The Genetics of Participation and Some Aspects of Genetic Data



Stefania Benonisdottir
Linacre College
Nuffield Department of Medicine
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2022

Abstract

One of the major challenges in human genetic studies is to differentiate between results that reflect biological mechanisms and those that reflect data specific properties. If genetics play a role in determining who would participate in a study that includes DNA collection, the genotypes of the participants would not be representative of the target population. By its very nature, this bias cannot be investigated by comparing genotypes of participants and non-participants. In this thesis I introduce a new method that enables one to investigate participation bias using only the genetic samples of the participants. The method involves comparing shared and non-shared alleles of first-degree relatives and is sensitive only to direct genetic effects, i.e. there is no confounding with population stratification or indirect genetic effects. Applying this method to perform a genome-wide association analysis using data of first-degree relative pairs in the UK Biobank, I demonstrate simultaneously that the method works and there is a very substantial genetic component underling participation in UK Biobank. The demonstration is achieved by showing that a polygenic score constructed from the association analysis can, in a non-overlapping sample, significantly predict traits and behaviour that include educational attainment, BMI, age at first birth and participation in follow-up studies. The method includes three separate test statistics that each captures real participation effects, but are affected by data artefacts in a very differ-

ent manner. Consequentially, data artefacts could potentially be identified through comparison of the three statistics, an alternative utility of the method. The thesis also includes investigating the effectiveness of using principal components to adjust for population stratification, and the many possible causes of deviation from Hardy-Weinberg equilibrium. Together with studying participation bias, these investigations have a common goal of potentially improving the many analyses that are, or can be, performed using a large dataset such as the UK Biobank.

Declaration

The work presented in this thesis are results from my own original work, except for as specified in the text.

The thesis does not exceed the word limit of 50,000 words.

Acknowledgements

I thank my supervisor Professor Augustine Kong for giving me the opportunity to work on exciting projects with him, supporting my research interests and giving me guidance, comments and great support. I also thank my supervisor Professor Cecilia Lindgren for providing me with support, guidance and advice throughout my studies and for introducing me to her excellent Lindgren group. I thank Laura Wittemans, Teresa Ferreira, Saskia Reibe-Pal, Christoffer Nellaker, Helen Long, Jenny Censin, Samvida Venkatesh, Emil Jørsboe and Duncan Palmer for many discussions about my projects. I thank Alexander Young for discussions on the material in chapter 3 and 4. I thank Snaedis Kristmundsdottir and Erna Ivarsdottir for our many discussions about statistical genetics. I thank Geoffrey Ferrari in the BMRC team for his technical assistance. Special thanks to Mummi, Benoni, Andri, my parents and my siblings for all their support.

During my studies I was supported by the Li Ka Shing Foundation and the Goodger and Schorstein scholarship. The research presented in this thesis has been conducted using the UK Biobank Resource under application 11867. The computational aspects were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

Publications or submitted publications arising from this D.Phil

Related to this thesis:

Benonisdottir, S. & Kong, A. (2022). The genetics of participation: method and analysis. *bioRxiv*, doi: 10.1101/2022.02.11.480067. *Under review.*

Young, A. I., **Benonisdottir, S.**, Przeworski, M., & Kong, A. (2019). Deconstructing the sources of genotype-phenotype associations in humans. *Science*, 365: 1396-1400, doi: 10.1126/science.aax3710.

Other:

Young, A. I., Nehzati, S. M., **Benonisdottir, S.**, Okbay, A., Jayashankar, H., Lee, C., ... & Kong, A. (2022). Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature Genetics* 54: 897-905, doi: 10.1038/s41588-022-01085-0

Contents

1 Introduction **1**

1.1 Background 1

1.2 Properties of Genetic Information 3

1.3 Genetic Data of Close Relatives 4

1.4 Genetic Datasets 7

1.5 Genome-Wide Association Studies (GWAS) 8

1.6 Participation Bias 12

2 Hardy-Weinberg Equilibrium **19**

2.1 Introduction 19

2.2 Methods 22

2.2.1 Data 22

2.2.2 HWE Statistic and Subgroups 24

2.3 Results 26

2.4 Conclusions 27

3 Behaviour of Principal Components **30**

3.1 Introduction 30

3.2 Methods 32

3.2.1 Data 32

3.2.2 Computing PCs and Eigenvalue-Equivalents 34

3.2.3 Computing Polygenic Scores and Scaled Variance 35

3.3 Results 37

3.4 Conclusions 39

4 The Genetics of Participation: Method and Analysis **41**

4.1 Introduction 41

4.2 The Three Principles of Participation Bias 43

4.2.1 The First Principle of Genetically Induced Participation Bias . . 44

4.2.1.1 Participating Parent-Offspring Pairs 46

4.2.1.2 Participating Sibling Pairs 47

4.2.2 The Second Principle of Genetically Induced Participation Bias 51

4.2.3 The Third Principle of Genetically Induced Participation Bias . 52

4.3 Primary Participation GWAS in UK Biobank 54

4.3.1 Methods 55

4.3.1.1 Genotype Data 55

4.3.1.2 Identifying Relatives 57

4.3.1.3 Inferring IBD Segments 59

4.3.1.4 Uncertainties Around Recombination Events 63

4.3.1.5 Inferring Shared Allele 64

4.3.1.6 Additional Quality Control 65

4.3.1.7 Expected IBD Sharing and Phasing Errors 66

4.3.1.8 IBD Fraction Threshold 68

4.3.1.9 Test Statistics 69

4.3.2 Detection of Data Artefacts 70

4.3.3 Results 79

4.3.3.1 Within-Sample Replication 83

4.3.3.2 Adjusting for Educational Attainment (EA) 87

4.3.3.3 Sex-Specific Analysis 88

4.3.3.4 Differences Between Siblings and the Unrelated 91

4.3.3.5 Genetic Correlations 91

4.4 Studying Relationships Between Various Frequency Differences by Simulations 93

4.5 Conclusions 103

5 Discussions **105**

A Supplementary Methods **133**

A.1 Phenotypes 133

A.1.1 Educational Attainment (EA) 134

A.1.2 Height 135

A.1.3 Body Mass Index (BMI) 136

A.1.4 Age at First Birth (AFB) 137

A.1.5 Number of Children (N.children) 138

A.1.6 Number of Siblings (N. siblings) 139

A.1.7 High-Density Lipoprotein (HDL) Cholesterol 140

A.1.8 Glycated Haemoglobin (HbA1c) 142

A.1.9 Vitamin D 143

A.1.10 Glucose 144

A.1.11 Grip Strength 145

A.1.12 Lipoprotein A 146

A.1.13 Sex Hormone Binding Globulin (SHBG) 148

A.1.14 Testosterone 149

A.1.15 Physical Activity Study Participation 150

A.1.16 Dietary Study Participation 151

A.2 Primary Participation Polygenic Scores 152

A.3 Within-Study Replication 153

A.4 LD Score Regression 153

B Supplementary Tables **156**

C Supplementary Figures **163**

List of Figures

1.1 IBD sharing of first-degree relatives 6

1.2 Participation bias in genetic studies. 13

1.3 Collider bias: Berkson's paradox 16

3.1 Behaviour of principal components. 38

4.1 The first principle of genetically induced participation bias 46

4.2 Parental transmission to sibling pairs 49

4.3 Flowchart summarizing the stepwise procedure for testing genetic variants for association with primary participation in UKBB 56

4.4 First-degree relatives included in analysis 58

4.5 Fraction of sibling pairs sharing sequence variants IBD=0, IBD=1 and IBD=2 62

4.6 Trimming sequence variants at the beginning and end of IBD regions. . 63

4.7 Inferring shared allele for IBD=1 when both individuals are heterozygous for the target sequence variant 64

4.8 Deviations from the expected fraction of the shared allele 68

4.9 Test statistics and allele frequency 73

4.10 Test statistic distribution 75

4.11 Variants that did not fulfil additional QC filters 78

4.12 Manhattan figures: Primary participation GWAS in the UKBB 81

4.13 QQ-plots: Primary participation GWAS in the UKBB 82

4.14 Primary participation polygenic scores ($pPGS$) computed with different weights 85

4.15 Primary participation polygenic score ($pPGS$) and educational attainment 89

4.16 Primary participation polygenic score ($pPGS$) and sex-specific analysis 90

4.17 Difference between IBD shared and IBD non-shared alleles 97

4.18 Average values: Difference between IBD shared and IBD non-shared alleles 98

4.19 Relative allele frequencies as a function of participation rate 100

4.20 Relative allele frequencies in different groups of individuals and genetic segments 102

A.1 Educational attainment (EA), year of birth (YOB) and sex 135

A.2 Height, age and sex 136

A.3 Body mass index (BMI), age and sex 137

A.4 Age at first birth (AFB) and year of birth (YOB) 138

A.5 Number of children (N. children), year of birth (YOB) and sex 139

A.6 Number of siblings (N. siblings), year of birth (YOB) and sex 140

A.7 High-density lipoprotein (HDL cholesterol), age and sex 141

A.8 Glycated haemoglobin (HbA1c), age and sex 143

A.9 Vitamin D, age and sex 144

A.10 Glucose, age and sex 145

A.11 Grip strength, age, height and sex 146

A.12 Lipoprotein A, age and sex 147

A.13 Sex hormone binding globulin (SHBG), age and sex 149

A.14 Testosterone, age and sex 150

C.1 Participants with a participating sibling and participants with a non-participating sibling 164

List of Tables

2.1 HWE statistics computed for two version of genotypes and 6 different subgroups 27

4.1 Brother-sister division 59

4.2 Test statistics and allele frequency 71

4.3 Primary participation polygenic score ($pPGS$) and phenotypes. 86

4.4 Genetic correlations between primary participation and other phenotypes 93

B.1 T^*_{PO} primary participation polygenic score ($pPGS$) and phenotypes . . 157

B.2 T^*_{SIB1} primary participation polygenic score ($pPGS$) and phenotypes . . 158

B.3 T_{SIB20} primary participation polygenic score ($pPGS$) and phenotypes . 159

B.4 T_{PO} primary participation polygenic score ($pPGS$) and phenotypes . . 160

B.5 T_{SIB1} primary participation polygenic score ($pPGS$) and phenotypes . . 161

B.6 Educational attainment and association with other phenotypes 162

Chapter 1

Introduction

1.1 Background

One of the major challenges in human genetic studies is to differentiate between results that reflect biological mechanisms in the underlying population and those that reflect data specific properties. Genetic studies require genetic data and that data need to be sampled, processed and genotyped. At each step, biases and errors might occur[119], and those can impact various downstream analysis in a different manner[12, 80]. Population inferences from genetic datasets rely on the assumption that the data is close enough to a random sample that it is broadly representative of the intended study population[119]. Often, the belief, or hope, is that the deviations from random sampling would have a negligible impact on the analyses performed[33]. With the latter, there is sometimes a lack of appreciation that some imperfections of the data set, whether it is related to sampling or data artefacts, could have little effect on some analyses, but would lead to seriously misleading results for other analyses if ignored. Random sampling is difficult in general and particularly so with genetics studies. Some people might be easier to invite than others and participation in genetic studies is usu-

ally a voluntary act[20, 81, 61].

It has been shown that participation bias can lead to serious problems in various sample surveys[8, 14]. To investigate the impact of such bias on inferences from genetic datasets, one first has to be able to identify the bias itself. That is a particularly challenging problem as genotypes of non-participants are by definition unavailable. This problem is addressed in chapter 4 of this thesis where I introduce a new method that enables one to investigate participation bias from the genetic samples alone. The method is based on an idea by Professor Augustine Kong, and involves comparing IBD shared alleles to IBD not-shared alleles of participating first-degree relatives. Chapters 2 and 3 are concerned with how testing for deviations from Hardy-Weinberg equilibrium (HWE) and computation of principal components can provide us with information about data specific properties and/or the underlying population.

The data applications presented in in this thesis are based on the UK Biobank (UKBB), a large cohort of ∼500,000 individuals from across the United Kingdom (UK)[20]. Uffelmann et al. (2021) note that the majority of human genetic studies are performed with publicly available datasets. Even when in-house data are available, they are usually co-analysed with publicly available resources[119]. The UKBB is a good example of this. It is currently being utilised by over 28 thousands of researchers from 98 countries[120] and increased understanding about its data specific properties is therefore of great relevance for the scientific community.

To provide a foundation for subsequent chapters, I first discuss specific properties of genetic information that make genetic datasets distinct from other sample surveys. I then discuss genome-wide association studies (GWAS) and explain how many sources

can contribute to the observed genotype-phenotype associations. Lastly, I discuss participation bias, previous attempts to quantify it in genetic studies and some of its consequences for genetic datasets.

1.2 Properties of Genetic Information

Genetic information is stored in a linear sequence of bases of deoxyribonucleic acid (DNA) and the DNA sequences are divided into chromosomes[76, p.8-9]. Each chromosome can be further divided into loci where a locus refers to a specific position on a chromosome. An allele is a base-pair sequence at a locus that is known to be polymorphic[136, p.10]. Polymorphic loci are often referred to as 'sequence variants' or 'variants' for short. In humans, the most common sequence variants are single base variations called single nucleotide polymorphisms (SNPs)[136, p.54]. Humans carry 23 chromosomes in pairs, 22 autosomal chromosomes and one sex chromosome[76, p.9]. The two alleles at a locus, one inherited from the mother and the other from the father, determine the genotype. A biallelic sequence variant has two observed alleles while a multiallelic sequence variant has three or more observed alleles. For biallelic sequence variants, the more common allele is referred to as the major allele and the less common allele is referred to as the minor allele. Individuals with two copies of the minor allele are said to be homozygous carriers of the minor allele, individuals with two copies of the major allele are homozygous carriers of the major allele and individuals with one copy of each allele are heterozygous carriers[136, p.10].

In genetic datasets, the process of acquiring the genotype information from a biological sample is called 'genotyping', and the samples are said to be 'genotyped'[64].

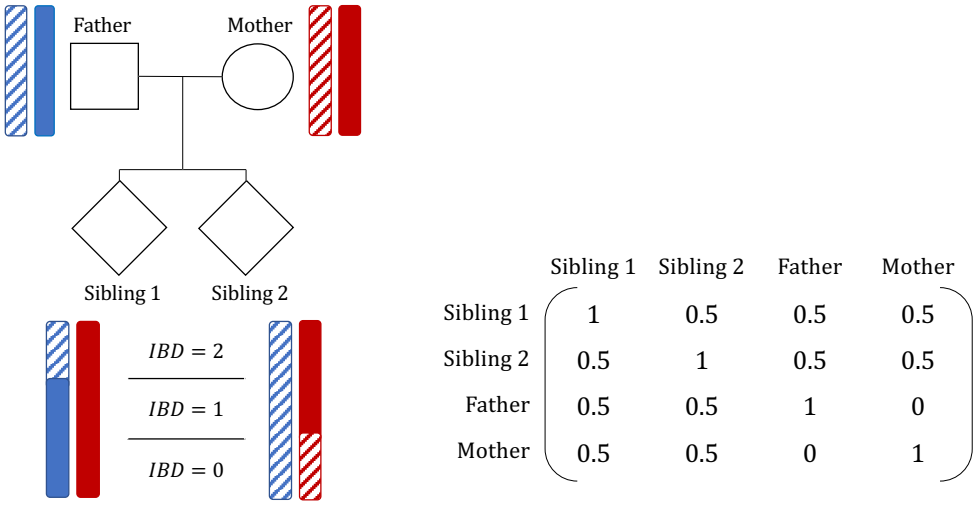
Human DNA is approximately 3.3 billion base-pairs in total length[136, p.2]. Genotyping is commonly performed for only a fraction of the genome using genotyping arrays for predefined polymorphic loci[122], however large-scale exome sequencing and whole-genome sequencing efforts are also ongoing[36, 35]. A good example of this is the UKBB. In the UKBB, the genotyping was initially performed with genotyping arrays of ∼800 thousand variants[20]. Later, exome-sequencing was added[121] and more recently, whole-genome sequencing as well[36]. Genotyping errors occur when the genotype determined by the molecular analysis, differs from the true genotype[12]. Data artefacts are a particular concern in genetic studies as they can easily lead to misleading results[103, 126]. High-profile examples include the longevity study of Sebastiani et al. (2010), in which reported associations were later found to be the result of an error-prone chip called 610-Quad[21, 103].

1.3 Genetic Data of Close Relatives

Although genetics is by definition the study of heredity and the variation of inherited characteristics[39, p.2], most studies published in the last fifteen years associating genetic variants with various human traits were performed with large sample of individuals who are not closely related to each other[119, 124]. Indeed, when the data set includes close relatives, it is often considered as a problem instead of an asset, and sometimes the close relatives are simply pruned away in the analyses[66]. By contrast, my research in general, and the method presented in chapter 4 in particular, often involves utilising information that is only available when one has close relatives within the dataset. For that reason, a brief introduction to genetic data of close relatives is given here.

As genetic information is transmitted from one generation to the other, close relatives are expected to have more similar DNA sequences than distant relatives. At a given locus, any two individuals may carry the same alleles, that is alleles identical by status (IBS), as there are many copies of an allele within a population. However, close relatives are more likely to do so as they may be carrying alleles lying on a long stretch of DNA that they both inherited from a recent common ancestor, that is alleles that are identical by descent (IBD) (see figure 1.1). As Thompson (2013) notes, '*there is no absolute measure of IBD*', IBD always has to be defined relative to a specified reference[113, p.301]. Palamara et al. (2015) define IBD segments '*as contiguous chromosomal regions in which the most recent common ancestor (MRCA) for two sampled chromosomes is unchanged*'[83, p.775]. The fewer the generations separating two IBD segments, the greater their expected length[59]. Usually IBD implies IBS, however mismatched sequences in IBD segments can occasionally occur due to rare events such as de-novo mutations[57]. As these sequence mismatches are related to meiotic events, they tend to be fewer in IBD segments of close relatives than distant relatives[83]. Relatives are individuals that share a not-too-distant common ancestor. Apart from identical twins, first-degree relatives, which include parent-offspring pairs and sibling pairs, are the most closely related. Ignoring inbreeding, first-degree relatives share, in expectation, one half of the autosomal genome[6, p.91]. Because there are two autosomal chromosomes, paternally and maternally inherited, for each individual, this means that a first-degree relative pair shares one copy of the chromosome IBD in expectation (see figure 1.1). Ignoring rare events, because a parent transmits one of his/her two chromosomes to a child, the parent-offspring pair always shares one allele IBD precisely without variation. The situation is more complicated for sibling pairs. At any particular locus, we can think of the number of alleles shared between siblings as determined

by a fair coin toss[134]. As a result, they share 0, 1, or 2 alleles IBD, referred to as the IBD state, with probabilities 1/4, 1/2, and 1/4, respectively.



	Sibling 1	Sibling 2	Father	Mother
Sibling 1	1	0.5	0.5	0.5
Sibling 2	0.5	1	0.5	0.5
Father	0.5	0.5	1	0
Mother	0.5	0.5	0	1

Figure 1.1: IBD sharing of first-degree relatives. The pedigree on the left shows how genetic segments are shared identical by descent (IBD) within a nuclear family. Sibling pairs can share 0, 1 or 2 alleles IBD while parent-offspring pairs always share 1 allele IBD. The correlation matrix on the right shows the expected genetic correlation between the members of the nuclear family, assuming random mating in the parents' generation.

Recombination occurs when there is an exchange of chromosomal segments between the two DNA strands of a given parent during the prophase of meiosis[136, p.9]. When there are no recombination events between two loci in both the paternal and maternal meiosis for both siblings, the two loci would have the same IBD state for that particular sibling pair. Locations where a recombination event occurs are where the IBD state changes (see figure 1.1). Because the recombination process involves a certain degree of randomness, there are two important consequences to note here. Firstly, even though sibling pairs share in expectation one-half of the autosomal genome, some pair would share more and some would share less. Secondly, for any specific locus, while the expected IBD sharing is one allele, the actual number is either 0, 1, or 2.

The method presented in chapter 4 requires determining the realised IBD state and the shared allele when IBD=1, for every locus in the genome for every sibling pair being studied, whenever possible. Inferring IBD shared alleles of sibling pairs has many implications beyond investigating the genetics of primary participation. For example, during my D.Phil I contributed to a paper by Young et al. (2022), '*Mendelian imputation of parental genotypes improves estimates of direct genetic effects*', in which information about IBD sharing between siblings was utilised to impute missing parental genotypes[135].

1.4 Genetic Datasets

Following a drastic decrease in genotyping costs, multiple genetic datasets of various design have been established [20, 81, 61]. Some of those datasets are disease-focused cohorts[41], some aim to represent the general population[81, 25] and other are sampled as a consumer product[68]. Many of the existing genetic datasets are large-scale in the sense that they include genetic samples from more than 100,000 individuals. Examples include the UKBB with ∼500 thousand samples[20], the Norwegian Mother and Child study (MoBA) with ∼285 thousand samples[69], the deCODE study with ∼166 thousand samples[49] and the 23andMe dataset with ∼2.5 million samples[91]. Meng (2018) proposes that the 'bigness of big data for population inferences' should be defined from the relative sample size rather than absolute values[74]. Most if not all genetic datasets are a subset of the intended study population. However, the relative sample size will depend on how the intended study population is defined, which is not always straight forward. Hartl and Jones (1998) define a population as a group of individuals that lives within a prescribed geographical area. They further note that

the area can be of any size but is usually considered to be the area in which individuals are likely to find mates[39, p.628]. For population based cohorts, one could for example define the intended study population as the individuals living within the country or region where the study takes place. E.g. participants in the UKBB are 0.75% of the total UK population[20, 26], participants in MoBA are 5% of the Norwegian population[69, 109], and participants in the deCODE study are 46% of the Icelandic population[5]. The intended study population could also be regarded as the group of individuals that were invited to participate. The UKBB invited approximately 9 million people from across the UK, and of those, 5.5% participated[33]. The MoBA study invited pregnant women attending a routine ultrasound examination in 1999-2009 and from those 41% participated[69]. For the China Kadoorie Biobank, a population cohort study with 500 thousand participants, the estimated participation rate was 30%[25]. For consumer-based datasets like the 23andMe dataset, it is however very challenging to quantify the population fraction and participation rates.

1.5 Genome-Wide Association Studies (GWAS)

In addition to genetic information, multiple large-scale genetic datasets also include extensive phenotype information about participants[25, 81, 20, 49]. Such datasets are commonly utilised to perform GWAS which have become the default for investigating genetic contribution to phenotypes[123, 77]. The term phenotype refers to any observable characteristics of an individual[76, p.6] such as height[131], BMI[107], disease status[102] or coffee consumption[108]. GWAS test each sequence variants for association with the phenotype of interest with the study unit being a genotyped individual with phenotype information. This is done simultaneously for hundreds of thousands or

millions of variants across the genome of multiple study units in a given dataset[119]. Since the first GWAS on age-related macular degeneration in 2005[56], almost 3,700 GWAS have been published and thousands of genotype-phenotype associations have been identified[77]. Uffelmann et al. (2021) note that previous GWAS have shown that most phenotypes are influenced by thousands of variants that individually have small effects[119].

The estimated GWAS effect of a sequence variant can be decomposed into direct effect, indirect effect and confounding effect[133]. Direct effects are causal effects of inherited variants. A genetic variant could for example affect adult height via the influence on growth hormone production. However, as Young et al. (2019) point out, many of the causal pathways 'are neither simple or direct'[133]. An indirect effect, which includes what some refer to as a genetic nurturing effect[60], occurs when a phenotypic value of a proband is affected by a genotype of another individual through shared environment[11]. Here the term 'proband' refers to an individual with phenotype information. As close relatives, such as parent-offspring pairs and siblings, are expected to share a substantial fraction of alleles IBD, a proband has a high probability of being a carrier of variants that have indirect effects through genotypes of close relatives. As a result, traditional GWAS estimates might incorporate indirect genetic effects of relatives when the genotyped proband is the study unit. An example could be a genetic variant that affects maternal uterine environment through the genotype of the mother. A heterozygous mother has a 50% chance of passing this variant on to her offspring and a homozygous mother has a 100% chance of passing the variant on to her offspring. As a result, the indirect effect of the variant could potentially be captured when performing GWAS with phenotyped newborns as study units.

If confounding has been completely accounted for, a typical GWAS estimates the sum of the direct and indirect effects[133]. Many sources can attribute to indirect effects, e.g. parents, grandparents and siblings. Kong et al. (2018) point out that the direct effect can be estimated by the difference between the effect of the transmitted parental alleles and non-transmitted parental alleles[60]. As sample sizes grow, genetic datasets will inevitably include a notable number of close relatives. This is already the case for the UKBB in which approximately 30% of the individuals have at least one close relative within the dataset[20]. This growth can and has facilitated family based association analyses such as sibling GWAS that compare within-family effects to between-family effects by defining the study unit as a pair of close relatives that are both genotyped and phenotyped[15]. Recently, Young et al. (2022) estimated direct effects by imputing missing parental genotypes for 39,614 individuals in the UKBB that had at least one genotyped sibling/parent. They estimated the educational attainment (EA) polygenic score (PGS) for non-transmitted alleles to have an effect on EA that was half of the effect of the transmitted PGS[135]. In comparison, by using genotyped probands with at least one genotyped parent in the deCODE study, Kong et al. (2018) demonstrated that the EA PGS for non-transmitted parental alleles had an estimated effect on EA that was roughly a third of the transmitted PGS in the Icelandic dataset[60]. Young et al. (2022) point out that the difference between those two estimates implies that indirect effects, population stratification and assortative mating contribute more to the correlation between the EA PGS and EA in the UK than in Iceland[135].

Linkage disequilibrium (LD), or genotype correlations, between physically distant sequence variants or variants on different chromosomes (non-local LD), can be indicative of factors such as population stratification and phenotypic assortative mating which

can cause confounding in GWAS. It is well established that population stratification is a major cause of confounding in GWAS[133]. In the presence of phenotypic differences between sub-populations, population stratification can lead to false positive GWAS results as variants of different sub-population frequencies may seem to associate with the phenotype independent of causality. Phenotypic assortative mating leads to non-local LD between variants with direct and indirect effects[133]. Phenotypic assortative mating may not directly induce false positive results but can inflate estimated effects when performing GWAS for the phenotype under assortative mating. Accounting for confounding effects is not a trivial task and there is a continuous discussion as to what extent existing methods are able to correct for population stratification and to what extent they are correcting out real phenotypic effects[40].

GWAS results can be used for various subsequent analysis such as for computing PGS or LD score regression. PGS is a single measure of an individual's genetic susceptibility to a phenotype[72] and are commonly computed by summing over allele counts weighted by GWAS effect sizes[115]. It has been suggested that PGS could have an useful application in precision medicine[54, 48] and perhaps more controversially, PGS have been discussed in the context of policy making[92, p.181] and embryo selection[52]. LD score regression involves regressing GWAS summary statistics on LD scores which are defined by Bulik-Sullivan et al. (2015) as single measures of the amount of genetic variation that a given sequence variant tags[17]. The estimated intercepts from LD score regression are often used to quantify inflation due to confounding effects such as population stratification and cryptic relatedness[17] and the estimated slopes are commonly used to investigate the shared genetic basis of different phenotypes[16].

1.6 Participation Bias

Participation bias refers to the phenomena when some individuals of the intended study population are more likely to be included in the sample than others. It can be caused by a passive component, that is when some individuals are more likely to be invited than other, or an active component, that is when some individuals are more likely to participate when invited than other. It goes without saying, that whatever the cause, passive, active or a combination of both, the end result will be a sample that is not representative of the intended study population (see figure 1.2). Participation bias goes by many names in the literature such as 'selection bias'[80, 81], 'sampling bias'[65] and 'ascertainment bias'[4]. The term 'participation bias' is chosen here but it should be emphasised that it encompasses both the passive component and the active component of the bias.

Some serious consequences of unrepresentative data have been well documented in the past. In his 1946 paper, *'Limitations of the application of fourfold table analysis to hospital data'*, statistician and physician Joseph Berkson demonstrated with a hypothetical example how diseases that were uncorrelated within the population could seem correlated in hospital data[10]. This observation has later been referred to as 'Berkson's paradox'[105]. Berkson explained in 1955[105] that the inspiration for his paper came from an autopsy-based study from 1929 in which the statistician and biologist Raymond Pearl observed an association between tuberculosis and cancer[88]. In his study, Pearl found that tuberculosis lesions occurred over twice as often in patients that had died from non-cancer causes than in patients that had died from cancer[88]. Later in the same year, Pearl noted that a potential cause for this observed negative

correlation was that people with cancer died before there was time for them to develop tuberculosis and vice versa[87]. More recent high-profile consequences of unrepresentative data include the misleading election poll results for the United States (US) 2016 presidential election and the European Union (EU) referendum vote[8]. Another recent example is reported in a study by Bradley et al. (2021) which demonstrated how unrepresentative big surveys significantly overestimated the COVID-19 vaccine uptake in the US. In the paper it was emphasised that not only did a survey of 250 thousand individuals significantly overestimate the US vaccine uptake, those inflated estimates also had very narrow confidence intervals due to the large sample size[14]. These results highlight what Meng (2018) refers to as the Big Data Paradox: '*the bigger the data, the surer we fool ourselves*'[74, p.702].

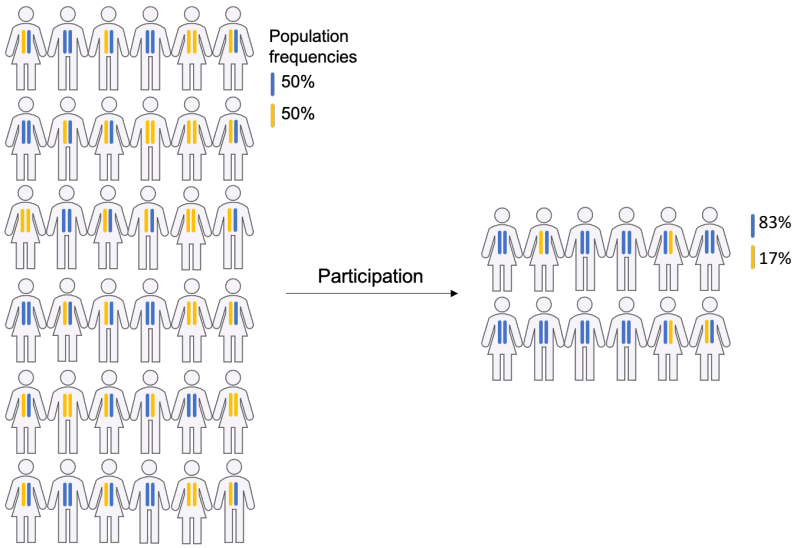


Figure 1.2: Participation bias in genetic studies. An illustrative example of participation bias leading to an unrepresentative genetic dataset.

It is widely acknowledged that most if not all genetic datasets are not randomly sampled[79] and there are multiple reports of them being unrepresentative. Nilsen

et al. (2009) report that young mothers and smokers are under-represented in the MoBA study while those who take multivitamin and folic acid are over-represented [81]. Krokstad et al. (2012) report that mortality rate is lower among participants in the Norwegian HUNT study and that participants tend to be of higher socio-economic status and less likely to suffer from diabetes mellitus and psychiatric disorders[61]. Fry et al. (2017) report that UKBB participants are more likely to be female, older and less likely to smoke[33]. While phenotypic differences might indicate genotypic differences, it is very difficult to quantify that relationship from phenotype data alone.

Multiple studies have investigated the genetic contribution to study engagement, which is referred to here as 'secondary participation'. In 2016, Martin et al. reported a positive correlation between a PGS for schizophrenia and not completing questionnaires and showing up for follow-up visits in the Avon Longitudinal Study of Parents and Children (ALSPAC)[73]. Two years later, Taylor et al. demonstrated how ALSPAC participants who engaged in follow-up assessments tended to have higher PGS for EA, agreeableness and openness and lower PGS for smoking initiation, BMI, neuroticism, ADHD and depression[111]. Many secondary participation GWAS have been performed in the UKBB[1, 118, 75]. Adams et al. (2020) identified multiple sequence variants associating with 'sharing e-mail information for follow-up' and 'participation in follow-up surveys of mental health'[1]. Mignogna et al. (2022) demonstrated how sequence variants are associated with non-response behaviour in survey questionnaires (answering 'I don't know' or 'Prefer not to answer')[75]. Tyrrell et al. (2021) performed a GWAS for 'filling out additional questionnaires' and 'agreeing to wear a physical activity monitor'. They identified 32 genome-wide significant signals and claim that their results suggest that educational attainment, menarche at an older age and taller

stature 'increase participation', while increased adiposity, neuroticism, dyslipidaemia, Alzheimer's and schizophrenia 'reduce participation'. They further demonstrated positive genetic correlations between secondary participation in the UKBB and secondary participation in the ALSPAC study[118]. In summary, these studies have shown that genetic factors play a part in secondary participation. However, they only include individuals that have already contributed a DNA sample and do therefore not address the potential genetic differences between those who participated in the first place and those who did not.

By its very nature, primary participation is difficult to investigate as genotypes of non-participants are unavailable. It is therefore striking to see that one can investigate the bias of sampled genotypes based on themselves alone. In a recent study, Pirastu et al. (2021) investigated sex-differential participation bias by performing GWAS on sex in a study of 3.3 million individuals. They propose that the genetic component of study participation can be investigated by performing a GWAS on non-heritable traits and as autosomal variants should be of the same population frequency among men and women, associations between genetic variants and sex are indicative of sex-differential participation bias[91]. A major strength of this approach is that it only requires the genotypes of participants. It cannot however identify variants that affect primary study participation of both sexes in a similar manner.

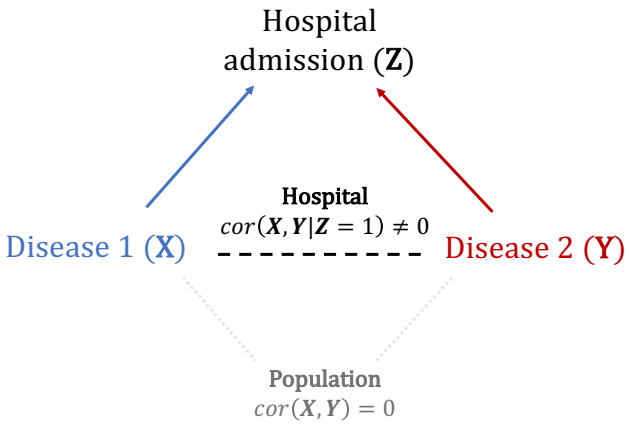


Figure 1.3: Collider bias: Berkson's paradox. Illustration of the hypothetical example demonstrated by Joseph Berkson (1946) in his paper *'Limitations of the application of fourfold table analysis to hospital data'* [10]. In his example, Berkson showed how two diseases that were uncorrelated in the population and influenced the probability of hospital admission independently, had a non-zero correlation within the hospital population.

If genetic factors are associated with primary participation, then genotypes of participants won't be representative of the population. This can have various consequences. For example, Albrechtsen et al. (2010) report that the choice of individuals used in the discovery process for SNP chips can influence measures of human diversity[4]. Recently, many have pointed out that PGS provide limited phenotype prediction in ancestry groups that are underrepresented in the GWAS from which the PGS weights originate from[72, 63]. Furthermore, by investigating PGS prediction within a group that shares a single common ancestry, Mostafavi et al. (2020) demonstrated that PGS prediction for a given individual can also depend on his/her similarities to the GWAS cohort in terms of sex, age and socioeconomic status[79].

Munafò et al. (2018) note that while it is widely appreciated that participation bias will impact prevalence estimates, it is often assumed that it does not have a strong effect on observed associations. They point out that such assumptions ignore the pos-

sibility of a collider bias[80]. Collider bias refers to the phenomena when two variables (X and Y) independently influence a third variable, (Z), and conditioning upon Z distorts the estimated association between X and Y[80]. Berkson's paradox is a special case of collider bias (see figure 1.3)[1][127]. Munafò et al. (2018) demonstrated with simulations that participation bias can lead to spurious genotype-phenotype association result[80]. Through the same mechanism, if a variant affects participation, collider bias could result in the sample correlation between maternally and paternally inherited alleles being different from the population correlation. In the same manner, sample correlations between different variants that influence participation could also be affected. That is, genetically induced participation bias has the potential of impacting the whole LD structure of a genetic dataset.

Instead of wondering what analyses could be affected by participation bias, one might consider asking what analyses would not be affected by it. I can hardly think of any. Indeed, it is more a question of the degree of the impact, i.e. some analyses might be minimally affected while others would be affected in a pivotal manner. For the latter, the potential problem has a very wide range, i.e. it is not necessarily limited to collider bias. Arguably, for a given dataset, understanding the magnitude and nature of the participation bias is a prerequisite to performing a proper analysis of its consequences. The research in chapter 4, the most important part of this thesis, is my effort to work towards that direction. In different ways, the research and results described in chapters 2 and 3 are also about learning more about the data, with the dual purpose of obtaining new results. Chapters 2 to 4 can be considered as separate research analysis and could be read independently of each other. However, while not planned, research in chapter 2 has helped me and Professor Kong in understanding

[1]One could also say that 'collider bias' is a general case of Berkson's paradox.

how genotyping artefacts could complicate the analyses of the method introduced in chapter 4. In the opposite direction, the research in chapter 4 leads to the recognition of that, in addition to data artefacts and population properties such as non-random mating, participation bias can contribute to deviations from HWE in the sample.

Chapter 2

Hardy-Weinberg Equilibrium

2.1 Introduction

When mating is random in an infinitely large population, the maternally inherited and paternally inherited alleles are uncorrelated. That zero correlation results in a greatly simplified relationship between genotype frequencies and allele frequencies as the expected genotype frequencies will be the same as what would be expected under random combination of alleles[27, p.34]. This principle was shown for any value of allele frequency by Hardy and Weinberg in 1908 and is known as the Hardy-Weinberg principle or the Hardy-Weinberg equilibrium (HWE)[27, p.35]. The Hardy-Weinberg principle can be easily verified by considering a biallelic SNP for which we code the two possible allele types as 0 and 1. Under random mating, the unordered genotypes can be regarded as the sum of two independent trials where the population allele frequency of the allele coded as 1, p, is the probability of 'success'. It follows that the expected genotype frequencies corresponding to genotypes 0, 1 and 2 will be $f_0 = (1-p)^2$, $f_1 = 2p(1-p)$ and $f_2 = p^2$. Furthermore, the expected genotype variance will be $2p(1-p)$. When mating choice depends on genotype (non-random mating)[27, p.34],

the maternally inherited and the paternally inherited alleles will be correlated. When the correlation is positive then the proportion of heterozygous carriers will be lower than expected under random combinations of alleles, $f_1 < 2p(1-p)$ and if the correlation is negative, the vice versa holds, $f_1 > 2p(1-p)$.

Crow and Kimura (1970) note that the main reasons for departures from HWE in populations are 1) inbreeding and 2) assortative mating[27, p.38]. Inbreeding refers to mate choice patterns where mates tend to be more related than expected under random pairing of individuals[6, p.70]. Assortative mating occurs when phenotypes of mates are more alike than what to be expected under random pairing of individuals[6, p.70]. Both inbreeding and assortative mating will lead to a positive correlation between maternally and paternally inherited alleles, $f_1 < 2p(1-p)$. However, inbreeding will affect all genetic variants in this manner while assortative mating only affects variants that are associated with the phenotypes underlying the mate choice[6, p.70]. Outbreeding and disassortative mating are the opposites of inbreeding and assortative mating. Outbreeding refers to mates being less related than expected by random pairing and disassortative mating refers to phenotypes of mates being less alike than expected by random. Both will lead to a negative correlation between maternally and paternally inherited alleles, $f_1 > 2p(1-p)$.

Selection can occur within populations when genetic variants affect survival rates[3], mating mechanisms or fertility rates[58]. If p is the population allele frequency in the reference generation in a randomly mating population, then selection favouring heterozygotes will lead to $f_1 > 2p(1-p)$ in the next generation while selection disfavouring heterozygotes will lead to $f_1 < 2p(1-p)$. A population that is in HWE at birth can show significant deviations from HWE at adult ages when survival rates depend on

genotype[27, p.38]. E.g. in the extreme case of a lethal allele type, the genotypes in the adult population will only include homozygous carriers of the non-lethal allele. HWE is not expected to hold in real populations[6, p.70] as real populations are never infinite[39, p.658]. It is however considered a good approximation for large randomly mating populations under no selection[6, p.70]. This is assuming no population structure. If there is subdivision within a population, then drift can lead to deviations from HWE even if there is random mating within each subpopulation[6, p.70]. In the absence of major forces influencing allele frequencies between generations (mutation, selection, drift, migration), HWE is expected to hold for an infinitely large randomly mating population from one generation to the other[39, p.635].

In genetic datasets, deviations from HWE can be indicative of an underlying biology, such as assortative mating or selection, but they can also be reflecting data specific properties, such as technical artefacts[130] or how the data was sampled[136, p.77]. Balding (2005) points out that genotyping errors are often the most probable culprit when SNPs deviate from HWE in genetic data[6, p.70] and Turner et al. (2011) report that an excess of heterozygous carriers 'typically' indicates assay problems[117]. Although not initially emphasised by Hardy and Weinberg[27, p.34-39], this quality control feature is widely utilised. Testing departures from HWE has become a routine procedure in genetic studies[117] and large-scale genetic datasets systematically investigate HWE with the sole aim of detecting problematic SNPs[20].

In this chapter I compute HWE statistics for different subgroups in the UK Biobank (UKBB) and investigate them jointly. In populations, deviations from HWE can occur in subgroups if probability of ascertainment depends on genotypes. For instance, a disease-causing SNP might conform to HWE when the dataset is examined as whole

but deviate from HWE within the group of patients[136, p.77]. The subgroup division that I explore is not based on disease status but 'number of close relatives within the dataset'. The aim of this investigation is to gain insight about how examining deviations from HWE can increase our knowledge about the underlying population and/or specific properties of the dataset.

2.2 Methods

2.2.1 Data

The UKBB is a prospective cohort that includes approximately 500 thousand genotyped individuals with extensive phenotype information from across the United Kingdom[20]. The UKBB data release includes array genotypes for 805,426 variants and imputed genotypes for 93,095,623 autosomal variants. The genotyping and imputation process is described in detail by Bycroft et al. (2018)[20]. In summary, genotypes were assayed using two similar genotyping arrays, the UK BiLEVE Axiom Array by Affymetrix and the Applied Biosystems UKBB Axiom Array. Haplotypes were estimated from genotypes of 670,739 autosomal variants that were present on both of these arrays and fulfilled various quality control criteria. After the pre-phasing step, haploid imputation followed[20].

The analysis of HWE was restricted to 407,152 individuals that are not closely related to each other (>3rd degree for all pairs) and 146,082 biallelic autosomal variants that Bycroft et al. (2018) described as 'high-quality' and used to attain the principal

components (PCs) provided in the UKBB data release[1][20]. The group of 407,152 individuals, which I refer to as *AAtotal* in this chapter, are defined by Bycroft et al. (2018) as the largest possible group of individuals that are not closely related to each other, ($> 3^{rd}$ degree for all possible pairs)[20]. Relatedness was solely inferred by Bycroft et al. (2018) from genotype information by estimating kinship coefficients with the program `KING`[71], as familial relationships of participants were not registered during data collection[20]. By using information about self-reported ethnic background and PCs, Bycroft et al. (2018) identified a subgroup of 'white British' (WB) individuals within the UKBB[20]. In the *AAtotal* group, 337,488 individuals are of WB ancestry and I refer to that group of individuals as *WBtotal* in this chapter.

The 146,082 biallelic sequence variants that I include in the HWE analysis, fulfilled various quality control criteria as described by Bycroft et al. (2018)[20]. In short, these SNPs fulfilled the following: 1) Present on both the UK BiLEVE and UKBB Axiom array, 2) missing rate <1.5%, 3) MAF>1%, 4) did not demonstrate batch effect, 5) did not fail quality control, (plate effect, deviations from HWE and sex effect), in all 106 batches of ∼4,700 individuals and 6) were not in extended linkage disequilibrium (LD) regions as reported by Price et al. (2008)[94]. Furthermore, to minimise local LD, the SNPs that passed those filters were pruned down to 146,082 SNPs such that pairwise LD corresponded to $r^2 < 0.1$ for all SNP pairs within each window of 1,000 variants[20]. There are two main reasons for restricting the analysis to this set of 'high-quality' SNPs. Firstly, and most importantly, the goal of the current HWE analysis is not to identify problematic variants but to shed light on collective characteristics

[1]The PCs computed by Bycroft et al. (2018) and provided in the UKBB data release, were based on 407,219 individuals and the direct genotyping calls of 147,604 variants. As some individuals had withdrawn consent, 407,152 of the 407,211 individuals were available for this analysis. Furthermore, in the data release, imputed biallelic genotypes are available for 146,082 of these variants and as I wanted to compare array genotypes to imputed genotypes, I restricted the analysis to those SNPs.

of the UKBB dataset and/or the underlying population. Secondly, as these 146,082 SNPs were used to compute the PCs provided in the UKBB data release, any increased understanding about their qualities should be useful for the scientific community.

2.2.2 HWE Statistic and Subgroups

I computed HWE statistics for the 146,082 'high-quality' SNPs in different subgroups of individuals. For each SNP j, the HWE statistic was computed as follows:

$$\omega_j = \frac{var(\boldsymbol{g_j})}{2p_j(1-p_j)} - 1 \tag{2.1}$$

with $\boldsymbol{g_j}$ denoting the vector of allele counts of genotypes for SNP j in a given subgroup and p_j denoting the allele frequency. Under HWE, the expected value of ω_j is 0 while a positive/negative ω_j indicates deficit/excess of heterozygotes.

By using information provided in the UKBB data release about relatedness and inferred WB ancestry, I attained the following subdivision of individuals:

- *AAtotal*: Largest possible subgroup of individuals of all ancestry (AA) that are not closely related to each other ($> 3^{rd}$ degree for all possible pairs, $N = 407,152$)
- *AAunr*: Individuals within *AAtotal* that have no close relatives in the full UKBB dataset ($N = 334,392$)
- *AArel*: Individuals within *AAtotal* that have at least one close relative in the full UKBB dataset but are not closely related to each other ($N = 72,760$)
- *WBtotal*: Individuals within *AAtotal* that have been concluded to be of white British (WB) ancestry ($N = 337,488$)

- *WBunr*: Individuals within *WBtotal* that have no close relatives in the full UKBB dataset ($N = 272,519$)
- *WBrel*: Individuals within *WBtotal* that have at least one close relative in the full UKBB dataset but are not closely related to each other ($N = 64,969$)

For each of the 146,082 SNPs, I computed 12 HWE statistics in total. One for each of the 6 groups for both array genotypes and imputed genotypes. Prior to the analysis, I converted the imputed genotypes to hard-call genotypes with `PLINK 1.90b`[95, 22] using the hard-call uncertainty threshold of 0.1. For each of the 12 groups, I examined the SNPs jointly by computing the mean HWE statistic and their standard error (SE). Even though the SNPs were pruned to be in low pairwise local LD, they cannot be assumed to be independent and therefore I estimated the SEs for the means with block jackknife[19] by dividing the 146,082 SNPs into blocks of 10 MB regions.

To put the group-wise aggregations into context with single SNP results, I also tested each of the 146,082 SNPs for deviations from HWE with an exact test by using the `--hardy` option in `PLINK 1.90b`[95, 129]. This is the same test as Bycroft et al. (2018) used in the UKBB quality control procedure[20]. However, that analysis only tested deviations from HWE for the array genotypes in 106 subgroups of ∼4,700 individuals. In their quality control procedure, a SNP was excluded from the subset of 'high-quality' SNPs if it deviated from HWE at a threshold of 10^{-12} in any of the 106 subgroups[20]. Notably, all the 146,082 SNPs that I investigate in the current analysis passed this HWE filter by Bycroft et al. (2018)[20].

2.3 Results

The results of the HWE analysis are summarised in table 2.1 which depicts the average HWE statistic of the 146,082 variants, $\overline{\omega}$, for each subgroup. Although the mean HWE statistics are small, all twelve of them are significantly different from zero given their 95% confidence intervals (see table 2.1). Focusing on the direct genotype calls, the HWE mean statistics are higher in the three *AA* sets than in the three *WB* sets. This result is consistent with the average genotype variance being lower in the more homogenous *WB* set than in the *AA* set which is expected to have a higher degree of sub-population division.

Another observation is that the mean HWE statistic is lower in the *rel* set than in the *unr* set for both *AA* and *WB*, suggesting that individuals that have a close relative within the UKBB set are less diverse as a group (less sub-population division) than those who don't have a relative. Whether the observed difference is due to the two groups having some subtle difference in genetic background, or is a result of a previously under-appreciated characteristic of the UKBB data is discussed in section 2.4 below.

A third observation is that the mean HWE statistic is significantly negative in all the three *WB* sets when the array genotypes are examined, indicating an excess of heterozygotes for a substantial number of sequence variants. Indeed when examined independently, 4,691 SNPs deviate significantly from HWE with an excess of heterozygotes in the *WBtotal* set for the array genotypes ($P < 10^{-12}$, exact test). Possible causes include selection favouring heterozygotes, outbreeding and disassortative mating[136, p.77]. However, genotyping artefacts are a more likely cause for this result, as an abundance of variants are affected.

This is supported by the results based on imputed genotype calls, for which 1,547 SNPs deviate significantly from HWE with an excess of heterozygotes in the *WBtotal* set. Furthermore, all of the mean HWE statistics based on imputed genotypes are significantly higher than the corresponding values based on direct genotype calls. Two out of the three values for the *WB* sets are positive, but notably the mean HWE statistic remains negative for the *WBrel* set. Thus excessive heterozygous calls are reduced but not fully eliminated by the imputed genotypes. Interpretation of these results are further addressed in section 2.4 below

Group	**N**	$\overline{\omega}_{direct}$ (95% CI)	$\overline{\omega}_{imp}$ (95% CI)
AAtotal	407,152	0.00639 (0.00629, 0.00650)	0.00802 (0.00791, 0.00812)
AAunr	334,392	0.00729 (0.00718, 0.00740)	0.00892 (0.00880, 0.00903)
AArel	72,760	0.00180 (0.00173, 0.00187)	0.00342 (0.00336, 0.00349)
WBtotal	337,488	-0.00144 (-0.00149, -0.00140)	7.67E-5 (3.86E-5, 1.15E-4)
WBunr	272,519	-0.00139 (-0.00144, -0.00134)	1.19E-4 (8.07E-5, 1.58E-4)
WBrel	64,969	-0.00166 (-0.00171, -0.00161)	-9.56E-5 (-1.38E-4, -5.28E-5)

Table 2.1: HWE statistics computed for two version of genotypes and 6 different subgroups. *AA* refers to all ancestry and *WB* refers to white British. The term *unr* refers to individuals that do not have a close relative within the total UKBB dataset. The term *rel* refers to individuals that have at least one close relative within the total UKBB dataset, but are not closely related to each other. $\overline{\omega}$ refers to the average HWE statistic computed from 146,082 SNPs in each subgroup of individuals. The subscripts *direct* and *imp* refer to direct call genotypes and imputed genotypes respectively. *CI* refers to confidence interval. The confidence intervals were computed with the normal approximation $x \pm 1.96 \cdot SE$, where the SEs were estimated with block jackknife by dividing the 146,082 SNPs into blocks of 10 MB regions[19].

2.4 Conclusions

Others have reported that SNP-chips are highly inaccurate for very rare variants in UKBB[132, 125]. Current results indicate that there are genotyping artefacts for low-frequency and common variants (MAF $> 1\%$) as well, and notable only 'high-quality'

SNPs were examined. This is further supported by a blog-post by Kanai et al. (2019) which points out that 3,987 common SNPs have no homozygous alternative genotypes among WB individuals in UKBB[51]. Genotyping artefacts in the UKBB have already had serious consequences. Wei and Nielsen (2019) retracted their paper when it became apparent that their main result, a common CCR5 $\sim \Delta 32$ deletion associating with increased mortality at a homozygous state, was generated by a genotyping artefact in the UKBB[126]. Maier et al. (2020) later reported that rs62625034, the SNP used as a proxy for the CCR5 $\sim \Delta 32$ deletion in the study by Wei and Nielsen, deviated from HWE in the UKBB data due to non-random missingness. In particular, they found that while the total missing rate for rs62625034 was 3.6%, when compared to the exome sequencing data, the genotype specific missingness for rs62625034 in the array data was 17.3%, 4.6% and 2.9%[70]. It is worth noting that rs62625034 is not among the 146,082 SNPs investigated in the current application.

With individual SNPs in the current HWE analysis, there are both those that have very significant excess of homozygotes and those that have very significant excess of heterozygotes. However, when evaluated as a group, for the direct call genotypes of the *WB* data, there is an obvious tilt towards an excess of heterozygotes. Strikingly, there is a clear difference with the imputed genotypes. The obvious conclusion is that there is a systematic artefact with the direct call genotypes. What is perhaps less obvious, but highlighted by the single variant example by Maier et al. (2020)[70], is that non-random missingness might play a part as well. For a large group of the SNPs examined here, the imputed genotypes are nearly identical to the direct call genotypes when the latter exist. In other words, the imputed genotypes are only filling in the missing genotypes and not 'correcting' the direct call genotypes. This means that the

observed phenomenon could largely be driven by non-random missing genotypes, i.e. homozygous genotypes are more likely to be missing genotypes, and does not necessarily mean that there are many errors in the direct genotype calls. Distinguishing between these two imperfections of the genotype data could be important for some applications and analyses.

While excess of homozygotes can be caused by population characteristics or data artefacts, most would believe that excess of heterozygotes, particularly with common SNPs, would nearly always be because of genotyping artefacts. While that might indeed be the case, my research on participation bias reveals that, in addition to outbreeding, disassortative mating, and survival that favours heterozygotes, that there is another non-genotyping-related phenomenon that could lead to a tilt towards more heterozygotes. In particular, for a SNP that is associated with participation in the UKBB, collider bias could have an effect that moves the correlation of the paternal and maternal alleles towards the negative direction. This adds another layer of complications to HWE analysis. Notably, the same SNP could also have a positive assortative mating effect, in which case the two effects would confound and partially cancel each other. Moreover, it is currently unclear whether participation bias could have contributed to the higher fraction of the heterozygotes observed with the *WBrel* set. All these highlight how complicated apparently 'simple' HWE analyses can be, and further investigation is warranted.

Chapter 3

Behaviour of Principal Components[1]

3.1 Introduction

Principal component analysis (PCA) aims to explain the variance-covariance structure of a multi-variable dataset with a small number of orthogonal vectors that are linear combinations of the variables in the original dataset[50, p.430-431]. Principal components (PCs) are commonly used in genome-wide association studies (GWAS) as additional explanatory variables in an attempt to account for population structure. The aim is usually to correct for population stratification[76, p.228] which refers to the phenomena when there are phenotypic differences between sub-populations[32, 44]. Population stratification can easily lead to spurious results when unaccounted for, as genetic variants of different sub-population frequencies might seem to associate with

[1]The findings presented in this chapter have also been published in figure 3 in: Young, A. I., **Benonisdottir, S.**, Przeworski, M., & Kong, A. (2019). Deconstructing the sources of genotype-phenotype associations in humans. *Science*, 365: 1396-1400, doi: 10.1126/science.aax3710. The figure that has been reused/modified from this publication is licensed under the American Association for the Advancement of Science and was authored by me and has been reused with the permission of the other authors of the paper.

the phenotype independent of causality[44]. For example, a genetic variant that is more common in Iceland than in the United Kingdom might seem to have an impact on the ability to speak Icelandic if one were to combine Iceland and the United Kingdom in one cohort.

In genetic studies, PCA involves eigenvalue decomposition of a genotype covariance matrix between samples[93, 43]. The resulting PCs are ranked by the amount of explained variance[76, p.228]. That is, the first PC is the linear combination that captures the greatest amount of genotypic variance, the second PC captures the second greatest amount of variance and so on[43]. Ideally, when the purpose is to capture population structure, the PCs are mainly capturing covariances between variants on different chromosomes, rather than covariances between variants within the same chromosome. The reason for that is simple. Within each chromosome, physically close variants are more likely to be inherited together than physically distant variants as the probability of meiotic recombination decreases with decreased physical distance[37]. It follows, that while within chromosome covariances could be reflecting population structure, they could also be reflecting low recombination between physically close variants. This is not the case with between chromosome covariances as they are unaffected by physical distance and recombination as chromosomes are transmitted independently during meiosis[136, p.8]. Variants on different chromosomes are therefore not expected to be correlated unless there are factors such as population structure or non-random mating at play.

Phenotypic differences between sub-populations can be attributed to environmental factors, genetic factors or both. If phenotypic differences between sub-populations are to any extent determined by genetic factors, then using PCs as additional explanatory

variables can potentially adjust away some of the direct genetic effects in addition to confounding effects. Phenotypic assortative mating causes correlations between causal variants that PCs capture. Thus, using PCs to correct for population stratification could in theory adjust away part of the direct genetic effect component.

In this chapter I perform PCA within the UK Biobank (UKBB) and investigate to what extent the PCs capture between chromosome covariances in an independent but comparable replication set. I also compute a height polygenic score (PGS) and investigate its relationship with the PCs as well as how between chromosome covariances of the PGS contribute to the variance.

3.2 Methods

3.2.1 Data

The following analysis is based on the UKBB, a prospective cohort that includes ∼500K genotyped individuals from across the United Kingdom[20] (see further information in chapter 2.2.1). The UKBB data release includes PCs that were computed by Bycroft et al. (2018)[20]. The PCs provided by UKBB were based on the largest possible set of individuals of all ancestry that are not closely related to each other (>3rd degree for all pairs)[20]. As a result, all individuals in the UKBB were either used to compute the provided PCs or are close relatives of an individual used to compute the PCs. To be able to examine to what extent PCs capture replicable population structure, an independent and comparable replication set was needed. For this purpose I defined the *WBunr* set in the HWE application in chapter 2 as 'the original set' and the *WBrel* set as 'the replication set' (see table 2.1). Individuals in the original set are not closely related

to anyone in the replication set, nor are individuals closely related to anyone within their own set. The present analysis only includes white British (WB) individuals[20] in order to have the same ancestral composition in the original and replication set. PCs were computed in the original set (N = 272,519) while conclusions about replicable behaviour of those PCs were drawn from the replication set (N = 64,969). The aim of this subdivision and relatedness filtering was to attain PCs that captured population stratification rather than clusters of close relatives as well as having an independent replication set. The sequence variants were restricted to the same 146,082 autosomal SNPs that were used to compute the PCs provided in the UKBB data release[2][20]. The 146,082 SNPs fulfilled various quality control criteria as described by Bycroft et al. (2018), had minor allele frequency (MAF) above 1% and were in low pairwise local linkage disequilibrium (LD) ($r^2 < 0.1$)[20]. Furthermore, this is the same set of SNPs that I investigated in the HWE analysis in chapter 2. Those results indicated that the imputed genotypes were less prone to data artefacts than the array genotypes and I therefore used the imputed genotype calls to compute the PCs. Prior the analysis, I converted the imputed genotypes of all the 146,082 variants to hard-call genotypes with `PLINK 1.90b`[95, 22] using the hard-call uncertainty threshold of 0.1. The average imputation information score, as reported by the UKBB, was 0.9997 (SD=0.0102) for these variants, with 99.8% of them having an imputation information score above 0.99.

[2]The PCs computed by Bycroft et al. (2018) and provided in the UKBB data release, were based on direct genotyping calls of 147,604 variants. In the data release, imputed biallelic genotypes are available for 146,082 of these variants and as I computed the PCs with the imputed genotypes, I restricted the analysis to those SNPs.

3.2.2 Computing PCs and Eigenvalue-Equivalents

The top 40 PCs were computed from the standardised genotype matrix of the individuals in the original set using the program `ProPCA`[2]. SNP-loadings, $w_1, \ldots, w_M$, were subject to $\sum_{j=1}^{M} w_j^2 = 1$. Genotypes were standardised using the expected genotype variance under HWE. That is, the standardised genotype of individual i for SNP j is:

$$z_{ij} = \frac{g_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}} \tag{3.1}$$

with g_{ij} denoting minor allele count and p_j denoting the MAF for SNP j in the original set. As a result, SNPs for which we have deficit of heterozygotes are expected to weigh heavier in the construction of the PCs than SNPs for which there is an excess of heterozygotes[50, p.437-439]. For the replication set, eigenvalue-equivalents, $\hat{\lambda}_{PC}$, were attained by computing variances of linear combinations of standardised genotypes weighted by the SNP loadings inferred from the original set. For this purpose, standardisation of genotypes was done with regard to MAF in the replication set rather than the original set. Let $\mathbf{z_j}$ denote a $N \times 1$ vector of standardised genotypes for SNP j with N being the number of individuals. In the original set and the replication set we have $\sum_{j=1}^{M} w_j^2 \cdot var(\mathbf{z_j}) \approx 1$. It follows that $\hat{\lambda}_{PC} > 1$ indicates that the inferred PC is capturing replicable correlations between SNP pairs on the same chromosome and/or SNP pairs on different chromosomes,

$$\begin{aligned}
\hat{\lambda}_{PC} &= var(\sum_{j=1}^{M} w_j \cdot \mathbf{z_j}) \\
&= \sum_{j=1}^{M} w_j^2 \cdot var(\mathbf{z_j}) + \sum_{j \neq s} w_j w_s \cdot cov(\mathbf{z_j}, \mathbf{z_s}) \\
&\approx 1 + \sum_{j \neq s}^{(j,s) \in W} w_j w_s \cdot cov(\mathbf{z_j}, \mathbf{z_s}) + \sum_{j \neq s}^{(j,s) \in B} w_j w_s \cdot cov(\mathbf{z_j}, \mathbf{z_s}),
\end{aligned} \tag{3.2}$$

with W denoting the set of SNP pairs on the same chromosome and B denoting the set of SNP pairs on different chromosomes.

I computed between chromosome eigenvalue-equivalents by ignoring covariances of within chromosome SNP pairs, W, and only summing over covariances for SNP pairs on different chromosomes, B. The between chromosome eigenvalue-equivalents were attained with the simple act of computing the variances of the chromosome specific linear combinations of the weighted standardised genotypes. Let K denote the set of variants on chromosome k, and let $\hat{\lambda}_{PC_k} = var(\sum_{j \in K} w_j \cdot \mathbf{z_j})$. By definition, all possible SNP pairs within K will belong to W. As a result we have

$$\sum_{k=1}^{22} \hat{\lambda}_{PC_k} \approx 1 + \sum_{j \neq s}^{(j,s) \in W} w_j w_s \cdot cov(\mathbf{z_j}, \mathbf{z_s}), \tag{3.3}$$

It follows that an eigenvalue-equivalent that ignores the contribution of within chromosome covariances can be written as

$$\hat{\lambda}_{PC_B} \approx 1 + \hat{\lambda}_{PC} - \sum_{k=1}^{22} \hat{\lambda}_{PC_k} \tag{3.4}$$

In summary, by appreciating the relationship between the overall variance and the chromosome specific variances, between chromosome eigenvalue-equivalents could be attained by only computing 22 variances. This approach unarguably required a lot less effort, both in terms of time and computing resources, than computing the covariances of the 553,170,297 within chromosome SNP pairs.

3.2.3 Computing Polygenic Scores and Scaled Variance

To examine to what extent PCs capture direct phenotypic effects, height PGSs were computed as sums of weighted allele counts of multiple SNPs, using height effects

reported by the GIANT Consortium (2014)[131] as weights. These summary statistics have previously been reported to be subject to residual population stratification[106]. Out of 2,553,374 SNPs available in both sets, 124,307 low-frequency and common SNPs (MAF> 1%) in low pairwise local LD ($r^2 < 0.1$) were extracted by choosing the most significant SNP in each LD-group given their P-value in the GIANT Consortium study. Here, LD computations were based on the 272,519 WB individuals in the original set. For this purpose, `PLINK 1.90b`[95] was used with the following flags: `--clump-kb 500 --clump-r2 0.1 --clump-p1 1 --clump-p2 1`. Height PGSs were computed for individuals in the original set and the replication set by using the `--score` flag in `PLINK 1.90b`[95]. This height PGS explains 14.5% of the variance of standardised height measurements in the original set ($P < 10^{-300}$, two-sided t-test, N=271,929). To compute the variance explained, height measurements were rank-based inverse normal transformed to a standard normal distribution separately for each sex and adjusted for age.

Scaled variance for the height PGS, $\hat{\lambda}_{PGS}$, was computed as following:

$$\hat{\lambda}_{PGS} = \frac{var(\sum_{j=1}^{M} b_j \cdot \boldsymbol{g_j})}{\sum_{j=1}^{M} b_j^2 \cdot 2p_j(1-p_j)} \tag{3.5}$$

with b_j denoting the reported height effect[131], p_j denoting the MAF, and $\boldsymbol{g_j}$ denoting a $N \times 1$ vector of allele counts of genotypes for SNP j with N being the number of individuals. The scaling factor is equal to the PGS variance if the 124,307 sequence variants were to be completely independent. Between chromosome scaled variance was computed as described in equation 3.4.

3.3 Results

The eigenvalues and eigenvalue-equivalents of the 40 PCs computed in the original and replication set, are depicted in figure 3.1. Except for the first PC, the eigenvalues computed in the original set are always higher than the eigenvalue-equivalents computed in the replication set. As the PCs were computed in the original set, this indicates that the attained PCs are at least partly overfitting noise. The mean eigenvalue for the last 32 PCs, decreases from 4.37 in the original set to 2.61 in the replication set and further to 1.03 for the between chromosome replication set. This suggests that only the first 8 PCs are capturing a substantial amount of population structure while the other 32 PCs are mostly capturing, although probably not entirely, within chromosome covariances and overfitting noise (see figure 3.1). It is worth noting that pairs on different chromosomes are 94.8% of total SNP pairs.

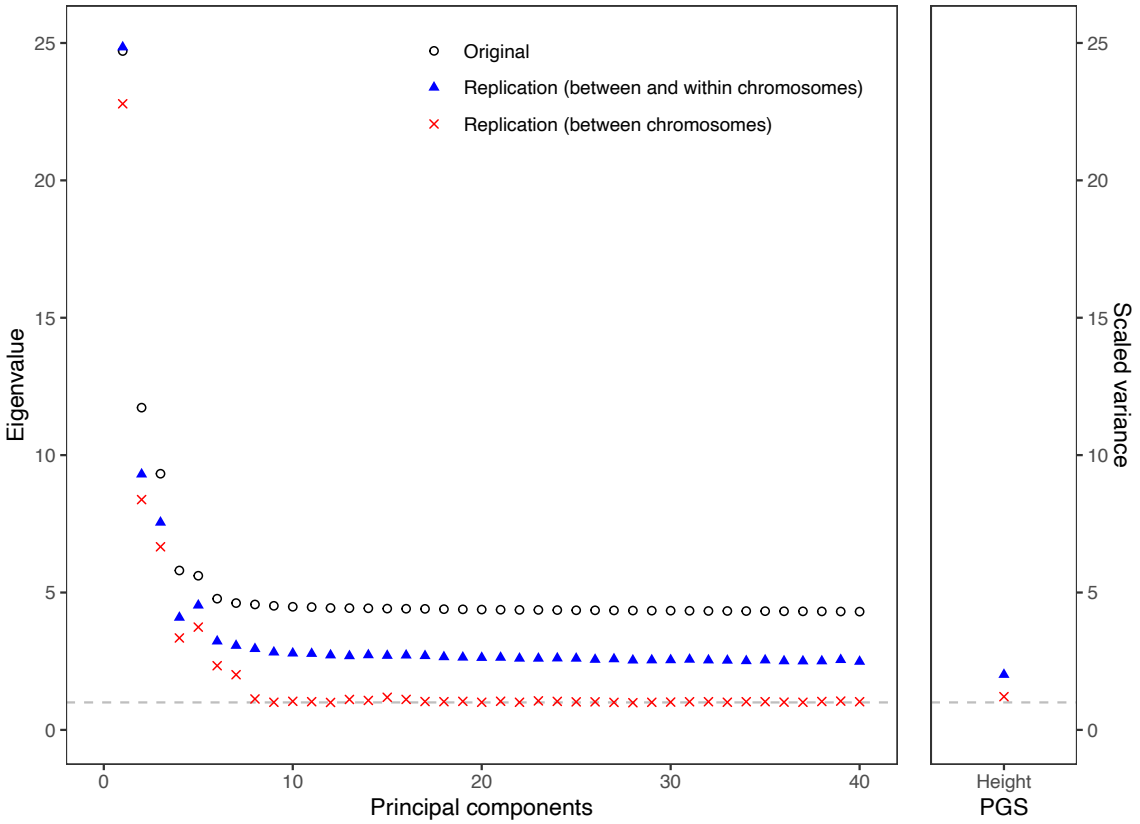


Figure 3.1: Behaviour of principal components. Figure depicting eigenvalues and eigenvalue-equivalents (y-axis on the left) computed from principal components (x-axis in the panel on the left) and scaled variance (y-axis on the right) computed from height polygenic scores (x-axis in the panel on the right) in different subgroups within the UKBB including and excluding the contribution of within chromosome covariances (colour and shape legend). The original set consists of 272,519 WB individuals without any close relatives (>3rd degree for all pairs) in the UKBB dataset. The replication set consists of 64,969 WB individuals with at least one close relative in the UKBB dataset but are not closely related to each other and not closely related to anybody in the original set. The dashed horizontal line in both panels (left and right) is set at y=1.

The height PGS scaled variance is 2.00 for the original set and 2.01 for the replication set and lowers to 1.21 when excluding covariances of SNP pairs within the same chromosome. These results indicate that the height PGS is capturing substantial amount of between chromosome covariances, the value of 1.21 only appears small when compared to the noise overfitting and within chromosome covariances that the PCs are capturing

(see Figure 3.1). In the replication set, the 40 PCs explain jointly 8.5% of the variance of the height PGS ($P < 10^{-300}$, F-test), indicating that the PCs could be capturing a small fraction of correlations between height variants.

3.4 Conclusions

Performing PCA and computing eigenvalue-equivalents in an independent replication set demonstrated that the PCs based on genotypes of 272,519 individuals within the UKBB, are capturing mostly within chromosome covariances and noise beyond the first 8 PCs. However this does not mean that the top PCs are capturing all relevant population structure. The within chromosome covariances and noise could be masking subtle sub-population patterns and as a result, residual population structure would remain even after correcting for 40 PCs. This result is supported by a study by Haworth et al. (2019), in which sequence variants associate with birth location after adjusting for PCs[40].

To have a better appreciation of the empirical results given here, some theoretical understanding would be beneficial. While the material has not been formally published as far as I know, a video provided by Chen and Bloemendal (2019) at the Broad Institute is highly recommended[23]. They point out that high-dimensional PCA is counterintuitive which can lead to unreliable results in genetic applications. That is because the number of SNPs used to derive the PCs is very large, i.e. in the current application the number of SNPs is 54% of the number of individuals. In such situations, there is a threshold value where many empirical PCs driven entirely by noise in the data would attain approximately as their apparent eigenvalue. Consequentially, the

true PCs that have eigenvalues above this threshold would show up in the calculations, while true PCs that have eigenvalues below this threshold value, which undoubtedly exist, would be suppressed. The empirical results shown here suggest that the threshold value in the WB subset of 'unrelated' individuals is approximately the eigenvalue of PC9. Patterson et al. (2006) refer to this threshold as the BBP threshold and discuss it in the context of what sample size would be required for PCA being able to detect a given level of genetic divergence[86]. However it is unclear how the contribution of within chromosome covariances could be incorporated into these mathematics.

Here it was demonstrated that the height PGS captures substantial part of covariances between variants on different chromosomes which could be a indication of assortative mating with regard to height. This could also be reflecting population stratification as it has been reported by Sohail et al. (2019) that the height effects reported by the GIANT Consortium[131] and used as PGS weights here, are subject to uncorrected population stratification[106]. Taken together, using these PCs as additional covariates in GWAS would probably only adjust out a small fraction of direct height effects at most since the within chromosome covariances and overfitting of noise is likely to overshadow most of the correlations between the height variants. Given reports of assortative mating with respect to height[104], somewhat ironically, the limitations of PCs in adjusting out population structure effects might be the main reason why PCs would not adjust out a substantial fraction of the effects of height variants.

Chapter 4

The Genetics of Participation: Method and Analysis[1]

4.1 Introduction

When genetic variants influence participation, genotypes of participants will not be representative of the intended study population. Unrepresentative data is a problem that affects all types of sample surveys as it can lead to seriously misleading conclusions[14]. It is widely acknowledged that many genetic datasets are subject to participation bias on a phenotype level[33, 81, 61]. However directly quantifying genetic differences between participants and non-participants is a challenging task as genotypes of non-participants are unavailable by definition. Previous studies of study engagement (secondary participation) have shown that among individuals that have

[1]The findings presented in this chapter are also presented in a paper that is currently under a review in a peer reviewed journal and is available as a preprint on bioRxiv: **Benonisdottir, S.** & Kong, A. (2022). The genetics of participation: method and analysis. *bioRxiv*, doi: 10.1101/2022.02.11.480067. The text, figures and tables that have been reused/modified from this preprint publication are licensed under CC.BY.NC.ND 4.0 and were authored by me and have been reused with the permission of the other author of the paper, Professor Augustine Kong. The method for testing variants for association with primary participation bias by comparing shared and non-shared alleles of first-degree relatives is based on an idea by Professor Augustine Kong.

provided a DNA sample, genetic factors play a role in participating in optional components such as filling out additional questionnaires, showing up for follow-up visits, not responding to questions or providing a valid email address[118, 111, 73, 1, 75]. However, as these studies only include individuals that have already contributed a DNA sample, they do not address the potential genetic difference between those who participated in the genetic study in the first place and those who did not.

Given this background, it is remarkable to see that one can investigate how the sampled genotypes are biased based on themselves alone. A recent study by Pirastu et al. (2021) identified single nucleotide polymorphisms (SNPs) that had significant allele frequency differences between male and female participants and proposed that those variants have differential participation effects for the sexes[91]. This approach has the advantage of only requiring genotypes of participants and relies on the assumption that genotypes of autosomal variants are of the same allele frequencies for both sexes in the population. However, this method cannot identify variants that affect study participation of both sexes in a similar manner, which are presumably more common and have broader impact. In this chapter I show a way to study these variants.

For binary phenotypes, identifying influential sequence variants traditionally involves comparing genotypes of cases to the genotypes of controls[18]. It goes with out saying that this approach is not applicable in a direct manner when the binary phenotype is 'participation in genetic studies'. In order to be a 'control', one cannot be a part of the genetic dataset and as a result, for this specific trait, the genetic datasets can only consist of cases. However, genetic data is unusually rich in information. In particular, all individuals are related to some degree, resulting in haplotype sharing that can be examined directly. As is shown in this chapter, the relatedness itself provides extra

information that enables us to study primary participation in genetic studies.

In the first part of this chapter, I describe various consequences of genetically induced participation bias and how they leave footprints within the sample itself. These consequences contain valuable information about the bias and enable the investigation of how genetic variants influence primary participation by only using the sample itself, something which has so far been impossible. In section 4.3, I use this method to perform a primary participation genome-wide association study (GWAS) with the UK Biobank (UKBB) data using only the genotypes of participants. Moving from theory to data application requires various data processing steps as well as careful considerations of data artefacts and their implications. The analysis presented in this chapter is a good example of this.

4.2 The Three Principles of Participation Bias

There are two properties that make genetic data distinct. Firstly, all individuals are genetically related by various degrees, and the relatedness is captured by the genotypes. Secondly, each individual has two copies of genetic segments on autosomal chromosomes, one inherited from each parent. Some of these segments are identical by descent (IBD), i.e. inherited from a recent common ancestor, with genetic segments in a relative. By regarding genetic data as a collection of genetic segments shared between relatives leads us to the three principles of genetically induced participation bias described below.

4.2.1 The First Principle of Genetically Induced Participation Bias

The key idea introduced here is that if an allele has a higher frequency in participants than non-participants, it would also have higher frequency in segments that are in two participants compared to segments that are only in one participant. Appreciating this detail allows us to perform a GWAS of 'primary participation in genetic studies' using only the genotypes of participants. Most importantly, this analysis is only sensitive to direct genetic effects[60, 135], and is immune to indirect genetic effects[133] and confounding effects such as those induced by population stratification[93].

The First Principle of Genetically Induced Participation Bias

On average, between two participants, genetic segments shared IBD, relative to segments that are not, are enriched with alleles that have positive direct effects on participation probability.

To understand why participation promoting alleles are expected to be enriched among IBD shared alleles compared to IBD non-shared alleles, one can consider the case of two siblings both having a rare disease. Assuming that the disease is caused by a rare mutation, the mutation is much more likely to be found in the IBD shared segments between siblings than in the non-shared segments as the latter requires at least two grandparents carrying the mutation while the former only requires one. Indeed, this is the principle underlying IBD-based affected sibling-pair linkage analysis[45]. As such, the method described below can be considered as an IBD-based IBS (identity by state) association analysis.

The first principle of genetically induced participation bias relies on the fact that all individuals are related to a certain degree and that genetic information captures those

relationships. In a large dataset, for a particular biallelic sequence variant, many pairs of individuals will share one allele inherited IBD from a not-so-distant common ancestor. For each of those pairs we observe three alleles distinct by descent. The shared allele is observed twice in the pair while the non-shared alleles are observed once (see figure 4.1a). While these alleles could sometimes be IBS, they are distinct by descent as, for a given degree of relatedness, they originate from three different ancestors. A case-control analysis can be performed by tabulating separately the allele frequencies (by state) among the IBD shared and the IBD non-shared alleles, and then comparing them. Under this framework, the shared alleles are the cases and the non-shared alleles are the controls and are matched to the extent that they are in the same individuals. However, this matching does not remove potential confounding completely as alleles in the population that were driven to higher frequency through natural selection in recent generations would also be shared by more individuals, both in the population and in the sample. Essentially, genetically induced participation bias is a form of selection and to cleanly distinguish it from other forms of selection requires more stringent matching of shared and non-shared alleles. We can achieve that by comparing shared and non-shared alleles of participating first-degree relatives.

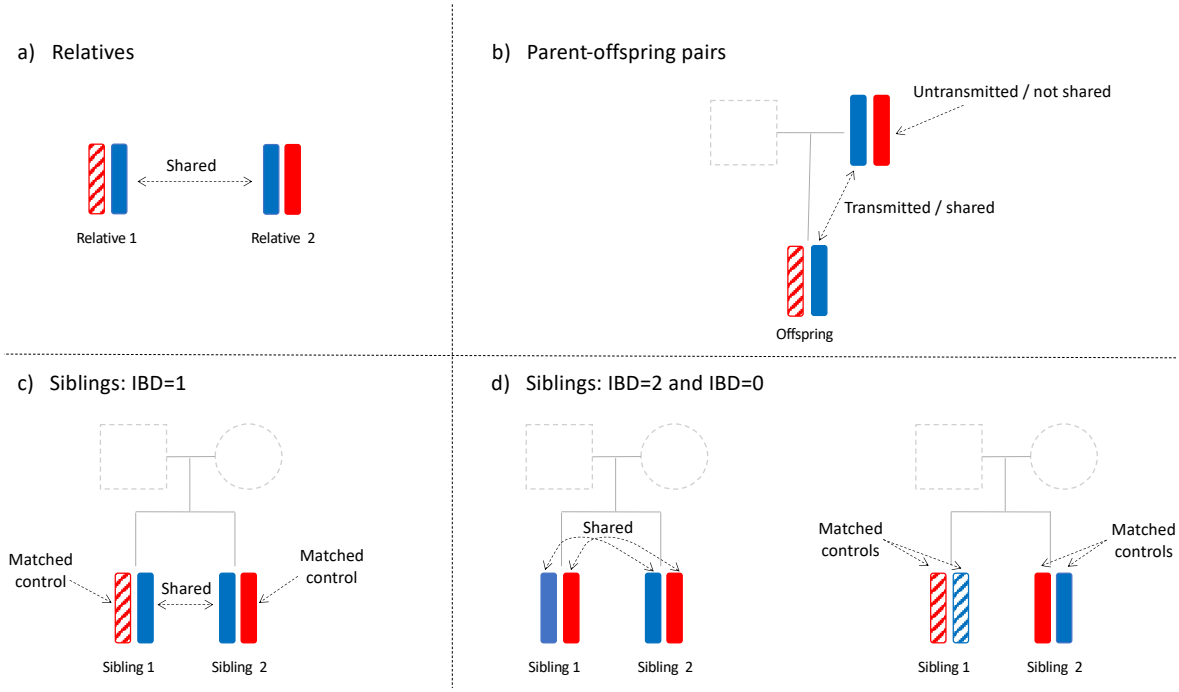


Figure 4.1: The first principle of genetically induced participation bias. Comparing IBD shared and non-shared alleles of a) General relative pairs. b) Parent-offspring pairs. c) Sibling pairs who share a locus IBD=1. d) Sibling pairs who share a locus IBD=2 and sibling pairs who share a locus IBD=0.

4.2.1.1 Participating Parent-Offspring Pairs

Parent-offspring pairs share one allele IBD for all sequence variants, unless the offspring carries de novo mutations (see figure 4.1b). For a participating parent-offspring pair where the other parent does not participate, we observe three alleles distinct by descent in the data. Firstly the allele transmitted from the participating parent to the participating offspring which we denote by a^T. Secondly is the allele of the participating parent that is not transmitted to the offspring which we denote by a^{NT}. Thirdly is the allele that the participating offspring inherited from the other parent which we denote by a^O. Thinking of the offspring as the proband, the non-transmitted allele is a perfect

match for the transmitted allele as they are both in the participating parent. The other allele of the participating offspring, a^O, is not as perfectly matched but could be useful for subsequent validation. As Mendelian rules dictate that each parental allele would have the same chance to be transmitted to the offspring, an allele frequency difference between transmitted and non-transmitted alleles within the sample would indicate an association with participation. Assuming we have N_{PO} participating parent-offspring pairs, the following frequency difference statistic can be used to test for association between a given sequence variant and participation:

$$\Delta_{PO} \stackrel{\text{def}}{=} \frac{\sum_j a_j^T - a_j^{NT}}{N_{PO}} = \frac{\sum_j a_j^T}{N_{PO}} - \frac{\sum_j a_j^{NT}}{N_{PO}} \stackrel{\text{def}}{=} p_T - p_{NT} \tag{4.1}$$

where j indexes the parent-offspring pairs. The principle behind comparing the transmitted and non-transmitted alleles to test for association with participation is similar to the principle underlying the transmission disequilibrium test[30]. In particular, conditional on the genotypes of the ascertained parents, Δ_{PO} has expectation zero under the null hypothesis of no association with participation and is not subject to population stratification induced bias as both alleles are in the parent. A participating trio can be treated as two parent-offspring pairs as the transmission from the two parents are independent under the null hypothesis.

4.2.1.2 Participating Sibling Pairs

Let N_{SIB} be the number of participating sibling pairs with non-participating parents. For a specific sequence variant, assuming random Mendelian transmission, a sibling pair is assumed to share 0, 1, or 2 alleles IBD with probabilities 0.25, 0.5 and 0.25 respectively. With data from dense sequence variants, the IBD state of a locus can usually be determined with high accuracy[71]. For a given sequence variant for which the IBD

sharing between participating siblings is known, two frequency-difference statistics are derived as follows.

1) **IBD=1: Shared vs non-shared**: For each participating sibling pair that shares a sequence variant in question IBD=1, we observe three alleles distinct by descent: the IBD shared allele, a^S, and the two distinct non-shared alleles, a^{NS1} and a^{NS2} (see figure 4.1c). Letting $j = 1, \ldots, N_{SIB1}$, index the sibling pairs that share the sequence variant in question IBD=1, we have the following frequency difference statistic:

$$\Delta_{SIB1} \stackrel{\text{def}}{=} \frac{\sum_j a_j^S - (a_j^{NS1} + a_j^{NS2})/2}{N_{SIB1}} = \frac{\sum_j a_j^S}{N_{SIB1}} - \frac{\sum_j (a_j^{NS1} + a_j^{NS2})/2}{N_{SIB1}} \stackrel{\text{def}}{=} p_{IBD1S} - p_{IBD1NS} \tag{4.2}$$

Note that for any one of these sibling pairs, if the shared allele is paternally inherited, then the two non-shared alleles are maternally inherited, and vice versa. Despite that, conditional on the genotypes of the two parents, without genetically induced participation bias, Δ_{SIB1} has expectation 0. This holds even in the most extreme case where fathers carry only allele 1 and mothers carry only allele 0. That can be demonstrated by considering the four parent-offspring transmissions jointly (see figure 4.2). Notably, when the IBD state is 1, the shared allele is equally likely to be paternal or maternal, and thus any systematic differences between fathers and mothers cancel in expectation.

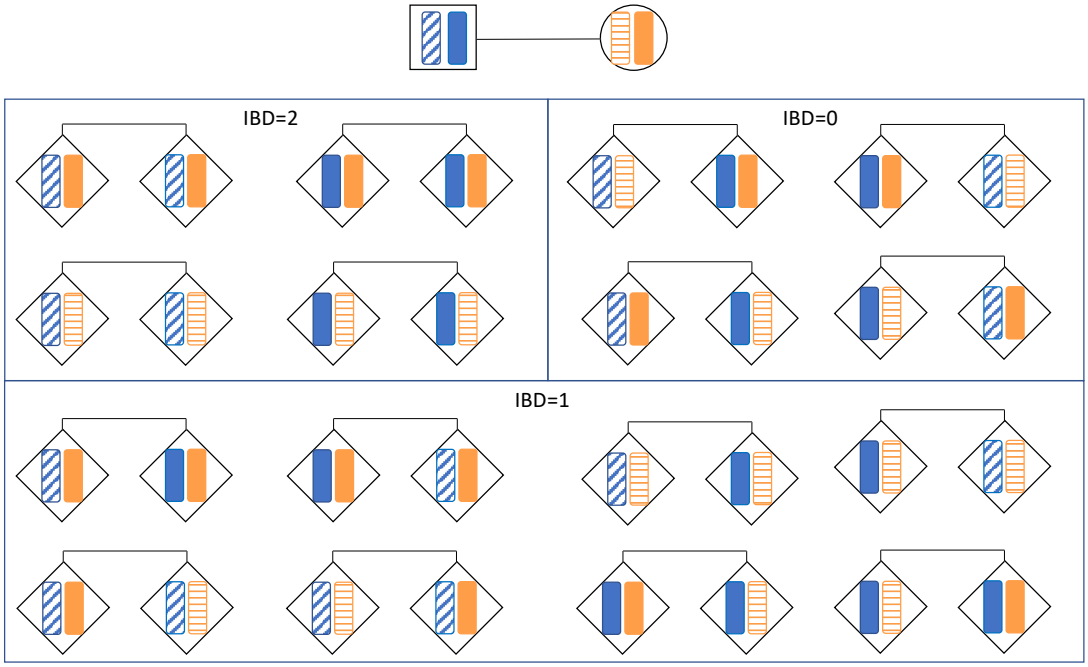


Figure 4.2: Parental transmission to sibling pairs. Figure showing the 16 possible combinations of ordered parental genetic segments for a sibling pair at a given locus. The father is shown as a rectangular shape with two blue distinct genetic segments and the mother is shown as a circular shape with two orange distinct genetic segments. Here, distinct refers to origin. That is, the four parental genetic segments could be identical by state but they are distinct with regard to grandparental origin. The sibling pairs are shown as diamond shapes, each carrying one blue genetic segment inherited from the father and one orange genetic segment inherited from the mother. The siblings share both segments identical by descent (IBD) for 1/4 of the combinations and the siblings share none of the segments IBD for 1/4 of the combinations. The siblings share one segment IBD for half of the combinations. For the $IBD = 1$ combinations, the shared segment is paternally inherited for half of them and maternally inherited for the other half.

2) **IBD=2 vs IBD=0**: For the approximately one-quarter of the participating sibling pairs (N_{SIB2}) that share two alleles IBD for the sequence variant in question, the average allele frequency is

$$p_{IBD2} = \frac{\sum_j g_{1j} + g_{2j}}{4 \cdot N_{SIB2}} = \frac{\sum_j g_{1j}}{2 \cdot N_{SIB2}} \tag{4.3}$$

where $j = 1, \ldots, N_{SIB2}$, indexes the participating sibling pairs with IBD state 2 and g_{1j} and g_{2j}, are the identical genotypes of sib 1 and sib 2 respectively in pair j. Similarly,

for the approximately one-quarter of the sibling pairs (N_{SIB0}) that share zero alleles IBD, we have

$$p_{SIB0} = \frac{\sum_j g_{1j} + g_{2j}}{4 \cdot N_{SIB0}} \tag{4.4}$$

where $j = 1, \ldots, N_{SIB0}$, indexes the participating sibling pairs with IBD state 0, and g_{1j} and g_{2j} are the genotypes of sibling 1 and sibling 2 respectively in pair j. The difference between p_{IBD2} and p_{IBD0} is another frequency difference statistic:

$$\Delta_{SIB20} \stackrel{\text{def}}{=} p_{IBD2} - p_{IBD0} \tag{4.5}$$

Here p_{IBD2} and p_{IBD0} are allele frequencies from different groups of sibling pairs. However, assuming random Mendelian transmission, the chance for a sibling pair to be in IBD state 0 or 2 is the same, and thus, without genetically induced participation bias, Δ_{SIB20}, has expectation zero. Notably, this is a comparison between different sibling pairs and, unlike the previous two comparisons, does not require splitting genotypes into shared non-shared alleles, the latter proved to be a very important detail in actual data analyses (see chapter 4.3). Also worth noting is that, in general, the same sibling pair would be in IBD state 2 for some regions, and in IBD state 0 (or 1) in other regions.

As the three statistics, Δ_{PO}, Δ_{SIB1}, and Δ_{SIB20}, are based on comparing IBD shared and IBD non-shared alleles, randomized by Mendelian transmissions, they are only sensitive to the direct effects[60, 135] of participation, and are unaffected by population stratification induced bias[93, 133] or indirect genetic effects from relatives[60, 135]. However, direct genetic effects can manifest in various ways, their relative contributions depending on the sampling scheme. For example, the effect can be 'voluntary', a person is more likely to participate when invited, or 'involuntary', e.g. a person is less likely to be invited because of its attributes. A t-statistic for testing can be computed for each

of the three frequency-difference statistics by dividing the latter by its standard error (SE) estimated from the sample. The three frequency-difference statistics or the three corresponding t-statistics can be combined, e.g. as a weighted linear combination, for testing, estimation or prediction purposes. However, having separate values of them is important because they can have different expectations depending on the nature of the participation bias, and they are impacted differently by genotyping and data processing errors. This is further discussed in chapter 4.3 which describes the data application to the UKBB.

4.2.2 The Second Principle of Genetically Induced Participation Bias

For a given genetic study and a given degree of relatedness, s, the intended study population can be divided into two groups 1) individuals with a participating relative of degree s or closer and 2) individuals with no participating relatives of degree s or closer. As genetic similarities increase with relatedness, the former group will usually be more likely to carry alleles that promote participation than the second group.

The Second Principle of Genetically Induced Participation Bias

When genetics influence participation, individuals with participating close relatives are usually more likely to harbour allele types that increase the probability of participation than individuals whose close relatives do not participate.

In genetic studies we only observe the genotypes of participants. Some of those participants will have close relatives within the dataset, others will not. Alleles in participants belonging to the latter group are by definition not IBD shared through a recent common ancestor with any other participant. By comparison, participants in the former

group have both shared and non-shared alleles. In the absence of confounding factors, comparing those two groups could yield information about the difference between participants and non-participants. However, unlike Δ_{PO}, Δ_{SIB1}, and Δ_{SIB20} described in 4.2.1, this comparison could be influenced by factors such as family size and population stratification. In particular, the probability of having a participating first-degree relative will partly depend on how many first-degree relatives an individual has to begin with. An only-child will not have a participating sibling, no matter how prone his or her genetic profile is to participation, and individuals with many siblings will have a higher probability of having a participating sibling than individuals with fewer siblings. Family structures differ between individuals, and, given previous reports[78], those differences are to some extent determined by genetic factors. That being said, dividing participants into cases and controls from their family history of participation could be useful for subsequent validation of results based on Δ_{PO}, Δ_{SIB1}, and Δ_{SIB20}. This is further investigated in the UKBB data application in section 4.3 and with simulations in section 4.4.

4.2.3 The Third Principle of Genetically Induced Participation Bias

If a participant has a higher than average expected participation genetic component value, so would his/her close relatives. It follows that the higher the genetic contribution to participation, the higher will be the number of close relatives within the dataset.

The Third Principle of Genetically Induced Participation Bias

When genetics contribute to participation, there are more close relative pairs among the participants than what to be expected when there is no genetic component at play.

This principle is a special case of one of the major building blocks of genetics, that heritable traits are correlated between relatives[114]. Indeed, before DNA information became available, genetic studies were performed by investigating phenotypic values within families[133]. However, genetic interpretation of phenotypic correlations is not straightforward because of the possibility of shared environmental influences[39, p.687-688]. As 'participation in genetic studies' can be regarded as a complex behavioural trait, environmental factors such as mutual consultation and the sampling scheme are likely to play a part.

Even though UKBB did not recruit families on purpose, the dataset contains more close relatives ($\leq$3rd degree) than what would be expected from random sampling, e.g. sibling pairs are twice as many as expected[20]. Bycroft et al. (2018) speculated that the cause of this overrepresentation of close relatives was mutual consultation and geographical clustering of close relatives[20]. By using Δ_{PO}, Δ_{SIB1}, and Δ_{SIB20} to test if genetic variants associate with participation in the UKBB, one can investigate if shared DNA between relatives also plays a role in this overrepresentation.

4.3 Primary Participation GWAS in UK Biobank

The UKBB is a large-scale database with genetic and phenotypic information of individuals from across the United Kingdom (UK)[20]. Invitations to participate were sent to approximately 9 million individuals who were aged between 40 and 69 years and lived within 25-mile radius of any of the 22 UKBB assessment centres[33]. In the end, 5.5% of those participated in the study (∼500,000 individuals) and went through baseline assessments that took place from 2006 to 2010[33]. In addition to phenotypic details collected at the baseline visit, information continued to be added, including follow-up studies for large subsets of the cohort[33, 29, 13].

It is widely acknowledged that the UKBB sample is not fully representative of the UK population[20, 33, 80]. The participants were more likely to be female, less likely to smoke and were older than non-participants[33]. Also, compared to the national average, participants were more educated[80], taller and had a lower BMI[33]. Assuming that these phenotypes are to some extent explained by genetic factors, these reported differences indicate that the genotypes of the UKBB sample are also not fully representative of the UK population. However, as genotypes of non-participants are not available and there is no UK national census for genotypes, this cannot be investigated directly with a traditional case-control comparison. Here I exploit the first principle of genetically induced participation bias to perform a primary participation GWAS within the UKBB, using only the genotypes of participants.

4.3.1 Methods

To perform a primary participation GWAS, the genotype data has to be transformed to a format that takes into account which alleles are IBD shared between pairs of relatives and which alleles are non-shared. This transformation requires multiple processing steps and quality control measures. Transforming the data to a IBD shared/non-shared format has many implications beyond performing participation GWAS and can elucidate certain characteristics of the dataset. A summary of the overall strategy for performing primary participation GWAS in the UKBB is shown in figure 4.3 and details of those steps are given in sections 4.3.1.1-4.3.2. After performing the primary participation GWAS, I validated the results with within-sample replications which are described in sections 4.3.3.1 and 4.3.3.4.

4.3.1.1 Genotype Data

I restricted the analysis to 402,199 individuals in the UKBB who had not withdrawn consent, did not have a duplicate/twin in sample and fulfilled the following criteria in the sample quality control file provided in the UKBB data, ukb_sqc_v2.txt: (1) Did not have an excess of third degree relatives, (2) was not a heterozygosity or missingness outlier (3), had missing rate below 2%, (4) was included in the kinship inference, (5) was included in the phasing input, (6) did not show potential sex chromosome aneuploidy, (7) did not demonstrate potential sex mismatch and (8) had been determined to be white British (WB) according to self-identified ancestry and PCA[20].

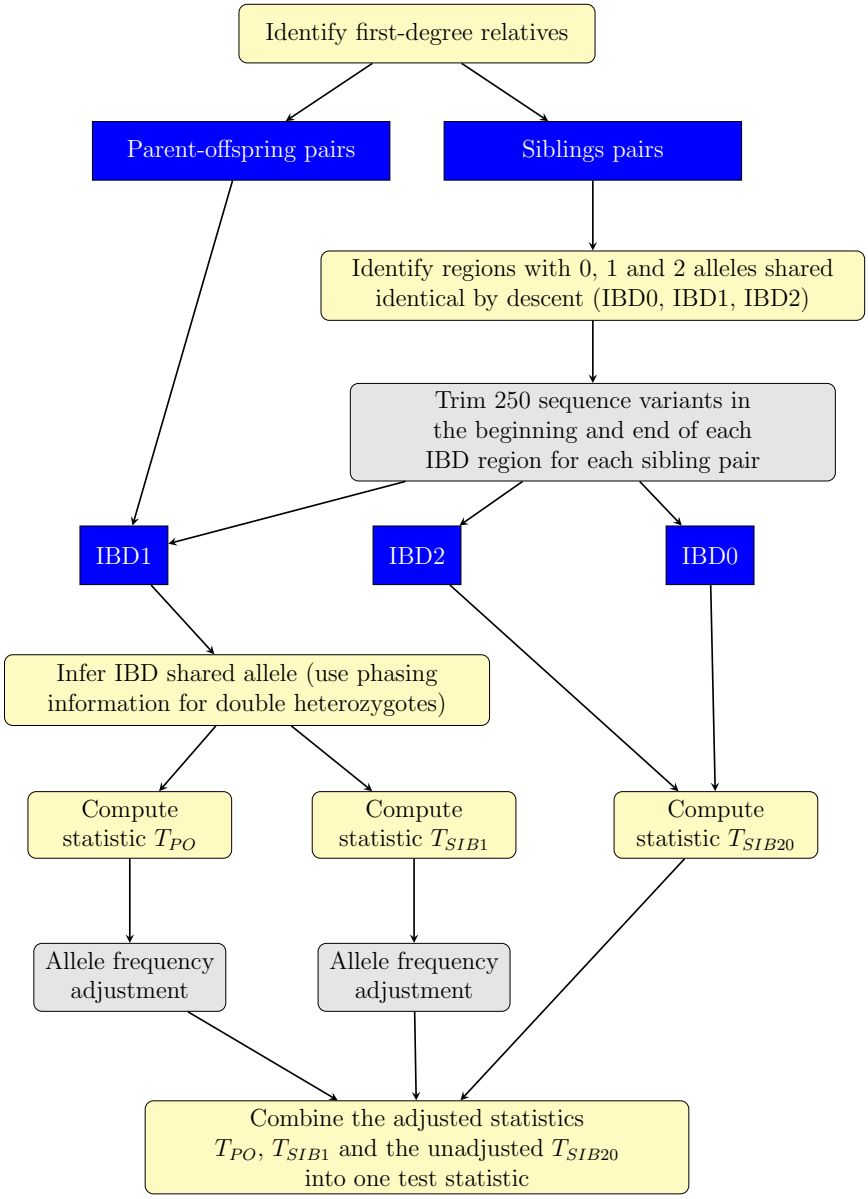


Figure 4.3: Flowchart summarizing the stepwise procedure for testing genetic variants for association with primary participation in UKBB. In the flowchart, core steps are shown as yellow rectangles with rounded corners. The procedure involves dividing the data into different data groups, which are shown here as blue rectangles. Grey rectangles with rounded corners show additional quality control steps, implemented to reduce or adjust for genotyping and data-processing errors. The genotype data was handled with `QCTOOL` version 2.0.1[7] and `PLINK` version 1.90b and 2.00[95, 96, 22]. First-degree relatives were identified with kinship coefficients provided in the UKBB data release. IBD segments of siblings were inferred with `snipar`[135]. Trimming of IBD regions and computations of allele frequencies and standard errors were done in `python` version 2.7.11[97]. Subsequent statistical analyses were performed in `R` version 3.4.3[98, 116, 128].

The application was restricted to 658,565 biallelic sequence variants that were available in the UKBB phased haplotype data[20]. The genotyping process and quality control for the sequence variants in the phasing input as well as the phasing process was described in detail by Bycroft et al. (2018). Briefly, `SHAPEIT3` was used to attain phased haplotypes for directly-genotyped sequence variants that fulfilled the following criteria: 1) Present on both the UK BiLEVE and UK Biobank Axiom array, 2) missing rate <5%, 3) MAF>0.0001, 4) did not demonstrate batch effect, 5) did not demonstrate array effect, 6) did not demonstrate discordance across control replicates and 7) did not fail quality control, (plate effect, deviation from HWE and sex effect), in more than one of the 106 batches of ∼4,700 individuals[20].

4.3.1.2 Identifying Relatives

The UKBB data release includes kinship coefficients, attained with the program `KING`[71], for pairs that were genetically inferred to be related of 3rd degree or closer. Among those pairs, I identified first-degree relatives of WB descent, 16,668 sibling pairs and 4,427 parent-offspring pairs, using kinship coefficient boundaries recommended by the authors of `KING`[71]. Pairs with kinship coefficient > 0.177 and < 0.354 were concluded to be first-degree relatives and, following Bycroft et al. (2018)[20], those who shared less than 0.12% IBD=0, were classified as parent-offspring pairs.

I chose the two groups, siblings and parent-offspring pairs, so that they did not overlap (figure 4.4). The parent-offspring pairs were made up of 739 trios, 635 father-offspring pairs and 2,314 mother-offspring pairs. The age-difference between the parents and their offspring ranged from 15 to 31 years with the mean being 23.43 years (SD=2.71).

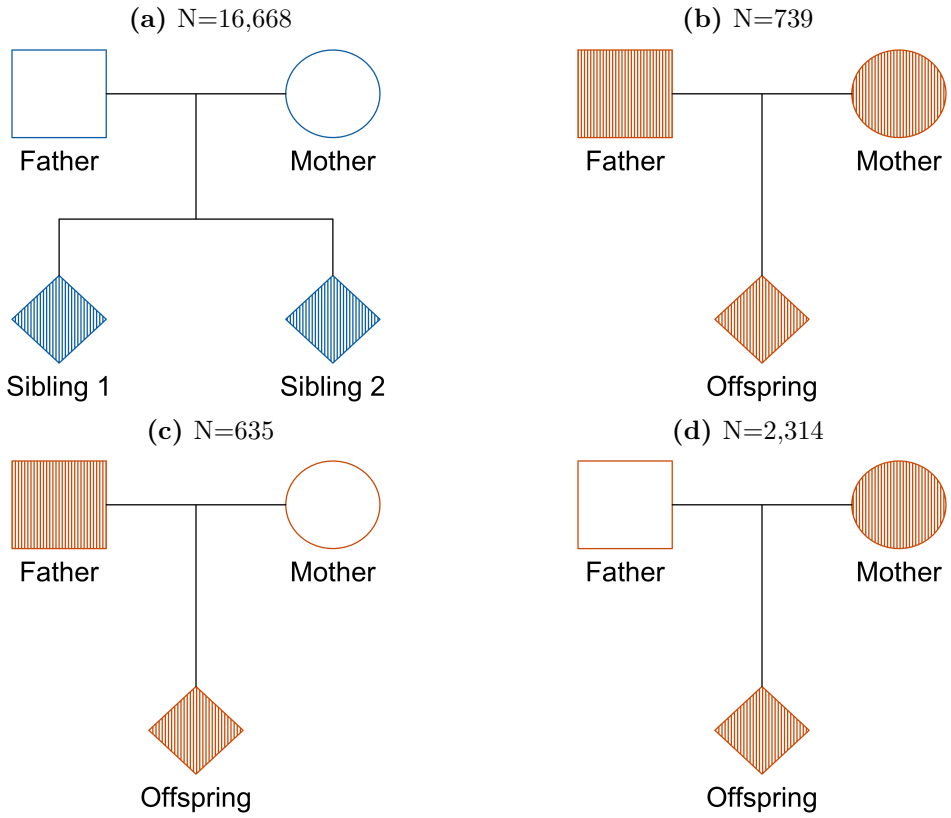


Figure 4.4: First-degree relatives included in analysis. The number of first-degree relatives in the UK Biobank that were selected for the current analysis. Shaded symbols indicate available genotype information. The 739 trios in b) count as 1,478 parent-offspring pairs and therefore the parent-offspring pairs are 4,427 in total. The groups of siblings and parent-offspring pairs were chosen so that they would not overlap.

For sibling-groups with more than 2 siblings in the dataset, I chose the first two participating siblings in the group. The male fraction was 43.4% among first participating siblings and 41.4% among second participating siblings (see table 4.1). Within the WB subset, 272,409 individuals had no close relatives within the UKBB (>3rd degree for all pairs). I refer to this set of individuals as the 'unrelated' set. The male fraction was 46.7% in the 'unrelated' set compared to 31.0% among the parents and 42.4% among the siblings. In addition, the 'unrelated' individuals were on average 5.6 months

younger than the sibling pairs.

		Second participating sibling	
		Brother	**Sister**
First participating sibling	**Brother**	3,307	3,920
	Sister	3,593	5,848

Table 4.1: Brother-sister division. Table showing how the 16,668 sibling pairs are divided into brother-brother, sister-sister, brother-sister and sister-brother pairs.

4.3.1.3 Inferring IBD Segments

I inferred IBD segments for the 16,668 sibling pairs with two different softwares, `KING` (version 2.2.4) [71] and `snipar`[135], and then compared their results. `KING` is a relationship inference software and was used by the UKBB to compute the kinship coefficients provided in the UKBB data release[20]. The authors of `KING` have not yet published a manuscript describing their IBD segment algorithm but claim in their online tutorial that it has been well tested[55]. The only input needed for inferring IBD segments with `KING` are genotypes for a dense set of sequence variants. `snipar` is a software package for imputing missing parental genotypes, performing family-based GWAS and PGS analysis and inferring IBD segments shared between siblings[135]. Young et al. (2022) compared the accuracy of `snipar` to `KING` and reported that in a subset of 31 nuclear families, `snipar` inferred the IBD state correctly 99.7% of the time while `KING` inferred the correct state 98.5% of the time[135]. When inferring IBD segments, `snipar` takes recombination probabilities into account[135]. For that reason, in addition to genotype data, information about genomic map positions are also needed as an input. For this analysis I used the genomic map provided with `BOLT-LMM (v.2.32)` [67]. Sibling genotypes for sequence variants in the UKBB haplotype data were used

as inputs for both `KING` and `snipar`. However, it is worth noting that neither program takes available phasing information into account in their IBD inferring algorithm.

To investigate and compare the results of `KING` and `snipar`, I computed, for each of the 658,565 biallelic sequence variants in the haplotype data, the fraction of sibling pairs sharing the sequence variant in question IBD=0, IBD=1 and IBD=2 as inferred by each of the two programs. The computed sibling fractions are depicted as dots in figure 4.5. The figure also depicts, with horizontal lines, the expected sibling fractions for each IBD status and their 95% confidence intervals (CI) computed with the normal approximation. Consistent with results previously reported by Young et al. (2022)[135], the figure shows how deviations from the expected fractions are much more likely to occur for IBD segments computed with `KING` than `snipar`, indicating that the results from `KING` are more prone to IBD calling errors. In particular, with `KING`, 69.6% of the 658,565 sequence variants have sibling fractions that are not within the 95% CI of the expected fraction for at least one IBD status, while with `snipar`, that is the case for 20.7% of the sequence variants. The most striking difference is seen for IBD=0, where the `KING` sibling fractions range from 0.166 to 0.421 while the `snipar` sibling fractions range from 0.238 to 0.265. `KING` reports which chromosomal regions are 'usable for IBD segment analysis' but segments with low marker density are excluded. E.g. for chromosome 1, `KING` infers IBD segments for the regions 0.8-121.3 MB and 143.7-249.2 MB separately, while the centromere, located at 121.5-142.5 MB in build 37[28], is excluded. As figure 4.5a shows, for the IBD segments inferred by `KING`, the deviations from the expected IBD fractions are most extreme at the edges of these chromosomal segments, where IBD=0 status appears to be systematically overcalled, and in long-range LD regions[94], where IBD=1 status appears to be systematically overcalled. In

the same time, `snipar` infers IBD segments within each chromosome and the fraction of sibling pairs sharing each variant IBD=0/1/2 are relatively close to the expected fraction from one chromosomal end to the other (see figure 4.5b).

`KING` is designed to infer IBD segments for relative pairs of 4th degree and closer while the IBD inferring algorithm of `snipar` is specifically designed for sibling pairs. Taking that into account, the superiority of `snipar` for this specific analysis is not surprising as we are only inferring IBD segments for sibling pairs. In light of these results, I used the IBD segments inferred by `snipar` for subsequent analysis.

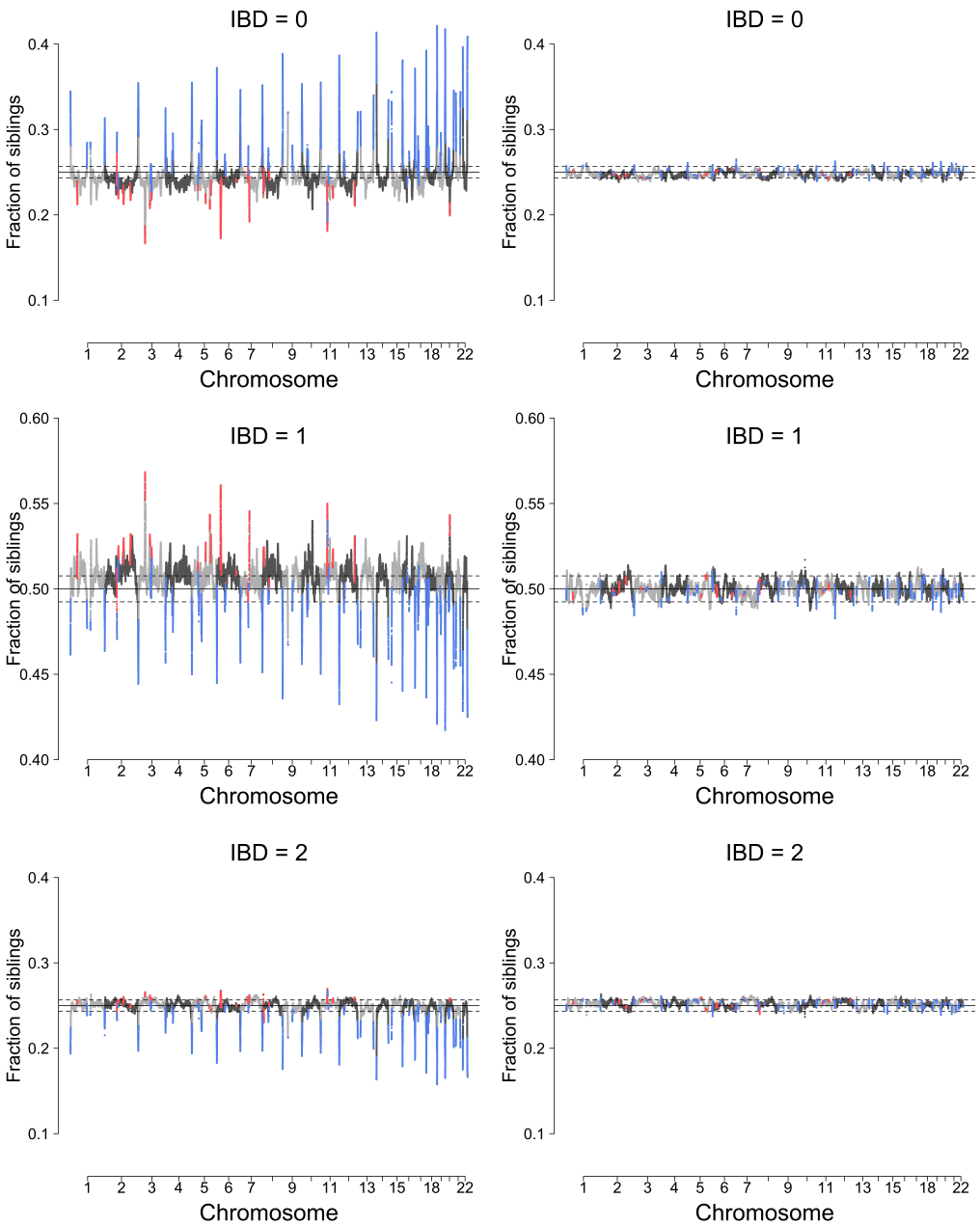


Figure 4.5: Fraction of sibling pairs sharing sequence variants IBD=0, IBD=1 and IBD=2. Each dot represents how many sibling pairs share a given sequence variant IBD=0, IBD=1 and IBD=2 divided by the total number of sibling pairs, 16,668. The y-axis corresponds to the fraction of siblings sharing the sequence variant IBD and the x-axis denotes the chromosomal position of each sequence variant (build 37). The black solid lines denote the expected fraction for each IBD scenario. The black dashed lines correspond to the 95% confidence interval for the expected fraction, $f \pm 1.96\sqrt{\frac{f(1-f)}{16{,}668}}$, with f denoting the expected fraction. Panel a) corresponds to IBD segments inferred with `KING` while b) corresponds to IBD segments computed with `snipar`. Blue dots denote sequence variants within +/- 2MB of the edges of the 40 regions that `KING` infers IBD for and red dots denote sequence variants in long-range LD regions.

4.3.1.4 Uncertainties Around Recombination Events

For each sibling pair, the error-rate of IBD segment placement is expected to be higher for sequence variants close to the edges of IBD regions, i.e. locations of recombination events where IBD state changes. To account for IBD inferring uncertainty at the edges of IBD segments, I trimmed away 250 sequence variants from the beginning and end of each identified IBD segment (contiguous sequence variants with the same IBD status called) for each sibling pair (see figure 4.6). This trimming removed 11,000 sequence variants all together ($2 \cdot 250$ sequence variants in each of the 22 chromosomes) and reduced the sample size (number of sibling pairs) for the remaining sequence variants (median reduction of 1,565 pairs).

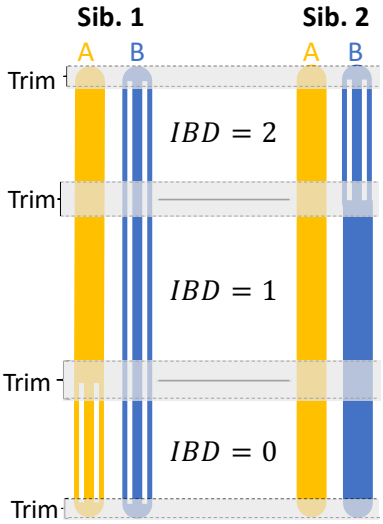


Figure 4.6: Trimming sequence variants at the beginning and end of IBD regions. To account for the error-rate of inferring IBD state being higher in the beginning and end of IBD regions, 250 sequence variants were trimmed from the beginning and end of each IBD specific region for each sibling pair.

4.3.1.5 Inferring Shared Allele

For a given biallelic sequence variant and a pair of first-degree relatives, there are 5 possible unordered combinations of genotypes when IBD=1. As IBD implies IBS, it is trivial to infer which allele is shared IBD when both individuals are homozygous for the same allele, or when one is heterozygous and the other is homozygous. For the double heterozygous scenario, I used the phasing information provided by the UKBB to infer which allele was shared. For a target sequence variant that both first-degree relatives were heterozygous for, I used the closest neighbouring sequence variant, for which one individual was heterozygous while the other was homozygous, to determine which haplotype was shared and subsequently which allele of the target sequence variant was shared (figure 4.7). This approach of resolving the IBD shared allele from phasing information was also used by Young et al. (2022) in their imputation of parental genotypes[135].

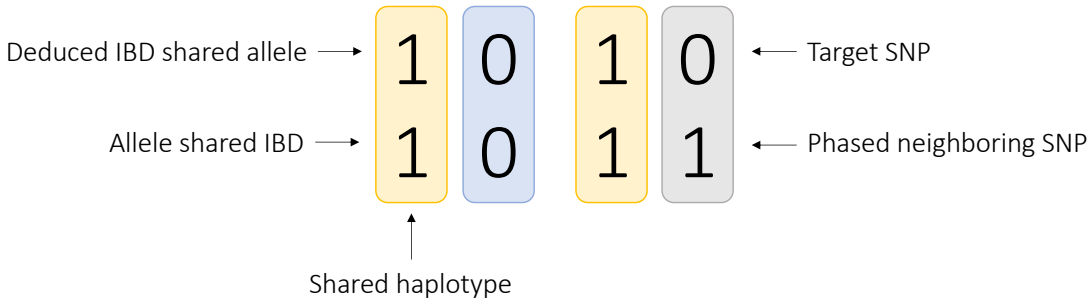


Figure 4.7: Inferring shared allele for IBD=1 when both individuals are heterozygous for target sequence variant. A neighbouring sequence variant, for which one individual is heterozygous while the other is homozygous, is informative of which allele is shared for the target sequence variant, for which both individuals are heterozygous, if those two sequence variants are phased together.

4.3.1.6 Additional Quality Control

To further reduce the impact of data artefacts on association analysis, I restricted the primary participation GWAS results to sequence variants in the UKBB haplotype data that fulfilled the following additional quality control criteria:

- Do not deviate significantly from HWE at $P < 10^{-5}$ in a group of 272,234 WB individuals with no close relatives in UK Biobank (exact test with `PLINK 1.90b`[95, 129]).
- Not among the 250 sequence variants in the beginning and end of each chromosome.
- Not in extended linkage disequilibrium (LD) regions as reported by Price et al. (2008) [94].
- With MAF>1% in the set of 16,668 sibling pairs
- With a rsname
- With non-ambiguous nucleotides
- Do not significantly deviate from shared allele expectations at $P < 10^{-6}$ in the group of double heterozygous parent-offspring pairs (see chapter 4.3.1.7).
- Observed fraction of sibling pairs sharing the sequence variant IBD 0, 1 or 2, no more than 3 standard deviations (SD) above the expected fraction (see chapter 4.3.1.8).
- Missing rate below 5% among the 16,668 sibling pairs in the genotyping array data.

After applying the additional filters described above, 500,632 sequence variants remained in the primary participation genome scan. In the following text I will refer to this reduced set of sequence variants as the 'subset of high-quality sequence variants'. The genetic variants in the UKBB haplotype data had previously been filtered based on various quality control criteria, including deviations from HWE. As the current analysis involves applying a new method to the UKBB data, an additional HWE filter was applied here. It is worth emphasising that the missing rate filter is not based on the haplotype data, in which there were no missing genotypes, but the genotype array data from which the haplotypes were estimated from. The main motivation behind this filter was that the array genotypes were not missing at random and a higher rate of missingness can be associated with lower data quality overall. Bycroft et al. (2018) performed quality control tests within 106 batches of individuals in the UKBB and if a variant failed quality control in a given batch, the corresponding genotypes were set as missing in that particular group of individuals[20]. All the genetic variants in the UKBB haplotype data had a missing rate below 5% in a set of 463,844 individuals within the UKBB dataset. However, as this non-random missingness and potential lower quality genotypes could have significant impact for certain analysis, I further omitted sequence variants that had a missing rate above 5% in the subsample of 16,668 sibling pairs.

4.3.1.7 Expected IBD Sharing and Phasing Errors

For each sequence variant, I investigated the accuracy of the phasing provided by the UKBB[20] by comparing the expected fraction of the shared allele to the observed fraction of the shared allele in the group of parent-offspring pairs that were heterozygous

for the sequence variant in question. A heterozygous parent passes either allele, 0 or 1, with an equal probability to his/her offspring. If the offspring is heterozygous as well, the probability of the shared allele being the one coded as 1 is simply the probability of the allele inherited from the other parent being the one coded as 0. Assuming random mating, the allele inherited from the other parent can be regarded as a randomly drawn allele from the population and hence, in the double heterozygous scenario, the probability of the shared allele being the one coded as 1 is the population allele frequency of the allele coded as 0. I removed 261 sequence variants for which the observed fraction of the shared allele, among the double heterozygous parent-offspring pairs, deviated from the expected fraction at a threshold of $P < 10^{-6}$ (two-sided binomial test in `R`). Of those 261 sequence variants, 79 sequence variants did not surpass the additional HWE filter described in section 4.3.1.6 (see figure 4.8).

For each sequence variant, I estimated the phasing error rate from trios where the offspring and one parent were heterozygous while the other parent was homozygous. In particular, I defined the phasing error rate as the fraction of wrongly inferred shared alleles of the double heterozygous parent-offspring pairs (from phasing information) given the genotype of the homozygous parent. For the subset of 500,632 high-quality sequence variants, the average error rate was 0.0054 with the number of trios underlying the statistic for each sequence variant ranging from 3 to 235.

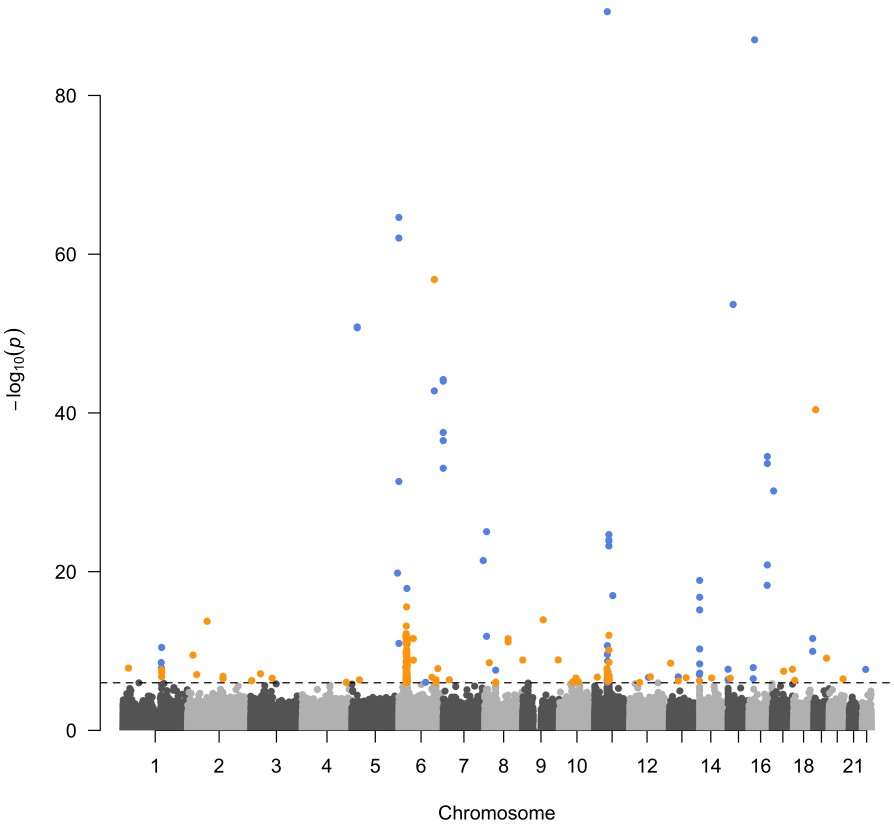


Figure 4.8: Deviations from the expected fraction of the shared allele. For each sequence variant, I tested if the expected fraction of the shared allele deviated from the observed fraction of the shared allele (inferred from phasing information) in the group of parent-offspring pairs that were heterozygous for the sequence variant in question. Each dot in the figure represents one sequence variant. The y-axis shows -$log_{10}(P)$ from a two-sided binomial test and the x-axis shows chromosomal position in build 37. The black dotted line is set at $P = 10^{-6}$. The 261 sequence variants that do not conform to the expected fraction of the shared allele at $P < 10^{-6}$ (two-sided binomial test in `R`) are coloured orange or blue in the figure. In particular, the 79 sequence variants that also deviated from HWE at $P < 10^{-5}$ (exact test in `PLINK 1.90b`) are coloured blue while the other 182 sequence variants are coloured orange.

4.3.1.8 IBD Fraction Threshold

To further account for uncertainty in inferring IBD status for siblings, I removed sequence variants for which there was strong evidence of IBD inferring problems. In particular, I removed all sequence variants for which the observed fraction of sibling

pairs sharing that sequence variant in question IBD = 0, 1 or 2 was more than 3 SD above the expected fraction, 0.25, 0.5 and 0.25 respectively in the non-trimmed data (see chapter 4.3.1.4). Note that if one IBD status (0, 1 or 2) is overcalled then at least one of the two other IBD statuses is undercalled for that specific sequence variant. Hence, thresholding the IBD fractions from above also affects the lower bounds. In total 2,453 sequence variants were more than 3 SD above the expected fraction for at least one IBD status.

4.3.1.9 Test Statistics

For each sequence variant I computed three separate t-statistics, T_{PO}, T_{SIB1} and T_{SIB20}, by dividing the frequency differences introduced in chapter 4.2.1, Δ_{PO}, Δ_{SIB1} and Δ_{SIB20}, by their SEs. The SEs were computed from the sample based on the variations of allele frequencies or allele frequency differences between individual relative pairs. For example, for a given sequence variant and parent-offspring pair j, we have $d_j^{PO} = a_j^T - a_j^{NT}$ where a_j^T denotes the transmitted parental allele and a_j^{NT} denotes the non-transmitted parental allele. Noting that $\Delta_{PO} = \frac{\sum_j d_j^{PO}}{N_{PO}}$, the SE for Δ_{PO} was computed as $\sqrt{\frac{s_{PO}^2}{N_{PO}}}$ with s_{PO}^2 denoting the sample variance of $\boldsymbol{d^{PO}}$ and N_{PO} denoting the number of parent-offspring pairs. In applications such as PCA, it is common procedure to standardise genotypes with the square root of the expected genotype variance under HWE[93]. By computing the SEs instead of using an expectation, I avoided making assumptions about HWE.

The T_{PO}-statistics were computed for each sequence variant in the UKBB haplotype data using 4,427 parent-offspring pairs. The sample size for the sibling specific t-statistics varied between sequence variants due to the variation in IBD sharing and

because of the trimming procedure described in chapter 4.3.1.4. In the subset of the 500,632 variants that surpassed various additional quality control filters (see chapter 4.3.1.6), the sample size for computing the T_{SIB1}-statistics ranged from 6,428 sibling pairs to 8,315 sibling pairs (mean=7,512; SD=279), The sample size for computing the IBD=0 specific allele frequencies ranged from 3,110 sibling pairs to 4,202 sibling pairs (mean=3,723; SD=143) while the sample size for computing the IBD=2 specific allele frequencies ranged from 3,137 sibling pairs to 4,246 siblings (mean=3,763; SD=159). When IBD=0, four distinct alleles are observed for a given sibling pair while when IBD=2, two distinct alleles are observed. In that context, the mean sample size for computing the IBD=0 specific allele frequencies was $14,892$ distinct alleles while the mean sample size for computing the IBD=2 specific allele frequencies was $7,526$ distinct alleles.

4.3.2 Detection of Data Artefacts

Initial examination revealed that when applied to the UKBB data, the IBD frequency difference method has a tendency to favour the major allele of a sequence variant as being positively associated with participation. When the three test statistics are investigated separately clear differences become apparent. In particular, while all three are positively correlated with allele frequency, the effect is much stronger for T_{PO} and T_{SIB1} than for T_{SIB20} (see table 4.2 and figure 4.9).

Test statistic	Effect	(SE)
T_{PO}	0.3781	(0.0064)
T_{SIB1}	0.6464	(0.0078)
T_{SIB20}	0.0344	(0.0050)

Table 4.2: Test statistics and allele frequency. This table depicts fitted slopes (Effect) and standard errors (SE) from three linear regressions through the origin with the corresponding test statistic as the response and centred allele frequency (AF-0.5) as an explanatory variable. The regression was performed with `R` (version 3.4.3) in the subset 500,632 sequence variants that passed additional quality control. As the sequence variants are not independent, standard errors were estimated with block jackknife[19] by dividing the 500,632 sequence variants into blocks of 10 MB regions.

The small major allele bias observed with T_{SIB20} could be real in the sense that it is not a result of data artefacts. In addition to that there is no reason to believe that the effects of the major and minor alleles would be perfectly balanced against each other, participation bias itself would generate a slight bias favouring the more common allele. Indeed, for a biallelic sequence variant where both alleles have frequency 0.5 in the population, the sample allele frequency would be > 0.5 for the participation promoting allele and < 0.5 for the other allele. However, the fact that the trend is so much stronger for T_{PO} and T_{SIB1} than for T_{SIB20}, indicates that the major allele tendency of T_{PO} and T_{SIB1} is largely due to data artefacts. Both T_{PO} and T_{SIB1} are allele comparison statistics while T_{SIB20} is a genotype comparison statistic. As a result, T_{PO} and T_{SIB1} would be sensitive to data processing errors that occur when genotypes are split into shared and non-shared alleles, something that does not affect T_{SIB20}.

In chapter 2, I demonstrated that an abundance of sequence variants in the UKBB are subject to genotyping errors that involve overcalling heterozygous genotypes. Such errors could generate a frequency dependent bias for T_{PO} and T_{SIB1}. That can heuristically be understood as follows. When sibling pairs or parent-offspring pairs are both

truly homozygous at an IBD=1 locus, they are, by definition, more likely to be homozygous for the major allele than the minor allele. Whenever one of the homozygous individuals is wrongly called heterozygous in the sample, the inferred IBD=1 shared allele will be the allele that both individuals are truly homozygous for, while the two inferred non-shared alleles would consist of one major and one minor allele. Thus, when aggregated over all pairs, random genotyping errors that involve homozygous carriers being wrongly called as heterozygous carriers would lead to a bias in undercalling the major allele as the non-shared allele when IBD=1. Such genotyping errors will add noise in the computation of T_{SIB20}, but will not bias the statistic towards the major allele.

As was noted in section 4.3.1.7, in a subset of parent-offspring trios where the offspring and one parent is heterozygous while the other parent is homozygous, the error rate of inferring the IBD shared allele of the heterozygous parent from phasing information was 0.54%. Since phasing information is used to infer the shared allele for double heterozygotes sharing a sequence variant IBD=1, phasing errors could also contribute to the major allele bias of T_{PO} and T_{SIB1}. When two heterozygous first-degree relatives share one allele IBD, the chance of allele 1 being the shared allele is one minus its frequency, i.e. the minor allele is more like to be the shared allele. It follows that random phasing errors would induce a negative bias on the minor allele being the shared allele for T_{PO} and T_{SIB1}.

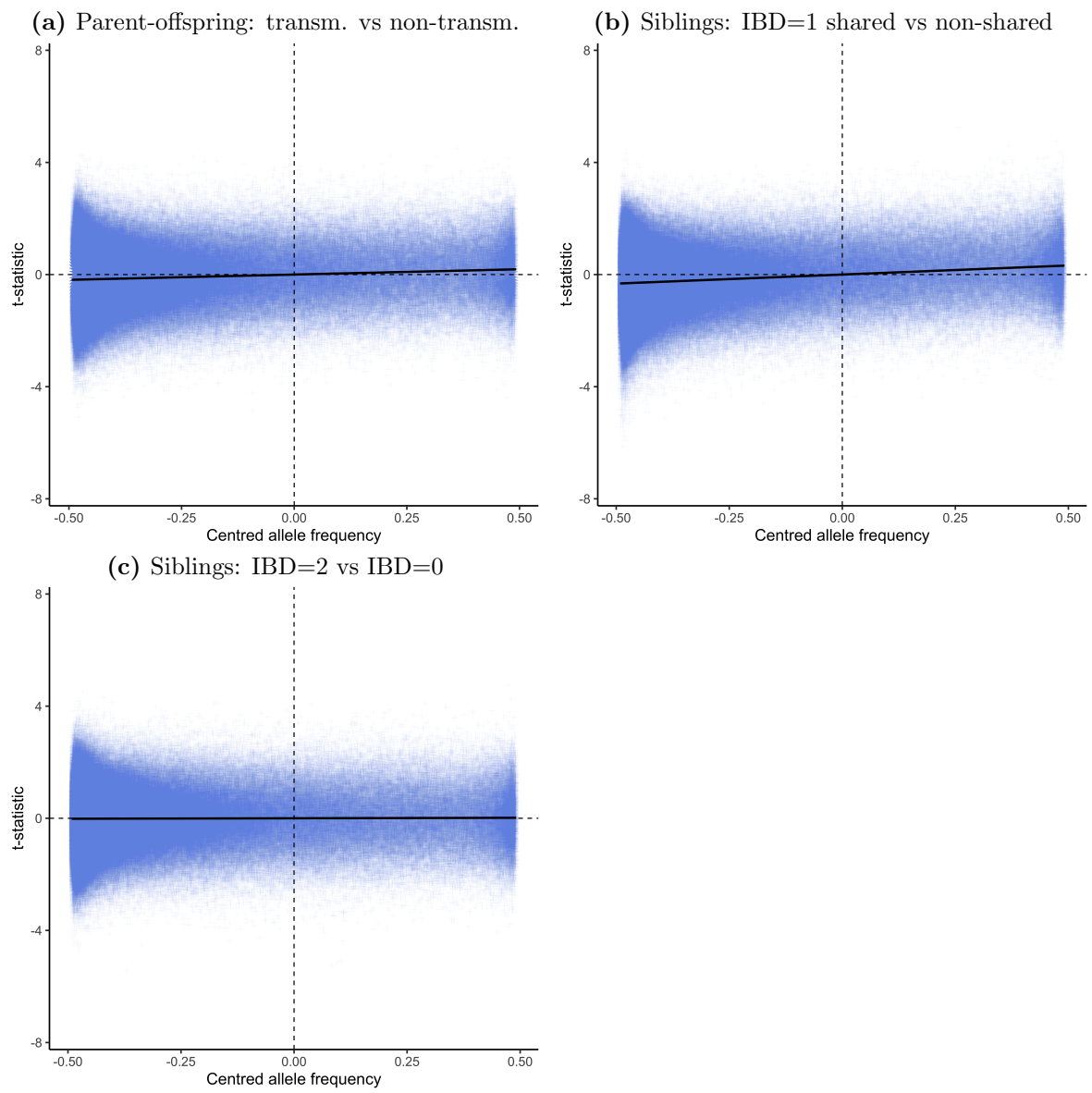


Figure 4.9: Test statistics and allele frequency. This figure depicts the relationship between the three different test statistics (y-axis) and centred allele frequency, $AF - 0.5$, (x-axis). The test statistics based on comparing transmitted and non-transmitted alleles of parent-offspring pairs are shown in figure a). The test statistics based on comparing shared and non-shared alleles of sibling pairs sharing a sequence variant in question IBD=1 are shown in figure b). The test statistic based on comparing allele frequency among sibling pairs sharing a sequence variant IBD=2 to allele frequency among sibling pairs sharing a sequence variant IBD=0 are shown in figure c). Each blue cross depicts a single sequence variant and the black solid lines depict the estimated linear relationship through the origin between centred allele frequency and each of the three specific test statistics as estimated with `R`. The grey dashed lines show the origins in the figure ($x = 0$ and $y = 0$)

I accounted for the major allele bias in subsequent analysis by adjusting T_{PO_k} and T_{SIB1_k} based on the coefficients attained from a linear regression through the origin with the corresponding test statistic as the response and centred allele frequency as an explanatory variable (see table 4.2). The estimated slope is bigger for T_{SIB1} than T_{PO} which is consistent with T_{SIB1} being based on a larger sample size. For each sequence variant k with an allele frequency of p_k, the resulting adjusted statistics were as follows:

$$\begin{aligned} T^*_{PO_k} &= T_{PO_k} - 0.3781 \cdot (p_k - 0.5) \\ T^*_{SIB1_k} &= T_{SIB1_k} - 0.6464 \cdot (p_k - 0.5) \end{aligned} \tag{4.6}$$

Regressing on centred allele frequency instead of allele frequency, ensures that the adjustment is invariant to which allele, major or minor, of a sequence variant is denoted as 1. In particular, the adjustment equals to 0 for sequence variants with $p_k = 0.5$. This does not mean that there are no data artefacts when the allele frequency is equal to 0.5. Rather, this means that when $p_k = 0.5$, the expected number of errors of calling 1 the shared/non-shared allele when it is not, is the same as the expected number of errors of calling 0 the shared/non-shared allele when it is not. Also, the adjustment is positive for $p_k < 0.5$ and negative for $p_k > 0.5$. The distribution of the test-statistics are shown in figure 4.10.

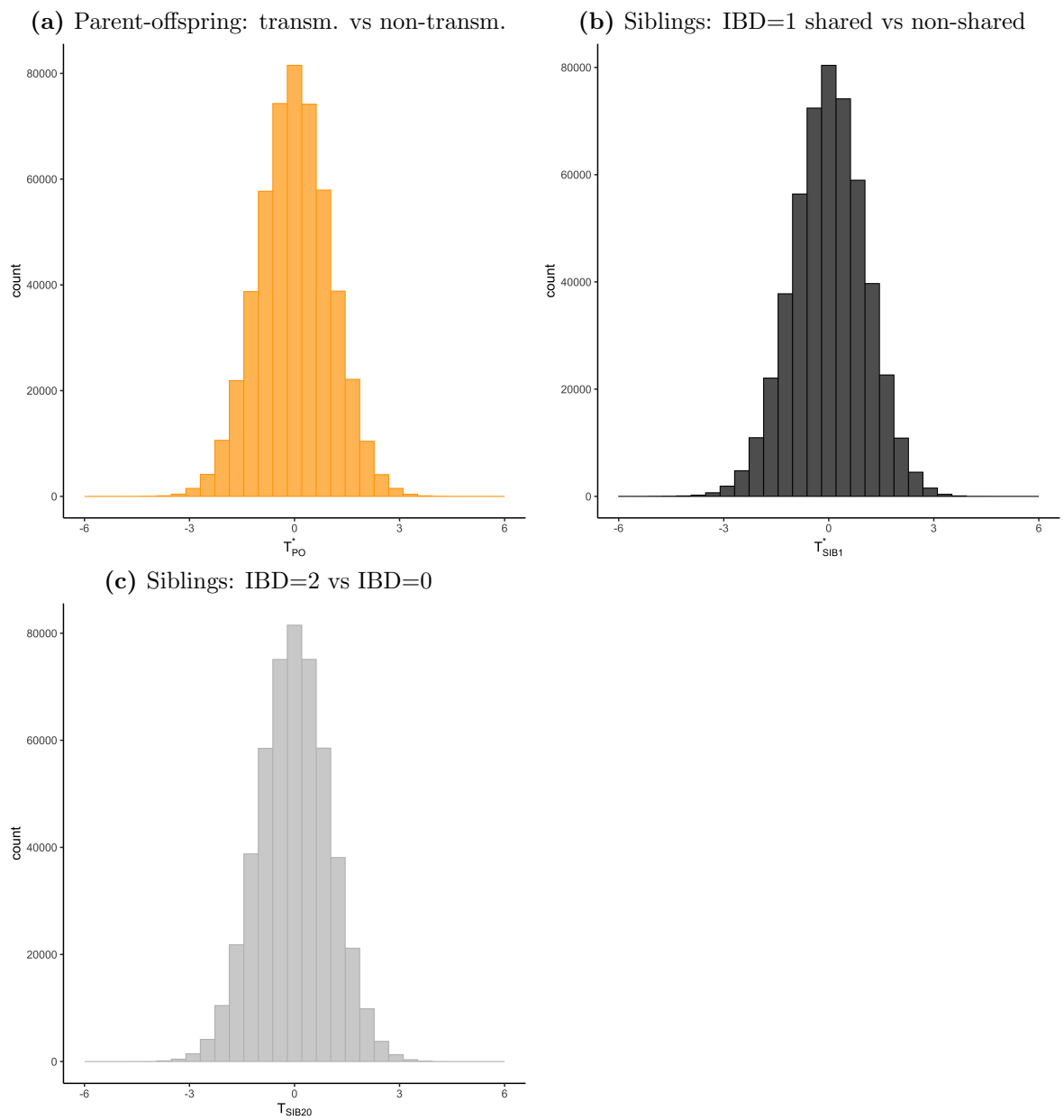


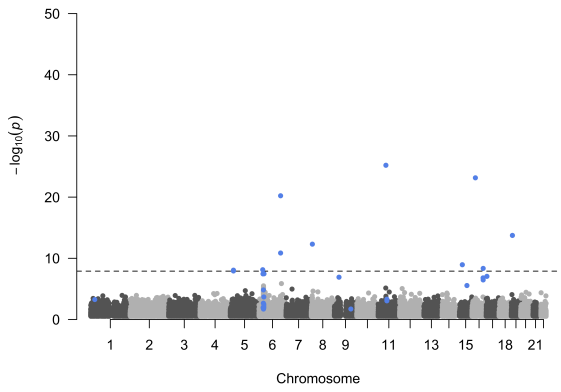
Figure 4.10: Test statistic distribution. This figure depicts histograms of a) the allele frequency adjusted T^*_{PO} statistics, b) the allele frequency adjusted T^*_{SIB1} statistics and c) the unadjusted T_{SIB20} statistics for the 500,632 high-quality biallelic sequence variants.

Because T_{SIB20} is not impacted by genotype splitting errors, results based on T_{SIB20} can be used in a straightforward manner, something that is supported by subsequent analyses. For T_{PO} and T_{SIB1}, the major allele bias and related problems are likely

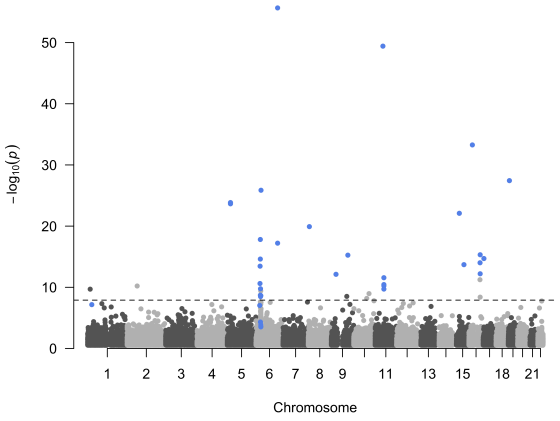
to affect different analyses in a different manner. The major allele bias adjustment described above might not be sufficient for certain types of applications and in those instances it could be better to rely only on the T_{SIB20} statistics. A good example of this is LD score regression. LD score regression involves regressing GWAS summary statistics on LD scores[17]. The intercepts from this regression are commonly used to measure inflation in the test-statistics due to confounding effects such as population stratification and cryptic relatedness[17]. When LD score regression is applied to the T_{SIB20} GWAS results using the program `LDSC` (v1.0.1)[17] (see section A.4), the fitted intercept is essentially equal to 1.0, indicating that the T_{SIB20} statistics are neither inflated because of data artefacts nor affected by issues such as population stratification, the latter consistent with the T_{SIB20} statistics only capturing direct effects. By contrast, when LD score regression is applied to the adjusted T^*_{SIB1} statistics, the fitted intercept is 1.09. Furthermore, the slope from the same regression, which are commonly used to estimate heritability[31], is negative, $-0.1608(SE = 0.0664)$. Bulik-Sullivan et al. (2015) note that although the slopes are usually positive, LD score regression can lead to a negative slope for phenotypes with 'extreme genetic architectures' such as when all causal sequence variants have a minor allele frequency below 1%[17]. However, the slope for the T_{SIB20} statistics is positive, and as T^*_{SIB1} is expected to have the same genetic architecture as T_{SIB20}, the negative slope for T^*_{SIB1} is nearly undoubtedly a consequence of genotype splitting errors. In particular, it can be shown that the errors that lead to the major allele bias would also tend to inflate the association χ^2-statistics by an amount that is negatively correlated with minor allele frequency, i.e. less inflation for sequence variants with close to 50/50 allele frequencies. Since LD scores are positively correlated with minor allele frequency, this explains the negative fitted slope observed for T^*_{SIB1}.

For most sequence variants individually, the bias induced by genotype splitting errors is small and only becomes apparent when the variants are analysed as a group. However, in the set of 157,933 sequence variants that were filtered out because they did not pass the additional quality control filters described in section 4.3.1.6, 35 have a significant association with primary participation at $P < 1.25 \cdot 10^{-8}$ for a combined test-statistic for T^*_{PO}, T^*_{SIB1} and T_{SIB20} (linear combination based on their SEs). These significant associations are solely driven by T^*_{PO} and T^*_{SIB1} as none of these 'low-quality' sequence variants are significant for T_{SIB20} (see figure 4.11). This provides further support that the results based on the T_{SIB20} statistics are very robust to data artefacts and emphasises the importance of examining the results of the three different statistics separately.

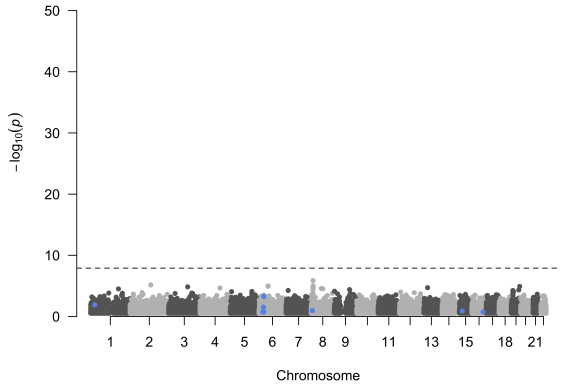
(a) Parent-offspring: transm. vs non-transm.



(b) Siblings: IBD=1 shared vs non-shared



(c) Siblings: IBD=2 vs IBD=0



(d) Combined test statistic

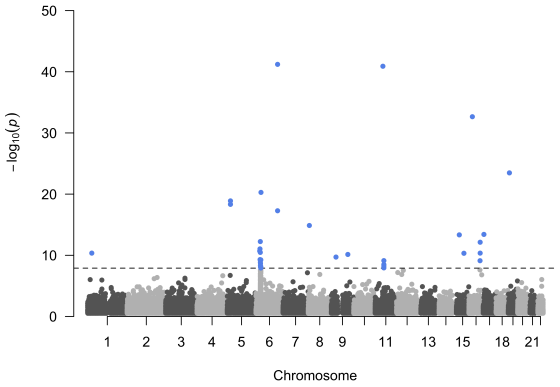


Figure 4.11: Variants that did not fulfil additional QC filters. The figure shows primary participation associations with $P < 0.3$ of the sequence variants that did not fulfil the additional quality control filters described in chapter 4.3.1.6. The black dashed line is set at $P = 1.25 \cdot 10^{-8}$ which is the traditional genome-wide significance threshold taking into account that there are four different test-statistics. The y-axis shows $-log_{10}(P)$ and the x-axis shows chromosomal position in build 37. The 35 sequence variants that surpass the significance threshold in the combined analysis are coloured blue in the figure. a) T^*_{PO}: test statistic based on comparing transmitted and non-transmitted alleles among parent offspring pairs. Adjusted for major allele bias (see chapter 4.3.2). b) T^*_{SIB1}: test statistic comparing shared and non-shared alleles among sibling pairs sharing the sequence variant $IBD = 1$. Adjusted for major allele bias (see chapter 4.3.2). c) T_{SIB20}: Test statistic based on comparing allele frequency among sibling pairs sharing a sequence variant IBD=2 to allele frequency among sibling pairs sharing a sequence variant IBD=0. d) Test statistic based on combining T^*_{PO}, T^*_{SIB1}, and T_{SIB20} into one test statistic using inverse standard errors as weights.

4.3.3 Results

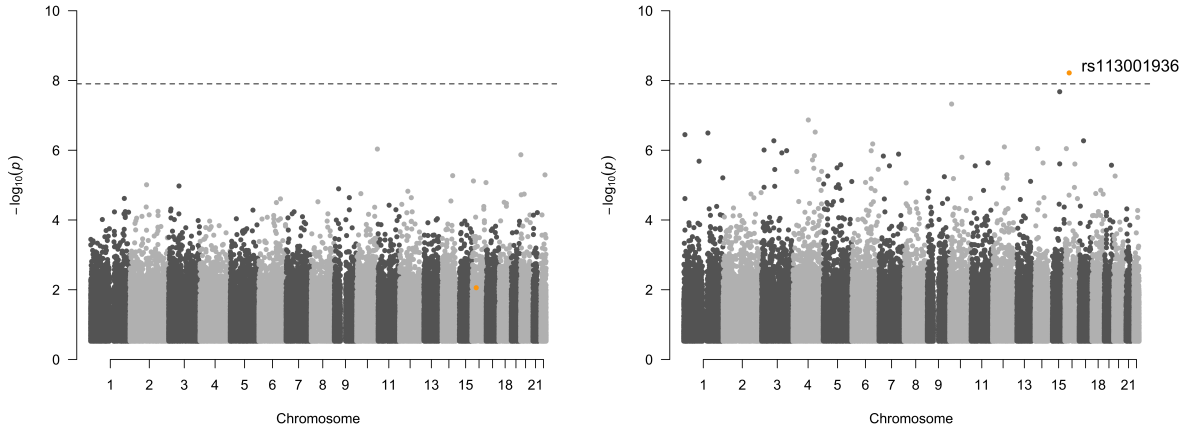
After quality filtering, I tested 500,632 sequence variants for association with primary participation in the UKBB by comparing shared and non-shared alleles of first-degree relatives. The Manhattan figures of the P-values show that for the combined test statistic, one sequence variant, rs113001936 at chr16:27327201 (build 37), is associated with primary participation with $P = 7.2 \cdot 10^{-10}$ (see figure 4.12). A literature search did not provide any obvious reasons for this SNP to have an effect on primary participation in genetic studies. The major allele of rs113001936, G, is the participation promoting allele in the UKBB. In the Genome Aggregation Database[53, 112], the reported minor allele frequency of rs113001936 ranges from 0% among East Asians to 7% among Ashkenazi Jewish while the minor allele frequency among European (non-Finnish) is reported to be 1.55%. rs113001936 is an intron variant that falls on an ensemble canonical transcript for the interleukin-4 receptor, *IL4R*[53, 112]. Interleukin-4 is important for immunoglobulin E production[62] and regulates proliferation and differentiation for a range of lymphoid and myeloid cells[47]. There are reports of variants at *IL4R* associating with eosinophil count [24, 101], asthma [38, 101] and C-reactive protein levels [100]. No phenotype associations are observed for rs113001936 at $P < 5.0 \cdot 10^{-8}$ among the thousands of phenotypes available in PheWeb [34, 89, 90, 46, 110], an online resource that shows GWAS results[2]. Furthermore, rs113001936 is only nominally significant when tested for association with primary participation with the T_{SIB20} statistic ($P = 0.048$). Taken together, even though I would keep an eye on it in future analysis, I currently believe it is not unlikely that this is a false positive, a result of data artefacts

[2]For the variant rs113001936, the PheWeb online resource[34], (`https://pheweb.org`, accessed Sept. 2022), includes association results for 1,419 EHR-derived phenotypes in the UKBB[90, 110] and association results from the Neale lab for 2,419 phenotypes in the UKBB[89, 46].

or sampling variation.

(a) Parent-Offspring: transm. vs non-transm.

(b) Siblings: IBD=1 shared vs non-shared



(c) Siblings: IBD=2 vs IBD=0

(d) Combined test statistic

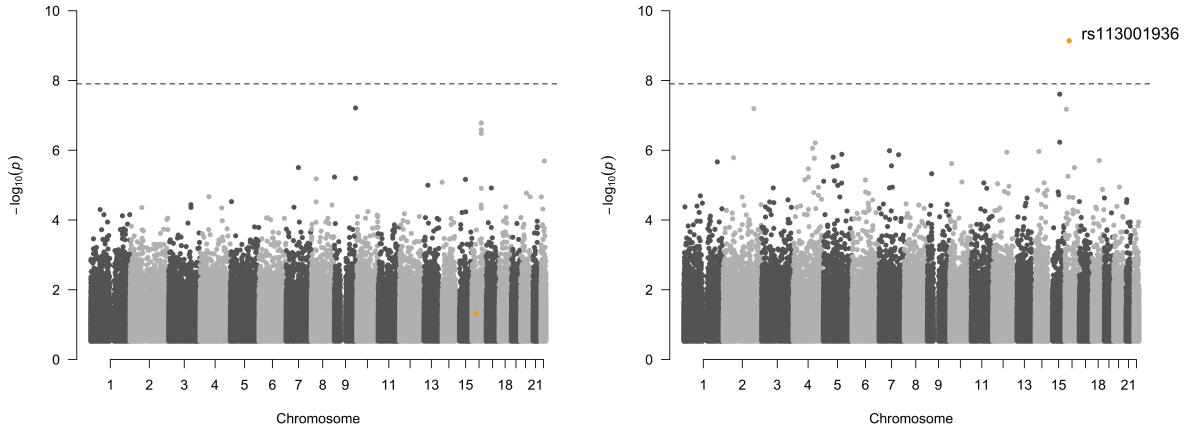
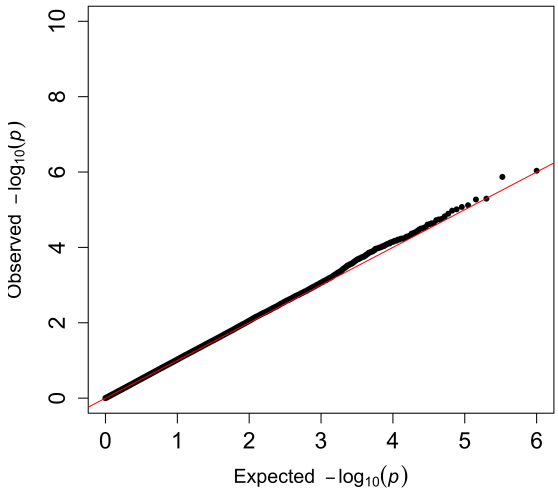
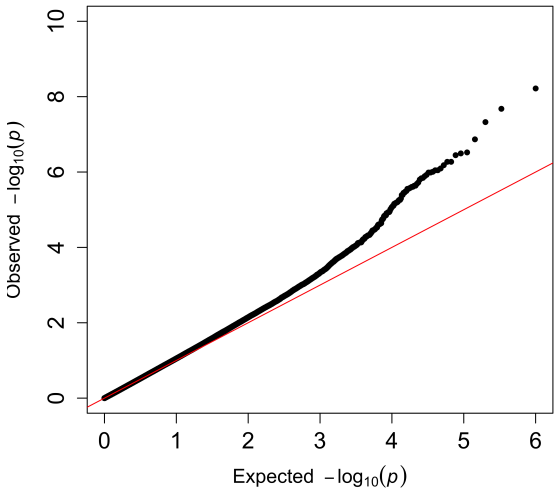


Figure 4.12: Manhattan figures: Primary participation GWAS in the UKBB. The figure depicts sequence variant associations with primary participation in UKBB. The figure shows associations with $P < 0.3$ among the 500,632 high-quality sequence variants that surpassed additional quality filters described in chapter 4.3.1.6. Each dot in the figure represents one sequence variant. The dot representing rs113001936 is coloured orange. The y-axis shows $-log_{10}(P)$ and the x-axis shows chromosomal position in build 37. The black dotted line is set at $P = 1.25 \cdot 10^{-8}$ which is the traditional genome-wide significance threshold taking into account that I'm testing 500,632 sequence variants in 4 different instances. a) T^*_{PO}: test statistic based on comparing transmitted and non-transmitted allele among parent offspring pairs. Adjusted for major allele bias (see chapter 4.3.2). b) T^*_{SIB1}: test statistic comparing shared and non-shared alleles among sibling pairs sharing the sequence variant $IBD = 1$. Adjusted for major allele bias (see chapter 4.3.2). c) T_{SIB20}: Test statistic based on comparing allele frequency among sibling pairs sharing a sequence variant IBD=2 to allele frequency among sibling pairs sharing a sequence variant IBD=0. d) Test statistic based on combining T^*_{PO}, T^*_{SIB1}, and T_{SIB20} into one test statistic using their inverse standard errors as weights.

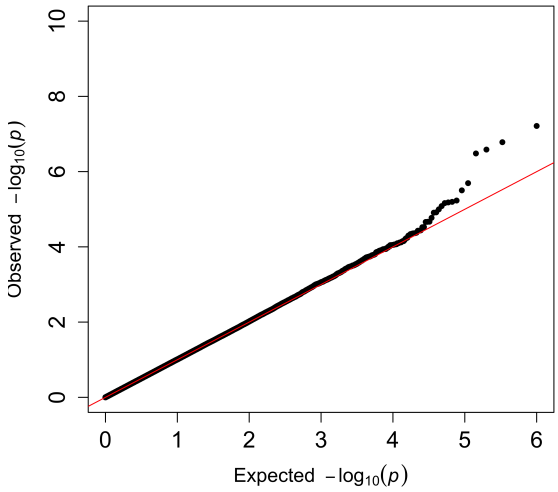
(a) Parent-Offspring: transm. vs non-transm.



(b) Siblings: IBD=1 shared vs non-shared



(c) Siblings: IBD=2 vs IBD=0



(d) Combined test statistic

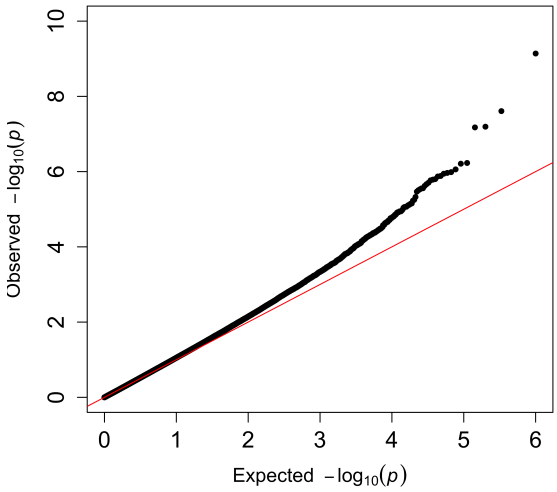


Figure 4.13: QQ-plots: Primary participation GWAS in the UKBB. The figure depicts qq-plots for sequence variant associations with primary participation in the UKBB. The figure shows the 500,632 high-quality sequence variants that surpassed additional quality filters described in chapter 4.3.1.6. Each dot in the figure represents one sequence variant. The y-axis shows the observed $-log_{10}(P)$ while the x-axis shows the expected $-log_{10}(P)$. a) T^*_{PO}: test statistic based on comparing transmitted and non-transmitted allele among parent offspring pairs. Adjusted for major allele bias (see chapter 4.3.2). b) T^*_{SIB1}: test statistic comparing shared and non-shared alleles among sibling pairs sharing the sequence variant $IBD = 1$. Adjusted for major allele bias (see chapter 4.3.2). c) T_{SIB20}: Test statistic based on comparing allele frequency among sibling pairs sharing a sequence variant IBD=2 to allele frequency among sibling pairs sharing a sequence variant IBD=0. d) Test statistic based on combining T^*_{PO}, T^*_{SIB1}, and T_{SIB20} into one test statistic using their inverse standard errors as weights.

QQ-plots of the P-values are shown in figure 4.13. For T^*_{PO}, depicted in panel a), an excess of sequence variants with $P < 0.001$ is observed, while an excess of sequence variants with $P < 0.01$ is observed for T^*_{SIB1} (panel b) and an excess of sequence variants with $P < 3.0 \cdot 10^{-5}$ is observed for T_{SIB20} (panel c). This difference further supports the robustness oft T_{SIB20} compared to T^*_{PO} and T^*_{SIB1}.

4.3.3.1 Within-Sample Replication

Even though there is no single sequence variant that shows convincing association with primary participation, that is not unexpected given that the effective sample size here is not that large for a complex trait. To validate that the primary participation GWAS results are capturing useful information about the genetics underlying participation in the UKBB, I computed three primary participation polygenic scores ($pPGS$), one for each of the three test statistic (T^*_{PO}, T^*_{SIB1} and T_{SIB20}). Values of the $pPGS$, standardized to have variance one, were computed for 272,409 WB individuals with no close relatives in the UKBB ($> 3rd$ degree for all pairs, see further information in section A.2). This group of individuals is referred to as the unrelateds here and do not overlap with the first-degree relatives used. I examined associations between each of the $pPGS$ and various phenotypes in the group of unrelateds (see section A.1 and A.3). Figure 4.14 and tables B.1, B.2 and B.3 show some of the strongest associations and include a few non-significant ones as well for reference. The T_{SIB20} $pPGS$ and the T^*_{SIB1} $pPGS$ are based on the same set of sibling pairs and are supposed to estimate essentially the same effects with comparable power. Thus it is reassuring to see that their associations with the various traits are compatible in general. In particular, with multiple-comparison adjustment, the difference between the T_{SIB20} $pPGS$ and

the T^*_{SIB1} *pPGS* associations is not significant for any of the traits. Considering that the T^*_{PO} *pPGS* is based on a smaller sample size, its association results are also generally in agreement. This shows that the genotype splitting errors impacting T^*_{PO} and T^*_{SIB1} only have a small effect on polygenic score prediction. Indeed, even without the allele frequency adjustment, the T_{PO} and T_{SIB1} *pPGSs* would have similar associations with the phenotypes (see tables B.4 and B.5). These empirical results are consistent with the theoretical understanding that the major allele bias is mainly adding a small amount of noise to polygenic score predictions.

Given that the individual *pPGSs* provide compatible predictions, I also constructed a *pPGS* based on the linear combination of all three statistics (see table 4.3 and figure 4.14). For the combined *pPGS* the strongest association is with educational attainment (EA) where the effect (correlation) is 0.0310 with $P = 3.6 \cdot 10^{-53}$. The next strongest association is with BMI ($P = 5.0 \cdot 10^{-22}$) where the effect is notably negative. The effect, 0.0300, is nearly as strong for age at first birth (AFB) of women as for EA ($P = 9.0 \cdot 10^{-21}$). (Note that the P-value is not as small as that for BMI due to a smaller sample size). These results, consistent with reported differences in EA and BMI between the UKBB sample and the UK population[33, 80], validate that the GWAS that I performed without any phenotype data can nonetheless capture genetic associations with primary participation.

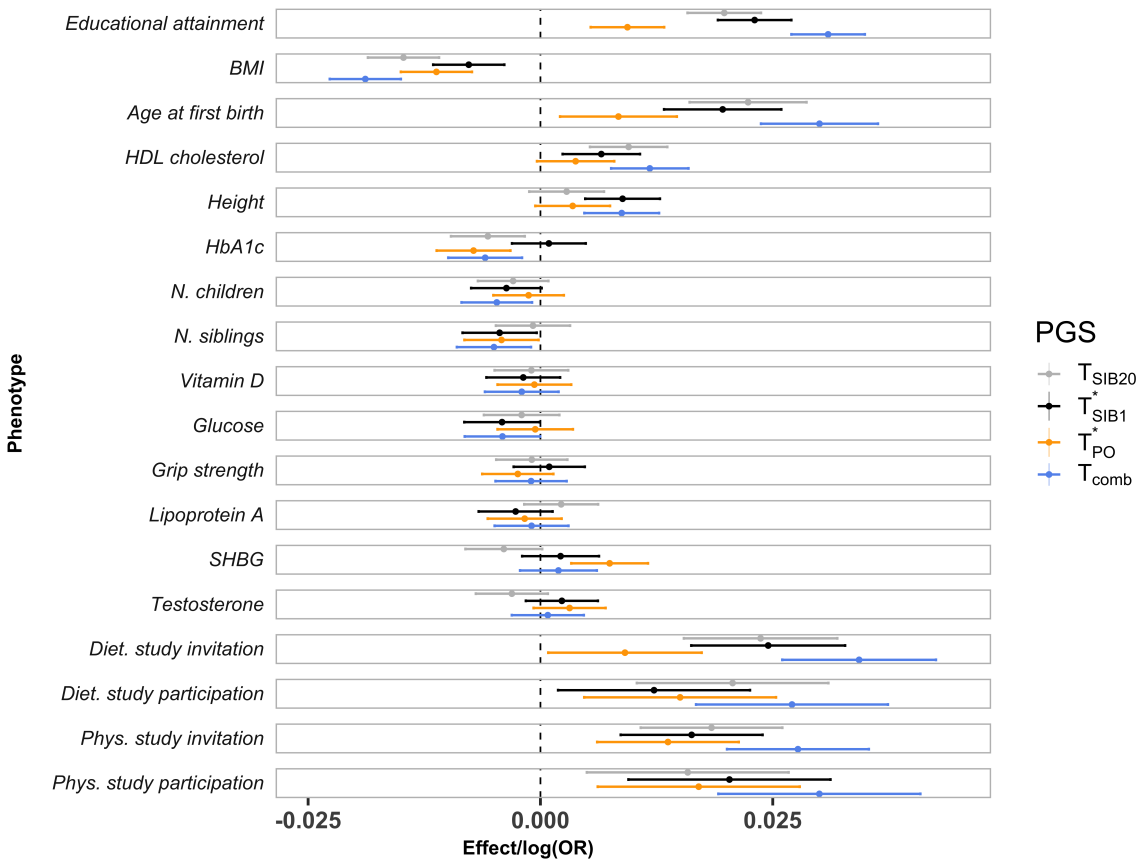


Figure 4.14: Primary participation polygenic scores (*pPGS*) computed with different weights. Each dot depicts the effect in SD or log(OR) of the *pPGS* (x-axis) on the corresponding phenotype (y-axis). The horizontal lines depict 95% confidence intervals taking the corresponding LD score regression intercept into account. The grey dots correspond to *pPGS* computed with T_{SIB20} weights, the black dots correspond to *pPGS* computed with T^*_{SIB1} weights, the orange dot correspond to *pPGS* computed with T^*_{PO} weights and the blue dots correspond to *pPGS* computed with weights that are a linear combination of the other three, T_{comb}. For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as the response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of WB descent who have no close relatives in the UKBB (>3rd degree for all pairs). Prior to linear regression, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex and then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). For the binary phenotypes, a logistic regression was performed in `R` including sex, YOB, age up to the order of three, 40 PCs, and genotyping array as additional covariates. Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. The sample size was 98,653 individuals for age at first birth, and ranged from 235,960 to 272,409 individuals for the other phenotypes (see table 4.3).

Quantitative traits	**Effect(SD)**	**P-value**	**N**
Educational attainment	0.0310	$3.6 \cdot 10^{-53}$	260,950
BMI	−0.0188	$5.0 \cdot 10^{-22}$	271,535
Age at first birth	0.0300	$9.0 \cdot 10^{-21}$	98,653
HDL cholesterol	0.0118	$3.0 \cdot 10^{-8}$	237,785
Height	0.0087	$2.0 \cdot 10^{-5}$	271,820
HbA1c	−0.0059	$3.3 \cdot 10^{-3}$	259,594
N. siblings	−0.0050	$1.4 \cdot 10^{-2}$	268,191
N. children	−0.0047	$1.5 \cdot 10^{-2}$	271,317
Glucose	−0.0041	$4.9 \cdot 10^{-2}$	237,629
Vitamin D	−0.0020	0.32	249,079
Lipoprotein A	−0.0010	0.64	251,698
SHBG	0.0019	0.36	235,960
Grip strength	−0.0010	0.61	270,525
Testosterone	0.0008	0.69	257,570
Binary traits	**log(OR)**	**P-value**	**Ncase/Ncontrol**
Diet. study invitation	0.0343	$4.5 \cdot 10^{-16}$	166,993/105,416
Diet. study participation	0.0271	$2.8 \cdot 10^{-7}$	54,124/112,869
Phys. study invitation	0.0277	$1.1 \cdot 10^{-12}$	132,633/139,776
Phys. study participation	0.0300	$6.3 \cdot 10^{-8}$	59,455/73,178

Table 4.3: Primary participation polygenic score (*pPGS*) and phenotypes. For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as a response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of white British descent who have no close relatives in UK Biobank (>3rd degree for all pairs). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex and then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated Haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Diet. study: Dietary study. Phys. study: Physical activity study. Information about age at first birth was only available for women. For the binary traits, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for the *pPGS* and the corresponding P-value. The P-values have been adjusted with the LD score regression intercept of the corresponding phenotype. This table shows results for the *pPGS* computed with weights that represent a weighted linear combination of the three test-statistics T^*_{PO}, T^*_{SIB1} and T_{SIB20}.

After the baseline assessments, subsets of the UKBB participants were invited to fur-

ther participate in a dietary study in 2011-2012 where invited participants were asked to answer a dietary questionnaire[13] and a physical activity study in 2013-2015 where invited participants were asked to wear an accelerometer for a week[29]. Not everybody was invited (the criteria for invitation included having a valid email address) and only a subset of those invited actually participated (see sections A.1.15 and A.1.16). I refer to participation in these follow-up studies as 'secondary participation'. For the dietary study, the estimated effect of the *pPGS* on being invited, in log(OR) is of 0.0343 ($P = 4.5 \cdot 10^{-16}$). For those invited, the estimated effect on actual participation in log(OR) is 0.0271 ($P = 2.8 \cdot 10^{-7}$). For the physical activity study, the corresponding effect estimates are 0.0277 ($P = 1.1 \cdot 10^{-12}$) and 0.0300 ($P = 6.3 \cdot 10^{-8}$) respectively.

4.3.3.2 Adjusting for Educational Attainment (EA)

Associations of the combined *pPGS* were further examined with adjustment for EA (figure 4.15). For phenotypes with unadjusted $P < 0.05/18$, the adjusted effects shrink but all associations remain significant at $P < 0.05$. Relatively, the effect on participating in the physical study when invited shrinks the least, by 18.0%, from 0.0300 to 0.0246, while the effect on dietary study invitation shrinks by 37.7%, from 0.0343 to 0.0214. This difference in shrinkage is partly because the estimated effect of EA on dietary study invitation is 0.4661, much larger than its effect on physical activity study participation, 0.2621 (see table B.6). From the EA effects alone, the participation bias would appear to be much stronger with dietary study invitation. By contrast, on the genetic level, the participation bias for physical study participation is stronger than that for dietary study invitation after adjustment for EA. Thus, while the genetic component to participation is associated with EA, its effects on other traits is

not manifested mainly through EA. Importantly, phenotypes known to correlate with participation do not fully capture the nature and magnitude of participation bias.

4.3.3.3 Sex-Specific Analysis

When males and females are analysed separately for the phenotypes in table 4.3, with multiple comparison adjustment, no significant difference is found for the estimated effects of the combined *pPGS*. Furthermore, while it has been reported that UKBB participation rates differ by sex and age[33], the *pPGS* is associated with neither in the group of 272,409 WB 'unrelated' individuals ($P > 0.05$). This implies that the effect of the *pPGS*, based on autosomal variants, is additive to the effects of sex and age on participation, with no detectable statistical interactions.

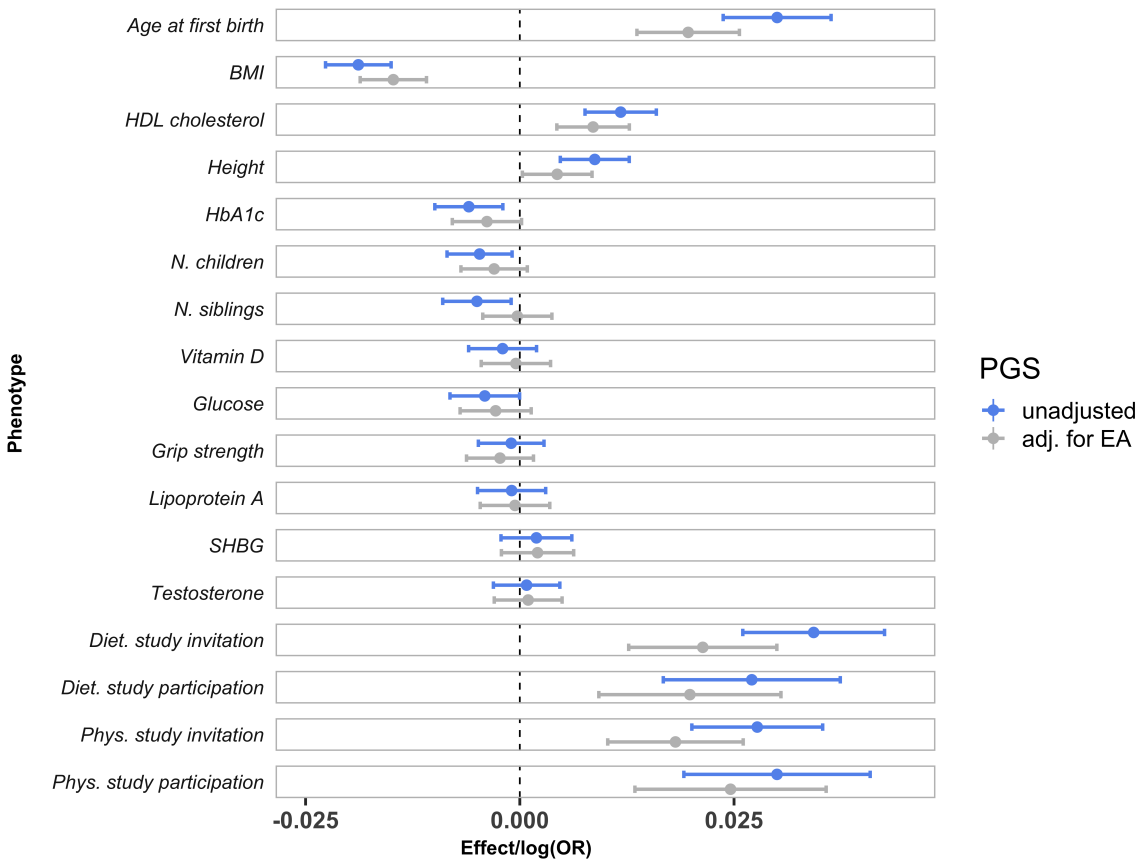


Figure 4.15: Primary participation polygenic score (*pPGS*) and educational attainment. Each dot depicts the effect in SD or log(OR) of the *pPGS* (x-axis) on the corresponding phenotype (y-axis). The horizontal lines depict 95% confidence intervals taking the corresponding LD score regression intercept into account. The blue dots correspond to regression not adjusted for educational attainment (EA) while the grey dots correspond to a regression including EA as an additional covariate. The sample size was 98,653 individuals for age at first birth and ranged from 235,960 to 272,409 individuals when not adjusted for EA. The sample size was 95,424 for age at first birth and ranged from 226,220 to 260,950 individuals when EA was added as an additional covariate. Genotyping array (BiLEVE or Axiom) was an additional covariate in both regressions. Sequence variant weights for the *pPGS* correspond to the combined test-statistic. Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex and then the residuals were rank-based inverse normalised except for EA for which the residuals were standardised (see section A.1). For the binary phenotypes, I performed a logistic regression in `R` including sex YOB, age up to the order of three, 40 PCs, and genotyping array as additional covariates. Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin.

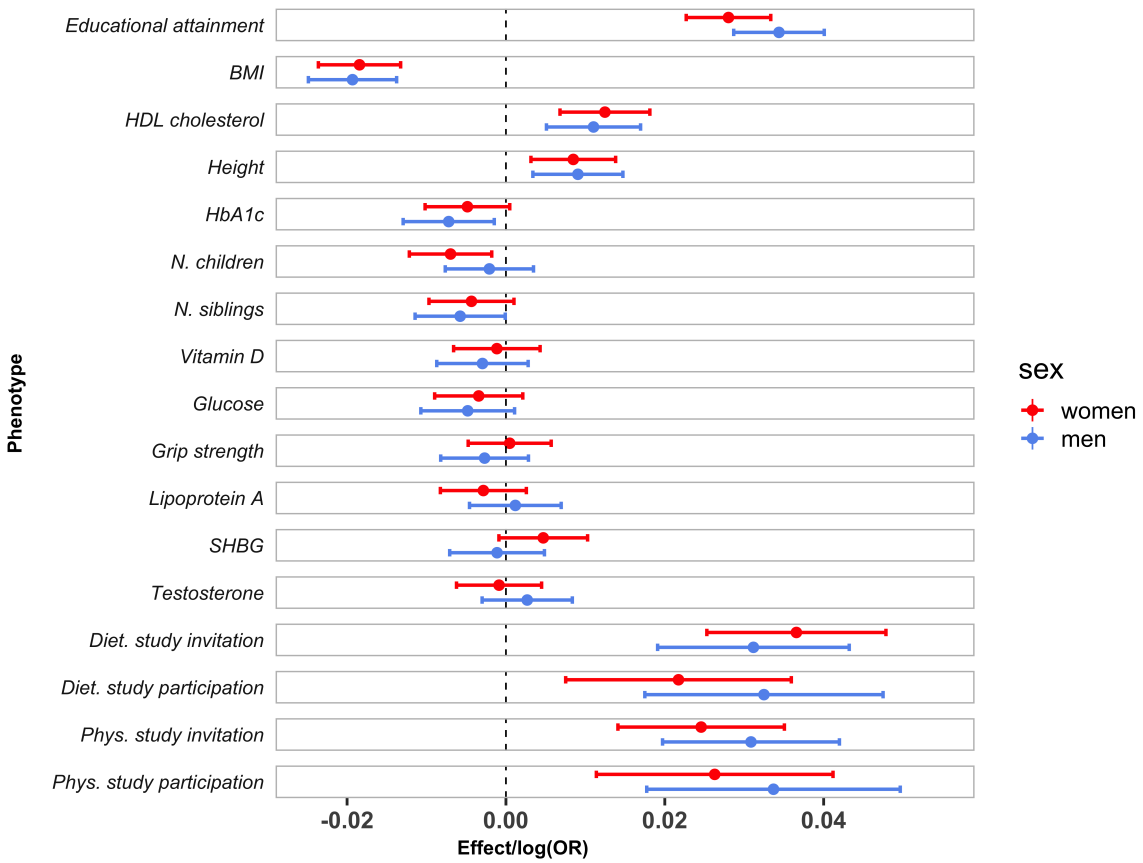


Figure 4.16: Primary participation polygenic score (*pPGS*) and sex-specific analysis. Each dot depicts the effect in SD or log(OR) of the *pPGS* (x-axis) on the corresponding phenotype (y-axis). The horizontal lines depict 95% confidence intervals taking the corresponding LD score regression intercept into account. The blue dots correspond to regression in the subset of men while the red dots correspond to regression in the subset of women. Genotyping array (BiLEVE or Axiom) was included as an additional covariate in the regressions. For men, the sample size ranged from 111,219 individuals to 127,186 individuals and for women the sample size ranged from 124,741 to 145,223 individuals. Sequence variant weights for the *pPGS* correspond to the combined test-statistic (not sex-specific weights). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex and then the residuals were rank-based inverse normalised, except for educational attainment for which the residuals were standardised, also separately for each sex (see section A.1). For the binary phenotypes, I performed a logistic regression in `R` separately for each sex, including YOB, age up to the order of three, 40 PCs, and genotyping array as additional covariates. Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin.

4.3.3.4 Differences Between Siblings and the Unrelated

The second principle of genetically induced participation bias describes how, ignoring confounding factors, individuals with participating close relatives are usually more likely to harbour allele types that increase the probability of participation than individuals whose close relatives do not participate. To examine this empirically using the UKBB data in a way that avoids over-fitting, I randomly partitioned the sibling pairs into two halves, derived a *pPGS* based on GWAS results performed with the first half (*pPGS1*), and compared the *pPGS1* values of sibling pairs in the second half with the unrelated group. The splitting reduces power. Nonetheless, the second half of the sibling pairs have average *pPGS1* values that is 0.044 SD higher than that of the unrelateds ($P = 5.0 \cdot 10^{-6}$, unpaired t-test). Switching the roles of the two halves of the sibling pairs, the first half of the sibling pairs have average *pPGS2* that is 0.026 SD higher than that of the unrelateds ($P = 8.5 \cdot 10^{-3}$, unpaired t-test). Importantly, these results provide further support that the method of comparing IBD shared and non-shared alleles of participating first-degree relatives is capturing real and significant information about the genetics underlying primary participation.

4.3.3.5 Genetic Correlations

I applied LD score regression with the program `LDSC`[17] to estimate the correlations between the genetic component underlying primary participation and the genetic components of the phenotypes depicted in table 4.3. Results for traits where the 95% confidence intervals of the correlation estimates do not include zero are presented in table 4.4. To avoid complications of genotyping splitting errors, the LD score regression analysis for primary participation was based on the between-sibling GWAS only

(T_{SIB20}). For the other traits, GWAS results are based on the same unrelateds used for the results in table 4.3 (see further information in A.4). Given the *pPGS* associations, it is not surprising that the primary participation genetic correlations with EA and the secondary participation traits are all significantly positive. However, while not insubstantial, those correlations are all below 0.5. That could be partly because, while T_{SIB20} only captures direct genetic effects, the other GWASs were performed in the 'standard' manner and thus capture population effects that include both direct and non-direct genetic effects[135]. To investigate this further, for EA, I also included in table 4.4 results based on a recently reported direct-effect-only GWAS (EA direct)[135] performed with UKBB first-degree relatives. The correlation estimate between primary participation and EA direct is indeed higher (0.59) although the SE is not small (0.28). Taken together, the genetic correlation estimates show that results obtained from studying EA, BMI, age at first birth or secondary participation are not adequate in providing a proper appreciation of the nature and magnitude of the genetic component underlying individuals' participation in UKBB.

Phenotype	rg(SE)
EA direct	0.5915 (0.2847)
EA	0.3689 (0.0886)
BMI	-0.2193 (0.0643)
Age at first birth	0.4520 (0.1161)
Dietary study invitation	0.4192 (0.1109)
Dietary study participation	0.3864 (0.1365)
Physical activity study invitation	0.4264 (0.1336)
Physical activity study participation	0.3646 (0.1393)

Table 4.4: Genetic correlations between primary participation and other phenotypes. This table depict genetic correlations between the primary participation GWAS results based on T_{SIB20} described in chapter 4.3.1.9 and a range of phenotypes (rg(SE)). These correlations were attained via LD score regression with the program `LDSC` (version 1.0.1)[17] using the European Ancestry LD scores computed by the Pan-UKB team[85]. EA: Educational attainment. EA direct refers to GWAS summary statistics from a recently published direct-effect-only GWAS on educational attainment[135]. The other GWAS summary statistics were attained by performing GWAS in the group of 272,409 'unrelated' WB UKBB individuals using `BOLT-LMM` (version 2.3)[66]. Results shown here correspond to a restricted intercept LD score regression (see section A.4).

4.4 Studying Relationships Between Various Frequency Differences by Simulations

Application to UKBB data demonstrated that by comparing IBD shared and non-shared alleles of participating first-degree relatives, we are able to study the genetic component of primary participation. Under very broad conditions, if an allele promotes participation, its frequency in the shared segments of participating first-degree-relative pairs would be higher than its frequency in the non-shared segments. One obvious question of interest is the quantitative relationship between this frequency difference and the difference of allele frequencies in participants and non-participants. To investigate this, I performed the following simulations described below. The simulations are based on a simplified model and can therefore not capture the real conditions perfectly.

However, they provide a reasonable guide to understanding the results attained from the application to the UKBB.

In the simulations, it was assumed that we had a population of sibling pairs in which the probability of participation was influenced by a standard normally distributed trait, X_{ij}, which was defined as the weighted sum of two uncorrelated variables, z_{ij} and B_{ij}. Here, the sibling pairs in the population are indexed with j and the siblings within each pair are indexed with $i = 1, 2$. z_{ij} denotes a standardised and mean centred genotype of a single SNP and B_{ij} denotes a standard normally distributed variable that captures all contribution to X_{ij} beyond z_{ij}, genetic and/or environmental. For sibling pair j we have,

$$\begin{aligned} X_{1j} &= w_z \cdot z_{1j} + w_B \cdot B_{1j} \\ X_{2j} &= w_z \cdot z_{2j} + w_B \cdot B_{2j}. \end{aligned} \tag{4.7}$$

It was assumed that

$$w_z^2 + w_B^2 = 1 \tag{4.8}$$

so that $var(X_{ij}) = 1$. The standardised genotype, z_{ij}, is capturing IBD shared and IBD not shared alleles between siblings, with the average being IBD=1. The contribution of B_{ij} could be shared and/or non-shared between siblings, with the sibling correlation within the population, ρ_B, defining how much of the B component is shared on average between siblings. The simulations were performed for a range of ρ_B (0.01 to 0.99), which allowed for different enrichment of participating sibling pairs. That is, the higher the ρ_B, the higher the sibling recurrence risk ratio (SRRR) and hence the higher fraction of participating sibling pairs. As was noted in section 4.2.3, multiple factors, both environmental and genetic, are likely to influence the relative recurrence risk ratio for

participation. It was assumed that the probability of participation was described with the logit model:

$$P(Z_{ij} = 1|X_{ij} = X) = \frac{1}{1 + e^{-\beta_0 - \beta_1 \cdot X}} \tag{4.9}$$

The logit model implies an additive effect on a log-odds scale, and was chosen for these simulations as various studies have indicated that a substantial fraction of genetic variance is explained by additive effects[84, 42].

The simulations were performed for two different values of w_z, $\sqrt{0.0010}$ and $\sqrt{0.0025}$. These choices of w_z corresponded to the sequence variant contributing 0.10% and 0.25% to the variance of X_{ij} respectively and B_{ij} explaining the other 99.90% and 99.75% of the variance. The overall participation rate within the population of sibling pairs, depends on β_0 and β_1. Six different participation rates were investigated, 0.055, 0.20, 0.40, 0.60, 0.80 and 0.945. A participation rate of 0.055 matches the participation rate in the UKBB[33] and a participation rate of 0.40 approximately matches the participation rate in the MoBA study[69] and the Icelandic population fraction in the deCODE study[5]. The simulations were performed assuming $\beta_1 = 1.2$ ($OR \approx 3.3$) and β_0 was made to take the value that resulted in the intended participation fraction. The choice of β_1 was loosely based on the UKBB. Bycroft et al. (2018) reported that twice as many siblings participated in the UKBB compared with what would have been expected under random sampling (SRRR=2)[20]. To attain this degree of sibling enrichment, β_1 had to be big enough to allow for a substantial participation correlation between siblings. The value $\beta_1 = 1.2$ was chosen so that given a participation rate of 0.055, the SRRR would be equal to 2 when $\rho_B = 0.70$. In comparison, when the overall participation rate is 0.40, $\beta_1 = 1.2$ corresponds to a SRRR$\approx$1.2 when $\rho_B = 0.70$.

I simulated trait values of 5 million sibling pairs in accordance to equation 4.7 and then allocated them into participants and non-participants assuming a participation probability described with equation 4.9. The simulation process was repeated 2,400 times for five different population allele frequency scenarios, p, for the single SNP (0.05, 0.25, 0.5, 0.75, 0.95). The allele frequency difference between participants and non-participants, $\Delta_{PART} = p_1 - p_0$, and the allele frequency difference between IBD shared and non-shared alleles of the participation sibling pairs, Δ_{IBD}, were then computed from the unstandardised genotypes underlying z_{ij}.

The results from the simulations are shown in figures 4.17, 4.18, 4.19, 4.20 and C.1. As the frequency difference from comparing IBD1 shared to IBD1 non-shared alleles were close to identical to the frequency difference from comparing sibling pairs sharing the SNP IBD=2 and IBD=0 respectively, I depict their average in the figures, Δ_{IBD}. Taken together, the simulations show that the quantitative relationship between Δ_{IBD} and Δ_{PART} depends on the total participation rate, enrichment of participating sibling pairs, allele frequency and effect size. The relative difference, $\Delta_{IBD}/\Delta_{PART}$, declines with sibling enrichment. When the total participation rate is 0.055 (figure 4.17a and 4.17b), for a sequence variant with a population allele frequency of 0.5, Δ_{IBD} ranges from being 94% of Δ_{PART} when SRRR=1 to being 82% when SRRR=2.5. When the total participation fraction is 0.40 (figure 4.17c and 4.17d), Δ_{IBD} ranges from being 60% of Δ_{PART} when SRRR=1 down to being 49% of Δ_{PART} for SRRR=1.3.

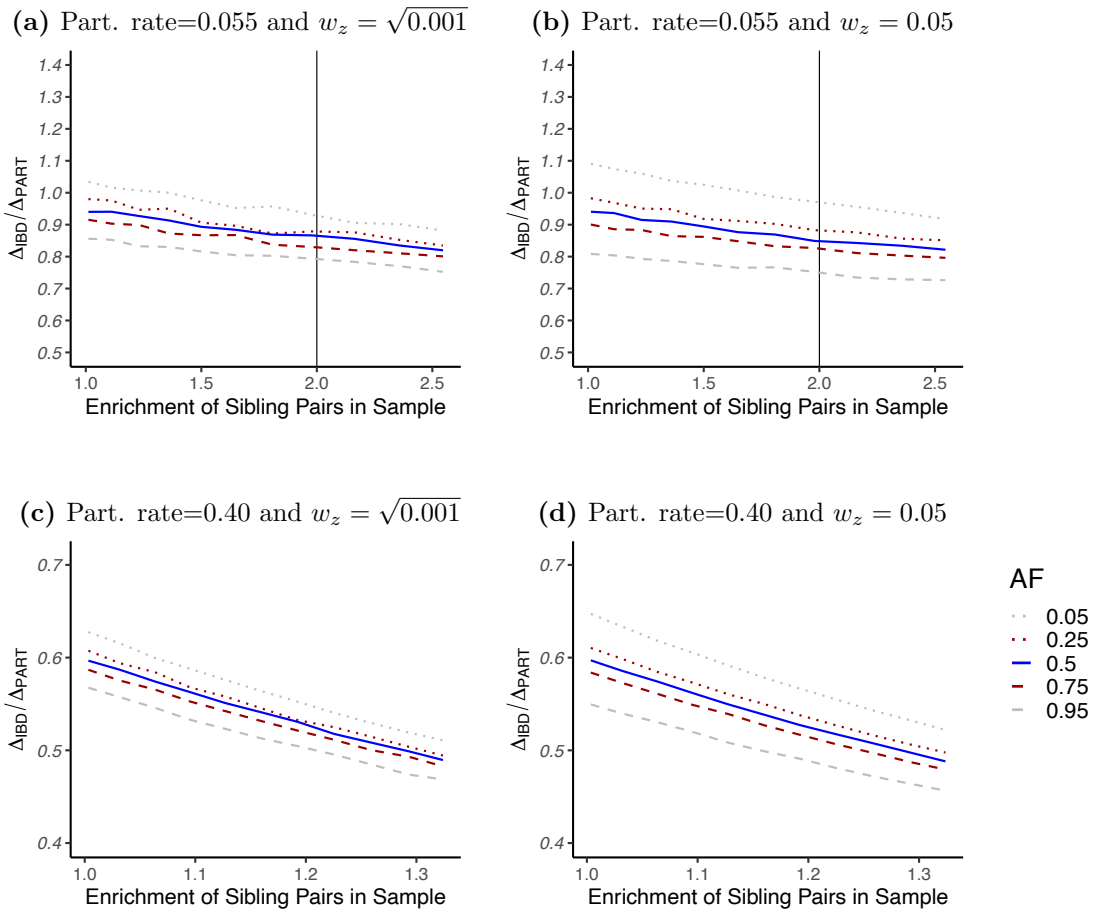


Figure 4.17: Difference between IBD shared and IBD non-shared alleles. Figures showing how the relative allele frequency difference $\Delta_{IBD}/\Delta_{PART}$ (y-axis) changes with the enrichment of sibling pairs in sample measured with the sibling recurrence risk ratio (x-axis) for different SNP effects and participation rates. Δ_{IBD} denotes the allele frequency difference between IBD shared and IBD non-shared alleles of participating siblings. Δ_{PART} denotes the allele frequency difference between participants and non-participants. These allele frequency differences were attained by simulating trait values, $X_{ij} = w_z \cdot z_{ij} + w_B \cdot B_{ij}$, of 5 million sibling pairs and then allocating them into participants and non-particiants assuming the logit model; $P(Z_{ij} = 1|X_{ij} = X) = \frac{1}{1+e^{-\beta_0 - 1.2 \cdot X}}$ with β_0 taking values corresponding to a participation rate of 0.055 (a and b) and 0.40 (c and d) respectively. z_{ij} denotes the standardised and mean centred genotype of a single biallelic SNP for individual i in sibling pair j. $B_{ij} \sim N(0, 1)$ denotes all other contributions to X_{ij}. z_{ij} and B_{ij} are uncorrelated while the sibling correlation for z_{ij} is 0.5 in the population. w_z was set at $\sqrt{0.001}$ (a and c) and 0.05 (b and d) respectively. The sibling correlation of B_{ij} in the population ranged from 0.01 to 0.99 in the simulations, allowing for different sibling enrichment. This was done for 5 different effect allele frequencies (AF) for the underlying genotype of z_{ij}. The simulation process was repeated 2,400 times and then the mean was computed from those replications. Δ_{PART} and Δ_{IBD} were computed for the unstandardised genotypes underlying z_{ij}. The solid, dotted and dashed curves depict the relative frequency difference for each AF scenario (legend at bottom right). The black solid vertical line in figures a) and b) is set at x=2, as this is the reported sibling enrichment in UKBB [20].

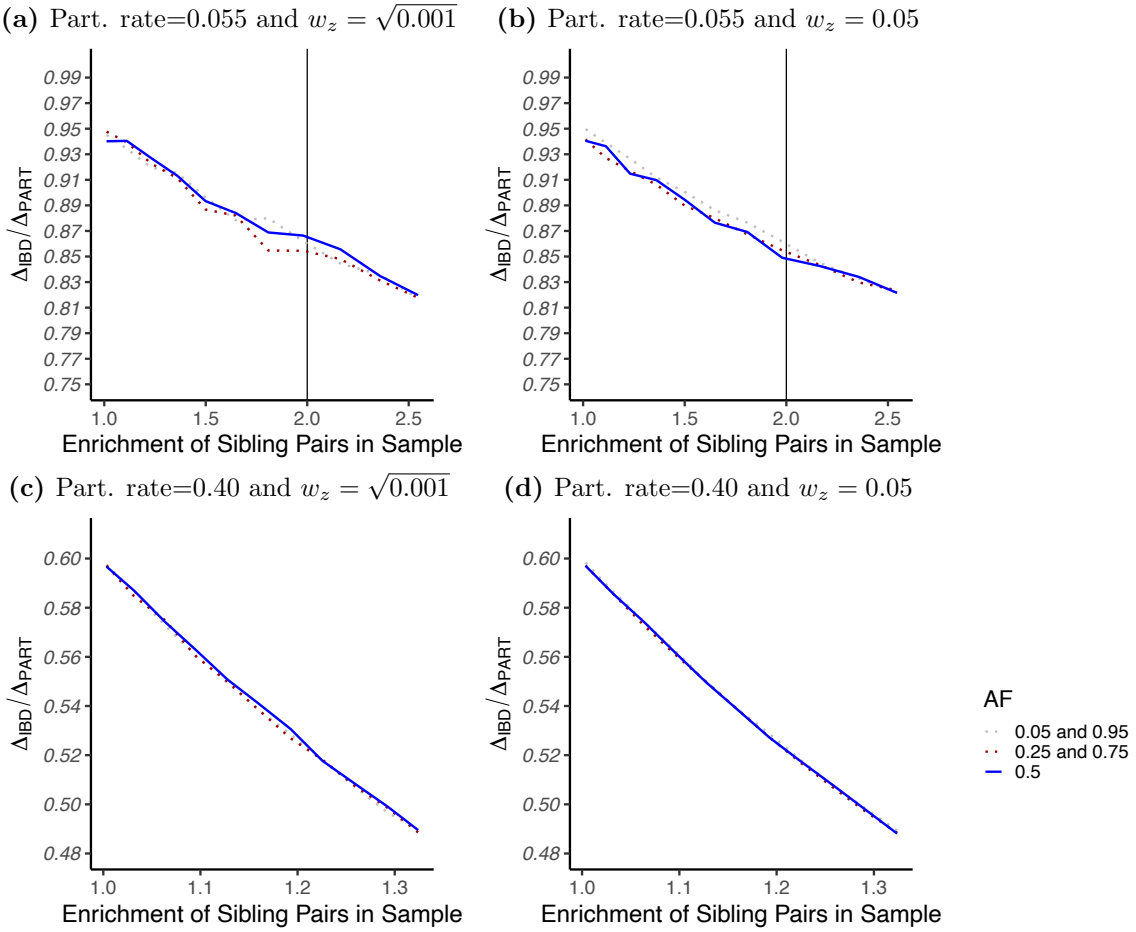


Figure 4.18: Average values: Difference between IBD shared and IBD non-shared alleles. Figures showing how the relative allele frequency difference $\Delta_{IBD}/\Delta_{PART}$ (y-axis) changes with the enrichment of sibling pairs in sample measured with the sibling recurrence risk ratio (x-axis) for different SNP effects and participation rates (a-d). Δ_{IBD} denotes the allele frequency difference between IBD shared and IBD non-shared alleles of participating siblings. Δ_{PART} denotes the allele frequency difference between participants and non-participants. Here, the average $\Delta_{IBD}/\Delta_{PART}$ has been computed for AF=0.25 and 0.75 and for AF=0.05 and 0.95. The allele frequency differences were attained by simulating trait values, $X_{ij} = w_z \cdot z_{ij} + w_B \cdot B_{ij}$, of 5 million sibling pairs and then allocating them into participants and non-participants assuming the logit model; $P(Z_{ij} = 1|X_{ij} = X) = \frac{1}{1+e^{-\beta_0 - 1.2 \cdot X}}$ with β_0 taking values corresponding to a participation rate of 0.055 (a and b) and 0.40 (c and d) respectively. z_{ij} denotes the standardised and mean centred genotype of a single biallelic SNP for individual i in sibling pair j. $B_{ij} \sim N(0, 1)$ denotes all other contributions to X_{ij}. z_{ij} and B_{ij} are assumed to be uncorrelated while the sibling correlation for z_{ij} is 0.5 in the population. w_z was set at $\sqrt{0.001}$ (a and c) and 0.05 (b and d) respectively. The sibling correlation of B_{ij} in the population ranged from 0.01 to 0.99 in the simulations, allowing for different sibling enrichment. The simulation process was repeated 2,400 times and then the mean was computed from those replications. Δ_{PART} and Δ_{IBD} were computed for the unstandardised genotypes underlying z_{ij}. The solid, dotted and dashed curves depict the relative frequency difference for each AF scenario (legend on the bottom right).

In the scenario that is loosely based on the UKBB, that is when the total participation rate is 5.5% and there are twice as many participating sibling pairs than expected under random sampling (SRRR=2), Δ_{IBD} ranges from being 75% to 97% of Δ_{PART} for the 5 different population allele frequency scenarios and the two different effect sizes. The relative difference $\Delta_{IBD}/\Delta_{PART}$ differs between allele frequency scenarios, but those differences become smaller as the SNP effects (w_z) decrease. The simulations demonstrate a mirror-image type relationship between the relative differences for scenarios when the minor allele is participation-promoting versus when the major allele is participation promoting. That is, for two causal SNPs with effect allele frequencies of p and $(1-p)$ respectively, their average $\Delta_{IBD}/\Delta_{PART}$ value is approximately equal to the ratio when $p = 0.5$. Figure 4.18 depicts those averages along with $\Delta_{IBD}/\Delta_{PART}$ when $p = 0.5$. The difference between the minor allele or the major allele having a positive effect can be partially understood as follows. When the minor allele has a positive effect, the allele frequency moves closer to 0.5 in the sample and the variance of alleles or genotypes would increase, whereas the opposite is true if the major allele is having a positive effect.

The relative difference $\Delta_{IBD}/\Delta_{PART}$ has a strong relationship with the total participation rate. When $p = 0.5$ and $\rho_B = 0.7$, $\Delta_{IBD}/\Delta_{PART}$ decreases from 0.85 to 0.05 when the participation rate increases from 5.5% to 94.5% (see figure 4.19a). It is worth pointing out that when Δ_{IBD} is examined relative to the difference between the sample allele frequency and the population allele frequency, $\Delta_{POP} = p_1 - p$, the relationship with participation rate is quite different. That is, for $p = 0.5$ and $\rho_B = 0.7$, $\Delta_{IBD}/\Delta_{POP}$ is relatively stable as the participation rate increases from 5.5% to 94.5% and is always within the range 0.86-0.95 (see figure 4.19b).

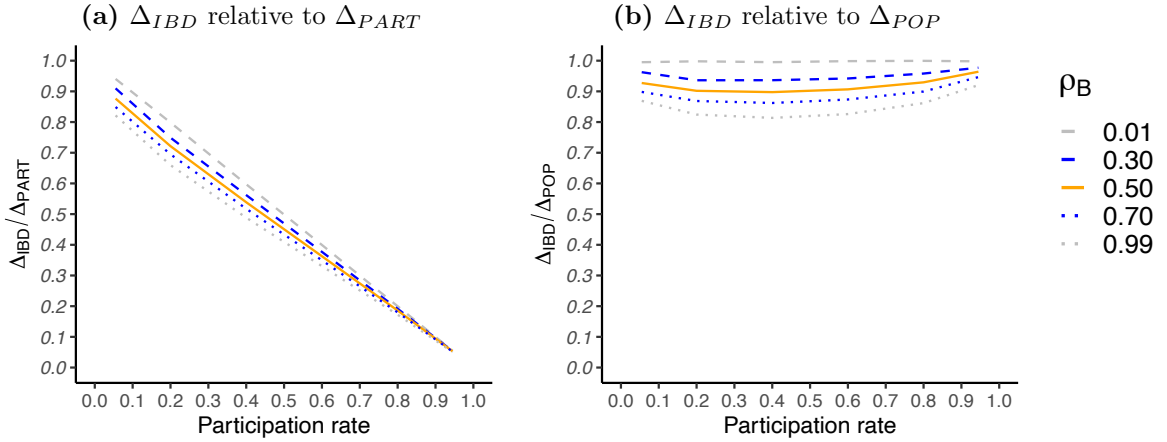


Figure 4.19: Relative allele frequencies as a function of participation rate. Figures showing how the relative allele frequency differences (a) $\Delta_{IBD}/\Delta_{PART}$ and (b) $\Delta_{IBD}/\Delta_{POP}$ (y-axis) change with the total participation rate (x-axis). Δ_{IBD} denotes the allele frequency difference between IBD shared and IBD non-shared alleles of participating siblings. Δ_{PART} denotes the allele frequency difference between participants and non-participants. Δ_{POP} denotes the allele frequency difference between participants and the population. The allele frequency differences were attained by simulating trait values, $X_{ij} = w_z \cdot z_{ij} + w_B \cdot B_{ij}$, of 5 million sibling pairs and then allocating them into participants and non-participants assuming the logit model; $P(Z_{ij} = 1|X_{ij} = X) = \frac{1}{1+e^{-\beta_0 - 1.2 \cdot X}}$ with β_0 taking values corresponding to a given participation rate. z_{ij} denotes the standardised and mean centred genotype of a single biallelic SNP for individual i in sibling pair j. $B_{ij} \sim N(0, 1)$ denotes all other contributions to X_{ij}, environmental and/or genetic. z_{ij} and B_{ij} are assumed to be uncorrelated while the sibling correlation for z is 0.5 in the population.This figure depicts results from simulations where w_z was set at 0.05 corresponding to w_B of $\sqrt{0.9975}$ and a population allele frequency of 0.5 for the single sequence variant. The sibling correlation of B_{ij} in the population, ρ_B, ranged from 0.01 and 0.99 in the simulations (see legend on the right), allowing for different sibling enrichment at a given participation rate. The simulation process was repeated 2,400 times and then the mean was computed from those replications. Δ_{IBD}, Δ_{PART} and Δ_{POP} were computed for the unstandardised genotypes underlying z_{ij}.

The quantitative relationship between Δ_{PART} and the allele frequency difference between participants with a participating sibling and participants with a non-participating sibling, Δ_{SIB}, was also investigated with the simulations (see figure 4.20 and figure C.1). Focusing on the scenario that is loosely based on the UKBB, participation rate of 0.055 and SRRR=2, when the single sequence variant has an allele frequency of 0.5

and contributes 0.25% to the variance of X, (corresponding to figure 4.20a), Δ_{SIB} is approximately 38% of Δ_{PART} and 45% of Δ_{IBD}. These simulation results are consistent with the observed *pPGS* difference between siblings and unrelateds in the UKBB for which the average *pPGS* was approximately 0.035 SD higher for UKBB participants with a participating sibling compared to the UKBB group of unrelateds.

In summary, the simulations demonstrate, that for alleles that have an effect on participation, relative allele frequencies in different groups of individuals and genetic segments will depend on many factors, most important of which are the overall participation rate and the participation rate of those with a close relative among the participants.

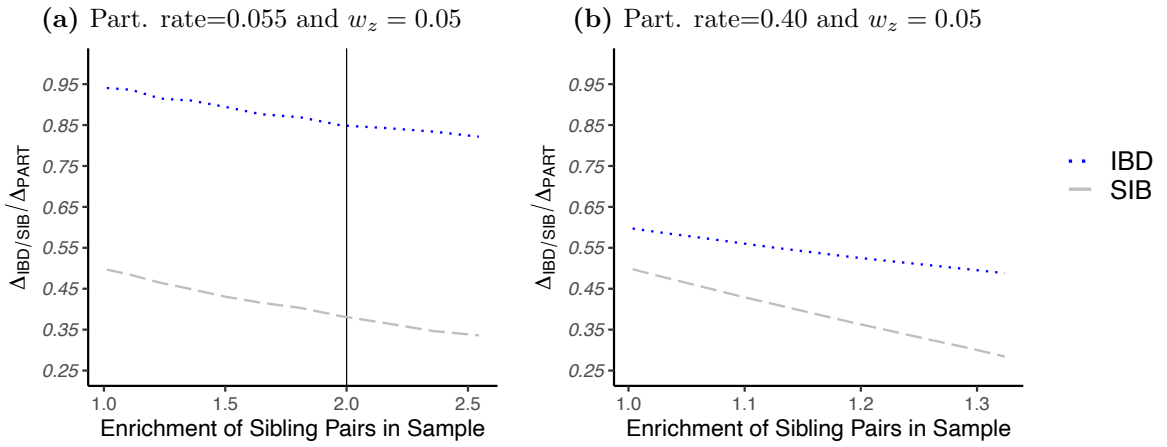


Figure 4.20: Relative allele frequencies in different groups of individuals and genetic segments. Figures showing how the relative allele frequency differences $\Delta_{IBD}/\Delta_{PART}$ and $\Delta_{SIB}/\Delta_{PART}$ (y-axis) change with the enrichment of sibling pairs in sample measured with the sibling recurrence risk ratio (x-axis) when participation rate is 0.055 (a) and 0.40 (b). Δ_{IBD} denotes the allele frequency difference between IBD shared and IBD non-shared alleles of participating siblings. Δ_{SIB} denotes the allele frequency difference between participants with a participating sibling and participants with a non-participating sibling. Δ_{PART} denotes the allele frequency difference between participants and non-participants. The allele frequency differences were attained by simulating trait values, $X_{ij} = w_z \cdot z_{ij} + w_B \cdot B_{ij}$, of 5 million sibling pairs and then allocating them into participants and non-participants assuming the logit model; $P(Z_{ij} = 1|X_{ij} = X) = \frac{1}{1+e^{-\beta_0 - 1.2 \cdot X}}$ with β_0 taking values corresponding to a participation rate of 0.055 (a) and 0.40 (b) respectively. z_{ij} denotes the standardised and mean centred genotype of a single biallelic SNP for individual i in sibling pair j. $B_{ij} \sim N(0, 1)$ denotes all other contributions to X_{ij}, environmental and/or genetic. z_{ij} and B_{ij} are assumed to be uncorrelated while the sibling correlation for z is 0.5 in the population.This figure depicts results from simulations where w_z was set at 0.05 corresponding to w_B of $\sqrt{0.9975}$ and a population allele frequency of 0.5 for the single sequence variant. The sibling correlation of B_{ij} in the population ranged from 0.01 and 0.99 in the simulations, allowing for different sibling enrichment at a participation rate of 0.055 and 0.40 respectively. The simulation process was repeated 2,400 times and then the mean was computed from those replications. Δ_{IBD}, Δ_{SIB} and Δ_{PART} were computed for the unstandardised genotypes underlying z_{ij}.

4.5 Conclusions

By exploiting the fact that participation bias leaves many footprints within genetic datasets, I was able to perform a GWAS on primary participation in the UKBB using only genetic data of the participants. Notably, each individual often harbours both 'case' and 'control' alleles, at different sites and the same sites. With nearly perfectly matched controls, the primary participation GWAS only captures direct effects and is unaffected by population stratification confounding.

One of the complications of studying the genetics of participation is the issue of replication. As study design, participation rate and the contribution of genetics can vary hugely from study to study, each study has to be evaluated on its own and the participation GWAS results cannot be expected to replicate across all studies. For this reason, I validated the participation GWAS through a special form of within-study replication. This was done by investigating the relationship between the *pPGS* and various phenotypes in a group of UKBB participants that do not overlap with the UKBB first-degree relatives that the participation GWAS is based on. Notably, the *pPGS* is constructed from a GWAS performed without phenotypes. Thus, its associations with phenotypes such as educational attainment and BMI, that have previously been reported to exhibit ascertainment bias[33, 80] is confirmation that the primary participation GWAS is indeed capturing the genetic effects of participation.

The data application to the UKBB data also shows that the method underlying the primary participation GWAS has an useful implication beyond investigating the genetic component of participation, the detection of data artefacts. That is, similar to the HWE test, the primary participation GWAS could also be a tool for detecting

problematic variants. Furthermore, the robustness of the T_{SIB20} statistics enables one to differentiate between significant results due to data artefacts and significant results due to true effect on participation.

The results of this study demonstrate that there can be a common component that underlies many different participation events. The *pPGS* captures primary participation, but it plays a role in both the passive (being invited) and active (deciding to participate when invited) phases of the secondary participation events. Thus, the relevance of a genetic component underlying participation should not be judged based on an individual selection event only. Its effect could accumulate through its impact on many participation events of a person's lifespan, or it can be magnified through nested participation, e.g. the participants in the dietary and physical activity studies have a higher average genetic propensity to participate than the other UKBB participants, who already have a higher average propensity than the population. In contrast to a common tendency to think of participation as a consequence of other characteristics and established traits, the propensity to participate in a whole range of events should rather be studied as a behavioural trait in its own right. Furthermore, it has been reported that the UKBB contains more close relatives than what would be expected by random sampling[20]. The results shown in this chapter demonstrate that genetic influences are likely to play a substantial role in that overrepresentation.

Chapter 5

Discussions

The most challenging part of investigating genetically induced participation bias is the issue of estimating the bias itself. In chapter 4, I described a new method that enables one to test sequence variants for an association with primary participation. The method involves comparing shared and non-shared alleles of first-degree relatives and one of its major strengths is that it only requires the genetic information of participants. Furthermore, it only captures direct effects of participation so the results are not confounded by factors such as population stratification and indirect effects of relatives. This is especially important when put into context with results from chapter 3 in which I demonstrated how principal component adjustment alone will most probably not capture all relevant population stratification. The overfitting of noise within the data and within-chromosome covariances are likely to mask subtle correlations between variants on different chromosomes.

Comparing shared and non-shared alleles of first-degree relatives yields three different test statistics for each biallelic sequence variant, T_{PO}, T_{SIB1} and T_{SIB20}. The first is based on comparing transmitted and non-transmitted parental alleles, the second is

based on comparing shared and non-shared alleles of sibling pairs sharing the sequence variant IBD=1 and the third is based on comparing alleles of sibling pairs sharing the sequence variant IBD=2 to alleles of sibling pairs sharing the sequence variant IBD=0. These three statistics are statistically independent and can therefore be compared for validation purposes or for learning more about the nature of the bias itself. In particular, T_{SIB1} and T_{SIB20} are based on the same sibling pairs and thus should have essentially the same (true) effect sizes, but they are utilizing separate information. Another advantage of having three statistics for each sequence variant, is that they are impacted by genotyping and data processing errors in a different manner. In particular, T_{PO} and T_{SIB1} involve splitting the genotypes into alleles, while T_{SIB20} does not. It turns out that this feature gives the method a dual purpose: 1) Investigate the genetic component of primary participation, 2) detect problematic sequence variants and provide a better understanding to the nature and magnitude of genotyping and phasing errors.

My application to the UK Biobank (UKBB) demonstrated that this new method can be easily applied to existing datasets without any additional data sampling as it does not require census data nor any specific disease or trait information about the participants. Within-sample replication revealed that this primary participation GWAS based on comparing shared and non-shared alleles of first-degree relatives is indeed capturing real and significant information about the genetics underlying primary participation. The application to the UKBB further demonstrated that T_{PO} and T_{SIB1} are sensitive to data processing errors while T_{SIB20} is very robust.

HWE testing is routinely used to detect genotyping errors. Comparing shared and non-shared alleles of first-degree relatives could further help with that. In the UKBB

application, abundance of variants demonstrated potential phasing problems but did not deviate from HWE. In a similar manner, there were also many sequence variants that deviated from HWE but would have not been flagged with T_{SIB1} or T_{PO}. It is worth noting that the test of HWE only has one degree of freedom and so has limitations in its utility to detect genotyping errors or to provide a full understanding of the nature of the genotyping errors underlying the deviation. Independent of phasing errors, genotyping errors that are not sensitive to HWE tests could also potentially affect T_{SIB1} or T_{PO}. Indeed, one of my plans for future research is to develop a data-quality evaluation method that utilises jointly the information captured by the HWE tests and the shared-versus-non-shared tests.

T_{SIB20} is much more simple to compute than T_{PO} and T_{SIB1} as it does not require phased data nor splitting up the genotypes and inferring the shared allele. Therefore, if time or computational resources are limited and the main aim is to investigate the genetic component of participation, I would recommend prioritising the T_{SIB20} analysis. However, while T_{PO} and T_{SIB1} are not as clean as T_{SIB20} as methods to detect participation bias, they can serve an important role in validating putative discoveries that are made based on T_{SIB20}. In a way, this is not that dissimilar to replicating associations in a different population.

Participation bias is a form of selection, and just like natural selection, it can influence the sample's covariance structure. That is, it can affect the sample correlation between maternally inherited and paternally inherited alleles as well as correlations between different sequence variants. The HWE analysis in chapter 2 was performed prior to the participation bias analysis. The HWE application to the UKBB demonstrated a difference between participants with at least one close relative within UKBB and par-

ticipants who had no close relatives within the dataset. In particular, when the HWE statistics of SNPs were examined jointly, it became evident that the mean statistic was lower for individuals with a close relative in sample compared to individuals with no close relatives in sample. For a given SNP, a negative HWE statistic indicates a negative correlation between maternally and paternally inherited alleles. When I initially performed the HWE analysis, my main conclusion was that the negative HWE statistics reflected genotyping errors. However, participation bias could play a part as well. In a randomly mating population, correlation between maternally and paternally inherited alleles will be zero. If a SNP has an additive effect on participation probability and is in HWE in the population, the maternally and paternally inherited alleles can become negatively correlated within the sample. In theory, in the absence of genotyping errors, participation bias could be investigated by testing for deviations from HWE within the sample. However, that would require prior knowledge about the degree of assortative mating within the population. If maternally and paternally inherited alleles are positively correlated within the population, then for a SNP with an additive effect on participation the sample correlation could be lower than the population correlation but it might not be negative.

There are many possibilities for future work involving the method underlying the primary participation GWAS. Firstly, the computation of IBD specific allele frequencies in the UKBB was based on scripts that I wrote in `python` and `R`. Transforming those scripts into a user-friendly package could make the method more accessible to other researchers. Another possibility for future work is to extend the method to more distant relative pairs. Such extension could increase power and provide additional opportunities for a within-sample replication. That could also increase the utility of the method

for others, especially those who work with datasets that do not contain many first-degree relatives. Thirdly, applying the method to other datasets and comparing those results to the UKBB application would be of great interest on multiple levels. Indeed, in this thesis I investigated the genetic component to 'participation in UK Biobank', a trait that could be very different from 'participation in 23andMe' or 'participation in the MoBA study'. Performing primary participation GWAS in other datasets would therefore be very interesting beyond the purpose of replication and/or meta-analysis.

Participation in genetic studies is a very compelling trait in itself. Not only does it have an active behavioural component, some individuals are more likely to participate than other, but also a passive component, some individuals are more likely to be invited than other. Although great care is often taken to attain representative samples, it is nearly impossible to establish sample protocols that do not reflect the humans that designed them to some degree. E.g. the decision about how to contact potential participants might reflect habits of those that design the study. Mills et al. (2019) note that if genetic discoveries are to play a part in personalised medicine, including a diverse set of participants is an important equality issue[77]. Individuals that belong to groups that are underrepresented in genetic datasets might not gain as much from the resulting scientific advances. An example of this are polygenic scores (PGS). PGS are predicted to have a future impact in personalised medicine but it has been shown that PGS have a significantly worse prediction power in populations that are underrepresented in GWAS[77]. Clinical sequencing is another example[5]. Richard et al. (2015) propose that population allele frequency should be taken into account when the potential pathogenicity of a sequence variant is assessed[99]. However, if a patient belongs to a group that is not represented in genetic datasets, defining what

constitutes as rare can be a difficult task.

In summary, increased information about genetically induced participation bias could be very valuable for those who design genetic studies and for those who draw inferences from the data. Furthermore, while the genetic component to participation can only be studied in genetic studies, the former can play a part in all types of datasets, genetic or otherwise. To end, I note that having close relatives in a large population-based genetic dataset should be treated as an asset instead of a nuisance. Some information that allows for doing the nearly impossible exists only because of the relatives.

Bibliography

[1] Adams, M. J., Hill, W. D., Howard, D. M., Dashti, H. S., Davis, K. A., Campbell, A., Clarke, T.-K., Deary, I. J., Hayward, C., Porteous, D. et al. (2020). Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. *International Journal of Epidemiology* 49: 410–421, doi:10.1093/ije/dyz134.

[2] Agrawal, A., Chiu, A. M., Le, M., Halperin, E. and Sankararaman, S. (2020). Scalable probabilistic PCA for large-scale genetic variation data. *PLoS Genetics* 16: e1008773, doi:10.1371/journal.pgen.1008773.

[3] Aidoo, M., Terlouw, D. J., Kolczak, M. S., McElroy, P. D., Ter Kuile, F. O., Kariuki, S., Nahlen, B. L., Lal, A. A. and Udhayakumar, V. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet* 359: 1311–1312, doi:10.1016/S0140-6736(02)08273-9.

[4] Albrechtsen, A., Nielsen, F. C. and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* 27: 2534–2547, doi:10.1093/molbev/msq148.

[5] Arnadottir, G. A., Norddahl, G. L., Gudmundsdottir, S., Agustsdottir, A. B., Sigurdsson, S., Jensson, B. O., Bjarnadottir, K., Theodors, F., Benonisdottir, S.,

Ivarsdottir, E. V. et al. (2018). A homozygous loss-of-function mutation leading to CYBC1 deficiency causes chronic granulomatous disease. *Nature Communications* 9: 4447, doi:10.1038/s41467-018-06964-x.

[6] Balding, D. J. (2005). *Weight-of-Evidence for Forensic DNA Profiles*. John Wiley & Sons.

[7] Band, G. and Marchini, J. (2022). QCTOOL 2.0.1. `https://www.well.ox.ac.uk/~gav/qctool_v2/`, (accessed Oct. 2022).

[8] Barnes, P. (2016). Reality Check: Should we give up on election polling? `https://www.bbc.com/news/election-us-2016-37949527`, (accessed Sept. 2022).

[9] Beasley, T. M., Erickson, S. and Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics* 39: 580–595, doi:10.1007/s10519-009-9281-0.

[10] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* 2: 47–53, doi:10.2307/3002000.

[11] Bijma, P. (2014). The quantitative genetics of indirect genetic effects: a selective review of modelling issues. *Heredity* 112: 61–69, doi:10.1038/hdy.2013.15.

[12] Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C. and Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* 13: 3261–3273, doi:10.1111/j.1365-294X.2004.02346.x.

[13] Bradbury, K. E., Young, H. J., Guo, W. and Key, T. J. (2018). Dietary assessment in UK Biobank: an evaluation of the performance of the touchscreen dietary questionnaire. *Journal of Nutritional Science* 7: e6, doi:10.1017/jns.2017.66.

[14] Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* 600: 695–700, doi:10.1038/s41586-021-04198-4.

[15] Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I. et al. (2020). Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications* 11: 3519, doi:10.1038/s41467-020-17117-4.

[16] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B. et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* 47: 1236–1241, doi:10.1038/ng.3406.

[17] 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: 291–295, doi:10.1038/ng.3211.

[18] Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology* 8: e1002822, doi:10.1371/journal.pcbi.1002822.

[19] Busing, F., Meijer, E. and Leeden, R. V. D. (1999). Delete-m jackknife for unequal m. *Statistics and Computing* 9: 3–8, doi:10.1023/A:1008800423698.

[20] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562: 203–209, doi: 10.1038/s41586-018-0579-z.

[21] Carmichael, M. (2010). The little flaw in the longevity-gene study that could be a big problem. `https://www.newsweek.com/little-flaw-longevity-gene-study-could-be-big-problem-74703`, (accessed Sept. 2022).

[22] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4: s13742–015, doi:10.1186/s13742-015-0047-8.

[23] Chen, C. and Bloemendal, A. (2019). Christina Chen, PCA and stratification in GWAS; Alex Bloemendal, primer on random matrix theory. `https://www.youtube.com/watch?v=B7ub92OLw1g`, (accessed Sept. 2022).

[24] Chen, M.-H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D. et al. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* 182: 1198–1213, doi:10.1016/j.cell.2020.06.045.

[25] Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F. and Li, L. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology* 40: 1652–1666, doi:10.1093/ije/dyr120.

[26] Coates, S., Tanna, P. and Scott-Allen, E. (2019). Overview of the UK population: August 2019. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/august2019 (accessed Nov. 2019).

[27] Crow, J. F. and Kimura, M. (1970). *An introduction to population genetics theory*. Harper & Row.

[28] Dimitromanolakis, A., Paterson, A. D. and Sun, L. (2019). Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data via TRUFFLE. *The American Journal of Human Genetics* 105: 78–88, doi:10.1016/j.ajhg.2019.05.007.

[29] Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., Van Hees, V. T., Trenell, M. I., Owen, C. G. et al. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS one* 12: e0169649, doi:10.1371/journal.pone.0169649.

[30] Ewens, W. J. and Spielman, R. S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics* 57: 455–464.

[31] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K. et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* 47: 1228–1235, doi:10.1038/ng.3404.

[32] Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N. et al. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics* 36: 388–393, doi:10.1038/ng1333.

[33] Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R. and Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology* 186: 1026–1034, doi: 10.1093/aje/kwx246.

[34] Gagliano Taliun, S. A., VandeHaar, P., Boughton, A. P., Welch, R. P., Taliun, D., Schmidt, E. M., Zhou, W., Nielsen, J. B., Willer, C. J., Lee, S. et al. (2020). Exploring and visualizing large-scale genetic associations by using PheWeb. *Nature Genetics* 52: 550–552, doi:10.1038/s41588-020-0622-5.

[35] Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E. et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* 47: 435–444, doi:10.1038/ng.3247.

[36] Halldorsson, B. V., Eggertsson, H. P., Moore, K. H., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., Palsson, G., Hardarson, M. T., Oddsson, A., Jensson, B. O. et al. (2022). The sequences of 150,119 genomes in the UK biobank. *Nature* 607: 732–740, doi:10.1038/s41586-022-04965-x.

[37] Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson, M. T., Eggertsson, H. P., Gunnarsson, B., Oddsson, A., Halldorsson, G. H.,

Zink, F. et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* 363: eaau1043, doi:10.1126/science.aau10.

[38] Han, Y., Jia, Q., Jahani, P. S., Hurrell, B. P., Pan, C., Huang, P., Gukasyan, J., Woodward, N. C., Eskin, E., Gilliland, F. D. et al. (2020). Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nature Communications* 11: 1776, doi:10.1038/s41467-020-15649-3.

[39] Hartl, D. L. and Jones, E. W. (1998). *Genetics: Principles and Analysis*. Jones and Bartlett Publishers, 4th ed.

[40] Haworth, S., Mitchell, R., Corbin, L., Wade, K. H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani, G., Paternoster, L., Smith, G. D. et al. (2019). Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature Communications* 10: 333, doi:10.1038/s41467-018-08219-1.

[41] Hébert, H. L., Shepherd, B., Milburn, K., Veluchamy, A., Meng, W., Carr, F., Donnelly, L. A., Tavendale, R., Leese, G., Colhoun, H. M. et al. (2018). Cohort profile: genetics of diabetes audit and research in Tayside Scotland (GoDARTS). *International Journal of Epidemiology* 47: 380–381, doi:10.1093/ije/dyx140.

[42] Hill, W. G., Goddard, M. E. and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics* 4: e1000008, doi:10.1371/journal.pgen.1000008.

[43] Hinrichs, A. L., Larkin, E. K. and Suarez, B. K. (2009). Population stratification and patterns of linkage disequilibrium. *Genetic epidemiology* 33: S88–S92, doi:10.1002/gepi.20478.

[44] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews genetics* 6: 95–108, doi: 10.1038/nrg1521.

[45] Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics* 52: 362–374.

[46] Howrigan, D., Abbott, L., Churchhouse, C. and Palmer, D. (2017). Details and considerations of the UK Biobank GWAS. `http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas`, (accessed Sept. 2022).

[47] Idzerda, R. L., March, C. J., Mosley, B., Lyman, S. D., Vanden Bos, T., Gimpel, S. D., Din, W. S., Grabstein, K. H., Widmer, M. B. and Park, L. S. (1990). Human interleukin 4 receptor confers biological responsiveness and defines a novel receptor superfamily. *The Journal of Experimental Medicine* 171: 861–873, doi: 10.1084/jem.171.3.861.

[48] Inshaw, J. R., Cutler, A. J., Burren, O. S., Stefana, M. I. and Todd, J. A. (2018). Approaches and advances in the genetic causes of autoimmune disease and their implications. *Nature Immunology* 19: 674–684, doi:10.1038/s41590-018-0129-8.

[49] Ivarsdottir, E. V., Holm, H., Benonisdottir, S., Olafsdottir, T., Sveinbjornsson, G., Thorleifsson, G., Eggertsson, H. P., Halldorsson, G. H., Hjorleifsson, K. E., Melsted, P. et al. (2021). The genetic architecture of age-related hearing impairment revealed by genome-wide association analysis. *Communications Biology* 4: 706, doi:10.1038/s42003-021-02224-9.

[50] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson, 6th ed.

[51] Kanai, M., Howrigan, D., Daly, M. and Finucane, H. (2019). Genotyped SNPs in UK Biobank failing Hardy-Weinberg equilibrium test. `http://www.nealelab.is/blog/2019/9/17/genotyped-snps-in-uk-biobank-failing-hardy-weinberg-equilibrium-test` (accessed Oct. 2022).

[52] Karavani, E., Zuk, O., Zeevi, D., Barzilai, N., Stefanis, N. C., Hatzimanolis, A., Smyrnis, N., Avramopoulos, D., Kruglyak, L., Atzmon, G. et al. (2019). Screening human embryos for polygenic traits has limited utility. *Cell* 179: 1424–1435, doi:10.1016/j.cell.2019.10.033.

[53] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P. et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581: 434–443, doi:10.1038/s41586-020-2308-7.

[54] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T. et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* 50: 1219, doi:10.1038/s41588-018-0183-z.

[55] KING (2022). KING Tutorial: Relationship Inference. `https://www.kingrelatedness.com/manual.shtml`, (accessed Sept. 2022).

[56] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T. et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308: 385–389, doi:10.1126/science.110955.

[57] Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A. et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488: 471–475.

[58] Kong, A., Frigge, M. L., Thorleifsson, G., Stefansson, H., Young, A. I., Zink, F., Jonsdottir, G. A., Okbay, A., Sulem, P., Masson, G. et al. (2017). Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences* 114: E727–E732, doi: 10.1073/pnas.1612113114.

[59] Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H. and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics* doi:10.1038/ng.216.

[60] Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., Benonisdottir, S., Oddsson, A., Halldorsson, B. V., Masson, G., Gudbjartsson, D. F., Helgason, A., Bjornsdottir, G., Thorsteinsdottir, U. and Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science* 359: 424–428, doi:10.1126/science.aan6877.

[61] Krokstad, S., Langhammer, A., Hveem, K., Holmen, T., Midthjell, K., Stene, T., Bratberg, G., Heggland, J. and Holmen, J. (2013). Cohort profile: the HUNT study, Norway. *International Journal of Epidemiology* 42: 968–977, doi:10.1093/ije/dys095.

[62] Kruse, S., Forster, J., Kuehr, J. and Deichmann, K. A. (1999). Characterization of the membrane-bound and a soluble form of human IL-4 receptor α produced by alternative splicing. *International Immunology* 11: 1965–1970, doi:10.1093/intimm/11.12.1965.

[63] Lam, M., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C. et al. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature Genetics* 51: 1670–1678, doi:10.1038/s41588-019-0512-x.

[64] Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics* 10: 387–406, doi:10.1146/annurev.genom.9.081307.164242.

[65] Liu, Q., Zhao, S., Shi, C.-M., Song, S., Zhu, S., Su, Y., Zhao, W., Li, M., Bao, Y., Xue, Y. et al. (2020). Population genetics of SARS-CoV-2: disentangling effects of sampling bias and infection clusters. *Genomics, Proteomics & Bioinformatics* 18: 640–647, doi:10.1016/j.gpb.2020.06.001.

[66] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature Genetics* 50: 906–908, doi:10.1038/s41588-018-0144-6.

[67] Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B. et al. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* 47: 284–290, doi:10.1038/ng.3190.

[68] Lu, M., Lewis, C. M. and Traylor, M. (2017). Pharmacogenetic testing through the direct-to-consumer genetic testing company 23andMe. *BMC Medical Genomics* 10: 47, doi:10.1186/s12920-017-0283-0.

[69] Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A. K., Handal, M., Haugen, M., Høiseth, G., Knudsen, G. P. et al. (2016). Cohort profile update: the Norwegian mother and child cohort study (MoBa). *International Journal of Epidemiology* 45: 382–388, doi:10.1093/ije/dyw029.

[70] Maier, R., Akbari, A., Wei, X., Patterson, N., Nielsen, R. and Reich, D. (2020). No statistical evidence for an effect of CCR5-Δ32 on lifespan in the UK Biobank cohort. *Nature Medicine* 26: 178–180, doi:10.1038/s41591-019-0710-1.

[71] Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M. and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26: 2867–2873, doi:10.1093/bioinformatics/btq559.

[72] Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M. and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* 51: 584–591, doi:10.1038/s41588-019-0379-x.

[73] Martin, J., Tilling, K., Hubbard, L., Stergiakouli, E., Thapar, A., Davey Smith, G., O'Donovan, M. C. and Zammit, S. (2016). Association of genetic risk

for schizophrenia with nonparticipation over time in a population-based cohort study. *American Journal of Epidemiology* 183: 1149–1158, doi:10.1093/aje/kww009.

[74] Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I) Law of Large Populations, Big Data Paradox and the 2016 US presidential election. *The Annals of Applied Statistics* 12: 685–726, doi:10.1214/18-AOAS1161SF.

[75] Mignogna, G., Carey, C. E., Wedow, R., Baya, N., Cordioli, M., Pirastu, N., Bellocco, R., Nivard, M. G., Neale, B. M., Walters, R. K. et al. (2022). Patterns of item nonresponse behavior to survey questionnaires are systematic and have a genetic basis. *bioRxiv,* doi:10.1101/2022.02.11.480140.

[76] Mills, M. C., Barban, N. and Tropf, F. C. (2020). *An Introduction to Statistical Genetic Data Analysis*. MIT Press.

[77] Mills, M. C. and Rahal, C. (2019). A scientometric review of genome-wide association studies. *Communications Biology* 2: 9, doi:10.1038/s42003-018-0261-x.

[78] Mills, M. C., Tropf, F. C., Brazel, D. M., Zuydam, N. van, Vaez, A., Pers, T. H., Snieder, H., Perry, J. R., Ong, K. K., Hoed, M. den et al. (2021). Identification of 371 genetic variants for age at first sex and birth linked to externalising behaviour. *Nature Human Behaviour* 5: 1717–1730, doi:10.1038/s41562-021-01135-3.

[79] Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K. and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* 9: e48376, doi:10.7554/eLife.48376.

[80] Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. and Davey Smith, G. (2018). Collider scope: when selection bias can substantially influence observed

associations. *International Journal of Epidemiology* 47: 226–235, doi:10.1093/ije/dyx206.

[81] Nilsen, R. M., Vollset, S. E., Gjessing, H. K., Skjaerven, R., Melve, K. K., Schreuder, P., Alsaker, E. R., Haug, K., Daltveit, A. K. and Magnus, P. (2009). Self-selection and bias in a large prospective pregnancy cohort in Norway. *Paediatric and Perinatal Epidemiology* 23: 597–608, doi:10.1111/j.1365-3016.2009.01062.x.

[82] Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S. M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T. et al. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics* 54: 437–449, doi:10.1038/s41588-022-01016-z.

[83] Palamara, P. F., Francioli, L. C., Wilton, P. R., Genovese, G., Gusev, A., Finucane, H. K., Sankararaman, S., Sunyaev, S. R., Bakker, P. I. de, Wakeley, J. et al. (2015). Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *The American Journal of Human Genetics* 97: 775–789, doi:10.1016/j.ajhg.2015.10.006.

[84] Palmer, D. S., Zhou, W., Abbott, L., Baya, N., Churchhouse, C., Seed, C., Poterba, T., King, D., Kanai, M., Bloemendal, A. et al. (2021). Analysis of genetic dominance in the UK Biobank. *bioRxiv,* doi:10.1101/2021.08.15.456387.

[85] Pan-UKBB team (2020). `https://pan.ukbb.broadinstitute.org`, (accessed Sept. 2022).

[86] Patterson, N., Price, A. L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* 2: e190, doi:10.1371/journal.pgen.0020190.

[87] Pearl, R. (1929). A note on the association of diseases. *Science* 70: 191–192, doi:10.1126/science.70.1808.191.b.

[88] Pearl, R. et al. (1929). Cancer and tuberculosis. *American Journal of Hygiene* 9: 97–159.

[89] PheWeb (2022). UKBiobank Neale v1. `https://pheweb.org/UKB-Neale/`, (accessed Sept. 2022).

[90] PheWeb (2022). UKBiobank TOPMed-imputed. `https://pheweb.org/UKB-TOPMed/`, (accessed Sept. 2022).

[91] Pirastu, N., Cordioli, M., Nandakumar, P., Mignogna, G., Abdellaoui, A., Hollis, B., Kanai, M., Rajagopal, V. M., Parolo, P. D. B., Baya, N. et al. (2021). Genetic analyses identify widespread sex-differential participation bias. *Nature Genetics* 53: 663–671, doi:10.1038/s41588-021-00846-7.

[92] Plomin, R. (2019). *Blueprint: How DNA makes us who we are*. MIT Press.

[93] 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: 904–909, doi:10.1038/ng1847.

[94] Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D. et al. (2008). Long-range

LD can confound genome scans in admixed populations. *The American Journal of Human Genetics* 83: 132–135, doi:10.1016/j.ajhg.2008.06.005.

[95] Purcell, S. and Chang, C. (2022). PLINK 1.90b2n. `www.cog-genomics.org/plink/1.9/`, (accessed Oct. 2022).

[96] Purcell, S. and Chang, C. (2022). PLINK 2.0. `www.cog-genomics.org/plink/2.0/`, (accessed Oct. 2022).

[97] Python Software Foundation (2022). Python. `http://www.python.org`, (accessed Sept. 2022).

[98] R Core Team (2017). R: A Language and Environment for Statistical Computing. `https://www.R-project.org/`, (accessed Sept. 2022).

[99] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E. et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* 17: 405–423, doi:10.1038/gim.2015.30.

[100] Said, S., Pazoki, R., Karhunen, V., Võsa, U., Ligthart, S., Bodinier, B., Koskeridis, F., Welsh, P., Alizadeh, B. Z., Chasman, D. I. et al. (2022). Genetic analysis of over half a million people characterises C-reactive protein loci. *Nature Communications* 13: 2198, doi:10.1038/s41467-022-29650-5.

[101] Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M. et al. (2021). A cross-

population atlas of genetic associations for 220 human phenotypes. *Nature Genetics* 53: 1415–1424, doi:10.1038/s41588-021-00931-x.

[102] Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., Pervjakova, N., Pers, T. H., Johnson, A. D., Eicher, J. D. et al. (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* 66: 2888–2902, doi:10.2337/db16-1253.

[103] Sebastiani, P., Solovieff, N., Puca, A., Hartley, S. W., Melista, E., Andersen, S., Dworkis, D. A., Wilk, J. B., Myers, R. H., Steinberg, M. H., Montano, M., Baldwin, C. T. and Perls, T. T. (2011). Retraction. *Science* 333: 404, doi:10.1126/science.333.6041.404-a.

[104] Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J. and Rose, R. J. (2003). Assortative mating by body height and BMI: Finnish twins and their spouses. *American Journal of Human Biology* 15: 620–627, doi:10.1002/ajhb.10183.

[105] Snoep, J. D., Morabia, A., Hernández-Díaz, S., Hernán, M. A. and Vandenbroucke, J. P. (2014). Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it. *International Journal of Epidemiology* 43: 515–521, doi:10.1093/ije/dyu026.

[106] Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W., Hirschhorn, J., Daly, M. J., Patterson, N. et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *ELife* 8: e39702, doi:10.7554/eLife.39702.

[107] Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R. et al. (2010).

Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* 42: 937–948, doi:10.1038/ng.686.

[108] Sulem, P., Gudbjartsson, D. F., Geller, F., Prokopenko, I., Feenstra, B., Aben, K. K., Franke, B., Heijer, M. den, Kovacs, P., Stumvoll, M. et al. (2011). Sequence variants at CYP1A1–CYP1A2 and AHR associate with coffee consumption. *Human Molecular Genetics* 20: 2071–2077, doi:10.1093/hmg/ddr086.

[109] Syse, A., Thomas, M. and Gleditsch, R. (2020). Norway's 2020 population projections. National level results, methods and assumptions. `https://www.ssb.no/en/befolkning/artikler-og-publikasjoner/norways-2020-population-projections`, (accessed Sept. 2022).

[110] Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M. et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590: 290–299, doi:10.1038/s41586-021-03205-y.

[111] Taylor, A. E., Jones, H. J., Sallis, H., Euesden, J., Stergiakouli, E., Davies, N. M., Zammit, S., Lawlor, D. A., Munafò, M. R., Davey Smith, G. et al. (2018). Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 47: 1207–1216, doi:10.1093/ije/dyy060.

[112] The Genome Aggregation Database (gnomAD) (2022). `https://gnomad.broadinstitute.org/variant/16-27315880-G-A?dataset=gnomad_r3`, (accessed Sept. 2022).

[113] Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* 194: 301–326, doi:10.1534/genetics.112.148825.

[114] Thompson, E. A. (2019). Correlations between relatives: From Mendelian theory to complete genome sequence. *Genetic Epidemiology* 43: 577–591, doi:10.1002/gepi.22206.

[115] Torkamani, A., Wineinger, N. E. and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* 19: 581–590, doi:10.1038/s41576-018-0018-x.

[116] Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software* doi:10.21105/joss.00731.

[117] Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., De Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G. et al. (2011). Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics* 68: 1.19.1–1.19.18, doi:10.1002/0471142905.hg0119s68.

[118] Tyrrell, J., Zheng, J., Beaumont, R., Hinton, K., Richardson, T. G., Wood, A. R., Davey Smith, G., Frayling, T. M. and Tilling, K. (2021). Genetic predictors of participation in optional components of UK Biobank. *Nature Communications* 12: 886, doi:10.1038/s41467-021-21073-y.

[119] Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. and Posthuma, D. (2021).

Genome-wide association studies. *Nature Reviews Methods Primers* 1: 59, doi: 10.1038/s43586-021-00056-9.

[120] UK Biobank (2022). Celebrating 15 years of UK Biobank. `https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/our-impact/celebrating-15-years-of-uk-biobank`, (accessed Oct. 2022).

[121] Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. D., Liu, D., Pandey, A. K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N. et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 586: 749–756, doi:10.1038/s41586-020-2853-0.

[122] Verlouw, J. A., Clemens, E., Vries, J. H. de, Zolk, O., Verkerk, A. J., Zehnhoff-Dinnesen, A. am, Medina-Gomez, C., Lanvers-Kaminsky, C., Rivadeneira, F., Langer, T. et al. (2021). A comparison of genotyping arrays. *European Journal of Human Genetics* 29: 1611–1624, doi:10.1038/s41431-021-00917-7.

[123] Vilhjálmsson, B. J. and Nordborg, M. (2013). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics* 14: 1–2, doi:10.1038/nrg3382.

[124] Visscher, P. M., Andrew, T. and Nyholt, D. R. (2008). Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *European Journal of Human Genetics* 16: 387–390, doi: 10.1038/sj.ejhg.5201990.

[125] Weedon, M. N., Jackson, L., Harrison, J. W., Ruth, K. S., Tyrrell, J., Hattersley, A. T. and Wright, C. F. (2019). Assessing the analytical validity of SNP-chips

for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing. *bioRxiv,* doi:10.1101/696799.

[126] Wei, X. and Nielsen, R. (2019). Retraction note: CCR5-Δ32 is deleterious in the homozygous state in humans. *Nature Medicine* 25, doi:10.1038/s41591-019-0637-6.

[127] Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology* 23: 159–164, doi:10.1097/EDE.0b013e31823b6296.

[128] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[129] Wigginton, J. E., Cutler, D. J. and Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *The American Journal of Human Genetics* 76: 887–893, doi:10.1086/429864.

[130] Wittke-Thompson, J. K., Pluzhnikov, A. and Cox, N. J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. *The American Journal of Human Genetics* 76: 967–986, doi:10.1086/430507.

[131] Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z. et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* 46: 1173–1186, doi:10.1038/ng.3097.

[132] Wright, C. F., West, B., Tuke, M., Jones, S. E., Patel, K., Laver, T. W., Beaumont, R. N., Tyrrell, J., Wood, A. R., Frayling, T. M. et al. (2019). Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants

in a population setting. *The American Journal of Human Genetics* 104: 275–286, doi:10.1016/j.ajhg.2018.12.015.

[133] Young, A. I., Benonisdottir, S., Przeworski, M. and Kong, A. (2019). Deconstructing the sources of genotype-phenotype associations in humans. *Science* 365: 1396–1400, doi:10.1126/science.aax3710.

[134] Young, A. I., Frigge, M. L., Gudbjartsson, D. F., Thorleifsson, G., Bjornsdottir, G., Sulem, P., Masson, G., Thorsteinsdottir, U., Stefansson, K. and Kong, A. (2018). Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics* 50: 1304–1310, doi:10.1038/s41588-018-0178-9.

[135] Young, A. I., Nehzati, S. M., Benonisdottir, S., Okbay, A., Jayashankar, H., Lee, C., Cesarini, D., Benjamin, D. J., Turley, P. and Kong, A. (2022). Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature Genetics* 54: 897–905, doi:10.1038/s41588-022-01085-0.

[136] Ziegler, A., König, I. R. and Pahlke, F. (2010). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. John Wiley & Sons.

Appendix A

Supplementary Methods

This appendix provides supplementary text about the UK Biobank (UKBB) data application presented in chapter 4.

A.1 Phenotypes

By using various data-fields in the UKBB data release, I constructed a range of phenotypic variables for 272,409 individuals that were not closely related to each other (>3rd degree for all pairs) and had been concluded to be white British (WB) by Bycroft et al. (2018)[20]. The phenotypes of interest were adjusted for year of birth (data-field 34), age at recruitment (data-field 21022), genotyping array (data-field 22000), and sex (data-field 31) when applicable. I also accounted for principal components (PCs) inferred with the program ProPCA[2]. Details about those PCs are described in chapter 3. Quantitative phenotypes were rank-based inverse normalised[9] except for educational attainment as it was not approximately normally distributed. Details about phenotype specific protocols and adjustments are provided below.

A.1.1 Educational Attainment (EA)

I constructed an educational attainment variable from the data-fields 'Qualifications' (data-field 6138) and 'Age completed full-time education' (data-field 845). Both of these data-fields contain answers to questions that participants were asked on a touch-screen in the initial assessment visit. The data-field 'Qualifications' denotes answers to the multiple-choice question: 'Which of the following qualifications do you have? (You can select more than one).' I mapped the multiple-choice options to years of schooling using the ISCED classification, except for the option 'NVQ or HND or HNC or equivalent'. Following Okbay et al. (2022), years of schooling of NVQ/HND/HNC holders was defined to be their answer in the data-field 'Age completed full-time education' minus five [82]. Individuals that answered 'Prefer not to answer' were mapped to 'Not applicable' and were excluded from subsequent analysis.

1) College or University degree: 20 years

2) A levels/AS levels or equivalent: 13 years

3) O levels/GCSEs or equivalent: 10 years

4) CSEs or equivalent: 10 years

5) NVQ or HND or HNC or equivalent: (Age completed full time education – 5) years

6) Other professional qualifications eg: nursing, teaching: 15 years

7) None of the above: 7 years

8) Prefer not to answer: Not applicable

Respondents who selected multiple options were assigned their highest category (maximum years of schooling). Individuals who had maximum year of schooling below 7 years were excluded from subsequent analysis. In the subset of 260,950 WB 'unrelated' individuals with applicable entries for year of schooling, I adjusted for year of birth

(YOB) up to the order of three, age at recruitment and 40 PCs separately for each sex. This was done with a linear regression in R where years of schooling equivalent was treated as a quantitative dependent variable. The mean centred and standardised residuals from these two regressions (men and women) were used as an approximation for educational attainment in subsequent analysis.

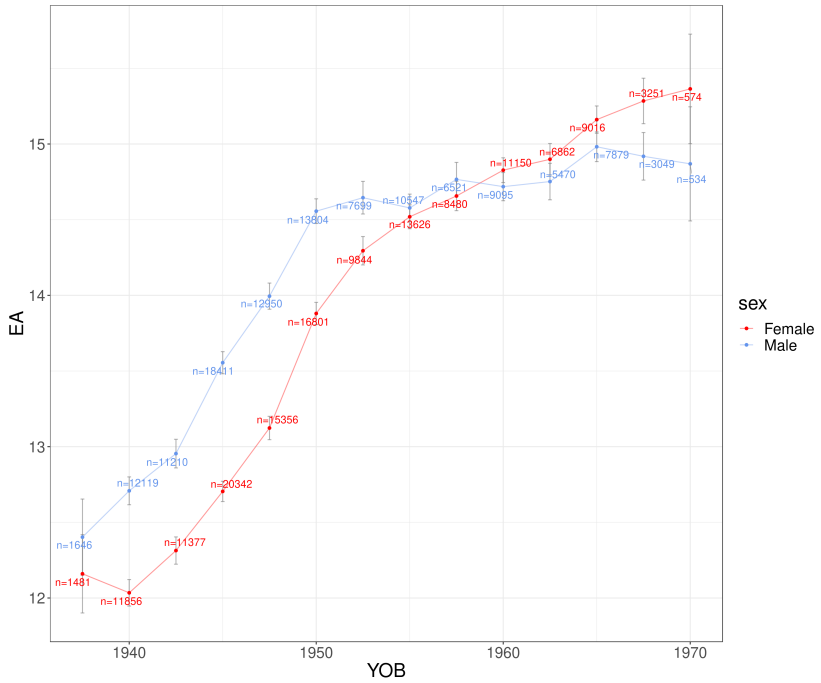


Figure A.1: Educational attainment (EA), year of birth (YOB) and sex. Average EA in years of schooling (y-axis) stratified by YOB (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per YOB and sex group, vertical lines depict 95% CI.

A.1.2 Height

The height variable was based on data-field 50-0-0: 'Standing height', which was measured in the initial assessment visit. In the subset of 271,820 WB 'unrelated' individuals with height measurements, I adjusted for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in

R treating rank-based inverse normalised height values as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

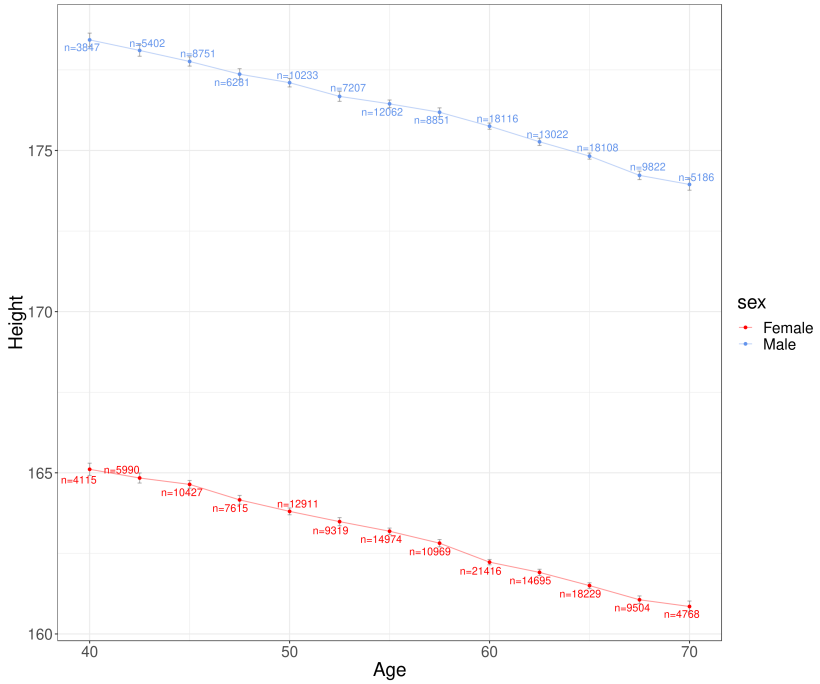


Figure A.2: Height, age and sex. Average height in cm (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.3 Body Mass Index (BMI)

The BMI variable was based on the data-field 21001-0-0: 'Body Mass index (BMI)', which was computed from height and weight measurements recorded in the initial assessment visit. In the subset of 271,535 WB 'unrelated' individuals with BMI values, I adjusted for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in R treating rank-based inverse normalised BMI values as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

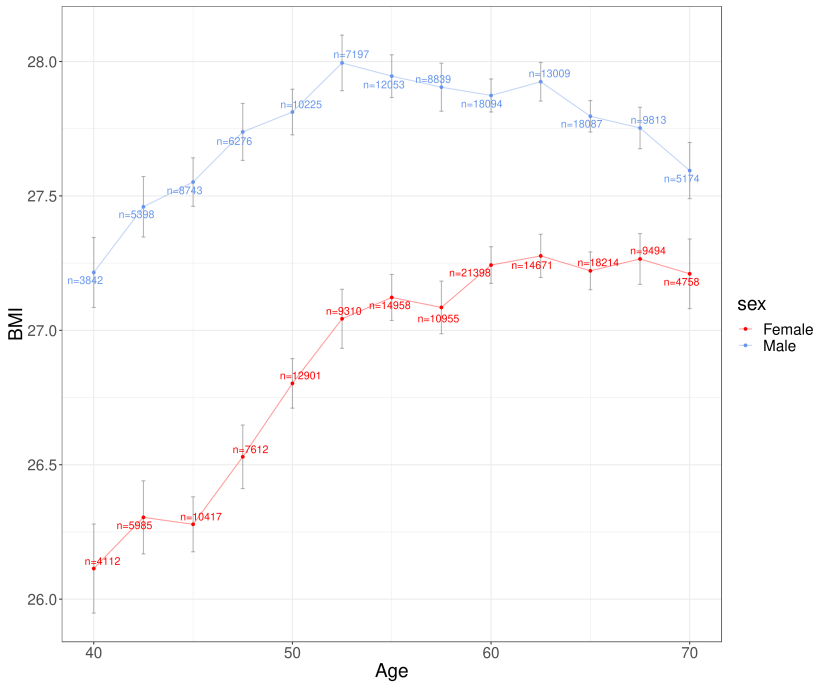


Figure A.3: Body mass index (BMI), age and sex. Average BMI (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.4 Age at First Birth (AFB)

The AFB variable was constructed from answers to the touchscreen question 'How old were you when you had your FIRST child?' which was asked in the initial assessment visit, (data-field 2754: 'Age at first live birth'). This information was collected from women who had indicated that they had given birth to a child in an answer to a previous question in the touchscreen questionnaire (data-field 2734: 'Number of live births'). Entries below 12 were excluded. In the subset of 98,653 WB 'unrelated' women with applicable AFB entries, I adjusted for year of birth up to the order of three, age at recruitment and 40 PCs by performing linear regression in `R`. The resulting residuals were then rank-based inverse normalised.

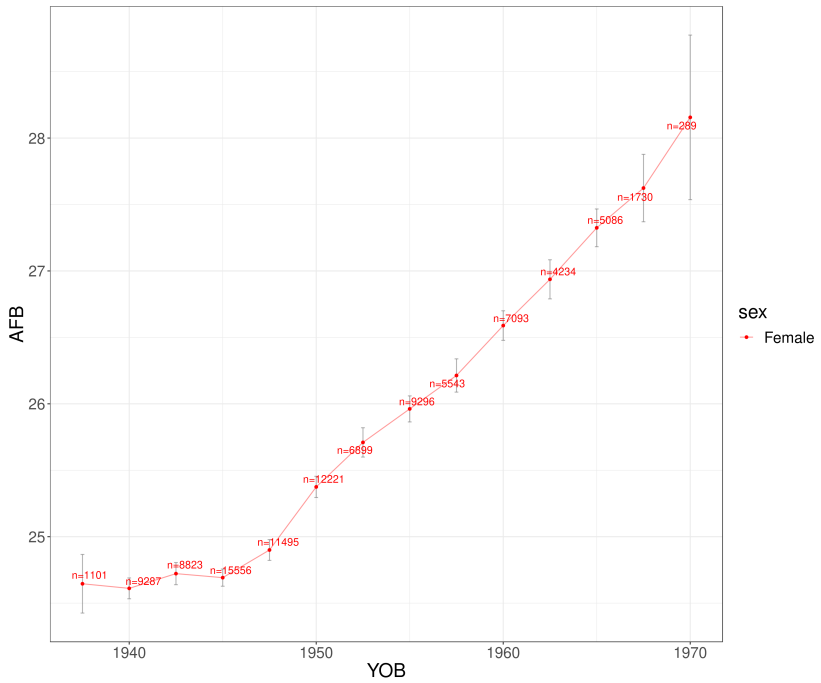


Figure A.4: Age at first birth (AFB) and year of birth (YOB). Average AFB (y-axis) stratified by YOB (x-axis) in the group of WB 'unrelated'. Each dot depicts the average per YOB group, vertical lines depict 95% CI.

A.1.5 Number of Children (N.children)

The N. children variable was constructed from answers to the touchscreen questions 'How many children have you fathered?' (data-field 2405: 'Number of children fathered') and 'How many children have you given birth to? (Please include live births only)' (data-field 2734: 'Number of live births'). Both questions were asked in the initial assessment visit. Data-field 2405 was only collected from males and data-field 2734 was only collected from women. In the subset of 271,317 WB 'unrelated' individuals with applicable N. children entries, I adjusted for year of birth up to the order of three, age at recruitment and 40 PCs separately for each sex by performing linear regression in `R`. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

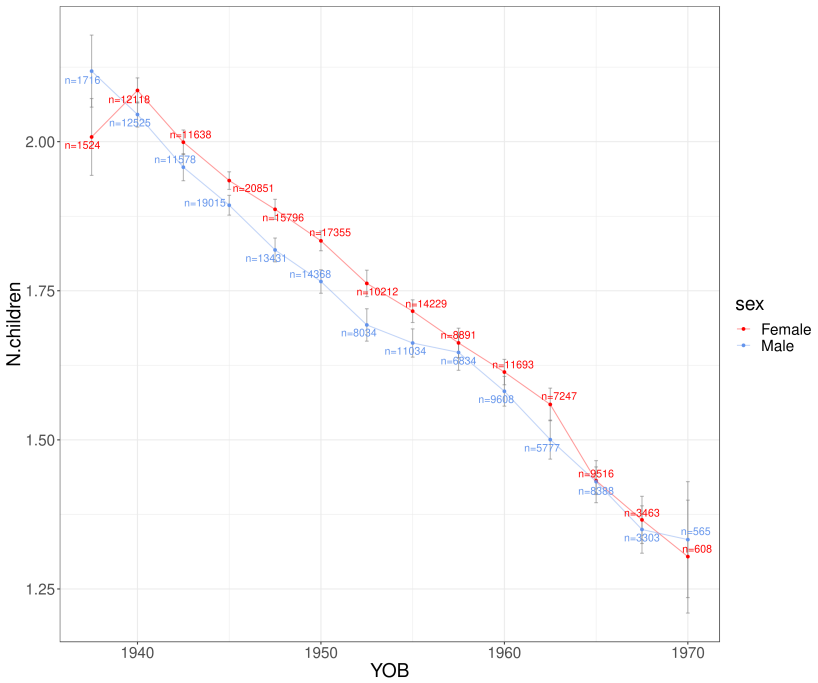


Figure A.5: Number of children (N. children), year of birth (YOB) and sex. Average N. children (y-axis) stratified by YOB (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per YOB and sex group, vertical lines depict 95% CI.

A.1.6 Number of Siblings (N. siblings)

The N. siblings variable was constructed from answers to the touchscreen questions 'How many sisters do you have? (Please include those who have died, and twin sisters. Do not include half-sisters, step-sisters or adopted sisters)' (data-field 1883: 'Number of full sisters') and 'How many brothers do you have? (Please include those who have died, and twin brothers. Do not include half-brothers, step-brothers or adopted brothers)' (data-field 1873: 'Number of full brothers'). Both questions were asked in the initial assessment visit. In the subset of 268,191 WB 'unrelated' individuals with applicable N. siblings entries, I adjusted for year of birth up to the order of three, age at recruitment and 40 PCs separately for each sex by performing linear regression

in R. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

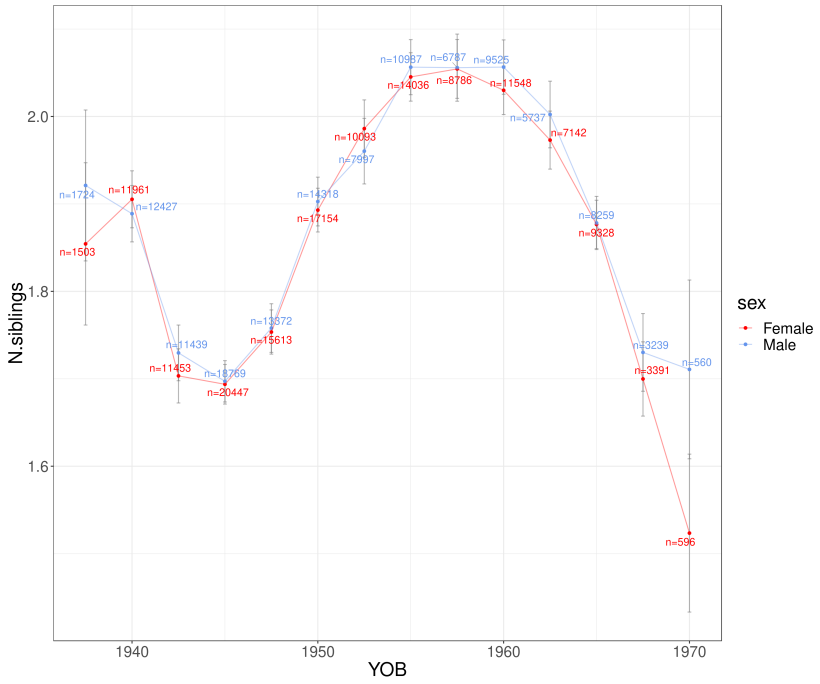


Figure A.6: Number of siblings (N. siblings), year of birth (YOB) and sex. Average N. siblings (y-axis) stratified by YOB (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per YOB and sex group, vertical lines depict 95% CI.

A.1.7 High-Density Lipoprotein (HDL) Cholesterol

The HDL cholesterol variable was based on data-field 30760-0-0: 'HDL Cholesterol', a biochemistry marker measured in blood samples collected at recruitment. In the group of WB 'unrelated' individuals, 34,628 individuals had missing HDL cholesterol values and, as indicated by data-field 30765: 'HDL cholesterol missing reason' and data-field 30766: 'HDL cholesterol missing reportability', 3 of those were missing because the values were below the reportable range and one was above the reportable range. I imputed the 'below reportable range' entries to the minimum reportable HDL cholesterol

measurement in the group of WB 'unrelated' individuals minus 0.001, that is to 0.218, and the 'above reportable range' entry to the maximum reportable HDL cholesterol measurement plus 0.001, that is to 4.402. Then, in the group of 237,785 WB 'unrelated' individuals with applicable values, I adjusted the HDL cholesterol values for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in `R` treating rank-based inverse normalised HDL cholesterol values (including the imputed ones) as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

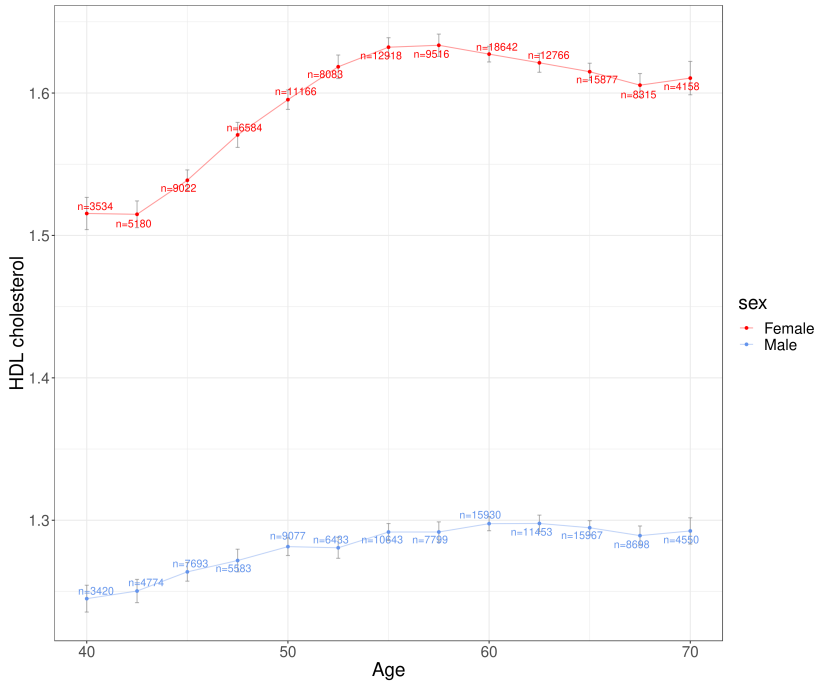


Figure A.7: High-density lipoprotein (HDL) cholesterol, age and sex. Average HDL cholesterol (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.8 Glycated Haemoglobin (HbA1c)

The HbA1c variable was based on data-field 30750-0-0: 'Glycated haemoglobin (HbA1c)', a biochemistry marker measured in blood samples collected at recruitment. In the group of WB 'unrelated' individuals, 12,926 individuals had missing HbA1c values. As indicated by data-field 30755: 'Glycated haemoglobin (HbA1c) missing reason' and data-field 30756 'Glycated haemoglobin (HbA1c) reportability', 111 of those were missing because the values were below reportable range. I imputed the 'below reportable range' entries to the minimum HbA1c measurement minus 0.001, that is to 15.299. In the group of 259,594 WB 'unrelated' individuals with applicable values, I adjusted HbA1c levels for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in `R` treating rank-based inverse normalised HbA1c values (including the imputed ones) as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

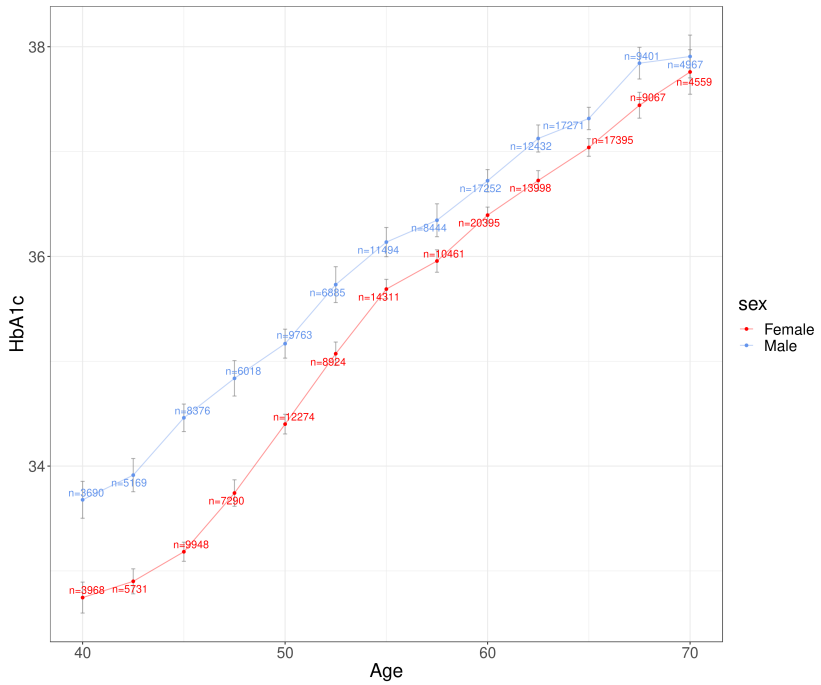


Figure A.8: Glycated haemoglobin (HbA1c), age and sex. Average HbA1c (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.9 Vitamin D

The vitamin D variable was based on data-field 30890-0-0: 'Vitamin D', a biochemistry marker measured in blood samples collected at recruitment. In the group of WB 'unrelated' individuals, 24,116 individuals had missing vitamin D values. As indicated by data-field 30895: 'Vitamin D missing reason' and data-field 30896 'Vitamin D reportability', 784 of those were missing because the values were below reportable range and 2 were missing because the values were above the reportable range. I imputed the 'below reportable range' entries to the minimum vitamin D measurement minus 0.001, that is to 9.999 and the 'above reportable range' entries to the maximum vitamin D measurement plus 0.001, that is to 340.001. In the group of 249,079 WB 'unrelated' individuals with applicable values, I adjusted vitamin D levels for year of birth, age at

recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in R treating rank-based inverse normalised vitamin D values (including the imputed ones) as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

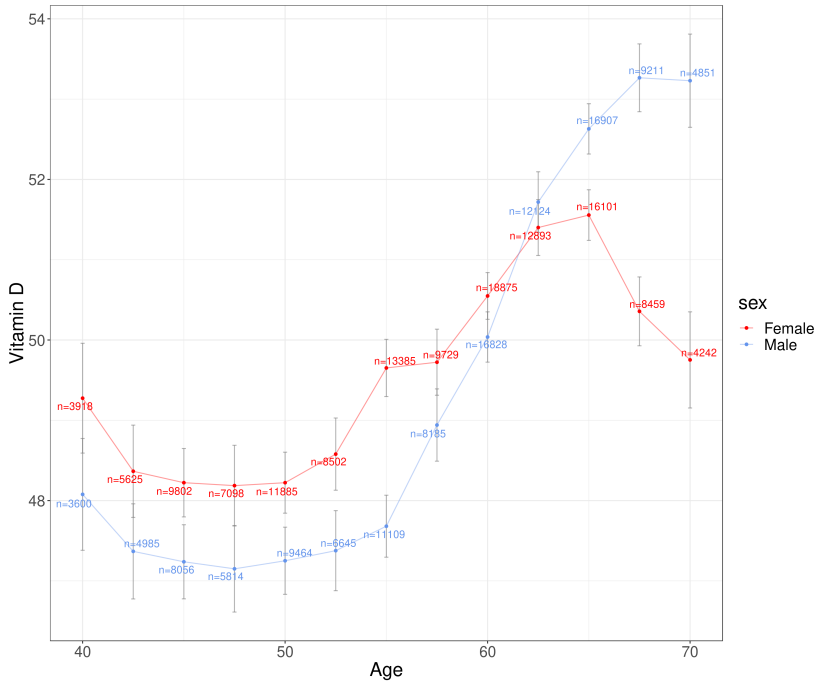


Figure A.9: Vitamin D, age and sex. Average vitamin D (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.10 Glucose

The glucose variable was based on data-field 30740-0-0: 'Glucose', a biochemistry marker measured in blood samples collected at recruitment. In the group of WB 'unrelated' individuals, 34,783 individuals had missing glucose values. As indicated by data-field 30745: 'Glucose missing reason' and data-field 30746 'Glucose reportability', three of those were missing because the values were below reportable range. I imputed the 'below reportable range' entries to the minimum glucose measurement minus 0.001,

that is to 1.004. In the group of 237,629 WB 'unrelated' individuals with applicable values, I adjusted glucose levels for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in `R` treating rank-based inverse normalised glucose values (including the imputed ones) as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

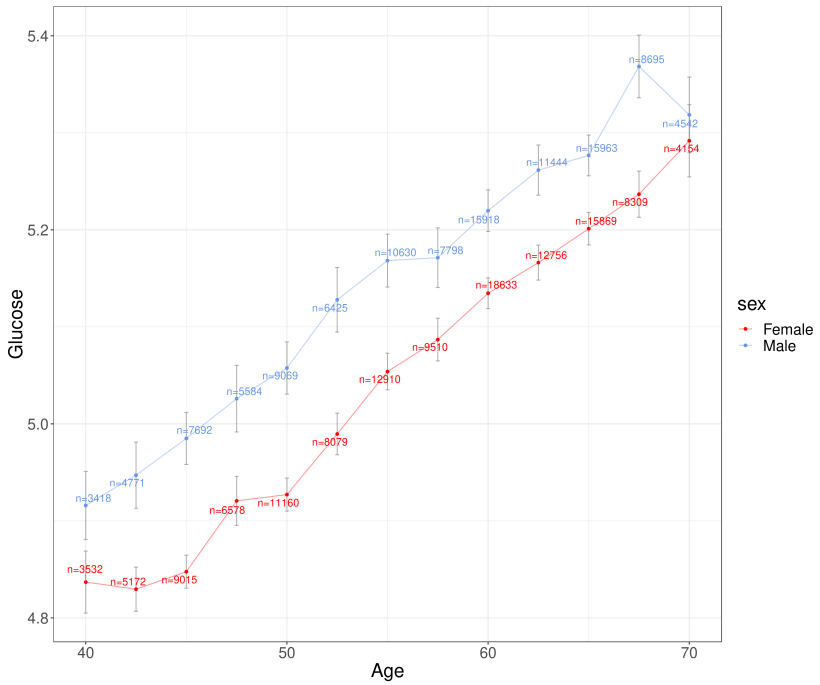


Figure A.10: Glucose, age and sex. Average glucose (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.11 Grip Strength

The grip strength variable was based on data-field 46 'Hand grip strength (left)' and data-field 47 'Hand grip strength (right)', physical measures attained in the initial assessment visit. For each individual I chose the maximum value (left, right) as their grip strength measurement. For the 270,525 WB 'unrelated' individuals with non-

missing grip strength and height values, I adjusted grip strength measurements for year of birth, age at recruitment up to the order of three, standing height and 40 PCs separately for each sex by performing linear regression in `R` treating rank-based inverse normalised grip strength values as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

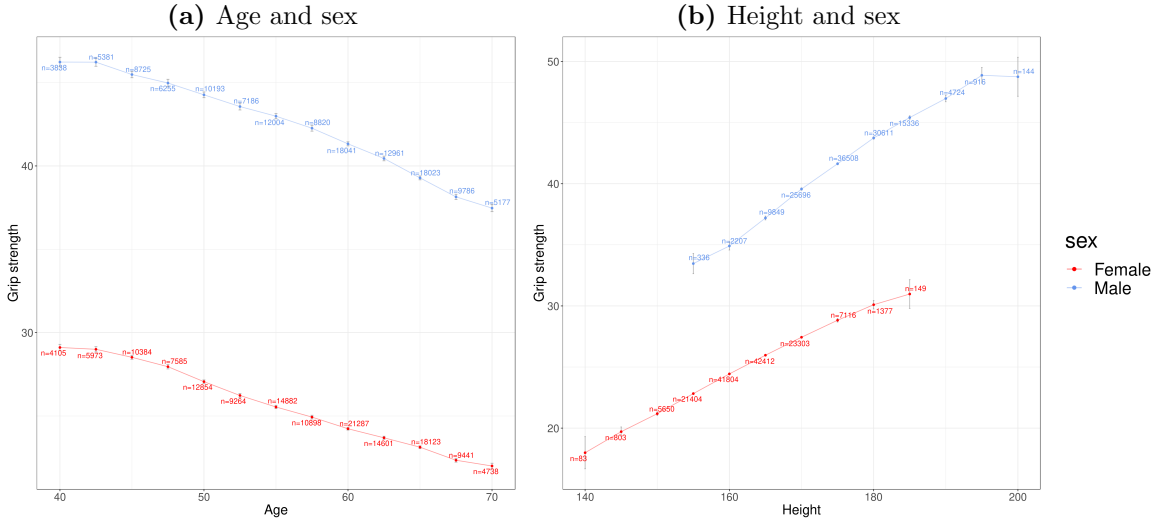


Figure A.11: Grip strength, age, height and sex. a) Average grip strength (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. b) Average grip strength (y-axis) stratified by height in cm (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Averages for groups with less than 50 individuals are not depicted. Each dot depicts the average per age/height and sex group, vertical lines depict 95% CI.

A.1.12 Lipoprotein A

The lipoprotein A variable was based on data-field 30790-0-0: 'Lipoprotein A', a biochemistry marker measured in blood samples collected at recruitment. In the group of WB 'unrelated' individuals, 65,941 individuals had missing lipoprotein A values. As indicated by data-field 30795: 'Lipoprotein A missing reason' and data-field 30796 'Lipoprotein A reportability', 26,943 of those were missing because the values were below reportable range and 18,287 were missing because the values were above the re-

portable range. I imputed the 'below reportable range' entries to the minimum lipoprotein A measurement minus 0.001, that is to 3.799 and the 'above reportable range' entries to the maximum lipoprotein A measurement plus 0.001, that is to 189.001. In the group of 251,698 WB 'unrelated' individuals with applicable values, I adjusted lipoprotein A levels for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in R treating rank-based inverse normalised lipoprotein A values (including the imputed ones) as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

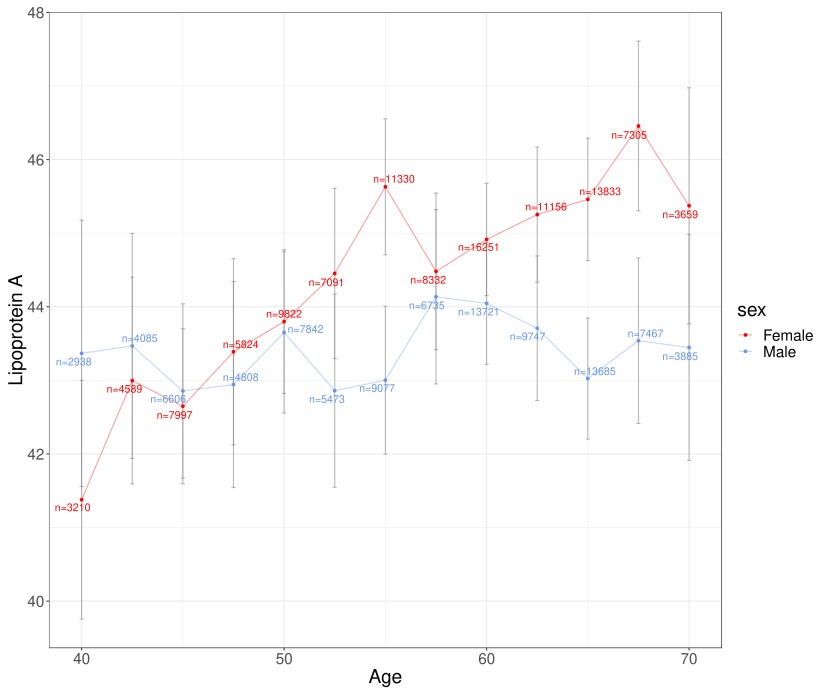


Figure A.12: Lipoprotein A, age and sex. Average lipoprotein A (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.13 Sex Hormone Binding Globulin (SHBG)

The SHBG variable was based on data-field 30830-0-0: 'SHBG', a biochemistry marker measured in blood samples collected at recruitment. In the group of WB 'unrelated' individuals, 36,843 individuals had missing SHBG values. As indicated by data-field 30835: 'SHBG missing reason' and data-field 30836 'SHBG reportability', 5 of those were missing because the values were below reportable range and 389 were missing because the values were above the reportable range. I imputed the 'below reportable range' entries to the minimum SHBG measurement minus 0.001, that is to 0.389 and the 'above reportable range' entries to the maximum SHBG measurement plus 0.001, that is to 241.921. In the group of 235,960 WB 'unrelated' individuals with applicable values, I adjusted SHBG levels for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in `R` treating rank-based inverse normalised SHBG values (including the imputed ones) as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

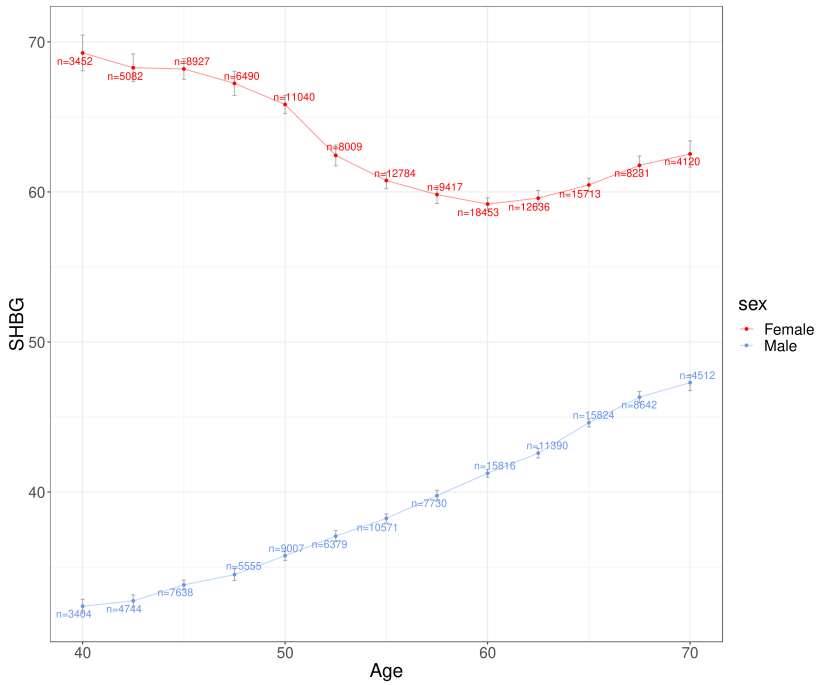


Figure A.13: Sex hormone binding globulin (SHBG), age and sex. Average SHBG (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.14 Testosterone

The testosterone variable was based on data-field 30850-0-0: 'Testosterone', a biochemistry marker measured in blood samples collected at recruitment. In the group of WB 'unrelated' individuals, 36,861 individuals had missing testosterone values. As indicated by data-field 30855: 'Testosterone missing reason' and data-field 30856: 'Testosterone reportability', 22,009 of those were missing because the values were below reportable range and 13 were missing because the values were above the reportable range. I imputed the 'below reportable range' entries to the minimum testosterone measurement minus 0.001, that is to 0.349 and the 'above reportable range' entries to the maximum testosterone measurement plus 0.001, that is to 54.343. In the group of 257,570 WB 'unrelated' individuals with applicable values, I adjusted testosterone lev-

els for year of birth, age at recruitment up to the order of three and 40 PCs separately for each sex by performing linear regression in `R` treating rank-based inverse normalised testosterone values (including the imputed ones) as a dependent variable. The resulting residuals, attained for males and females separately, were then rank-based inverse normalised.

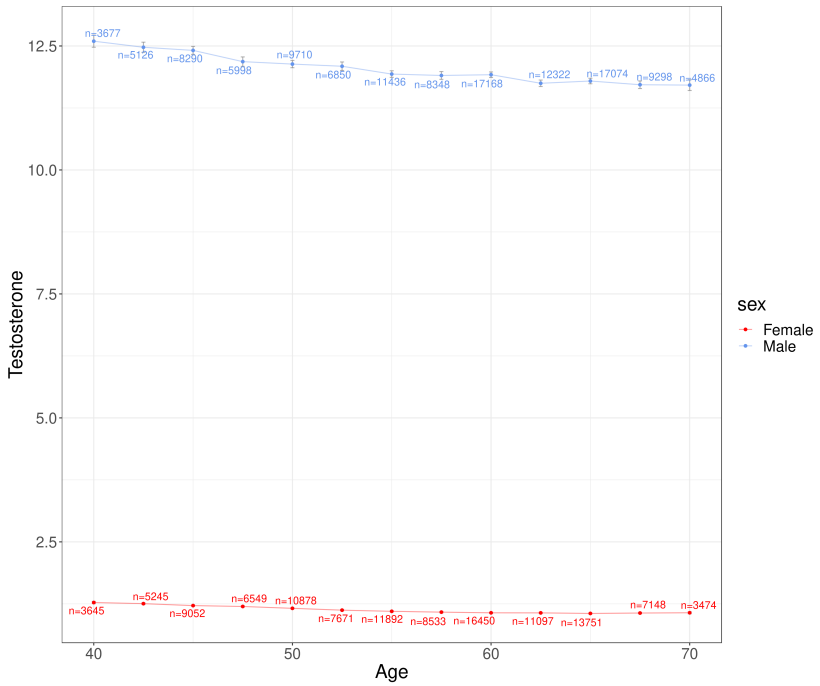


Figure A.14: Testosterone, age and sex. Average testosterone (y-axis) stratified by age (x-axis) and sex (colour legend) in the group of WB 'unrelated'. Each dot depicts the average per age and sex group, vertical lines depict 95% CI.

A.1.15 Physical Activity Study Participation

The physical activity study invitation variable was based on data-field 11005: 'Invitation to physical activity study, acceptance'. The 132,633 WB 'unrelated' individuals with the data-field values 'No response', 'Partial' or 'Completed' were defined as cases. That is, they had been invited to participate in the study. The 139,776 WB 'unrelated' individuals with non-applicable (NA) values in this field were defined as controls. That

is, they had not been invited to the study.

The physical activity study participation variable was based on data-field 11005: 'Invitation to physical activity study, acceptance'. The 59,455 WB 'unrelated' individuals with the data-field values 'Partial' or 'Completed' were defined as cases. The 73,178 WB 'unrelated' individuals with the data-field value 'No response' were defined as controls.

A.1.16 Dietary Study Participation

The dietary study invitation variable was based on data-field 11001: 'Invitation to complete online 24-hour recall dietary questionnaire, acceptance'. The 166,993 WB 'unrelated' individuals with the data-field values 'No response', 'Partial' or 'Completed' were defined as cases. That is, they had been invited to participate in the study. The 105,416 WB 'unrelated' individuals with non-applicable (NA) values in this field were defined as controls. That is, they had not been invited to the study.

The dietary study participation variable was based on data-field 11001: 'Invitation to complete online 24-hour recall dietary questionnaire, acceptance'. The 54,124 WB 'unrelated' individuals with the values 'Partial' or 'Completed' in the field were defined as cases. The 112,869 WB 'unrelated' individuals with the data-field value 'No response' were defined as controls.

A.2 Primary Participation Polygenic Scores

I computed the participation polygenic scores (*pPGS*) with `PLINK 1.90b` [95] by summing over weighted alleles of the 500,632 common high-quality biallelic sequence variants that had surpassed various additional quality control filters (see chapter 4.3.1.6). Each sequence variant was weighted by its corresponding t-statistic from the participation GWAS and the allele corresponding to a positive t-statistic was regarded as the effect-allele. This was done by using the `--score` flag in `PLINK`[95]. I computed separate *pPGS* for each of the four test-statistics described in chapter 4.3.2 (T_{PO}, T_{SIB1}, T_{SIB20} and a linear combination of those three, T_{comb}). Furthermore, to be able to investigate *pPGS* values of individuals with close relatives in the dataset, I also divided the 16,668 sibling pairs into two groups by random, derived primary participation t-statistics from each group and computed additional *pPGS* using those weights.

pPGS based on sequence variants in low pairwise LD were computed as well. In particular, I extracted 189,174 sequence variants from the group of 500,632 high-quality variants by choosing the most significant sequence variant in each LD-group given their T_{comb} *P*-value. LD computations were based on 10,000 WB individuals with no close relatives in the UK Biobank dataset. For this purpose, `PLINK 1.90b`[95] was used with the following flags: `--clump-kb 500 --clump-r2 0.1 --clump-p1 1 --clump-p2 1`. The *pPGS* based on all 500,632 sequence variants was chosen for subsequent analysis as it had a better prediction power for EA in the subset WB individuals with no close relatives within UKBB ($r^2 = 0.0010$ ($P = 1.5 \cdot 10^{-56}$) vs. $r^2 = 0.0006$ ($P = 1.4 \cdot 10^{-37}$), N=260,950). To compute the variance explained, EA values were transformed as described in section A.1.1.

A.3 Within-Study Replication

By using various data-fields in the UKBB data release, I constructed a range of phenotypic variables (see table 4.3) and investigated their relationships with the *pPGS*. For quantitative phenotypes, this was done by performing linear regression in `R` (version 3.4.3) with the corresponding phenotype as the response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of 272,409 WB individuals with no close relatives in UKBB (>3rd degree for all pairs). Prior to regression analysis, the quantitative phenotypes were transformed as described in section A.1. The *pPGS* was computed as described in section A.2. For the binary phenotypes, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. For each *pPGS*-phenotype association result, I accounted for population stratification and cryptic relatedness within the group of 272,409 individuals by dividing the chi-square statistic from the regression with the LD score regression intercept of the corresponding phenotype[17] (see section A.4).

A.4 LD Score Regression

I performed LD score regression with the program `LDSC` (version 1.0.1)[17] using the European Ancestry LD scores computed by the Pan-UKB team[85] (downloaded on April 7th 2021). The analysis was restricted to the 500,632 variants that passed the additional quality control filters described in 4.3.1.6. LD score regression intercepts and genetic correlations were estimated for the primary participation test statistics described in chapter 4.3.1.9, direct-effect-only EA as reported by Young et al. 2022[135]

and the phenotypes shown in table 4.3. I attained summary statistics for the phenotypes in table 4.3 by running GWAS in the group of WB 'unrelated' using `BOLT-LMM` (version 2.3)[66]. For quantitative phenotypes, which had been transformed as described in appendix A.1, genotyping array was added as an additional covariate in the GWAS. For the categorical phenotypes, YOB, age at recruitment up to the order of three, 40 PCs, sex and genotyping array were added as covariates. For the LD score regression analysis I used the standard linear regression $P-$values, `P_LINREG`, from the `BOLT` output.

Genetic correlations between T_{SIB20} and other phenotypes were attained with a constrained-intercept LD score regression using the flags: `--intercept-h2 1,b1 --intercept-gencov 0,0`. The h2 intercept for T_{SIB20} was restricted to 1. The h2 intercept for direct EA was set at `b1=1` and the intercept for other phenotypes was restricted to their fitted intercept from unrestricted LD score regression (e.g. the fitted intercept for AFB was `b1=1.02`). While we have a strong prior for restricting the h2 intercepts for T_{SIB20} and direct EA to 1, the rationale for restricting the intercepts for the other phenotypes to their fitted value is not as obvious. However, this was considered appropriate for this analysis as these fitted intercepts are estimated from well-powered GWASs. Furthermore, genetic correlation analysis with unrestricted h2 intercepts resulted in genetic correlations that were consistent with the restricted analysis and standard errors that were on average only 10% higher for the traits depicted in table 4.4. T_{SIB20} is not based on a traditional genotype-phenotype association analysis and as a result there is no overlap between the T_{SIB20} GWAS and the other GWASs. For this reason, the intercept for the genetic covariance was restricted to zero in all analysis. Furthermore, except for the direct EA GWAS, the T_{SIB20} GWAS is based on a group of individuals

that does not overlap with the group of individuals that the other GWASs are based on.

Appendix B

Supplementary Tables

Quantitative traits	**Effect(SD)**	**P-value**	**N**
Educational attainment	0.0094	$3.5 \cdot 10^{-6}$	260,950
BMI	−0.0112	$1.0 \cdot 10^{-8}$	271,535
Age at first birth	0.0084	$8.9 \cdot 10^{-3}$	98,653
HDL cholesterol	0.0038	$7.5 \cdot 10^{-2}$	237,785
Height	0.0035	$9.1 \cdot 10^{-2}$	271,820
HbA1c	−0.0072	$3.8 \cdot 10^{-4}$	259,594
N. siblings	−0.0042	$4.0 \cdot 10^{-2}$	268,191
N. children	−0.0013	0.51	271,317
Glucose	−0.0006	0.79	237,629
Vitamin D	−0.0007	0.75	249,079
Lipoprotein A	−0.0017	0.40	251,698
SHBG	0.0074	$4.1 \cdot 10^{-4}$	235,960
Grip strength	−0.0024	0.21	270,525
Testosterone	0.0031	0.11	257,570
Binary traits	**log(OR)**	**P-value**	**Ncase/Ncontrol**
Diet. study invitation	0.0091	$3.1 \cdot 10^{-2}$	166,993/105,416
Diet. study participation	0.0150	$4.3 \cdot 10^{-3}$	54,124/112,869
Phys. study invitation	0.0137	$4.2 \cdot 10^{-4}$	132,633/139,776
Phys. study participation	0.0170	$2.1 \cdot 10^{-3}$	59,455/73,178

Table B.1: T^*_{PO} primary participation polygenic score (*pPGS*) and phenotypes. For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as a response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of white British descent who have no close relatives in UKBB (>3rd degree for all pairs). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex. Then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Diet. study: Dietary study. Phys. study: Physical activity study. Information about age at first birth was only available for women. For the binary traits, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for the *pPGS* and the corresponding P-value. The P-values have been adjusted with the LD score regression intercept of the corresponding phenotype. This table shows results for the *pPGS* computed with T^*_{PO} weights which are based on comparing transmitted and untransmitted alleles among 4,427 parent-offspring pairs in the UKBB and then adjusted for major allele bias.

Quantitative traits	**Effect(SD)**	**P-value**	**N**
Educational attainment	0.0230	$3.2 \cdot 10^{-30}$	260,950
BMI	−0.0077	$7.9 \cdot 10^{-5}$	271,535
Age at first birth	0.0196	$1.1 \cdot 10^{-9}$	98,653
HDL cholesterol	0.0066	$2.1 \cdot 10^{-3}$	237,785
Height	0.0088	$1.7 \cdot 10^{-5}$	271,820
HbA1c	0.0009	0.65	259,594
N. siblings	−0.0044	$3.1 \cdot 10^{-2}$	268,191
N. children	−0.0037	$5.9 \cdot 10^{-2}$	271,317
Glucose	−0.0041	$4.6 \cdot 10^{-2}$	237,629
Vitamin D	−0.0018	0.36	249,079
Lipoprotein A	−0.0027	0.19	251,698
SHBG	0.0022	0.30	235,960
Grip strength	0.0009	0.63	270,525
Testosterone	0.0023	0.24	257,570
Binary traits	**log(OR)**	**P-value**	**Ncase/Ncontrol**
Diet. study invitation	0.0245	$6.5 \cdot 10^{-9}$	166,993/105,416
Diet. study participation	0.0122	$2.0 \cdot 10^{-2}$	54,124/112,869
Phys. study invitation	0.0163	$3.0 \cdot 10^{-5}$	132,633/139,776
Phys. study participation	0.0203	$2.5 \cdot 10^{-4}$	59,455/73,178

Table B.2: T^*_{SIB1} primary participation polygenic score (*pPGS*) and phenotypes. For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as a response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of white British descent who have no close relatives in UKBB (>3rd degree for all pairs). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex. Then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Diet. study: Dietary study. Phys. study: Physical activity study. Information about age at first birth was only available for women. For the binary traits, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for the *pPGS* and the corresponding P-value. The P-values have been adjusted with the LD score regression intercept of the corresponding phenotype. This table show results for the *pPGS* computed with T^*_{SIB1} weights which are based on comparing shared and non-shared alleles among sibling pairs in UKBB that share a variant in question IBD=1 and then adjusted for major allele bias.

Quantitative traits	**Effect(SD)**	**P-value**	**N**
Educational attainment	0.0198	$1.1 \cdot 10^{-22}$	260,950
BMI	−0.0147	$4.5 \cdot 10^{-14}$	271,535
Age at first birth	0.0223	$3.4 \cdot 10^{-12}$	98,653
HDL cholesterol	0.0095	$7.8 \cdot 10^{-6}$	237,785
Height	0.0028	0.17	271,820
HbA1c	−0.0057	$5.2 \cdot 10^{-3}$	259,594
N. siblings	−0.0008	0.69	268,191
N. children	−0.0029	0.13	271,317
Glucose	−0.0020	0.33	237,629
Vitamin D	−0.0010	0.63	249,079
Lipoprotein A	0.0022	0.27	251,698
SHBG	−0.0039	$6.2 \cdot 10^{-2}$	235,960
Grip strength	−0.0009	0.63	270,525
Testosterone	−0.0031	0.12	257,570
Binary traits	**log(OR)**	**P-value**	**Ncase/Ncontrol**
Diet. study invitation	0.0237	$1.9 \cdot 10^{-8}$	166,993/105,416
Diet. study participation	0.0207	$8.4 \cdot 10^{-5}$	54,124/112,869
Phys. study invitation	0.0184	$2.2 \cdot 10^{-6}$	132,633/139,776
Phys. study participation	0.0159	$4.2 \cdot 10^{-3}$	59,455/73,178

Table B.3: T_{SIB20} primary participation polygenic score (*pPGS*) and phenotypes. For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as a response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of white British descent who have no close relatives in UKBB (>3rd degree for all pairs). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex. Then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Diet. study: Dietary study. Phys. study: Physical activity study. Information about age at first birth was only available for women. For the binary traits, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for the *pPGS* and the corresponding P-value. The P-values have been adjusted with the LD score regression intercept of the corresponding phenotype. This table shows results for the *pPGS* computed with T_{SIB20} weights which are based on comparing allele frequency among sibling pairs in UKBB sharing a variant in question IBD=2 to sibling pairs sharing the variant IBD=0.

Quantitative traits	**Effect(SD)**	**P-value**	**N**
Educational attainment	0.0104	$2.5 \cdot 10^{-7}$	260,950
BMI	−0.0114	$5.3 \cdot 10^{-9}$	271,535
Age at first birth	0.0093	$3.7 \cdot 10^{-3}$	98,653
HDL cholesterol	0.0034	0.11	237,785
Height	0.0033	0.11	271,820
HbA1c	−0.0074	$2.6 \cdot 10^{-4}$	259,594
N. siblings	−0.0043	$3.3 \cdot 10^{-2}$	268,191
N. children	−0.0011	0.56	271,317
Glucose	−0.0010	0.64	237,629
Vitamin D	−0.0007	0.75	249,079
Lipoprotein A	−0.0021	0.29	251,698
SHBG	0.0072	$6.0 \cdot 10^{-4}$	235,960
Grip strength	−0.0023	0.24	270,525
Testosterone	0.0031	0.12	257,570
Binary traits	**log(OR)**	**P-value**	**Ncase/Ncontrol**
Diet. study invitation	0.0103	$1.4 \cdot 10^{-2}$	166,993/105,416
Diet. study participation	0.0153	$3.6 \cdot 10^{-3}$	54,124/112,869
Phys. study invitation	0.0153	$8.8 \cdot 10^{-5}$	132,633/139,776
Phys. study participation	0.0183	$1.0 \cdot 10^{-3}$	59,455/73,178

Table B.4: T_{PO} primary participation polygenic score (*pPGS*) and phenotypes. For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as a response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of white British descent who have no close relatives in UKBB (>3rd degree for all pairs). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex. Then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Diet. study: Dietary study. Phys. study: Physical activity study. Information about age at first birth was only available for women. For the binary traits, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for the *pPGS* and the corresponding P-value. The P-values have been adjusted with the LD score regression intercept of the corresponding phenotype. This table shows results for the *pPGS* computed with unadjusted T_{PO} weights which are based on comparing transmitted and untransmitted alleles among 4,427 parent-offspring pairs in the UKBB.

Quantitative traits	**Effect(SD)**	**P-value**	**N**
Educational attainment	0.0244	$1.4 \cdot 10^{-33}$	260,950
BMI	−0.0081	$3.6 \cdot 10^{-5}$	271,535
Age at first birth	0.0208	$1.0 \cdot 10^{-10}$	98,653
HDL cholesterol	0.0058	$6.7 \cdot 10^{-3}$	237,785
Height	0.0083	$5.1 \cdot 10^{-5}$	271,820
HbA1c	0.0004	0.83	259,594
N. siblings	−0.0046	$2.4 \cdot 10^{-2}$	268,191
N. children	−0.0034	$8.3 \cdot 10^{-2}$	271,317
Glucose	−0.0047	$2.3 \cdot 10^{-2}$	237,629
Vitamin D	−0.0018	0.37	249,079
Lipoprotein A	−0.0034	0.10	251,698
SHBG	0.0019	0.37	235,960
Grip strength	0.0011	0.57	270,525
Testosterone	0.0022	0.26	257,570
Binary traits	**log(OR)**	**P-value**	**Ncase/Ncontrol**
Diet. study invitation	0.0261	$6.4 \cdot 10^{-10}$	166,993/105,416
Diet. study participation	0.0126	$1.7 \cdot 10^{-2}$	54,124/112,869
Phys. study invitation	0.0187	$1.70 \cdot 10^{-6}$	132,633/139,776
Phys. study participation	0.0221	$6.80 \cdot 10^{-5}$	59,455/73,178

Table B.5: T_{SIB1} primary participation polygenic score (*pPGS*) and phenotypes For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as a response and the *pPGS* and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of white British descent who have no close relatives in UKBB (>3rd degree for all pairs). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age and 40 PCs separately for each sex. Then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Diet. study: Dietary study. Phys. study: Physical activity study. Information about age at first birth was only available for women. For the binary traits, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for the *pPGS* and the corresponding P-value. The P-values have been adjusted with the LD score regression intercept of the corresponding phenotype. This table shows results for the *pPGS* computed with unadjusted T_{SIB1} weights which are based on comparing shared and not-shared alleles of sibling pairs in UKBB that share the given variant IBD=1.

Quantitative traits	**Effect(SD)**	**P-value**	**N**
BMI	-0.1395	$< 1.0 \cdot 10^{-300}$	$260,118$
Age at first birth	0.3500	$< 1.0 \cdot 10^{-300}$	$95,424$
HDL cholesterol	0.0993	$< 1.0 \cdot 10^{-300}$	$227,964$
Height	0.1458	$< 1.0 \cdot 10^{-300}$	$260,389$
HbA1c	-0.0728	$1.1 \cdot 10^{-272}$	$248,889$
N. children	-0.0550	$3.2 \cdot 10^{-170}$	$260,070$
N. siblings	-0.1454	$< 1.0 \cdot 10^{-300}$	$256,956$
Vitamin D	-0.0283	$9.8 \cdot 10^{-43}$	$238,755$
Glucose	-0.0260	$1.5 \cdot 10^{-34}$	$227,812$
Grip strength	0.0468	$2.1 \cdot 10^{-121}$	$259,141$
Lipoprotein A	-0.0099	$1.7 \cdot 10^{-6}$	$241,318$
SHBG	0.0348	$9.1 \cdot 10^{-59}$	$226,220$
Testosterone	0.0078	$1.1 \cdot 10^{-4}$	$246,979$
Binary traits	**log(OR)**	**P-value**	**Ncase/Ncontrol**
Diet. study invitation	0.4661	$< 1.0 \cdot 10^{-300}$	$159,956/100,994$
Diet. study participation	0.2621	$< 1.0 \cdot 10^{-300}$	$51,891/108,065$
Phys. study invitation	0.3262	$< 1.0 \cdot 10^{-300}$	$127,024/133,926$
Phys. study participation	0.1826	$2.5 \cdot 10^{-212}$	$56,817/70,207$

Table B.6: Educational attainment and association with other phenotypes. For the quantitative traits, I performed linear regression in `R` with the corresponding phenotype as a response and the educational attainment and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of white British descent who have no close relatives in UKBB (>3rd degree for all pairs). Prior to regression analysis, the quantitative traits, including educational attainment, were adjusted for year of birth (YOB), age and 40 PCs separately for each sex. Then the residuals were rank-based inverse normalised except for educational attainment for which the residuals were standardised (see section A.1). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Diet. study: Dietary study. Phys. study: Physical activity study. Information about age at first birth was only available for women. For the binary traits, I performed a logistic regression in `R` including YOB, age up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for educational attainment and the corresponding P-value. The P-values have been adjusted with the LD score regression intercept of the corresponding phenotype.

Appendix C

Supplementary Figures

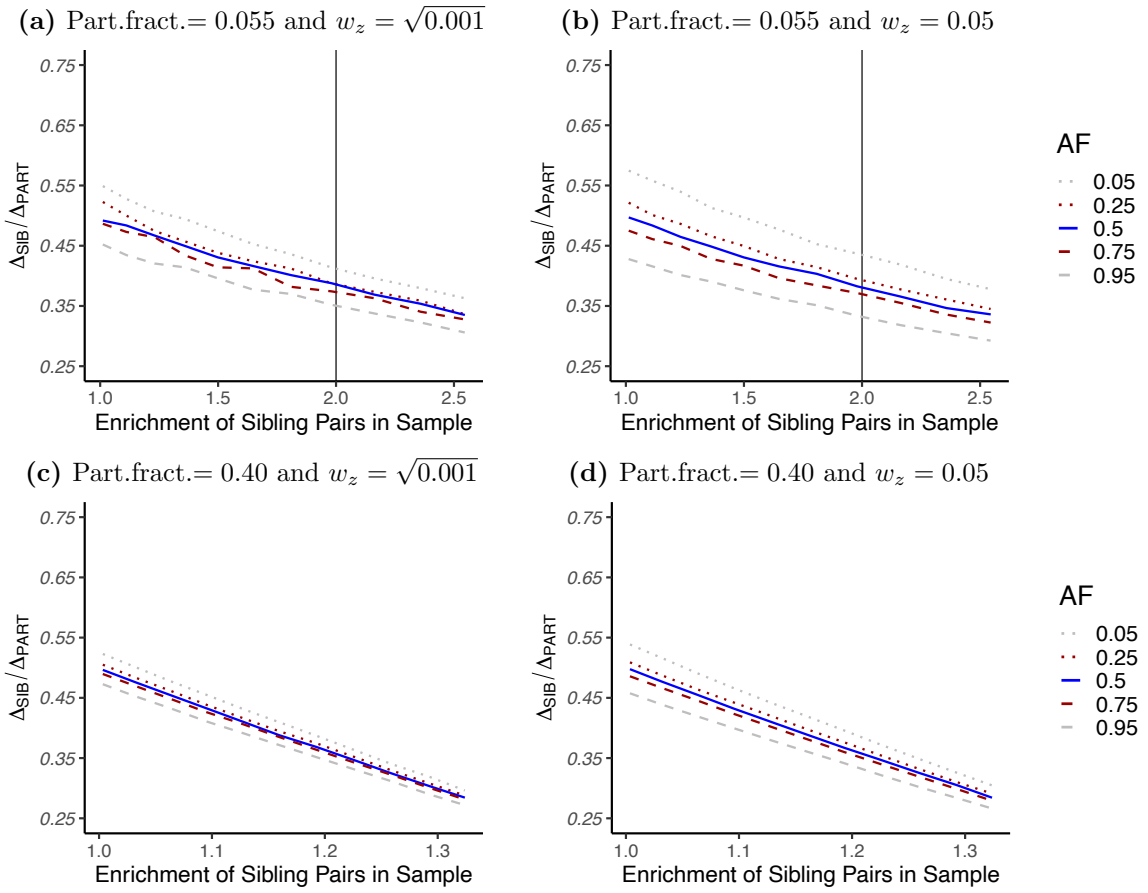


Figure C.1: Participants with a participating sibling and participants with a non-participating sibling. Figures showing how the relative allele frequency difference $\Delta_{SIB}/\Delta_{PART}$ (y-axis) changes with the enrichment of sibling pairs in sample (x-axis) for different participation fractions (a-d). Δ_{SIB} denotes the allele frequency difference between participants with a participating sibling and participants with a non-participating sibling. Δ_{PART} denotes the allele frequency difference between participants and non-participants. These allele frequency differences were attained by simulating trait values, $X_{ij} = w_z \cdot z_{ij} + w_B \cdot B_{ij}$, of 5 million sibling pairs and then allocating them into participants and non-particiants assuming the logit model; $P(Z_{ij} = 1|X_{ij} = X) = \frac{1}{1+e^{-\beta_0 - 1.2 \cdot X}}$. z_{ij} denotes the standardised and mean centred genotype of a single biallelic SNP for individual i in sibling pair j. $B_{ij} \sim N(0, 1)$ denotes all contributions to X_{ij}, environmental and/or genetic, beyond z_{ij}. z_{ij} and B_{ij} are assumed to be uncorrelated while the sibling correlation for z_{ij} is 0.5. In the simulations, w_z was set at $w_z = \sqrt{0.001}$ (panel a and c) and 0.05 (panel b and d). The sibling correlation of B_{ij} ranged from 0.01 and 0.99, allowing for different sibling enrichment. This was done for 5 different allele frequencies (AF) for the underlying genotype of z_{ij}. The simulation process was repeated 2,400 times and then the mean was computed from those replications. Δ_{PART} and Δ_{SIB} were computed for the unstandardised genotype underlying z_{ij}. The dotted and dashed curves depict the relative frequency difference for each AF scenario (legend on the right). The black solid vertical line in figure a) and b) is set at x=2, as this is the reported sibling enrichment in UKBB [20].