam(.*)csv$")
# read in each  file named in the list and create list of dataframes or #files
myAncCHMlist <- lapply(myAncCHMNames, read.table, header = TRUE, sep = "")
# Read in each file in the list of files and save as R dataframe
 myAncCHMNames <- file_path_sans_ext(myAncCHMNames, compression = FALSE)
 names(myAncCHMlist) <- myAncCHMNames
 names(myAncCHMlist)
 lapply(names(myAncCHMlist), function(i) {
  assign(i, myAncCHMlist[[i]])
  save(list=i, file=paste0(i, ".Rdata"))
 })
 #Read in each file in the list of files and save as R dataframe
 #and load files into the  R environment
 names(myAncCHMlist) <- myAncCHMNames
 names(myAncCHMlist)
  lapply(names(myAncCHMlist), function(i) {
  assign(i, myAncCHMlist[[i]], envir= .GlobalEnv)
  save(list=i, file=paste0(i, ".Rdata"))
 })

#####################################################################
#To create add ancestry of chromosome 1 to simulated samples function
#####################################################################
library("abind")
 myFunCHROM1ProG <- function(x, z=myAncNames,k=o2_1,...) {
  lapply(x, function(x){
    lapply(k, function(k){
      x <- array(rbinom(n= sum(!is.na(x)), size = 1, prob=k),
   dim = c(sum(!is.na(x)), 1))
       return(data.frame(x))
  })
 })
}

#####################################################################
#To run add ancestry of chromosome 1 to simulated samples function after
#creating it and to save results
set.seed(4101)
NewAncC1 <- myFunCHROM1ProG(myAncCHMlist)
#####################################################################
myAncCHMNames <- file_path_sans_ext(myAncCHMNames,
compression = FALSE)
names(NewAncC1) <- names(mget(myAncCHMNames))
names(NewAncC1)

#####################################################################
# To save files in a file list as csv files after making changes to data frames
# Extracting column one scenario one
```

```
library("writexl")
lapply(names(NewAncC1), function(d) {
 write_xlsx(NewAncC1[[d]][[1]],
        path=paste0("CH1", d, ".xlsx",sep= " "),
        col_names = TRUE)
})

#####################################################################
# To save files in a file list as R files after making changes to data frames
# Extracting column one scenario one
lapply(names(NewAncC1), function(i,k=o2_1,...) {
 assign(i, NewAncC1[[i]] [1])
 save(list=i, file=paste0("CH1",i,".Rdata"))
})

#####################################################################
#Save path to folder that holds multiple .csv files
folder <- paste("C: /FOLDER ADDRESS PART 1",
          "/ FOLDER ADDRESS PART 2/Data Generation",
          "/Datasets/S1Datasets/",sep="")
#Get list of file names in from directory folder containing files that will be
#used to create list
myAncCHM1Names <- list.files(path=folder, pattern="^CH1(.*)xlsx$")
library("readxl")
#Read in each  file named in the list and create list of dataframes or files
myAncCHM1list <- lapply(myAncCHM1Names, read_excel, col_names = TRUE)
#Read in each file in the list of files and save as R dataframe
#and load files into the  R environment
myAncCHM1Names <- file_path_sans_ext(myAncCHM1Names, compression = FALSE)
names(myAncCHM1list) <- myAncCHM1Names
names(myAncCHM1list)
lapply(names(myAncCHM1list), function(i) {
 assign(i, myAncCHM1list[[i]], envir= .GlobalEnv)
 save(list=i, file=paste0(i, ".Rdata"))
})
#####################################################################
```

D.1.2: R syntax used to generate ancestry of paternal chromosome associated with a tested causal SNP at specified locus

```
#################################################################
### Simulating ancestry of chromosome 2 at a specified locus associated with a
### causal SNP from an admixed individual assumed to originate from two
### ancestral populations based on a given ancestry proportion
#################################################################
# Create working directories
#################################################################
#Create main directory
mainDir <- paste("C: /FOLDER ADDRESS PART 1",
          "/ FOLDER ADDRESS PART 2",
          "/Data Generation/Datasets", sep="")
#Create sub directory
subDir <- "S1Datasets"
dir.create(file.path(mainDir, subDir))
setwd(file.path(mainDir, subDir))
# get Working Directory
getwd()
#################################################################
# Save path to folder that holds the simulated csv files
folder <- paste("C: /FOLDER ADDRESS PART 1",
          "/ FOLDER ADDRESS PART 2",
          "/Datasets/S1Datasets/",sep="")

# Get list of file names in from directory folder containing files that
# will be used to create list
library("tools")
myAncCHM1Names <- list.files(path=folder, pattern="^CH1(.*)xlsx$")
library("readxl")
#Read in each  file named in the list and create list of dataframes or files
myAncCHM1list <- lapply(myAncCHM1Names, read_excel, col_names = TRUE)

# Read in each file in the list of files and save as R dataframe
#library("tools")
myAncCHM1Names <- file_path_sans_ext(myAncCHM1Names, compression = FALSE)
names(myAncCHM1list) <- myAncCHM1Names
names(myAncCHM1list)
lapply(names(myAncCHM1list), function(i) {
 assign(i, myAncCHM1list[[i]])
 save(list=i, file=paste0(i, ".Rdata"))
})
```

```
#####################################################################
#Specification of scenario data values – ancestry proportions
#####################################################################
# Three main scenarios were considered for the ancestry proportion
# (scenario 1; o1_1 <- 0.1 (population 1)   and o2_1 <- 0.9 (population 2);
# scenario 2; o1_1 <- 0.3 (population 1)   and o2_1 <- 0.7 (population 2);
# and scenario 3; o1_1 <- 0.5 (population 1)   and o2_1 <- 0.5 (population 2);
#####################################################################
o1_1 <-  0.1  #Ancestry proportions for population 1
o2_1 <-  0.9  #Ancestry proportions for population 2

#####################################################################
#To create add ancestry of chromosome 2 to simulated samples function
#####################################################################
library("abind")
#####################################################################
myFunCHROM2ProG <- function(x, z,CHROM2, k=o2_1,...) {
lapply(names(x), function(i) {
  CHROM2  <- 8
  z <- 4
  CHROM2  <- array(rbinom(n= sum(!is.na(x[[i]][1])), size = 1, prob=k), dim =
c(sum(!is.na(x[[i]][1])), 1))
  addCHROM2 <- CHROM2
  addCHROM1 <- as.numeric(x[[i]][[1]])
  z <- abind(addCHROM1, addCHROM2=addCHROM2, along=2)
   colnames(z)[1] <- "CHROM1"
   colnames(z)[2] <- "CHROM2"
  return(data.frame(z))
  })
}

#####################################################################
#To run add ancestry of chromosome 2 to simulated samples function after
#creating it and to save results
set.seed(2391)
NewAncC2 <- myFunCHROM2ProG(myAncCHM1list)
myAncCHM1Names <- file_path_sans_ext(myAncCHM1Names, compression = FALSE)
names(NewAncC2) <- names(mget(myAncCHM1Names))
names(NewAncC2)

#####################################################################
# To save files in a file list as csv files after making changes to data frames
# Extracting column one scenario one
library("writexl")
```

```
lapply(names(NewAncC2), function(d) {
 write_xlsx(NewAncC2[[d]],
        path=paste0("CH2", d, ".xlsx",sep= " "),
        col_names = TRUE)
})

#####################################################################
# To save files in a file list as R files after making changes to data frames
# Extracting column one scenario one
lapply(names(NewAncC2), function(i,k=o2_1,...) {
 assign(i, NewAncC2[[i]])
 save(list=i, file=paste0("CH2",i,".Rdata"))
})

#####################################################################
#Save path to folder that holds multiple .csv files
folder <- paste("C: /FOLDER ADDRESS PART 1",
          "/ FOLDER ADDRESS PART 2",
          "/Datasets/S1Datasets/",sep="")
#Get list of file names in from directory folder containing files that will be
#used to create list
myAncCHM2Names <- list.files(path=folder, pattern="^CH2(.*)xlsx$")
library("readxl")
#Read in each  file named in the list and create list of dataframes or files
myAncCHM2list <- lapply(myAncCHM2Names, read_excel, col_names = TRUE)
#Read in each file in the list of files and save as R dataframe
#and load files into the  R environment
myAncCHM2Names <- file_path_sans_ext(myAncCHM2Names, compression = FALSE)
names(myAncCHM2list) <- myAncCHM2Names
names(myAncCHM2list)
lapply(names(myAncCHM2list), function(i) {
 assign(i, myAncCHM2list[[i]], envir= .GlobalEnv)
 save(list=i, file=paste0(i, ".Rdata"))
})

#####################################################################
```

D.1.3: Setting used for detailed assessment of ancestry proportion

```
#################################################################
#Specification of scenario data values - ancestry proportion
#################################################################
o1_1 <-  c(0.1,0.15,0.2,0.25,0.3,0.35,0.4,0.45,0.5)  #Ancestry proportions for population 1
o2_1 <-  c(0.9,0.85,0.8,0.75,0.7,0.65,0.6,0.55,0.5)  #Ancestry proportions for population 2

#################################################################
# Create empty dataframes for simulated samples
#################################################################
Population <- 0
reps = 1000 # Number of datasets to be generated in each condition
N_list = c(1000) # Population size
for (N in N_list) {
 for (i in 1:reps) {
  for (j in length(Population)) {
   Population <- matrix(data=5,nrow = N, ncol = 1)
   colnames(Population) <-  "ID"
   file=as.character(paste("Sample",N,"_",i,".csv", sep=" "))
   write.table(Population,file,row.names = FALSE)

  }
 }
}

#################################################################
# Save path to folder that holds the .csv files
folder <- paste("C: /FOLDER ADDRESS PART 1",
"/ FOLDER ADDRESS PART 2",
"/Datasets/S1Datasets/",sep="")
# Get list of file names in from directory folder containing files that will be
# used to create list
library("tools")
myAncNames <- list.files(path=folder, pattern="^Sample")
# read in each file named in the list and create list of dataframes or files
 myAnclist <- lapply(myAncNames, read.table, header = TRUE, sep = "")
# Read in each file in the list of files and save as R dataframe
 myAncNames <- file_path_sans_ext(myAncNames, compression = FALSE)
 names(myAnclist) <- myAncNames
 names(myAnclist)
 lapply(names(myAnclist), function(i) {
  assign(i, myAnclist[[i]])
  save(list=i, file=paste0(i, ".Rdata"))
 })
```

```
#Read in each file in the list of files and save as R dataframe
#and load files into the  R environment
names(myAnclist) <- myAncNames
names(myAnclist)
 lapply(names(myAnclist), function(i) {
  assign(i, myAnclist[[i]], envir= .GlobalEnv)
  save(list=i, file=paste0(i, ".Rdata"))
 })

#####################################################################
# To create function to add ancestry to simulated samples
#####################################################################
library("abind")

 myFunCHROM1ProG <- function(x, z=myAncNames,k=o2_1,...) {
   lapply(x, function(x){
     lapply(k, function(k){
    z <- 5

    addCHROM1 <- array(rbinom(n= sum(!is.na(x)), size = 1, prob=k), dim = c(sum(!is.na(x)),
1))
    addCHROM2 <- array(rbinom(n= sum(!is.na(x)), size = 1, prob=k), dim = c(sum(!is.na(x)),
1))
    z <- abind(CHROM1=addCHROM1,  CHROM2=addCHROM2, along=2)
    return(data.frame(z))
   })
 })
}

#####################################################################
#To run add ancestry of chromosomes to simulated samples function after
#creating it and to save results
set.seed(8897)
NewAncC1 <- myFunCHROM1ProG(myAnclist)

myAncNames <- file_path_sans_ext(myAncNames, compression = FALSE)
names(NewAncC1) <- names(mget(myAncNames))
names(NewAncC1)

#####################################################################
# To save files in a file list as csv files after making changes to data frames
# Extracting column one scenario one
library("writexl")
lapply(names(NewAncC1), function(d) {
 write_xlsx(NewAncC1[[d]][[1]],
         path=paste0("A1CHM", d, ".xlsx",sep= " "),
         col_names = TRUE)
```

```
})
lapply(names(NewAncC1), function(d) {
  write_xlsx(NewAncC1[[d]][[2]],
          path=paste0("A2CHM", d, ".xlsx",sep= " "),
          col_names = TRUE)
})
```

D.2: R syntax used to generate allele and genotype of causal SNP

```
#################################################################
###The main excerpt of the R syntax used to simulate the genotype of the causal SNP
###for each individual in the subpopulation samples are included.  Each subpopulation
###genotypes were simulated based on population specific RAF. Included also are
###excerpts of the syntax used to simulate the allele of each chromosome based of the RAF
###of the ancestral populations.
#################################################################
```

D.2.1: R syntax used to generate allele of chromosomes associated with tested causal SNP at specified locus

```
#################################################################
#Specification of Scenario Data Values - alleles
#################################################################
q1_1 <- 0.1   #Risk allele frequency for population 1
q2_1 <- 0.5   #Risk allele frequency for population 2

p1_1 <-  c(1-q1_1)      #Second allele frequency for population 1
p2_1 <-  c(1-q2_1)      #Second allele frequency for population 2

#################################################################
#To calculate SNP frequencies for each population
#################################################################
library(plyr)

List1 <- list(A=p1_1, a=q1_1)
List1m <- do.call(cbind,  List1)
List2 <-list(A=p2_1, a=q2_1)
List2m <- do.call(cbind,  List2)

ConVList1 <- apply(List1m, 1, function(x) list(c(x[1], x[2])))
ConVList2 <- apply(List2m, 1, function(x) list(c(x[1], x[2])))
names(ConVList1) <- "rep1"
names(ConVList2) <- "rep1"

GenotypeStatus1 <- function (j,  ConVList1, ConVList2, z,...)  {
 Pro1 <- c(ConVList1[[j]][[1]][["A"]], ConVList1[[j]][[1]][["a"]])
 Pro2 <- c(ConVList2[[j]][[1]][["A"]], ConVList2[[j]][[1]][["a"]])
 return(Pro1)
}
Prob1 <- lapply(names(ConVList1), GenotypeStatus1, ConVList1=ConVList1,
ConVList2=ConVList2)
```

```
GenotypeStatus2 <- function (j,  ConVList1, ConVList2, z,...)  {
 Pro1 <- c(ConVList1[[j]][[1]][["A"]],ConVList1[[j]][[1]][["a"]])
 Pro2 <- c(ConVList2[[j]][[1]][["A"]],ConVList2[[j]][[1]][["a"]])
 return(Pro2)
}
Prob2 <- lapply(names(ConVList1), GenotypeStatus2, ConVList1=ConVList1,
ConVList2=ConVList2)

print(Prob1)
print(Prob2)
Pr1 <-Prob1[[1]]
Pr2 <-Prob2[[1]]
print(Pr1)
print(Pr2)

#####################################################################
# Add SNP column to dataframes for each population
#####################################################################
library("abind")

myFunAlleleProG <- function(x, z, g, y1=SNPs1, y2=SNPs2, Pro1=Pr1,  Pro2=Pr2,...) {
 lapply(names(x), function(i) {
  ALLELE1 <- 4
  ALLELE2 <- 5
  ALE1 <- 8
  ALE2 <- 6
  z <- 4
  ALE1 <- array(sample(x=0:1, size= sum(!is.na(x[[i]][1])), replace=T, prob=Pro1), dim =
c(sum(!is.na(x[[i]][1])), 1))
  addALE1 <- ALE1

  ALE2 <- array(sample(x=0:1, size= sum(!is.na(x[[i]][1])), replace=T, prob=Pro2), dim =
c(sum(!is.na(x[[i]][1])), 1))
  addALE2 <- ALE2
  addCHROM1 <- as.numeric(x[[i]][["CHROM1"]])
  addCHROM2 <- as.numeric(x[[i]][["CHROM2"]])

  z <- abind(addCHROM1,addCHROM2, ALLELE1=ifelse(addCHROM1== 0, addALE1,
addALE2),
      ALEID1=ifelse(addCHROM1 == 0, "ALEp1", "ALEp2"),
       ALLELE2=ifelse(addCHROM2== 0, addALE1, addALE2),
      ALEID2=ifelse(addCHROM2 == 0, "ALEp1", "ALEp2"),  along=2)
```

```
  colnames(z)[1] <- "CHROM1"
  colnames(z)[2] <- "CHROM2"
  colnames(z)[3] <- "ALLELE1"
  colnames(z)[5] <- "ALLELE2"

  return(data.frame(z))

 })

}
#####################################################################
#To run add SNP function after creating it and to save results of add SNP function
set.seed(3939)
NewGeno1 <- myFunAlleleProG(myAncCHM2list)
print(NewGeno1)
#####################################################################
# To save files in a file list as csv files after adding SNP column to dataframes
#####################################################################
# To save files in a file list as csv files after making changes to data frames
# Extracting columns for scenario one
#####################################################################

myFunGenotypeProG <- function(x, v, i,...) {
 lapply(x, function(i,...) {

  addALLELE1 <- (i[["ALLELE1"]])
  addALLELE2 <- (i[["ALLELE2"]])
  v <- 4
  v <- abind(i, GENOTYPE= paste(addALLELE1,addALLELE2,sep=""), along=2)

  return(data.frame(v))
 })
}

#####################################################################
NewGeno2 <- myFunGenotypeProG(NewGeno1)
print(NewGeno2)

#####################################################################
myFunRecodeSNP <- function(x, z, i,...) {

 lapply(x, function(i,...) {

  i[["GENOTYPE_F"]] <- factor(i[["GENOTYPE"]],
                levels=c("00","01","10","11"),
                labels=c("AA","Aa","aA" ,"aa"))
```

```
  i[["SNPs"]] <- revalue(i[["GENOTYPE"]],
                  c("00"="0","01"="1","10"="1","11"="2"))

  i[["SNPs_F"]] <- factor(i[["SNPs"]],
                      levels=c(0,1,2),
                      labels=c("AA","Aa","aa"))
  return(data.frame(i))
 })
}

NewGeno3 <- myFunRecodeSNP(NewGeno2)
print(NewGeno3)

#####################################################################
myFunCHROMANC <- function(x, z, i,...) {
lapply(x, function(i,...) {
addSample <- i
checkCHROM1 <- (i[["CHROM1"]])
checkCHROM2 <- (i[["CHROM2"]])

CHROMANC <- ifelse(checkCHROM1== 0 & checkCHROM2 == 0,
            CHROMANC <- 2,
            ifelse(checkCHROM1== 1 & checkCHROM2 == 1,
                CHROMANC <- 0,
                ifelse(checkCHROM1== 0 & checkCHROM2 == 1, CHROMANC <- 1, CHROMANC
<- 1)
                    )
                )

z <- abind(addSample, CHROMANC=CHROMANC, along=2)

return(data.frame(z))
})
}
NewGeno4 <- myFunCHROMANC(NewGeno3)
print(NewGeno4)

#####################################################################
```

D.2.2: Setting used for detailed assessment of ancestry specific RAF

To assess further the impact of population specific RAF on power nine different scenarios were considered. The first scenario (shown below) allowed the RAF to vary in population 1 while RAF was held fixed at 0.1 in population 2. In the second scenario the RAF was held fixed at 0.2 in the second population. A similar process was followed through to scenario nine where RAF in the first population was varied but held fixed at 0.9 in the second population.

```
#################################################################
#Specification of scenario data values - RAF
#################################################################

q1_1 <- c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)  #Risk allele frequency for population 1
q2_1 <- c(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1)  #Risk allele frequency for population 2
p1_1 <-  c(1-q1_1)      #Second allele frequency for population 1
p2_1 <-  c(1-q2_1)      #Second allele frequency for population 2

#################################################################
#Specification of scenario data values - RAF
#################################################################
q1_1 <- c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)  #Risk allele frequency for population 1
q2_1 <- c(0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2)  #Risk allele frequency for population 2
p1_1 <-  c(1-q1_1)      #Second allele frequency for population 1
p2_1 <-  c(1-q2_1)      #Second allele frequency for population 2
#################################################################
```

D.3: R syntax used to generate allele and genotype of tag SNP

```
#################################################################
###The main excerpt of the R syntax used to simulate the genotype of  a tag SNP
###for each individual in the subpopulation samples are included.  Each subpopulation
###genotypes were simulated based on population specific RAF. Included also are
###excerpts of the syntax used to simulate the allele of maternal and paternal
###chromosome based of the RAF of the ancestral populations.
#################################################################
```

D.3.1: R syntax used to generate allele of chromosomes associated with tested tag SNP at specified locus

```
#################################################################
#Specification of scenario data values used to simulate alleles of tag SNP
#################################################################
LD <- c(0, 0.05,0.15,0.25,0.5,0.75,0.85,0.95,1)
q1_1 <- 0.1   #P(a) Risk allele frequency for population 1 Casual SNP
q2_1 <- 0.5   #P(a) Risk allele frequency for population 2 Casual SNP
g1_1 <- 0.1   #P(b) Risk allele frequency for population 1 Tag SNP
g2_1 <- 0.5   #P(b) Risk allele frequency for population 2 Tag SNP
p1_1 <-  c(1-q1_1)      #P(A) Second allele frequency for population 1 Casual SNP
p2_1 <-  c(1-q2_1)      #P(A)  Second allele frequency for population 2 Casual SNP
t1_1 <-  c(1-g1_1)      #P(B)Second allele frequency for population 1 Tag SNP
t2_1 <-  c(1-g2_1)      #P(B)Second allele frequency for population 2 Tag SNP

#haplotype frequencies for two loci with two alleles
#assuming linkage disequilibrium
#P(BA) haplotype frequencies for x11 Population 1
A1B1_1d <- (((p1_1)*(t1_1)) + LD*(sqrt(t1_1*g1_1*p1_1*q1_1)))
#P(Ba) haplotype frequencies for x12 Population 1
A1b2_1d <- p1_1 - A1B1_1d
#P(bA) haplotype frequencies for x21 Population 1
a2B1_1d <- t1_1 - A1B1_1d
#P(ba) haplotype frequencies for x22 Population 1
a2b2_1d <- (1 - (A1B1_1d +  A1b2_1d + a2B1_1d))

#haplotype frequencies for two loci with two alleles
# assuming linkage disequilibrium
#P(BA)   haplotype frequencies for x11 Population 2
A1B1_2d <- (((p2_1)*(t2_1)) + LD*(sqrt(t2_1*g2_1*p2_1*q2_1)))
#P(Ba)  haplotype frequencies for x12 Population 2
A1b2_2d <- p2_1 - A1B1_2d
#P(bA)  haplotype frequencies for x21 Population 2
a2B1_2d <- t2_1 - A1B1_2d
```

```
#P(ba)  haplotype frequencies for x22 Population 2
a2b2_2d <- (1 - (A1B1_2d +  A1b2_2d + a2B1_2d))

#Probability of seeing a Tag SNP allele given the casual allele
B_A1 <- A1B1_1d/p1_1
b_A1 <- A1b2_1d/p1_1
B_a1 <- a2B1_1d/q1_1
b_a1 <- a2b2_1d/q1_1

#Probability of seeing a Tag SNP allele given the casual allele
B_A2 <- A1B1_2d/p2_1
b_A2 <- A1b2_2d/p2_1
B_a2 <- a2B1_2d/q2_1
b_a2 <- a2b2_2d/q2_1
#################################################################
#To calculate Allele frequencies for each population
#################################################################
library(plyr)

List1 <- list(A=B_A1, a=b_A1)
List1m <- do.call(cbind,  List1)
List2 <-list(A=B_a1, a=b_a1)
List2m <- do.call(cbind,  List2)
List3 <- list(A=B_A2, a=b_A2)
List3m <- do.call(cbind,  List3)
List4 <-list(A=B_a2, a=b_a2)
List4m <- do.call(cbind,  List4)

ConVList1 <- apply(List1m, 1, function(x) list(c(x[1], x[2])))
ConVList2 <- apply(List2m, 1, function(x) list(c(x[1], x[2])))
ConVList3 <- apply(List3m, 1, function(x) list(c(x[1], x[2])))
ConVList4 <- apply(List4m, 1, function(x) list(c(x[1], x[2])))

names(ConVList1) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")
names(ConVList2) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")
names(ConVList3) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")
names(ConVList4) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")

GenotypeStatus1 <- function (j,  ConVList1, ConVList2, ConVList3, ConVList4, z,...)  {
  Pro1A <- c(ConVList1[[j]][[1]][["A"]], ConVList1[[j]][[1]][["a"]])
  Pro1a <- c(ConVList2[[j]][[1]][["A"]], ConVList2[[j]][[1]][["a"]])
  Pro2A <- c(ConVList3[[j]][[1]][["A"]], ConVList3[[j]][[1]][["a"]])
  Pro2a <- c(ConVList4[[j]][[1]][["A"]], ConVList4[[j]][[1]][["a"]])
```

```
 return(Pro1A)
}
Prob1A  <- lapply(names(ConVList1), GenotypeStatus1, ConVList1=ConVList1,
ConVList2=ConVList2,
            ConVList3=ConVList3, ConVList4=ConVList4)
#repeated to abstract (Pro1a)
GenotypeStatus2 <- function (j,  ConVList1, ConVList2, ConVList3, ConVList4, z,...)  {
 Pro1A <- c(ConVList1[[j]][[1]][["A"]], ConVList1[[j]][[1]][["a"]])
 Pro1a <- c(ConVList2[[j]][[1]][["A"]], ConVList2[[j]][[1]][["a"]])
 Pro2A <- c(ConVList3[[j]][[1]][["A"]], ConVList3[[j]][[1]][["a"]])
 Pro2a <- c(ConVList4[[j]][[1]][["A"]], ConVList4[[j]][[1]][["a"]])
 return(Pro1A)
}
Prob1a  <- lapply(names(ConVList1), GenotypeStatus2, ConVList1=ConVList1,
ConVList2=ConVList2,
            ConVList3=ConVList3, ConVList4=ConVList4)

#repeated to abstract (Pro2A)
GenotypeStatus3 <- function (j,  ConVList1, ConVList2, ConVList3, ConVList4, z,...)  {
 Pro1A <- c(ConVList1[[j]][[1]][["A"]], ConVList1[[j]][[1]][["a"]])
 Pro1a <- c(ConVList2[[j]][[1]][["A"]], ConVList2[[j]][[1]][["a"]])
 Pro2A <- c(ConVList3[[j]][[1]][["A"]], ConVList3[[j]][[1]][["a"]])
 Pro2a <- c(ConVList4[[j]][[1]][["A"]], ConVList4[[j]][[1]][["a"]])
 return(Pro2A)
}
Prob2A  <- lapply(names(ConVList1), GenotypeStatus3, ConVList1=ConVList1,
ConVList2=ConVList2,
            ConVList3=ConVList3, ConVList4=ConVList4)
#repeated to abstract (Pro2a)
GenotypeStatus4 <- function (j,  ConVList1, ConVList2, ConVList3, ConVList4, z,...)  {
 Pro1A <- c(ConVList1[[j]][[1]][["A"]], ConVList1[[j]][[1]][["a"]])
 Pro1a <- c(ConVList2[[j]][[1]][["A"]], ConVList2[[j]][[1]][["a"]])
 Pro2A <- c(ConVList3[[j]][[1]][["A"]], ConVList3[[j]][[1]][["a"]])
 Pro2a <- c(ConVList4[[j]][[1]][["A"]], ConVList4[[j]][[1]][["a"]])
 return(Pro2a)
}
Prob2a  <- lapply(names(ConVList1), GenotypeStatus4, ConVList1=ConVList1,
ConVList2=ConVList2,
            ConVList3=ConVList3, ConVList4=ConVList4)
#################################################################
print(Prob1A)
print(Prob1a)
print(Prob1A)
```

```
print(Prob2a)
Pr1 <-Prob1A
Pr2 <-Prob1a
Pr3 <-Prob2A
Pr4 <-Prob2a
print(Pr1)
print(Pr2)
print(Pr3)
print(Pr4)
names(Pr1) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")
names(Pr2) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")
names(Pr3) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")
names(Pr4) <- c("rep1","rep2","rep3","rep4","rep5","rep6","rep7","rep8","rep9")

#####################################################################
# To create add genotype of SNP column to dataframes for each population funcion
#####################################################################
library("abind")
#####################################################################
myFunTagProG1 <- function(x, z, y1=ALLEP1A, y2=ALLEP1a, y3=ALLEP2A,  y4=ALLEP2a,
                rr=Pr1, ss=Pr2, tt=Pr3,uu=Pr4,...) {

 lapply(names(x), function(i) {
  addSample <- (x[[i]])
  checkCHR1  <- as.numeric(x[[i]][["CHROM1"]])
  checkCHR2  <- as.numeric(x[[i]][["CHROM1"]])
  checkALLELE1 <- as.numeric(x[[i]][["ALLELE1"]])
  checkALLELE2 <- as.numeric(x[[i]][["ALLELE2"]])
   ######
  TagALLEB1 <- 8
  TagALLEB2 <- 8
  ALLEP1A <- 8
  ALLEP1a <- 8
  ALLEP2A <- 8
  ALLEP2a <- 8
   z <- 9
   ALLEP1A <- array(sample(x=0:1, size= sum(!is.na(x[[i]][1])), replace=T,
                    prob=rr[[1]]), dim = c(sum(!is.na(x[[i]][1])), 1))
     addALLEP1A <- ALLEP1A
  ALLEP1a <- array(sample(x=0:1, size= sum(!is.na(x[[i]][1])), replace=T,
                    prob=ss[[1]]), dim = c(sum(!is.na(x[[i]][1])), 1))
     addALLEP1a <- ALLEP1a
  ALLEP2A <- array(sample(x=0:1, size= sum(!is.na(x[[i]][1])), replace=T,
                    prob=tt[[1]]), dim = c(sum(!is.na(x[[i]][1])), 1))
```

```
    addALLEP2A <- ALLEP2A
  ALLEP2a <- array(sample(x=0:1, size= sum(!is.na(x[[i]][1])), replace=T,
                  prob=uu[[1]]), dim = c(sum(!is.na(x[[i]][1])), 1))
    addALLEP2a <- ALLEP2a


#################################################################
   TagALLEB1 <- ifelse(checkCHR1== 0 & checkALLELE1 == 0,
               TagALLEB1 <- addALLEP1A,   # add Tag SNP genotype BB for population 1 Tag
SNPsB <- "BB",
               ifelse(checkCHR1== 1 & checkALLELE1 == 0,
                    TagALLEB1 <- addALLEP2A,  # add Tag SNP genotype BB for population 2
                      ifelse(checkCHR1== 0 & checkALLELE1 == 1,TagALLEB1 <-
addALLEP1a,TagALLEB1 <- addALLEP2a)
                    )
               )

#################################################################
    TagALLEB2 <- ifelse(checkCHR2== 0 & checkALLELE2 == 0,
                TagALLEB2 <- addALLEP1A,   # add Tag SNP genotype BB for population 1 Tag
SNPsB <- "BB",
                ifelse(checkCHR2== 1 & checkALLELE2 == 0,
                     TagALLEB2 <- addALLEP2A,  # add Tag SNP genotype BB for population 2
                      ifelse(checkCHR2== 0 & checkALLELE2 == 1,TagALLEB2 <-
addALLEP1a,TagALLEB2 <- addALLEP2a)
                    )
               )
  z <- abind(addSample, TagALLE1=TagALLEB1, TagALLE2=TagALLEB2, along=2)
  return(data.frame(z))
   })
}
#################################################################
#To run add SNP function after creating it and to save results of add SNP function
set.seed(7787)
NewTag SNP1 <- myFunTagProG1(myTaglist)
print(NewTag SNP1)


#################################################################
#To create function used to add labels to tag SNP genotype
#################################################################
myFunGenotypeProG1 <- function(x, v, i,...) {
 lapply(x, function(i,...) {
```

```
  addALLELEB1 <- (i[["TagALLE1"]])
  addALLELEB2 <- (i[["TagALLE2"]])
  v <- 4
  v <- abind(i, GENOTYPEb= paste(addALLELEB1,addALLELEB2,sep=""), along=2)

  return(data.frame(v))
 })
}

#####################################################################
```
 # To run add labels to tag SNP genotype function and save results
```
NewTag SNP1_2 <- myFunGenotypeProG1(NewTag SNP1)
print(NewTag SNP1_2)

#####################################################################
```
#To create function used to recode tag SNP
```
#####################################################################
myFunRecodeSNP1 <- function(x, z, i,...) {

 lapply(x, function(i,...) {
    i[["GENOTYPE_Fb"]] <- factor(i[["GENOTYPEb"]],
                  levels=c("00","01","10","11"),
                  labels=c("AA","Aa","aA" ,"aa"))
  i[["Tag SNPs"]] <- revalue(i[["GENOTYPEb"]],
                 c("00"="0","01"="1","10"="1","11"="2"))
    i[["Tag SNPs_F"]] <- factor(i[["Tag SNPs"]],
                  levels=c(0,1,2),
                  labels=c("AA","Aa","aa"))
  return(data.frame(i))
 })
}

#####################################################################
```
 # To run recode tag SNP function and save results
```
NewTag SNP1_3 <- myFunRecodeSNP1(NewTag SNP1_2)
print(NewTag SNP1_3)
#####################################################################
```

D.3.2: Setting where LD is assumed different among ancestral populations

The scenarios where LD was assumed to be different among populations incorporated the scenarios where LD was held fixed in population 1while LD varied in the second population. Nine different scenarios were considered, the first scenario (shown below) assumed LD in population 1 was fixed at 0 while LD varied in population. In the second scenario LD in population was fixed at 0.05 while LD in the second population. A similar process was followed through to scenario nine where LD in the first population was fixed at 1 but varied in the second population.

```
#################################################################
#Specification of scenario data values used to simulate alleles and genotype of tag SNP
#################################################################
#scenario 1
LDr1 <- c(0,0,0,0,0,0,0,0,0)
LDr2 <- c(0, 0.05,0.15,0.25,0.5,0.75,0.85,0.95,1)
q1_1 <- 0.1   #P(a) Risk allele frequency for population 1 Casual SNP
q2_1 <- 0.5   #P(a) Risk allele frequency for population 2 Casual SNP
g1_1 <- 0.1   #P(b) Risk allele frequency for population 1 Tag SNP
g2_1 <- 0.5   #P(b) Risk allele frequency for population 2 Tag SNP
p1_1 <-  c(1-q1_1)      #P(A) Second allele frequency for population 1 Casual SNP
p2_1 <-  c(1-q2_1)      #P(A)  Second allele frequency for population 2 Casual SNP
t1_1 <-  c(1-g1_1)      #P(B)Second allele frequency for population 1 Tag SNP
t2_1 <-  c(1-g2_1)      #P(B)Second allele frequency for population 2 Tag SNP

#################################################################
#Specification of scenario data values used to simulate alleles and genotype of tag SNP
#################################################################
#scenario 2
LDr1 <- c(0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05)
LDr2 <- c(0,0.05,0.15,0.25,0.5,0.75,0.85,0.95,1)
q1_1 <- 0.1   #P(a) Risk allele frequency for population 1 Casual SNP
q2_1 <- 0.5   #P(a) Risk allele frequency for population 2 Casual SNP
g1_1 <- 0.1   #P(b) Risk allele frequency for population 1 Tag SNP
g2_1 <- 0.5   #P(b) Risk allele frequency for population 2 Tag SNP
p1_1 <-  c(1-q1_1)      #P(A) Second allele frequency for population 1 Casual SNP
p2_1 <-  c(1-q2_1)      #P(A)  Second allele frequency for population 2 Casual SNP
t1_1 <-  c(1-g1_1)      #P(B)Second allele frequency for population 1 Tag SNP
t2_1 <-  c(1-g2_1)      #P(B)Second allele frequency for population 2 Tag SNP
```

D.4: R syntax used to generate AOO of disease

```
#############################################################
###The main excerpt of the R syntax used to simulate the AOO of disease for both
### the subpopulation and admixed population simulated data are included.
###Simulated AOO is based primarily on the Cox model, however, to facilitate
###comparison, AOO based on the Weibull model is included.  The settings used
### to simulate different censoring rates is also included.
#############################################################
```

D.4.1: R syntax used to generate AOO of disease associated with a tested causal SNP within an admixed population based on the Cox model

```
#############################################################
#Specification of scenario data values for Cox models
#############################################################
#The original setting used for lambdaT <- 15 and lambdaC <- 0.000003125 was updated
#to lambdaT <- 18 and lambdaC <- 0.001 for the admixed population simulations to
#allow for 5% censoring including due to dropout
#############################################################
Study_tC <- 50              # Study time in years
lambdaT <-  18              # baseline hazard rate (ho(t))
lambdaC <- 0.001            # hazard rate of censoring
beta_G_C <- c(0,0.025,0.05,0.075,0.10,0.125,0.15,0.175)       #  log hazard ratio associated with
#genotype of causal SNP
#############################################################
# To create function to Simulate survival time (AOO) based on Cox model
#############################################################
library("abind")
myFunTTEaddModelCox <- function(x, z, T0, T1, T2, C0, C1, C2, time_OT,
                    eventStatus, b1=beta_G_C,,...) {
# to use lapply to apply x over all the different datasets in the file list of datasets applied
 lapply(names(x), function(i) {


   z <-0
   T0 <-0
   C0 <-0
   time_OT <- 0
   eventStatus <-0
# to use lapply to apply b1 over the different (beta_G_C <-
c(0,0.025,0.05,0.075,0.10,0.125,0.15,0.175) ) values
   lapply(b1, function(k) {
    # to add the SNP variable to the time to event (T) equation
    add_X1_G_C <- as.numeric(x[[i]][["SNPs"]])
    # true event time - simulating true event time based on Weibull distribution
    T0 <- rweibull((n=sum(!is.na(x[[i]][["SNPs"]]))), shape =1,
              (scale =lambdaT*exp(-k*(as.numeric(add_X1_G_C))) ))
```

```
    # to change event times over 50 to 50
    T1 <- ifelse(T0 >= 50, Study_tC, T0)
    # to change event times less than zero (negative values) to 0
    T2 <- ifelse(T0 < 0, 0, T1)

    # censoring time - simulating censoring time based on the exponential distribution
    C0 <- rexp((n=sum(!is.na(x[[i]][["SNPs"]]))), (rate =lambdaC))
    # to change censoring times over 50 to 50
    C1  <- ifelse(C0 >= 50, Study_tC, C0)
    # to change censoring times less than zero (negative values) to 0
    C2  <- ifelse(C0 < 0, 0, C1)
    # Observed time is minimum of censored and true event time
    time_OT <- pmin(T2,C2)
    # set to (1) if event is oberved
    eventStatus1 <- time_OT ==T2   # set to (1) if event is oberved
    eventStatus2 <-  ifelse(T0 > 50, "FALSE", eventStatus1)
    eventStatus3 <-  ifelse((T0 == 50 & C0 >= 50), "TRUE", eventStatus2)
    eventStatus4 <-  ifelse((T0 == C0 & T0 <= 50), "TRUE", eventStatus3)
    eventID <- ifelse(eventStatus4=="TRUE", 1, 0)
    # Use array bind to add columns T,  C, time_OT, eventStatus to datasets
    z <- abind(x[[i]], ET=T2, CT=C2, OT=time_OT, eventStatus=eventStatus4,
           eventID=eventID, along=2)
    return(data.frame(z))
  })
 })
 }

#################################################################
#To run Cox model function after creating it and save results of Cox model function
set.seed(5845)
addModelCox <- myFunTTEaddModelCox(myModellist)
```

D.4.2: R syntax used to generate AOO of disease associated with a tested causal SNP with an admixed population based on the Weibull model

```
#################################################################
# Specification of scenario data values for Weibull model
Study_tC <- 50                # Study time in years
lambdaT <-  30                # baseline hazard rate (ho(t))
lambdaC <- 0.001          # hazard rate of censoring
beta_G_C <- c(0,0.0125,0.025,0.0375,0.05,0.0625,0.075,0.0875)       #  log hazard ratio
associated with
 # genotype of causal SNP
#################################################################
# To create function to simulate survival time (AOO) based on Weibull model
#################################################################
library("abind")
myFunTTEaddModelWEI <- function(x, z, T0, T1, T2, C0, C1, C2, time_OT,
                      eventStatus, b1=beta_G_C,...) {
# to use lapply to apply x over all the different datasets in the file list of datasets applied
  lapply(names(x), function(i) {
    z <-0
   T0 <-0
   C0 <-0
   time_OT <- 0
   eventStatus <-0
# to use lapply to apply b1 over the different (beta_G_C <-
c(0,0.0125,0.025,0.0375,0.05,0.0625,0.075,0.0875) ) values
   lapply(b1, function(k) {
    # to add the SNP variable to the time to event (T) equation
    add_X1_G_C <- as.numeric(x[[i]][["SNPs"]])
    # true event time - simulating true event time based on Weibull distribution
    T0 <- rweibull((n=sum(!is.na(x[[i]][["SNPs"]]))), shape =2,
              (scale =lambdaT*exp(-k*(as.numeric(add_X1_G_C))) ))
    # to change event times over 50 to 50
    T1 <- ifelse(T0 >= 50, Study_tC, T0)
    # to change event times less than zero (negative values) to 0
    T2 <- ifelse(T0 < 0, 0, T1)
    # censoring time - simulating censoring time based on the exponential distribution
    C0 <- rexp((n=sum(!is.na(x[[i]][["SNPs"]]))), (rate =lambdaC))
    # to change censoring times over 50 to 50
    C1  <- ifelse(C0 >= 50, Study_tC, C0)
    # to change censoring times less than zero (negative values) to 0
    C2  <- ifelse(C0 < 0, 0, C1)
    # Observed time is minimum of censored and true event time
    time_OT <- pmin(T2,C2)
    # set to (1) if event is oberved
    eventStatus1 <- time_OT ==T2   # set to (1) if event is oberved
    eventStatus2 <-  ifelse(T0 > 50, "FALSE", eventStatus1)
```

```
    eventStatus3 <-  ifelse((T0 == 50 & C0 >= 50), "TRUE", eventStatus2)
    eventStatus4 <-  ifelse((T0 == C0 & T0 <= 50), "TRUE", eventStatus3)
    eventID <- ifelse(eventStatus4=="TRUE", 1, 0)
    # Use array bind to add columns T,  C, time_OT, eventStatus to datasets
    z <- abind(x[[i]], ET=T2, CT=C2, OT=time_OT, eventStatus=eventStatus4,
           eventID=eventID, along=2)
    return(data.frame(z))
  })
 })
 }

#####################################################################
#To run Weibull model function after creating it and save results of Weibull
#model function
set.seed(3427)
addModelWEI <- myFunTTEaddModelWEI(myModellist)
#####################################################################
```

Appendix E: R syntax used to conduct data analysis of the simulated AOO of disease data

Table of contents

Appendix E: R syntax used to conduct data analysis of the simulated AOO of disease data

E.1: R syntax used to undertake Cox analysis in an admixed population

E.1.1: R syntax used to undertake Cox analysis of simulated data in an admixed population

```
#####################################################################
#Data analysis based on Cox model
#####################################################################
library(survival)
args(coxph)

# Function to run Cox PH model based on Additive SNPs variable
myResultsModelCoxS <- function(x,...) {

 lapply(names(x), function(i) {

   snptemp <- as.numeric(x[[i]][["SNPs"]])
   timetemp <- as.numeric(x[[i]][["OT"]])
   eventID  <-  as.numeric(x[[i]][["eventID"]])

   CphModel_M2 <- coxph(Surv(timetemp, eventID)~ snptemp)

  sumCoxM2 <- summary(CphModel_M2)
  print(sumCoxM2)
   return(sumCoxM2)

 })
}

#####################################################################
#Function to abstract and store model results
#####################################################################
library("abind")
mySaveModelCoxS <- function(x,...) {
 lapply(names(x), function(i) {

   coefv<- coef(x[[i]])[1,1]

   #to extract hazard ratio
   expv<- coef(x[[i]])[1,2]
```

```
  SEv<- coef(x[[i]])[1,3]
  Zv <- coef(x[[i]])[1,4]
  #to extract p.value
  Pv <- coef(x[[i]])[1,5]
  ID <- length(x[i])

  # to combine dataset results into one dataframe(or array) by row
  w <- abind(ID=ID, coef=coefv, HR=expv, SE=SEv, Z_test=Zv, P_value=Pv, along=2)
  return(data.frame(w))
 })
}

#####################################################################
#To run Cox model function after creating it and save results of Cox model function
ResCoxB1S <-mySaveModelCoxS(ResultsCoxB1S)
ResCoxB1S_DF <- do.call(rbind,ResCoxB1S)
print(ResCoxB1S_DF)
ResCoxB1S_DF$ID <- seq_len(nrow(ResCoxB1S_DF))
save(ResCoxB1S_DF, file=paste("C: /FOLDER ADDRESS PART 1",
                    "/ FOLDER ADDRESS PART 2",
                    "/Output Analysis/Tables/S1Tables/ResCoxB1S_DF.Rda",sep=""))
#####################################################################
```

E.1.2: Specification for different forms of the Cox model in an admixed population

```
#Cox analysis based on genotype SNP
snptemp <- as.numeric(x[[i]][["SNPs"]])
timetemp <- as.numeric(x[[i]][["OT"]])
eventID  <-  as.numeric(x[[i]][["eventID"]])
CphModel_M2 <- coxph(Surv(timetemp, eventID)~ snptemp)

#Cox analysis based on ancestry of genotype SNP
anctemp<- as.numeric(x[[i]][["CHROMANC"]])
timetemp <- as.numeric(x[[i]][["OT"]])
eventID  <-  as.numeric(x[[i]][["eventID"]])
CphModelA_M2 <- coxph(Surv(timetemp, eventID)~ anctemp)

#Cox analysis based on genotype SNP with ancestry as covariate
anctemp<- as.numeric(x[[i]][["CHROMANC"]])
snptemp <- as.numeric(x[[i]][["SNPs"]])
timetemp <- as.numeric(x[[i]][["OT"]])
eventID  <-  as.numeric(x[[i]][["eventID"]])
CphModel_M2 <- coxph(Surv(timetemp, eventID)~ snptemp + anctemp)
```

E.2: R syntax used to undertake Weibull analysis in an admixed population

E.2.1: R syntax used to undertake Weibull analysis of simulated data in an admixed population

```
#################################################################
#Data analysis based on general Weibull model
#################################################################

library(survival)
library(eha)


# Function to run Weibull model based on Additive SNPs variable

myResultsModelWeiS <- function(x,...) {

 lapply(names(x), function(i) {
  snptemp <- as.numeric(x[[i]][["SNPs"]])
  timetemp <- as.numeric(x[[i]][["OT"]])
  eventID  <-  as.numeric(x[[i]][["eventID"]])
  WphModel_M2S <- survreg(Surv(timetemp, eventID )~ snptemp, dist="weibull")

  sumWeiM2S <- summary(WphModel_M2S)
  print(sumWeiM2S)
  return(sumWeiM2S)
 })
}


#################################################################
#Function to abstract and store model results
#################################################################
library("abind")
library("mlr")

mySaveModelWeiS <- function(x,...) {
 lapply(names(x), function(i) {
  coefv<-  x[[i]][["table"]][[2,1]]

  #to extract hazard ratio
  expv<-  exp( coefv)
  SEv<-  x[[i]][["table"]][[2,2]]
  Zv<-  x[[i]][["table"]][[2,3]]

  #to extract p.value
  Pv<- x[[i]][["table"]][[2,4]]
```

```
   ID <- length(x[i])
   Scale1  <-  x[[i]][["scale"]][1]

   # to combine dataset results into one dataframe(or array) by row
   w <- abind(ID=ID, coef=coefv, HR=expv, SE=SEv, Z_test=Zv, P_value=Pv, Scale=Scale1,
along=2)
   return(data.frame(w))
 })
}

#####################################################################
#To run Weibull model function after creating it and save results of Weibull model
function
ResWeiB1S <-mySaveModelWeiS(ResultsWeiB1S)
ResWeiB1S_DF <- do.call(rbind,ResWeiB1S)
print(ResWeiB1S_DF)
ResWeiB1S_DF$ID <- seq_len(nrow(ResWeiB1S_DF))

save(ResWeiB1S_DF, file=paste("C: /FOLDER ADDRESS PART 1",
                    "/ FOLDER ADDRESS PART 2",
                    "/Output Analysis/Tables/S1Tables/ResWeiB1S_DF.Rda",sep=""))
#####################################################################
```

E.2.2: Specification for different forms of the Weibull model in an admixed population.

```
#Weibull analysis based on genotype SNP
snptemp <- as.numeric(x[[i]][["SNPs"]])
timetemp <- as.numeric(x[[i]][["OT"]])
eventID  <-  as.numeric(x[[i]][["eventID"]])
WphModel_M2S <- survreg(Surv(timetemp, eventID )~ snptemp, dist="weibull")

#Weibull analysis based on ancestry of genotyped SNP
anctemp <- as.numeric(x[[i]][["CHROMANC"]])
timetemp <- as.numeric(x[[i]][["OT"]])
eventID  <-  as.numeric(x[[i]][["eventID"]])
WphModel_M2A <- survreg(Surv( timetemp, eventID)~ anctemp, dist="weibull")

#Weibull analysis based on genotype SNP with ancestry as covariate
anctemp<- as.numeric(x[[i]][["CHROMANC"]])
snptemp <- as.numeric(x[[i]][["SNPs"]])
timetemp <- as.numeric(x[[i]][["OT"]])
eventID  <-  as.numeric(x[[i]][["eventID"]])
WphModel_M2S <- survreg(Surv(timetemp, eventID )~ snptemp + anctemp, dist="weibull")
```

Appendix F: R syntax used for generation and data analysis of GRS simulated data

Table of Contents

Appendix F: R syntax used for generation and data analysis of GRS simulated data

F.1: R syntax used for generating GRS simulated data

```
set.seed(1364)
#####################################################################
nsim <- 1000     #Number of simulations/samples
maf <- 0.05      #Minor allele frequency
beta <- 0.05     #Log hazard ratio of GRS
nind <- 1000     #Number of individuals in the sample
ngen <- 1              #Number of SNPs in the GRS
time <- 50             #Study period
base <- -50/log (0.5)  #baseline hazard
#####################################################################
#simulating AOO of disease conditional on GRS
#####################################################################
library("abind")
library(MASS)
library(survival)
#creating empty datsets
zstat <- matrix(nrow=nsim,ncol=3,0)
cstat <- matrix(nrow=nsim,ncol=3,0)
for(n in 1:nsim){
xgen <- matrix(nrow=nind,ncol=ngen+1,0)
xphen <- matrix(nrow=nind,ncol=4,0)
#simulating genotype and GRS data
for(i in 1:nind){
for(j in 1:ngen){
#generating genotype of individual SNPs for each individual
xgen[i,j] <- rbinom(1,2,maf)

#generating GRS for each individual in sample
xgen[i,ngen+1] <- xgen[i,ngen+1]+xgen[i,j]
}

#rescaling GRS to have mean of zero
xgen[i,ngen+1]<- xgen[i,ngen+1]-2*ngen*maf
#generating AOO of disease conditional on GRS
xphen[i,1] <- rweibull(1,1,base*exp(-xgen[i,ngen+1]*beta))   #event time
xphen[i,2] <- xphen[i,1]                    #censoring time
xphen[i,3] <- 1                             #event status
xphen[i,4] <- 1                             #ordered event status
if(xphen[i,1]>time) xphen[i,2] = time
if(xphen[i,1]>time) xphen[i,3] = 0
```

```
if(xphen[i,1]>time) xphen[i,4] = 0
if(xphen[i,1] < 25) xphen[i,4] = 2

}

Pheno <- abind(xphen, xgen[,ngen+1], along=2)
Pheno <- as.data.frame(Pheno)
```

F.2: R syntax used for data analysis simulated GRS data

```
#################################################################
#analysis simulated based on the proportional odds model
Polrx <- polr(as.ordered(xphen[,4]) ~ xgen[,ngen+1],method ="logistic")
zstat[n,1] <- summary(Polrx)$coefficients[1,3]

#################################################################
#analysis simulated based on the logistic model
glmx <- glm(xphen[,3]~xgen[,ngen+1],family="binomial")
zstat[n,2] <- summary(glmx)$coefficients[2,3]

#################################################################
#analysis simulated based on the Cox PH model
coxx <- coxph(Surv(xphen[,2],xphen[,3])~xgen[,ngen+1])
zstat[n,3] <- summary(coxx)$coefficients[,4]

print(Polrx)
PolrxA <- print(Polrx)
PolrxB <- print(summary(Polrx))

print(glmx)
glmxA <- print(glmx)
glmxB <- print(summary(glmx))

print(coxx)
coxxA <- print(coxx)
coxxB <- print(summary(coxx))

}
#################################################################
#################################################################
#Saving Z-value results
write.table(zstat,'zstatout1',row.names=F,col.names=F)
zstatout1 <- read.table("zstatout1")
zstatO1 <- zstatout1

#################################################################
#calculating P-values for each model based on Z values
#################################################################
pstat <- matrix(nrow=nsim,ncol=3,0)
#calculate P-value for proportional odds model
PolrP <- pnorm(abs(zstatO1[, 1]), lower.tail=FALSE) * 2
pstat[,1] = PolrP

#calculate P-value for binary logistic model
glmP <- pnorm(abs(zstatO1[, 2]), lower.tail=FALSE) * 2
```

```
pstat[,2]=glmP
#calculate P-value for Cox PH model
coxP <- pnorm(abs(zstatO1[, 3]), lower.tail=FALSE) * 2
pstat[,3]=coxP
#####################################################################
#saving P-value results
write.table(pstat,'pstatout1',row.names=F,col.names=F)
pstatout1 <- read.table("pstatout1")
pstatO1 <- pstatout1
#####################
options(scipen=20)
#####################################################################
#PROPORTIONAL ODDS MODEL
#count number of samples with significant P-values
#####################################################################
NumSamples <- length(pstatO1[,1])
NumPRS <- ngen
SigPvalG <- length(pstatO1[,1] [pstatO1[,1] <= 5*10^-8])
SigPvalS <- length(pstatO1[,1] [pstatO1[,1] <= 0.05])
PowerpG <- SigPvalG/NumSamples
PowerpS <- SigPvalS/NumSamples
powPRSPO1 <- list(NumSamples=NumSamples, NumPRS=NumPRS, SigPvalG=SigPvalG,
   SigPvalS=SigPvalS, Power_G=PowerpG, Power_S=PowerpS)

#####################
powPRSPO1 <- do.call(rbind,powPRSPO1)
powPRSPO1 <- data.frame(Nam=row.names(powPRSPO1), powPRSPO1, row.names=NULL)
names(powPRSPO1)
print(powPRSPO1)
write.table(powPRSPO1,'powPRSPO1.csv',row.names=F,col.names=F)

#####################################################################
#LOGISTIC MODEL
#count number of samples with significant P-values
#####################################################################
NumSamples <- length(pstatO1[,2])
NumPRS <- ngen
SigPvalG <- length(pstatO1[,2] [pstatO1[,2] <= 5*10^-8])
SigPvalS <- length(pstatO1[,2] [pstatO1[,2] <= 0.05])
PowerpG <- SigPvalG/NumSamples
PowerpS <- SigPvalS/NumSamples
powPRSLR1 <- list(NumSamples=NumSamples, NumPRS=NumPRS, SigPvalG=SigPvalG,
          SigPvalS=SigPvalS, Power_G=PowerpG, Power_S=PowerpS)

#####################
powPRSLR1 <- do.call(rbind,powPRSLR1)
powPRSLR1 <- data.frame(Nam=row.names(powPRSLR1), powPRSLR1, row.names=NULL)
```

```
names(powPRSLR1)
print(powPRSLR1)
write.table(powPRSLR1,'powPRSLR1.csv',row.names=F,col.names=F)

#################################################################
#Cox PH model
#count number of samples with significant P-values
#################################################################
NumSamples <- length(pstatO1[,3])
NumPRS <- ngen
SigPvalG <- length(pstatO1[,3] [pstatO1[,3] <= 5*10^-8])
SigPvalS <- length(pstatO1[,3] [pstatO1[,3] <= 0.05])
PowerpG <- SigPvalG/NumSamples
PowerpS <- SigPvalS/NumSamples
powPRSCX1 <- list(NumSamples=NumSamples, NumPRS=NumPRS, SigPvalG=SigPvalG,
          SigPvalS=SigPvalS, Power_G=PowerpG, Power_S=PowerpS)
######################
powPRSCX1 <- do.call(rbind,powPRSCX1)
powPRSCX1 <- data.frame(Nam=row.names(powPRSCX1), powPRSCX1, row.names=NULL)
names(powPRSCX1)
print(powPRSCX1)
write.table(powPRSCX1,'powPRSCX1.csv',row.names=F,col.names=F)
#################################################################
```

Appendix G: R syntax used for used for constructing T2D GRS

Table of Contents

Appendix G: R syntax used for used for constructing T2D GRS

G.1: R syntax used to import genotype probabilities or genotype dosage

G.1.1: R syntax used to import genotype dosage values

```
#################################################################
cd /DIRECTORY ADDRESS/FOLDER ADDRESS
nano bashrc
source bashrc
bcftools query T2D_DATASET_DOS.vcf -f
'%CHROM\t%POS\t%ID\t%REF\t%ALT\t%QUAL\t%FILTER [\t%DS]\n' -H >
T2D_DATASET_GRS.vcf
cp T2D_DATASET_GRS.vcf T2D_DATASET_GRS.csv


#################################################################
R
#################################################################
#DATASET 1 – SAMPLES - importing dataset with SNP genotype information
#pertaining to each individual included in the genotype sample.
R_T2D_DATASET_GRS <- read.csv("/DIRECTORY ADDRESS/FOLDER ADDRESS/
T2D_DATASET_GRS.csv", header=TRUE, sep="\t")
save(R_T2D_DATASET_GRS, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS/
R_T2D_DATASET_GRS.Rda",sep=""))
load("R_T2D_DATASET_GRS.Rda")
View(R_T2D_DATASET_GRS)
#################################################################
```

G.1.2: R syntax used to import genotype probabilities values

```
#################################################################
cd /DIRECTORY ADDRESS/FOLDER ADDRESS
nano bashrc
source bashrc
bcftools query T2D_DATASET_DOS.vcf -f
'%CHROM\t%POS\t%ID\t%REF\t%ALT\t%QUAL\t%FILTER [\t%GP{0}]\n' -H >
T2D_DATASET_GRS0.vcf
bcftools query T2D_DATASET_DOS.vcf -f
'%CHROM\t%POS\t%ID\t%REF\t%ALT\t%QUAL\t%FILTER [\t%GP{1}]\n' -H >
T2D_DATASET_GRS1.vcf
bcftools query T2D_DATASET_DOS.vcf -f
'%CHROM\t%POS\t%ID\t%REF\t%ALT\t%QUAL\t%FILTER [\t%GP{2}]\n' -H >
T2D_DATASET_GRS2.vcf

awk '{print NR,$1,$2,$3,$4,$5}' T2D_DATASET_GRS0.vcf

cp T2D_DATASET_GRS0.vcf T2D_DATASET_GRS0.csv
cp T2D_DATASET_GRS1.vcf T2D_DATASET_GRS1.csv
cp T2D_DATASET_GRS2.vcf T2D_DATASET_GRS2.csv

#################################################################
R
#################################################################
#DATASET 1 – SAMPLES - importing dataset with SNP genotype information
#pertaining to each individual included in the genotype sample.
#DATASET 1a - SAMPLES - Genotype Probability 1
R_T2D_DATASET_GRS0 <- read.csv("/DIRECTORY ADDRESS/FOLDER ADDRESS/
WTCCC_GRS0.csv", header=TRUE, sep="\t")
library("stringr")
R_T2D_DATASET_GRS0$X.3.ID  <- str_replace_all(R_T2D_DATASET_GRS0$X.3.ID, '_(.*)_(.*)$',
'')

save(R_T2D_DATASET_GRS0, file=paste("/DIRECTORY ADDRESS/ FOLDER ADDRESS
/R_T2D_DATASET_GRS0.Rda",sep=""))
load("R_T2D_DATASET_GRS0.Rda")
View(R_T2D_DATASET_GRS0)
#############
#############
#DATASET 1b - SAMPLES - Genotype Probability 2
R_T2D_DATASET_GRS1 <- read.csv("/DIRECTORY ADDRESS/ FOLDER ADDRESS
/WTCCC_GRS1.csv", header=TRUE, sep="\t")
library("stringr")
R_T2D_DATASET_GRS1$X.3.ID  <- str_replace_all(R_T2D_DATASET_GRS1$X.3.ID, '_(.*)_(.*)$',
'')

save(R_T2D_DATASET_GRS1, file=paste("/ph-users/odessica/T2D-
Dataset2/R_T2D_DATASET_GRS1.Rda",sep=""))
```

```
load("R_T2D_DATASET_GRS1.Rda")
View(R_T2D_DATASET_GRS1)
##############
##############
#DATASET 1c - SAMPLES - Genotype Probability 3
R_T2D_DATASET_GRS2 <- read.csv("/DIRECTORY ADDRESS/ FOLDER ADDRESS/
WTCCC_GRS2.csv", header=TRUE, sep="\t")
library("stringr")
R_T2D_DATASET_GRS2$X.3.ID  <- str_replace_all(R_T2D_DATASET_GRS2$X.3.ID, '_(.*)_(.*)$',
'')

save(R_T2D_DATASET_GRS2, file=paste("/DIRECTORY ADDRESS/ FOLDER ADDRESS/
R_T2D_DATASET_GRS2.Rda",sep=""))
load("R_T2D_DATASET_GRS2.Rda")
View(R_T2D_DATASET_GRS2)

#####################################################################
#dataset to check probability totals
library("dplyr")
R_T2D_DATASET_GRS_check <- R_T2D_DATASET_GRS0 %>%  mutate_at(.vars=
vars(matches("^X.(.*)WTCCC(.*).GP",
       ignore.case =FALSE)), .funs=funs(. +R_T2D_DATASET_GRS1$. +
R_T2D_DATASET_GRS2$.))
save(R_T2D_DATASET_GRS_check, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS/
R_T2D_DATASET_GRS_check.Rda",sep=""))
#####################################################################

#calculating Dosage for alternative allele
library("dplyr")
R_T2D_DATASET_GRS00 <- R_T2D_DATASET_GRS0 %>%  mutate_at(.vars=
vars(matches("^X.(.*)WTCCC(.*).GP",
       ignore.case =FALSE)), .funs=funs(. *0))
save(R_T2D_DATASET_GRS00, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS/
R_T2D_DATASET_GRS00.Rda",sep=""))

R_T2D_DATASET_GRS11 <- R_T2D_DATASET_GRS1 %>%  mutate_at(.vars=
vars(matches("^X.(.*)WTCCC(.*).GP",
       ignore.case =FALSE)), .funs=funs(. *1))
save(R_T2D_DATASET_GRS11, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS/
R_T2D_DATASET_GRS11.Rda",sep=""))

R_T2D_DATASET_GRS22 <- R_T2D_DATASET_GRS2 %>%  mutate_at(.vars=
vars(matches("^X.(.*)WTCCC(.*).GP",
       ignore.case =FALSE)), .funs=funs(. *2))
save(R_T2D_DATASET_GRS22, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS/
R_T2D_DATASET_GRS22.Rda",sep=""))
```

```
#####################################################################
#Final Dosage value
#Dosage=probability of allele 1 *0) + (probability of allele 2*1) + (probability of #allele
3 * 2))
library("dplyr")
R_T2D_DATASET_GRS <- R_T2D_DATASET_GRS00 %>%  mutate_at(.vars=
vars(matches("^X.(.*)WTCCC(.*).GP",
        ignore.case =FALSE)), .funs=funs(. +R_T2D_DATASET_GRS11$. +
R_T2D_DATASET_GRS22$.))

save(R_T2D_DATASET_GRS, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS/
R_T2D_DATASET_GRS.Rda",sep=""))

#####################################################################
load("R_T2D_DATASET_GRS.Rda")
View(R_T2D_DATASET_GRS)
```

G.2: R syntax used to adjust genotype dosage values in line with EA

```
#################################################################
#DATASET 2 - DISCOVERY SNPs – importing dataset with information pertaining #to
each T2D SNP from base GWAS included in the construction of T2D GRS
#################################################################
library(readxl)
Dis_GRS_All <- read_excel("/DIRECTORY ADDRESS/ FOLDER ADDRESS/Discovery SNP.xlsx")
save(Dis_GRS_All, file=paste("/DIRECTORY ADDRESS/ FOLDER ADDRESS/
Dis_GRS_All.Rda",sep=""))
load("Dis_GRS_All.Rda")
View(Dis_GRS_All)

#################################################################
#DATASET 3 - merging SNP genotype data from target and discovery T2D GWAS
#################################################################
T2D_DATASET_GRS_Check_M <- merge(Dis_GRS_All,R_T2D_DATASET_GRS, by.x =
"Position_b37", by.y ="X.2.POS")
View(T2D_DATASET_GRS_Check_M)
#remove SNP with info score less than 0.4
T2D_DATASET_GRS_Check_M <-  subset(T2D_DATASET_GRS_Check_M,
Position_b37!=127631181)
T2D_DATASET_GRS_Check_M <-
T2D_DATASET_GRS_Check_M[order(T2D_DATASET_GRS_Check_M$Chr,T2D_DATASET_GRS_C
heck_M$Position_b37),]
length(unique(T2D_DATASET_GRS_Check_M$Position_b37))

save(T2D_DATASET_GRS_Check_M, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS /
T2D_DATASET_GRS_Check_M.Rda",sep=""))

load("T2D_DATASET_GRS_Check_M.Rda")
View(T2D_DATASET_GRS_Check_M)
length(T2D_DATASET_GRS_Check_M$X.4.REF)

#################################################################
#select SNPs with differences in reference allele assignment
#want effect allele in base GWAS dataset to be same as alternative allele
# as dosage in the sample dataset is for the alternative allele
#################################################################
library("dplyr")
T2D_DATASET_GRS_Check_M$REF_DIff <- (T2D_DATASET_GRS_Check_M$NEA
==T2D_DATASET_GRS_Check_M$X.4.REF)
T2D_DATASET_GRS_Check_M$ALT_DIff <- (T2D_DATASET_GRS_Check_M$EA
==T2D_DATASET_GRS_Check_M$X.5.ALT)
save(T2D_DATASET_GRS_Check_M, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS /
T2D_DATASET_GRS_Check_M.Rda",sep=""))

T2D_DATASET_GRS_ALT_DIFF <- subset(T2D_DATASET_GRS_Check_M, ALT_DIff =="FALSE")
T2D_DATASET_GRS_ALT_SAME <- subset(T2D_DATASET_GRS_Check_M, ALT_DIff =="TRUE")
```

```
T2D_DATASET_GRS_ALT_DIFF <- as.data.frame(T2D_DATASET_GRS_ALT_DIFF)
T2D_DATASET_GRS_ALT_SAME <- as.data.frame(T2D_DATASET_GRS_ALT_SAME)
#################
save(T2D_DATASET_GRS_ALT_DIFF, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS /
T2D_DATASET_GRS_ALT_DIFF.Rda",sep=""))
save(T2D_DATASET_GRS_ALT_SAME, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS
/ T2D_DATASET_GRS_ALT_SAME.Rda",sep=""))
load("T2D_DATASET_GRS_ALT_DIFF.Rda")
View(T2D_DATASET_GRS_ALT_DIFF)
load("T2D_DATASET_GRS_ALT_SAME.Rda")
View(T2D_DATASET_GRS_ALT_SAME)
#################
library("plyr")
count(T2D_DATASET_GRS_Check_M$REF_DIff)
count(T2D_DATASET_GRS_Check_M$ALT_DIff)
length(unique(T2D_DATASET_GRS_Check_M$Position_b37))
length(unique(T2D_DATASET_GRS_ALT_DIFF$Position_b37))
length(unique(T2D_DATASET_GRS_ALT_SAME$Position_b37))
#####################################################################



#####################################################################
# adjusting dosage values in sample of individuals dataset
#####################################################################
library("dplyr")
T2D_DATASET_GRS_ALT_DIFFa <- T2D_DATASET_GRS_ALT_DIFF %>% mutate_at(.vars=
vars(matches("^(.*)PT(.*)SM(.*)",
       ignore.case =FALSE)), .funs=funs(2- .))

save(T2D_DATASET_GRS_ALT_DIFFa, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS
/ T2D_DATASET_GRS_ALT_DIFFa.Rda",sep=""))

#####################################################################
#append or merge to original dataset
#####################################################################
T2D_DATASET_GRS_Adjust_M <- rbind(T2D_DATASET_GRS_ALT_DIFFa,
T2D_DATASET_GRS_ALT_SAME)
length(unique(T2D_DATASET_GRS_Adjust_M$Position_b37))
```

G.3: R syntax used to calculate weighted GRS

```
#####################################################################
#Calculation of weighted score for each SNP
#####################################################################
T2D_DATASET_GRSw_Adjust_Mr <- T2D_DATASET_GRS_Adjust_M %>% mutate_at(.vars=
vars(matches("^(.*)PT(.*)SM(.*)",
             ignore.case =FALSE)), .funs=funs((log(OR)) *.))

save(T2D_DATASET_GRSw_Adjust_Mr, file=paste("/DIRECTORY ADDRESS/FOLDER
ADDRESS / T2D_DATASET_GRSw_Adjust_Mr.Rda",sep=""))

#library("dplyr")
load("T2D_DATASET_GRSw_Adjust_Mr.Rda")
View(T2D_DATASET_GRSw_Adjust_Mr)

#####################################################################
#Calculation of overall weighted GRS value for each individual in the sample
#####################################################################
T2D_DATASET_GRSw_Adjust_All <- T2D_DATASET_GRSw_Adjust_Mr %>%
summarise_at(.vars= vars(matches("^(.*)PT(.*)SM(.*)",
              ignore.case =FALSE)), sum, na.rm = TRUE)
View(T2D_DATASET_GRSw_Adjust_All)
save(T2D_DATASET_GRSw_Adjust_All, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS
/ T2D_DATASET_GRSw_Adjust_All.Rda",sep=""))
load("T2D_DATASET_GRSw_Adjust_All.Rda")
#####################################################################
library("dplyr")
T2D_DATASET_GRSw_Adjust_All_t <-  t(T2D_DATASET_GRSw_Adjust_All)

T2D_DATASET_GRSw_Adjust_All_t1 <- data.frame(names =
row.names(T2D_DATASET_GRSw_Adjust_All_t),
                    T2D_DATASET_GRSw_Adjust_All_t, row.names = NULL)
colnames(T2D_DATASET_GRSw_Adjust_All_t1) <- c("SampleID","GRS_wei")
View(T2D_DATASET_GRSw_Adjust_All_t1)


#removing extra characters from ID numbers
library("stringr")
T2D_DATASET_GRSw_Adjust_All_t1$SampleID  <-
str_replace_all(T2D_DATASET_GRSw_Adjust_All_t1$SampleID, '^X.*.PT.', 'PT-')
T2D_DATASET_GRSw_Adjust_All_t1$SampleID  <-
str_replace_all(T2D_DATASET_GRSw_Adjust_All_t1$SampleID, '.SM.', '-SM-')
T2D_DATASET_GRSw_Adjust_All_t1$SampleID  <-
str_replace_all(T2D_DATASET_GRSw_Adjust_All_t1$SampleID, '.DS$', '')

save(T2D_DATASET_GRSw_Adjust_All_t1, file=paste("/DIRECTORY ADDRESS/FOLDER
ADDRESS / T2D_DATASET_GRSw_Adjust_All_t1.Rda",sep=""))
```

```
load("T2D_DATASET_GRSw_Adjust_All_t1.Rda")
View(T2D_DATASET_GRSw_Adjust_All_t1)
head(T2D_DATASET_GRSw_Adjust_All_t1,10)
```

G.4: R syntax used to calculate unweighted GRS

```
############################################################
#Calculation of overall unweighted GRS value for each individual in the sample
############################################################
T2D_DATASET_GRSu_Adjust_All <- T2D_DATASET_GRS_Adjust_M %>% summarise_at(.vars=
vars(matches("^(.*)PT(.*)SM(.*)",
             ignore.case =FALSE)), sum, na.rm = TRUE)
View(T2D_DATASET_GRSu_Adjust_All)
save(T2D_DATASET_GRSu_Adjust_All, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS
/ T2D_DATASET_GRSu_Adjust_All.Rda",sep=""))
load("T2D_DATASET_GRSu_Adjust_All.Rda")
###################################
library("dplyr")
T2D_DATASET_GRSu_Adjust_All_t <-  t(T2D_DATASET_GRSu_Adjust_All)

T2D_DATASET_GRSu_Adjust_All_t1 <- data.frame(names =
row.names(T2D_DATASET_GRSu_Adjust_All_t),
                    T2D_DATASET_GRSu_Adjust_All_t, row.names = NULL)
colnames(T2D_DATASET_GRSu_Adjust_All_t1) <- c("SampleID","GRS_unw")
View(T2D_DATASET_GRSu_Adjust_All_t1)
#################################################################
#removing extra characters from ID numbers
library("stringr")
T2D_DATASET_GRSu_Adjust_All_t1$SampleID  <-
str_replace_all(T2D_DATASET_GRSu_Adjust_All_t1$SampleID, '^X.*.PT.', 'PT-')
T2D_DATASET_GRSu_Adjust_All_t1$SampleID  <-
str_replace_all(T2D_DATASET_GRSu_Adjust_All_t1$SampleID, '.SM.', '-SM-')
T2D_DATASET_GRSu_Adjust_All_t1$SampleID  <-
str_replace_all(T2D_DATASET_GRSu_Adjust_All_t1$SampleID, '.DS$', '')

save(T2D_DATASET_GRSu_Adjust_All_t1, file=paste("/DIRECTORY ADDRESS/FOLDER
ADDRESS / T2D_DATASET_GRSu_Adjust_All_t1.Rda",sep=""))

load("T2D_DATASET_GRSu_Adjust_All_t1.Rda")
View(T2D_DATASET_GRSu_Adjust_All_t1)
head(T2D_DATASET_GRSu_Adjust_All_t1,10)
```

G.5: R syntax used to merge GRS value to phenotype data

```
#####################################################################
#loading phenotype dataset and nominal significant GRS datasets
#####################################################################
load("T2D_Dataset_pheno_adj_IND.Rda")
View(T2D_Dataset_pheno_adj_IND)
head(T2D_Dataset_pheno_adj_IND, 10)

load("T2D_DATASET_GRSw_Adjust_All_t1.Rda")
View(T2D_DATASET_GRSw_Adjust_All_t1)
head(T2D_DATASET_GRSw_Adjust_All_t1,10)

load("T2D_DATASET_GRSu_Adjust_All_t1.Rda")
View(T2D_DATASET_GRSu_Adjust_All_t1)
head(T2D_DATASET_GRSu_Adjust_All_t1,10)

#####################################################################
#To add weighted GRS to T2D sample dataset
T2D_sample_GRS <- merge(T2D_Dataset_pheno_adj_IND, T2D_DATASET_GRSw_Adjust_All_t1,
        by.x = "ID_2", by.y = "SampleID")

save(T2D_sample_GRS, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS /
T2D_sample_GRS.Rda",sep=""))
load("T2D_sample_GRS.Rda")
View(T2D_sample_GRS)
head(T2D_sample_GRS)

#####################################################################
#To add unweighted GRS to T2D sample dataset
T2D_sample_GRS <- merge(T2D_sample_GRS, T2D_DATASET_GRSu_Adjust_All_t1,
        by.x = "ID_2", by.y = "SampleID")

save(T2D_sample_GRS, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS /
T2D_sample_GRS.Rda",sep=""))
load("T2D_sample_GRS.Rda")
View(T2D_sample_GRS)



#####################################################################
#loading phenotype dataset and genome-wide significant GRS datasets
#####################################################################
load("T2D_sample_GRS.Rda")
View(T2D_sample_GRS)
head(T2D_sample_GRS, 10)

load("T2D_DATASET_GRSw_Adjust_Allg_t1.Rda")
View(T2D_DATASET_GRSw_Adjust_Allg_t1)
```

```
head(T2D_DATASET_GRSw_Adjust_Allg_t1,10)

load("T2D_DATASET_GRSu_Adjust_Allg_t1.Rda")
View(T2D_DATASET_GRSu_Adjust_Allg_t1)
head(T2D_DATASET_GRSu_Adjust_Allg_t1,10)


#####################################################################
#To add weighted GRS to T2D sample dataset
T2D_sample_GRS <- merge(T2D_sample_GRS, T2D_DATASET_GRSw_Adjust_Allg_t1,
        by.x = "ID_2", by.y = "SampleID")

save(T2D_sample_GRS, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS /
T2D_sample_GRS.Rda",sep=""))

load("T2D_sample_GRS.Rda")
View(T2D_sample_GRS)
#####################################################################
#To add unweighted GRS to T2D sample dataset
T2D_sample_GRS <- merge(T2D_sample_GRS, T2D_DATASET_GRSu_Adjust_Allg_t1,
        by.x = "ID_2", by.y = "SampleID")

save(T2D_sample_GRS, file=paste("/DIRECTORY ADDRESS/FOLDER ADDRESS /
T2D_sample_GRS.Rda",sep=""))

load("T2D_sample_GRS.Rda")
View(T2D_sample_GRS)
head(T2D_sample_GRS, 10)
#####################################################################
```

Appendix H: R syntax used to conduct data analysis of T2D GRS AOO data

Table of Contents

Appendix H: R syntax used to conduct data analysis of T2D GRS AOO data

H.1: R syntax used to undertake Cox PH analysis

```
############################################################
#weighted GRS (nominal significance level)
############################################################
library(survival)
args(coxph)
Cox_Model_1 <- coxph(Surv(timeAGE_O, eventID)~ GRSwei,
data=T2D_Dataset_R)
SumCox_Model_1 <- summary(Cox_Model_1)
print(SumCox_Model_1)

library(survival)
args(coxph)
Cox_Model_2 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI,
data=T2D_Dataset_R)
SumCox_Model_2 <- summary(Cox_Model_2)
print(SumCox_Model_2)

library(survival)
args(coxph)
Cox_Model_3 <- coxph(Surv(timeAGE_O, eventID)~ SEX + C1 + C2 + GRSwei,
data=T2D_Dataset_R)
SumCox_Model_3 <- summary(Cox_Model_3)
print(SumCox_Model_3)

library(survival)
args(coxph)
Cox_Model_3.2 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI + C1 + C2 + GRSwei,
data=T2D_Dataset_R)
SumCox_Model_3.2 <- summary(Cox_Model_3.2)
print(SumCox_Model_3.2)

############################################################
#unweighted GRS (nominal significance level)
############################################################
library(survival)
args(coxph)
Cox_Model_4 <- coxph(Surv(timeAGE_O, eventID)~ GRSunw,
data=T2D_Dataset_R)
SumCox_Model_4 <- summary(Cox_Model_4)
print(SumCox_Model_4)

library(survival)
args(coxph)
```

```
Cox_Model_5 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI,
data=T2D_Dataset_R)
SumCox_Model_5 <- summary(Cox_Model_5)
print(SumCox_Model_5)


############################################################
library(survival)
args(coxph)
Cox_Model_6 <- coxph(Surv(timeAGE_O, eventID)~ SEX + C1 + C2 + GRSunw,
data=T2D_Dataset_R)
SumCox_Model_6 <- summary(Cox_Model_6)
print(SumCox_Model_6)

library(survival)
args(coxph)
Cox_Model_6.2 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI + C1 + C2 + GRSunw,
data=T2D_Dataset_R)
SumCox_Model_6.2 <- summary(Cox_Model_6.2)
print(SumCox_Model_6.2)

############################################################
#weighted GRS (genome-wide significance level)
############################################################

library(survival)
args(coxph)
Cox_Model_7 <- coxph(Surv(timeAGE_O, eventID)~  GRSwei_G,
data=T2D_Dataset_R)
SumCox_Model_7 <- summary(Cox_Model_7)
print(SumCox_Model_7)

library(survival)
args(coxph)
Cox_Model_8 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI,
data=T2D_Dataset_R)
SumCox_Model_8 <- summary(Cox_Model_8)
print(SumCox_Model_8)

library(survival)
args(coxph)
Cox_Model_9 <- coxph(Surv(timeAGE_O, eventID)~ SEX + C1 + C2 + GRSwei_G,
data=T2D_Dataset_R)
SumCox_Model_9 <- summary(Cox_Model_9)
print(SumCox_Model_9)

library(survival)
args(coxph)
Cox_Model_9.2 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI + C1 + C2 + GRSwei_G,
data=T2D_Dataset_R)
```

```
SumCox_Model_9.2 <- summary(Cox_Model_9.2)
print(SumCox_Model_9.2)
############################################################
#unweighted GRS (genome-wide significance level)
############################################################
library(survival)
args(coxph)
Cox_Model_10 <- coxph(Surv(timeAGE_O, eventID)~ GRSunw_G,
data=T2D_Dataset_R)
SumCox_Model_10 <- summary(Cox_Model_10)
print(SumCox_Model_10)

library(survival)
args(coxph)
Cox_Model_11 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI,
data=T2D_Dataset_R)
SumCox_Model_11 <- summary(Cox_Model_11)
print(SumCox_Model_11)

library(survival)
args(coxph)
Cox_Model_12 <- coxph(Surv(timeAGE_O, eventID)~ SEX + C1 + C2 + GRSunw_G,
data=T2D_Dataset_R)
SumCox_Model_12 <- summary(Cox_Model_12)
print(SumCox_Model_12)

library(survival)
args(coxph)
Cox_Model_12.2 <- coxph(Surv(timeAGE_O, eventID)~ SEX + BMI + C1 + C2 + GRSunw_G,
data=T2D_Dataset_R)
SumCox_Model_12.2 <- summary(Cox_Model_12.2)
print(SumCox_Model_12.2)
```

H.2: R syntax used to undertake logistic analysis

```
############################################################
#weighted GRS (nominal significance level)
############################################################
library("DescTools")
Ltc_Model_1 <- glm(eventID ~ GRSwei,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_1 <- summary(Ltc_Model_1)
ANOVALtc_Model_1 <- anova(Ltc_Model_1, test = "Chisq")
R2Ltc_Model_1 <- PseudoR2(Ltc_Model_1, which="all")
print(SumLtc_Model_1)
print(ANOVALtc_Model_1)
print(R2Ltc_Model_1)

Ltc_Model_2 <- glm(eventID ~ SEX + BMI,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_2 <- summary(Ltc_Model_2)
ANOVALtc_Model_2 <- anova(Ltc_Model_2, test = "Chisq")
R2Ltc_Model_2 <- PseudoR2(Ltc_Model_2, which="all")
print(SumLtc_Model_2)
print(ANOVALtc_Model_2)
print(R2Ltc_Model_2)

Ltc_Model_3 <- glm(eventID ~ SEX + C1 + C2 + GRSwei,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_3 <- summary(Ltc_Model_3)
ANOVALtc_Model_3 <- anova(Ltc_Model_3, test = "Chisq")
R2Ltc_Model_3 <- PseudoR2(Ltc_Model_3, which="all")
print(SumLtc_Model_3)
print(ANOVALtc_Model_3)
print(R2Ltc_Model_3)

Ltc_Model_3.2 <- glm(eventID ~ SEX + BMI + C1 + C2 + GRSwei,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_3.2 <- summary(Ltc_Model_3.2)
ANOVALtc_Model_3.2 <- anova(Ltc_Model_3.2, test = "Chisq")
R2Ltc_Model_3.2 <- PseudoR2(Ltc_Model_3.2, which="all")
print(SumLtc_Model_3.2)
print(ANOVALtc_Model_3.2)
print(R2Ltc_Model_3.2)

############################################################
#unweighted GRS (nominal significance level)
############################################################
Ltc_Model_4 <- glm(eventID ~ GRSunw,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_4 <- summary(Ltc_Model_4)
ANOVALtc_Model_4 <- anova(Ltc_Model_4, test = "Chisq")
```

```
R2Ltc_Model_4 <- PseudoR2(Ltc_Model_4, which="all")
print(SumLtc_Model_4)
print(ANOVALtc_Model_4)
print(R2Ltc_Model_4)

Ltc_Model_5 <- glm(eventID ~ SEX + BMI,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_5 <- summary(Ltc_Model_5)
ANOVALtc_Model_5 <- anova(Ltc_Model_5, test = "Chisq")
R2Ltc_Model_5 <- PseudoR2(Ltc_Model_5, which="all")
print(SumLtc_Model_5)
print(ANOVALtc_Model_5)
print(R2Ltc_Model_5)

#################################################################
Ltc_Model_6 <- glm(eventID ~ SEX + C1 + C2 + GRSunw,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_6 <- summary(Ltc_Model_6)
ANOVALtc_Model_6 <- anova(Ltc_Model_6, test = "Chisq")
R2Ltc_Model_6 <- PseudoR2(Ltc_Model_6, which="all")
print(SumLtc_Model_6)
print(ANOVALtc_Model_6)
print(R2Ltc_Model_6)

#################################################################
Ltc_Model_6.2 <- glm(eventID ~ SEX + BMI + C1 + C2 + GRSunw,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_6.2 <- summary(Ltc_Model_6.2)
ANOVALtc_Model_6.2 <- anova(Ltc_Model_6.2, test = "Chisq")
R2Ltc_Model_6.2 <- PseudoR2(Ltc_Model_6.2, which="all")
print(SumLtc_Model_6.2)
print(ANOVALtc_Model_6.2)
print(R2Ltc_Model_6.2)

#################################################################
#weighted GRS (genome-wide significance level)
#################################################################
Ltc_Model_7 <- glm(eventID ~ GRSwei_G,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_7 <- summary(Ltc_Model_7)
ANOVALtc_Model_7 <- anova(Ltc_Model_7, test = "Chisq")
R2Ltc_Model_7 <- PseudoR2(Ltc_Model_7, which="all")
print(SumLtc_Model_7)
print(ANOVALtc_Model_7)
print(R2Ltc_Model_7)

Ltc_Model_8 <- glm(eventID ~ SEX + BMI,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_8 <- summary(Ltc_Model_8)
ANOVALtc_Model_8 <- anova(Ltc_Model_8, test = "Chisq")
```

```
R2Ltc_Model_8 <- PseudoR2(Ltc_Model_8, which="all")
print(SumLtc_Model_8)
print(ANOVALtc_Model_8)
print(R2Ltc_Model_8)

Ltc_Model_9 <- glm(eventID ~ SEX +C1 + C2 + GRSwei_G,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_9 <- summary(Ltc_Model_9)
ANOVALtc_Model_9 <- anova(Ltc_Model_9, test = "Chisq")
R2Ltc_Model_9 <- PseudoR2(Ltc_Model_9, which="all")
print(SumLtc_Model_9)
print(ANOVALtc_Model_9)
print(R2Ltc_Model_9)

Ltc_Model_9.2 <- glm(eventID ~ SEX + BMI +C1 + C2 + GRSwei_G,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_9.2 <- summary(Ltc_Model_9.2)
ANOVALtc_Model_9.2 <- anova(Ltc_Model_9.2, test = "Chisq")
R2Ltc_Model_9.2 <- PseudoR2(Ltc_Model_9.2, which="all")
print(SumLtc_Model_9.2)
print(ANOVALtc_Model_9.2)
print(R2Ltc_Model_9.2)

############################################################
#unweighted GRS (genome-wide significance level)
############################################################
Ltc_Model_10 <- glm(eventID ~ GRSunw_G,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_10 <- summary(Ltc_Model_10)
ANOVALtc_Model_10 <- anova(Ltc_Model_10, test = "Chisq")
R2Ltc_Model_10 <- PseudoR2(Ltc_Model_10, which="all")
print(SumLtc_Model_10)
print(ANOVALtc_Model_10)
print(R2Ltc_Model_10)

Ltc_Model_11 <- glm(eventID ~ SEX + BMI,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_11 <- summary(Ltc_Model_11)
ANOVALtc_Model_11 <- anova(Ltc_Model_11, test = "Chisq")
R2Ltc_Model_11 <- PseudoR2(Ltc_Model_11, which="all")
print(SumLtc_Model_11)
print(ANOVALtc_Model_11)
print(R2Ltc_Model_11)

Ltc_Model_12 <- glm(eventID ~ SEX +C1 + C2 + GRSunw_G ,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_12 <- summary(Ltc_Model_12)
ANOVALtc_Model_12 <- anova(Ltc_Model_12, test = "Chisq")
R2Ltc_Model_12 <- PseudoR2(Ltc_Model_12, which="all")
print(SumLtc_Model_12)
```

```
print(ANOVALtc_Model_12)
print(R2Ltc_Model_12)

Ltc_Model_12.2 <- glm(eventID ~ SEX + BMI +C1 + C2 + GRSunw_G ,
family = binomial(link = 'logit'),data=T2D_Dataset_R)
SumLtc_Model_12.2 <- summary(Ltc_Model_12.2)
ANOVALtc_Model_12.2 <- anova(Ltc_Model_12.2, test = "Chisq")
R2Ltc_Model_12.2 <- PseudoR2(Ltc_Model_12.2, which="all")
print(SumLtc_Model_12.2)
print(ANOVALtc_Model_12.2)
print(R2Ltc_Model_12.2)
```

H.3: R syntax used to undertake proportional odds analysis

```
############################################################
#weighted GRS (nominal significance level)
############################################################
library(MASS)
library("DescTools")
Polr_Model_1 <- polr(T2D_ord ~ GRSwei,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_1 <- summary(Polr_Model_1)
ptable_1 <- coef(summary(Polr_Model_1))
P <- pnorm(abs(ptable_1[, "t value"]), lower.tail=FALSE) * 2
ptable_1 <- cbind(ptable_1, "P value" =P)
R2Polr_Model_1 <- PseudoR2(Polr_Model_1, which="all")
print(Polr_Model_1)
print(SumPolr_Model_1)
print(ptable_1)
print(R2Polr_Model_1)

library(MASS)
Polr_Model_2 <- polr(T2D_ord ~ SEX + BMI,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_2 <- summary(Polr_Model_2)
ptable_2 <- coef(summary(Polr_Model_2))
P <- pnorm(abs(ptable_2[, "t value"]), lower.tail=FALSE) * 2
ptable_2 <- cbind(ptable_2, "P value" =P)
R2Polr_Model_2 <- PseudoR2(Polr_Model_2, which="all")
print(Polr_Model_2)
print(SumPolr_Model_2)
print(ptable_2)
print(R2Polr_Model_2)

library(MASS)
Polr_Model_3 <- polr(T2D_ord ~  SEX + C1 + C2 + GRSwei,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_3 <- summary(Polr_Model_3)
ptable_3 <- coef(summary(Polr_Model_3))
P <- pnorm(abs(ptable_3[, "t value"]), lower.tail=FALSE) * 2
ptable_3 <- cbind(ptable_3, "P value" =P)
R2Polr_Model_3 <- PseudoR2(Polr_Model_3, which="all")
print(Polr_Model_3)
print(SumPolr_Model_3)
print(ptable_3)
print(R2Polr_Model_3)

library(MASS)
Polr_Model_3.2 <- polr(T2D_ord ~  SEX + BMI + C1 + C2 + GRSwei,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_3.2 <- summary(Polr_Model_3.2)
```

```
ptable_3.2 <- coef(summary(Polr_Model_3.2))
P <- pnorm(abs(ptable_3.2[, "t value"]), lower.tail=FALSE) * 2
ptable_3.2 <- cbind(ptable_3.2, "P value" =P)
R2Polr_Model_3.2 <- PseudoR2(Polr_Model_3.2, which="all")
print(Polr_Model_3.2)
print(SumPolr_Model_3.2)
print(ptable_3.2)
print(R2Polr_Model_3.2)
############################################################
#unweighted GRS (nominal significance level)
############################################################
library(MASS)
Polr_Model_4 <- polr(T2D_ord ~ GRSunw,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_4 <- summary(Polr_Model_4)
ptable_4 <- coef(summary(Polr_Model_4))
P <- pnorm(abs(ptable_4[, "t value"]), lower.tail=FALSE) * 2
ptable_4 <- cbind(ptable_4, "P value" =P)
R2Polr_Model_4 <- PseudoR2(Polr_Model_4, which="all")
print(Polr_Model_4)
print(SumPolr_Model_4)
print(ptable_4)
print(R2Polr_Model_4)

library(MASS)
Polr_Model_5 <- polr(T2D_ord ~ SEX + BMI,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_5 <- summary(Polr_Model_5)
ptable_5 <- coef(summary(Polr_Model_5))
P <- pnorm(abs(ptable_5[, "t value"]), lower.tail=FALSE) * 2
ptable_5 <- cbind(ptable_5, "P value" =P)
R2Polr_Model_5 <- PseudoR2(Polr_Model_5, which="all")
print(Polr_Model_5)
print(SumPolr_Model_5)
print(ptable_5)
print(R2Polr_Model_5)

############################################################
library(MASS)
Polr_Model_6 <- polr(T2D_ord ~ SEX + C1 + C2 + GRSunw,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_6 <- summary(Polr_Model_6)
ptable_6 <- coef(summary(Polr_Model_6))
P <- pnorm(abs(ptable_6[, "t value"]), lower.tail=FALSE) * 2
ptable_6 <- cbind(ptable_6, "P value" =P)
R2Polr_Model_6 <- PseudoR2(Polr_Model_6, which="all")
print(Polr_Model_6)
print(SumPolr_Model_6)
print(ptable_6)
print(R2Polr_Model_6)
```

```
############################################################
library(MASS)
Polr_Model_6.2 <- polr(T2D_ord ~ SEX + BMI + C1 + C2 + GRSunw,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_6.2 <- summary(Polr_Model_6.2)
ptable_6.2 <- coef(summary(Polr_Model_6.2))
P <- pnorm(abs(ptable_6.2[, "t value"]), lower.tail=FALSE) * 2
ptable_6.2 <- cbind(ptable_6.2, "P value" =P)
R2Polr_Model_6.2 <- PseudoR2(Polr_Model_6.2, which="all")
print(Polr_Model_6.2)
print(SumPolr_Model_6.2)
print(ptable_6.2)
print(R2Polr_Model_6.2)
############################################################
#weighted GRS (genome-wide significance level)
############################################################
library(MASS)
Polr_Model_7 <- polr(T2D_ord ~ GRSwei_G,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_7 <- summary(Polr_Model_7)
ptable_7 <- coef(summary(Polr_Model_7))
P <- pnorm(abs(ptable_7[, "t value"]), lower.tail=FALSE) * 2
ptable_7 <- cbind(ptable_7, "P value" =P)
R2Polr_Model_7 <- PseudoR2(Polr_Model_7, which="all")
print(Polr_Model_7)
print(SumPolr_Model_7)
print(ptable_7)
print(R2Polr_Model_7)

library(MASS)
Polr_Model_8 <- polr(T2D_ord ~ SEX + BMI,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_8 <- summary(Polr_Model_8)
ptable_8 <- coef(summary(Polr_Model_8))
P <- pnorm(abs(ptable_8[, "t value"]), lower.tail=FALSE) * 2
ptable_8 <- cbind(ptable_8, "P value" =P)
R2Polr_Model_8 <- PseudoR2(Polr_Model_8, which="all")
print(Polr_Model_8)
print(SumPolr_Model_8)
print(ptable_8)
print(R2Polr_Model_8)

library(MASS)
Polr_Model_9 <- polr(T2D_ord ~ SEX + C1 + C2 + GRSwei_G,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_9 <- summary(Polr_Model_9)
ptable_9 <- coef(summary(Polr_Model_9))
P <- pnorm(abs(ptable_9[, "t value"]), lower.tail=FALSE) * 2
ptable_9 <- cbind(ptable_9, "P value" =P)
```

```
R2Polr_Model_9 <- PseudoR2(Polr_Model_9, which="all")
print(Polr_Model_9)
print(SumPolr_Model_9)
print(ptable_9)
print(R2Polr_Model_9)

library(MASS)
Polr_Model_9.2 <- polr(T2D_ord ~ SEX + BMI + C1 + C2 + GRSwei_G,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_9.2 <- summary(Polr_Model_9.2)
ptable_9.2 <- coef(summary(Polr_Model_9.2))
P <- pnorm(abs(ptable_9.2[, "t value"]), lower.tail=FALSE) * 2
ptable_9.2 <- cbind(ptable_9.2, "P value" =P)
R2Polr_Model_9.2 <- PseudoR2(Polr_Model_9.2, which="all")
print(Polr_Model_9.2)
print(SumPolr_Model_9.2)
print(ptable_9.2)
print(R2Polr_Model_9.2)

############################################################
#unweighted GRS (genome-wide significance level)
############################################################
library(MASS)
Polr_Model_10 <- polr(T2D_ord ~ GRSunw_G,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_10 <- summary(Polr_Model_10)
ptable_10 <- coef(summary(Polr_Model_10))
P <- pnorm(abs(ptable_10[, "t value"]), lower.tail=FALSE) * 2
ptable_10 <- cbind(ptable_10, "P value" =P)
R2Polr_Model_10 <- PseudoR2(Polr_Model_10, which="all")
print(Polr_Model_10)
print(SumPolr_Model_10)
print(ptable_10)
print(R2Polr_Model_10)

library(MASS)
Polr_Model_11 <- polr(T2D_ord ~ SEX + BMI,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_11 <- summary(Polr_Model_11)
ptable_11 <- coef(summary(Polr_Model_11))
P <- pnorm(abs(ptable_11[, "t value"]), lower.tail=FALSE) * 2
ptable_11 <- cbind(ptable_11, "P value" =P)
R2Polr_Model_11 <- PseudoR2(Polr_Model_11, which="all")
print(Polr_Model_11)
print(SumPolr_Model_11)
print(ptable_11)
print(R2Polr_Model_11)

library(MASS)
Polr_Model_12 <- polr(T2D_ord ~ SEX + C1 + C2 + GRSunw_G,
```

```
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_12 <- summary(Polr_Model_12)
ptable_12 <- coef(summary(Polr_Model_12))
P <- pnorm(abs(ptable_12[, "t value"]), lower.tail=FALSE) * 2
ptable_12 <- cbind(ptable_12, "P value" =P)
R2Polr_Model_12 <- PseudoR2(Polr_Model_12, which="all")
print(Polr_Model_12)
print(SumPolr_Model_12)
print(ptable_12)
print(R2Polr_Model_12)
```

library(MASS)

```
Polr_Model_12.2 <- polr(T2D_ord ~ SEX + BMI + C1 + C2 + GRSunw_G,
method ="logistic",data=T2D_Dataset_R)
SumPolr_Model_12.2 <- summary(Polr_Model_12.2)
ptable_12.2 <- coef(summary(Polr_Model_12.2))
P <- pnorm(abs(ptable_12.2[, "t value"]), lower.tail=FALSE) * 2
ptable_12.2 <- cbind(ptable_12.2, "P value" =P)
R2Polr_Model_12.2 <- PseudoR2(Polr_Model_12.2, which="all")
print(Polr_Model_12)
print(SumPolr_Model_12)
print(ptable_12)
print(R2Polr_Model_12)
```

H.4: R syntax used to undertake meta-analysis

H.4.1: Syntax used to undertake meta-analysis based on P-values from Cox PH model HR

```
############################################################
#weighted GRS (nominal significance level)
############################################################
library(metap)
rMeta_cTTEb1 <- sumz(mPV_W05, MEweight, data=MEb_CModel_ALL_MER)
print(rMeta_cTTEb1)

############################################################
#unweighted GRS (nominal significance level)
############################################################
library(metap)
rMeta_cTTEb2 <- sumz(mPV_uW05, MEweight, data=MEb_CModel_ALL_MER)
print(rMeta_cTTEb2)

############################################################
#weighted GRS (genome-wide significance level)
############################################################
library(metap)
rMeta_cTTEb3 <- sumz(mPV_W08, MEweight, data=MEb_CModel_ALL_MER)
print(rMeta_cTTEb3)

############################################################
#unweighted GRS (genome-wide significance level)
############################################################
library(metap)
rMeta_cTTEb4 <- sumz(mPV_uW08, MEweight, data=MEb_CModel_ALL_MER)
print(rMeta_cTTEb4)
```

H.4.2: Syntax used to undertake meta-analysis based on OR from the logistic model

```
############################################################
library(rmeta)
rMeta_bLOGb1 <- meta.summaries(Est_W05, SE_W05, method="fixed",
names=Study_ID, logscale=FALSE, data=MEb_LModel_ALL_MER)
SUMrMeta_bLOGb1  <- summary(rMeta_bLOGb1)

############################################################
library(rmeta)
```

```
rMeta_bLOGb2 <- meta.summaries(Est_uW05, SE_uW05, method="fixed",
names=Study_ID, logscale=FALSE, data=MEb_LModel_ALL_MER)
SUMrMeta_bLOGb2  <- summary(rMeta_bLOGb2)

############################################################
library(rmeta)
rMeta_bLOGb3 <- meta.summaries(Est_W08, SE_W08, method="fixed",
names=Study_ID, logscale=FALSE, data=MEb_LModel_ALL_MER)
SUMrMeta_bLOGb3  <- summary(rMeta_bLOGb3)

############################################################
library(rmeta)
rMeta_bLOGb4 <- meta.summaries(Est_uW08, SE_uW08, method="fixed",
names=Study_ID, logscale=FALSE, data=MEb_LModel_ALL_MER)
SUMrMeta_bLOGb4  <- summary(rMeta_bLOGb4)
```