Quantitative trait variation and adaptation in contemporary humans

Hakhamanesh Mostafavi

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019
ABSTRACT

Quantitative trait variation and adaptation in contemporary humans

Hakhamanesh Mostafavi

Human genomic data sets are now reaching sample sizes on the order of hundreds of thousands and soon exceeding millions, providing unprecedented opportunities to understand human evolution. Most studies of human adaptation so far have focused on selection that has acted over the past million to few thousand years. However, powered by large data sets, it is now feasible to study allele frequency changes that occur within the short timescale of a few generations, directly observing selection acting in contemporary humans. I take this approach in the work presented in Chapter 1 of this thesis, where we performed a genome-wide scan to identify a set of genetic variants that influence age-specific mortality in present-day samples. Our findings include two variants in the APOE and CHRNA3 loci, as well as sets of variants contributing to a number of traits, including coronary artery disease and cholesterol levels, and intriguingly, to timing of puberty and child birth.

New research directions have also opened up with the advent of large-scale genome-wide association studies (GWAS), which have begun to uncover genetic variants underlying a number of human traits, ranging from disease susceptibility to social and behavioral traits such as educational attainment and neuroticism. One
such direction is the use of polygenic scores (PGS), which aggregate GWAS findings into one score as a measure of genetic propensity for traits, for phenotypic prediction. A major obstacle to this application is that the prediction accuracy of PGS drops in samples that have a different genetic ancestry than the GWAS sample. Our work, presented in Chapter 2, demonstrates that PGS prediction accuracy is also variable within genetic ancestries depending on factors such as age, sex, and socioeconomic status, as well as GWAS study design. These findings have important implications for the increasing use of these measures in diverse disciplines such as social sciences and human genetics.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>iii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 1  Identifying genetic variants that affect viability in large cohorts</td>
<td>9</td>
</tr>
<tr>
<td>1.1 Abstract</td>
<td>9</td>
</tr>
<tr>
<td>1.2 Introduction</td>
<td>10</td>
</tr>
<tr>
<td>1.3 Results</td>
<td>13</td>
</tr>
<tr>
<td>1.3.1 A method for testing for differences in allele frequencies across age bins</td>
<td>13</td>
</tr>
<tr>
<td>1.3.2 Testing for changes in allele frequency at individual genetic variants</td>
<td>18</td>
</tr>
<tr>
<td>1.3.3 Testing for changes in allele frequency at trait-associated variants</td>
<td>25</td>
</tr>
<tr>
<td>1.4 Discussion</td>
<td>33</td>
</tr>
<tr>
<td>1.5 Materials and Methods</td>
<td>37</td>
</tr>
</tbody>
</table>
Chapter 2 Variable prediction accuracy of polygenic scores within an ancestry group

2.1 Abstract ................................................. 84
2.2 Introduction ........................................... 85
2.3 Results .................................................. 90
   2.3.1 Sample characteristics of the GWAS and prediction set can influence prediction accuracy even within a single ancestry ... 90
   2.3.2 Potential portability obstacles explored through a comparison of standard and family-based GWAS ...................... 98
2.4 Implications .......................................... 105
2.5 Materials and Methods ............................ 108
2.6 Supplementary Materials .......................... 117

Future directions ........................................ 157

Bibliography ............................................. 161

Appendix ................................................... 177
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Power of the model to detect changes in allele frequency with age</td>
<td>14</td>
</tr>
<tr>
<td>1.2</td>
<td>Testing for the influence of single genetic variants on age-specific mortality in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort</td>
<td>20</td>
</tr>
<tr>
<td>1.3</td>
<td>Testing for the influence of single genetic variants on age-specific mortality in the UK Biobank</td>
<td>23</td>
</tr>
<tr>
<td>1.4</td>
<td>Testing for the influence of sets of trait-associated variants on survival of the fathers of UK Biobank participants</td>
<td>29</td>
</tr>
<tr>
<td>1.5</td>
<td>Testing for the influence of sets of trait-associated variants on survival of the mothers of UK Biobank participants</td>
<td>32</td>
</tr>
<tr>
<td>1.6</td>
<td>Results of principal component analysis (PCA)</td>
<td>50</td>
</tr>
<tr>
<td>1.7</td>
<td>Age distribution of the GERA individuals at the time of the survey, year 2007</td>
<td>51</td>
</tr>
<tr>
<td>1.8</td>
<td>Comparison of trends in allele frequency with age and age at death</td>
<td>52</td>
</tr>
<tr>
<td>1.9</td>
<td>Validation of European ancestry in GERA</td>
<td>53</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>1.10</td>
<td>Distribution of GERA individuals for PCs inferred from 1000 Genome Project phase 3 data</td>
<td></td>
</tr>
<tr>
<td>1.11</td>
<td>Quantile-quantile plots for model results for individual variants in GERA</td>
<td></td>
</tr>
<tr>
<td>1.12</td>
<td>Frequency of the G allele of rs4988235 with age of the GERA participants</td>
<td></td>
</tr>
<tr>
<td>1.13</td>
<td>Regional plot for the APOE locus</td>
<td></td>
</tr>
<tr>
<td>1.14</td>
<td>Frequency of rs6857 genotypes with age in GERA</td>
<td></td>
</tr>
<tr>
<td>1.15</td>
<td>Frequency of rs6857 genotypes with age among males and females in GERA</td>
<td></td>
</tr>
<tr>
<td>1.16</td>
<td>Frequency of the APOE gene alleles with age in GERA</td>
<td></td>
</tr>
<tr>
<td>1.17</td>
<td>Enrollments of individuals in the Kaiser Permanente Medical Care Plan</td>
<td></td>
</tr>
<tr>
<td>1.18</td>
<td>Testing for the influence of single genetic variants on age-specific mortality in the GERA cohort</td>
<td></td>
</tr>
<tr>
<td>1.19</td>
<td>Quantile-quantile plots for model results for individual variants in the UK Biobank</td>
<td></td>
</tr>
<tr>
<td>1.20</td>
<td>Effect of rs1051730 (CHRNA3) on survival in GERA</td>
<td></td>
</tr>
<tr>
<td>1.21</td>
<td>Allele frequencies of variants in the MEOX2 locus with mother’s age at death in the UK Biobank</td>
<td></td>
</tr>
<tr>
<td>1.22</td>
<td>No significant effect of rs4721453 (near MEOX2) on survival in GERA</td>
<td></td>
</tr>
<tr>
<td>1.23</td>
<td>Ascertainment bias towards older participants introduced by using parental ages at death in the UK Biobank</td>
<td></td>
</tr>
<tr>
<td>1.24</td>
<td>Testing for a significant age by sex effect of individual genetic variants in the UK Biobank</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>Heat map showing interdependence between the age effects of pairs of trait-associated variants in the UK Biobank</td>
<td></td>
</tr>
</tbody>
</table>

iv
1.26 Testing for age effect of sets of trait-associated variants in the UK Biobank, treating age variables as ordinal ........................................... 70

1.27 Testing for age by sex effect of sets of trait-associated variants in the UK Biobank .............................................................. 71

1.28 Trajectories of polygenic scores with father’s age for traits associated with paternal survival in the UK Biobank ........................................... 72

1.29 Trajectories of polygenic scores with mother’s age for traits associated with maternal survival in the UK Biobank ........................................... 73

1.30 Protective effect of later predicted puberty timing on survival in GERA .............................................................. 74

1.31 Testing for stabilizing selection on traits in the UK Biobank ................................................................................. 75

1.32 Power of the model to detect changes in allele frequency with age .............................................................. 76

1.33 Association between variants influencing age at first birth and apparent fertility in the UK Biobank .................................................. 77

1.34 Ascertainment bias towards older participants for older parental age at death categories .................................................. 78

2.1 Variable prediction accuracy within an ancestry group ................................................................................. 92

2.2 Differences in environmental variance alone do not explain the variable prediction accuracy .................................................. 97

2.3 Comparison of prediction accuracy of standard and sib-based polygenic scores ................................................................................. 102

2.4 Variable prediction accuracy (measured as $R^2$) even within an ancestry group ................................................................................. 145
2.5 Dependence on the number of SNPs included in the polygenic score

2.6 Estimating mean effect size across strata

2.7 Variable prediction accuracy within an ancestry: sensitivity to controls for population structure

2.8 Comparison of siblings and unrelated individuals in the UK Biobank with respect to age, SES, and sex ratio

2.9 Comparison of siblings and unrelated individuals in the UK Biobank with respect to population structure

2.10 Simulation results for polygenic scores based on standard GWAS and sib-GWAS in the presence of indirect effects

2.11 Simulation results for polygenic scores based on standard GWAS and sib-GWAS in the presence of assortative mating

2.12 Comparison of prediction accuracies of polygenic scores based on standard and sib-GWAS matched for sex ratio

2.13 Prediction accuracy of polygenic scores based on sib- and standard GWAS for a range of traits

2.14 Variable prediction accuracy within an ancestry group for binary education phenotypes

2.15 Comparison of prediction accuracies of polygenic scores ($R^2$) based on standard and sib-GWAS
## List of Tables

1.1  Associations between sets of trait-associated variants and paternal and maternal survival among the UK Biobank participants of British ancestry under the Cox model.  ............................................. 29

1.2  List of phenotypes and abbreviations. The numbers of loci passing quality control measures are shown for each data set. .................................................. 79

1.3  Results of the Cox model for association of polygenic scores for 42 traits with survival of parents of the UK Biobank participants.  ...................... 80

1.4  Age dependency of hazard ratios for the top associations with parental survival in the UK Biobank under the Cox model.  .............................. 81

1.5  Replication of associations in the discovery panel (UK Biobank individuals of British ancestry) in the UK Biobank participants of non-British ancestry.  82

1.6  Testing for change in polygenic scores with age of the GERA participants.  83

2.1  Traits analyzed and their corresponding data fields in UK Biobank.  ... 141

2.2  Genetic correlations across samples that vary by a study characteristic.  142

2.3  Sample sizes for siblings and unrelated sets  .......................... 143

2.4  Academic degree to years of schooling conversion table.  ............. 144
Acknowledgements

First, I would like to express my most sincere gratitude to my advisor Molly Przeworski. I am forever grateful to her for her trust in me, for the opportunity to join her lab, and the unwavering support every step of the way. She always provided the perfect mix of mentorship and academic freedom I needed to grow as a person, and a scientist.

I owe many thanks to Joe Pickrell for co-supervising the work presented in Chapter 1 of this dissertation, and Guy Sella for always providing insightful guidance and encouragement. I am much indebted to my incredible friend, Arbel Harpak, who jointly authored the work presented in Chapter 2. Knowing him has made me more of a learned person. Collaborating with Jeremy Berg and Zach Fuller on their amazing work was a privilege. Many thanks go to all my other collaborators, specially Jonathan Pritchard, Graham Coop, Dalton Conley, and Tomaz Berisa, and to my committee members, Itsik Pe’er and Peter Andolfatto. My gratitude is extended to John Hunt, without whom I would have not been able to join the Department of Biological Sciences.

The quality discussions with the members of Przeworski and Sella labs: Ipsita,
Felix, David, Laura, Guy A., Yuval, Zach B., Carla, Kristen, Ziyue, Priya, Eduardo, Molly S., Sonal, Alva, Chen, Yuki, and Egill, and their friendship were priceless. Also, thanks to Ipsita Agarwal for her comments on this dissertation, and Zach Fuller and Guy Amster for the technical support.

I would like to acknowledge my family: my brother Nariman, and cousins, Siavosh, Soroush, Sorena and Alireza. Mom, this, and any other achievements of mine, are indeed yours. Ava, nothing in my life would be meaningful without you.
To my mother.
Introduction

Since emerging \(~200,000\) years ago in Africa, modern humans have populated the globe in a relatively short amount of time, which required being able to adapt to a diverse range of environments (e.g., with respect to sunlight exposure, diet, climate). Furthermore, changes in life-style brought about new challenges, e.g., spread of infectious diseases with increased population densities facilitated by the transitioning from hunter-gathering to farming. Understanding to what extent the genetic variation segregating in humans is shaped by such selection pressures versus neutral evolutionary processes has been a long-standing goal of human evolutionary biology.

With the availability of large human data sets made possible by the advances in the genome sequencing and genotyping technologies over the past twenty years, several statistical methods have been developed to use polymorphism data to detect the footprint of past adaption events in the genome of present-day humans. Almost all early efforts focused on strong positive selection at a single locus, namely a “selective sweep”, where an advantageous mutation rapidly rises in frequency in the population. Compelling examples of human adaption were identified by searching the
genome for signatures of selective sweeps, notably for mutations underlying adaptation to diet, altitude, skin pigmentation, and pathogen resistance (e.g., 3, 4, 7, 8). However, later studies demonstrated that strong selection at single-locus level did not make a large contribution in shaping the genetic variation landscape in modern humans 9, 10. These findings led to the realization that selection on polygenic traits acting simultaneously at several loci, i.e., “polygenic selection”, would induce small shifts in allele frequency at trait-associated loci that are missed by methods designed to detect large allele frequency changes at a single locus, and motivated a shift in focus to understanding and detecting this form of selection 11.

With the advent of large-scale genome-wide association studies (GWAS), we are now beginning to uncover the genetic architecture of many human traits 12. GWAS scan the genome for association with a trait by regressing phenotypic values on genotypes, typically one locus at a time 13. A major finding is that most traits of interest are highly polygenic, influenced by hundreds to thousands of loci distributed across the genome 14, 15, 16. For example, Yengo et al. identified 1,185 SNPs (single nucleotide polymorphisms) associated with human height 17; Boyle et al. estimated that there are >100,000 SNPs in the genome with causal effect on height 14.

The first use of GWAS findings in order to test for polygenic selection in humans was reported for height in Europe, where Turchin et al. showed that height increasing alleles have systematically higher frequencies in Northern Europeans compared to Southern Europeans that is not explained by genetic drift alone 18. Signature of polygenic selection for height was also detected with more sophisticated approaches, notably by Berg and Coop, who formulated a null model for the dispersion of poly-
genic scores (PGS, which for a given trait represent the genetic predisposition by aggregating the effect of many trait-associated loci into one score, see below) across populations under genetic drift [19], and by Field et al., who developed a method based on a signature of genetic diversity around trait-associated SNPs that is sensitive to more recent selection events (during the past 2,000-3,000 years) [20]. Moreover, signals were detected for other traits, such as BMI, age at menarche, and infant head circumference, suggesting that polygenic selection may have been the abundant mode of adaptation in humans [20, 19, 21, 22]. The validity of these signals, however, have been recently revisited (see below).

Studies of past adaptation using present-day populations suffer from two major limitations: (i) they are by design indirect and based on specific models of adaptation (e.g., selective sweeps) and therefore may provide a limited and biased view of human adaptation. (ii) They do not provide information about the component of evolutionary fitness that loci underlying adaptation affected; even for well-studied examples, such as LCT mutations associated with lactose tolerances, it remains unclear whether they were selected because they provided an advantage for survival, fecundity, or inclusive fitness (the ability of individuals to impact the fitness of their kin) [23]. With the advent of large genomic data sets, however, it is now becoming feasible to study natural selection directly in present-day samples, with the further ability to disentangle the effects on fitness components (by collecting information on number of children, lifespan, number of grandchildren, etc.).

Remarkably, the first known target of selection in humans, the sickle cell allele, was identified by Allison through direct observation of its effect on survival in malaria-
endemic regions [24]. The sickle cell allele is an extreme example however, and most effects on fitness are likely much smaller, notably for variants underlying polygenic selection. Directly observing small evolutionary change within a few generations thus requires huge sample sizes which were not available until the last few years.

In the work presented in Chapter 1, we focused on the viability component of fitness in contemporary samples that are directly observable [25]. We used two large data sets, the GERA cohort in California [26, 27] and the UK Biobank [28], to look for variants that affect age-specific mortality. We tested whether the frequency of genetic variants changed with age of the individuals in these data sets (or by proxy with the age at death of their parents) more than expected by chance, after controlling for ancestry differences across age cohorts. We found two variants across the genome: the $\epsilon 4$ allele of the APOE gene, a major risk factor for Alzheimer’s disease, and a variant in the CHRNA3 locus previously shown to be associated with smoking behavior. To observe only two common variants with large effect on survival even when they exert their effect late in life (despite having enough power to detect such variants) was surprising to us, considering that prevalent theories of aging predict an accumulation of harmful genetic effects post reproductive age. Our findings suggest that either (i) late-onset variants have been kept at low frequency by selection, and so are not neutrally evolving as typically assumed, or (ii) the mutational target size for variants with only late-acting effects is tiny. If (i) is true, it raises the question of why APOE and CHRNA3 variants are common. We hypothesize that they either became harmful recently (e.g., for CHRNA3 variants as a result of increased cigarette smoking prevalence), or they provided a fecundity or early-onset survival advantage.
in ancestral environments.

Considering the polygenic nature of most human traits, we extended our approach to test for the collective effect of sets of genetic variants previously identified in GWAS. To this end, we constructed PGS for 42 traits and investigated whether these scores vary by age. We found associations for a number of traits, e.g., for coronary artery disease (CAD), BMI, cholesterol levels. Intriguingly, we found a negative association between longer lifespan and variants delaying the age at puberty and age at first child birth, possibly reflecting a trade-off between longevity and fecundity. These findings underscore the importance of studying fitness components separately (e.g., this work), complementing the approaches that aim to infer the net effect of genetic variants (e.g., by looking at lifetime reproductive success, e.g., \[29, 30, 31\]). We also identified age effects that were different in males and females, notably for variants associated with CAD and cholesterol levels. These results further suggest that human viability is extremely polygenic and potentially more variants will be found with increasing sample sizes; for instance, a recent GWAS of parental lifespan with about 1 million samples identified 12 genome-wide significant loci (compared to two in our study) \[32\].

Our results presented in Chapter 1 serve as a proof-of-concept that genetic data sets are becoming large enough such that we can begin to study the ongoing natural selection in humans. Such large samples present unique challenges, however. For instance, two recent studies, Berg et al. (on which I am a coauthor, see Appendix) \[33\] and Sohail et al. \[34\], discovered that the signals of polygenic selection previously detected for height have been inflated by population structure confounding in the
GWAS (a spurious correlation between a genotype and a trait caused if the genotype covaries with causal loci or with environmental factors contributing to the trait). Previous polygenic selection analysis on height relied on the GWAS performed by the GIANT consortium \[35, 36\]. However, polygenic adaptation signals were diminished or largely attenuated when based on GWAS results from the UK Biobank, which included individuals of relatively homogenous ancestry compared to GIANT which was a meta-analysis of independent GWAS with samples from across Europe. Further analyses pointed to biases in GIANT effect size estimates along the geographical axis where signals of polygenic selection were previously detected. Without access to individual-level data, it is difficult to assess the source of the problem in GIANT. One possibility is that current state-of-the-art methods to deal with population structure are not sufficient, especially when meta-analyzing results across different sources.

These findings further raise the more general question of the reliability of GWAS results. Of course the answer depends on the application. Systematic biases in GWAS would likely have major consequences for polygenic analyses, e.g., polygenic adaptation as discussed above or any application that uses PGS \[37, 38\], which aggregate subtle biases across many loci. In their simplest form, PGS are calculated as the number of trait-increasing alleles carried by an individual weighted by their effect size estimated in GWAS. In addition to evolutionary analyses, PGS are increasingly being used in fields such as human genetics and social sciences. In human genetics the goal is to identify individuals at risk of disease in the tail of the disease liability PGS distribution (e.g., \[39, 40, 41, 42\]). In social sciences, PGS are used to control for genetic variation in order to gain resolution on the effect of an environmental
factor or policy (e.g., schooling reform) on an outcome (e.g., BMI) \[43, 44, 45\].

One major obstacle to the broad use of PGS is that PGS lose their predictive ability (typically measured by how much of the trait variation they explain in a sample not included in the GWAS) in samples that have different ancestry than the GWAS sample \[46, 47, 48\]. In other words, PGS are not “portable” across ancestry groups. For example, Martin et al. reported that averaged across a number of traits in the UK Biobank, the prediction accuracy of PGS based on European ancestry individuals (reflecting the state of the field that available GWAS are predominantly performed in European ancestry individuals) drops by a factor of \(~2\) in samples with East Asian ancestry and by a factor of \(~5\) in samples with African ancestry (relative to European ancestry samples) \[47\].

Recent discussion about this challenge has focused primarily on the impact of differences in linkage disequilibrium patterns and allele frequencies across human populations that arose from their distinct demographic and recombination histories \[46, 49, 50, 51, 52, 53\]. However, factors other than ancestry could in principle affect PGS portability, e.g., indirect genetic effects ("genetic nurture"), assortative mating, population structure confounding, gene-by-environment interaction (GxE) \[54, 55, 56, 57, 58, 59, 60\]. I discuss these factors in more detail in Chapter 2, and present results illustrating that these considerations are not hypothetical \[61\].

To this end, we focused on a subset of the UK Biobank with relatively homogenous ancestry \[28\]. With three example traits, diastolic blood pressure, BMI and years of schooling, we show that within this homogenous group, the portability of PGS depends on the characteristics of the GWAS and prediction sets, such as sex, age and
socioeconomic status (SES). Such characteristics inevitably vary across data sets. We compare PGS based on standard GWAS and PGS based on sibling-GWAS, i.e., by detecting association by regressing phenotypic differences between siblings on their genetic differences. In so doing, we demonstrate that processes such as parenting behavior and assortative mating may have a large contribution to the predictive power of PGS based on standard GWAS, notably for behavioral and social traits such as years of schooling, smoking behavior and age at first sexual intercourse. As these processes are socially and culturally mediated, it is likely that they vary between as well as within ancestry groups. Taken together these results show that (i) PGS prediction accuracy can vary within samples of similar genetic ancestry, and (ii) the variable performance of PGS across ancestries cannot be predominantly attributed to population genetic parameters.

In summary, the work presented in this thesis introduces two new types of studies that are now possible with the advent of large genetic data sets, e.g., study of natural selection in contemporary humans, and making phenotypic prediction using genetic data, but also illustrates the new challenges involved with such opportunities, notably with regard to population structure confounding and the variable prediction accuracy of PGS.
Chapter 1

Identifying genetic variants that affect viability in large cohorts


1.1 Abstract

A number of open questions in human evolutionary genetics would become tractable if we were able to directly measure evolutionary fitness. As a step towards this goal, we developed a method to examine whether individual genetic variants, or sets of genetic variants, currently influence viability. The approach consists in testing whether the frequency of an allele varies across ages, accounting for variation in ancestry. We applied it to the Genetic Epidemiology Research on Adult Health and
Aging (GERA) cohort and to the parents of participants in the UK Biobank. Across the genome, we found only a few common variants with large effects on age-specific mortality: tagging the APOE $\epsilon 4$ allele and near CHRNA3. These results suggest that when large, even late-onset effects are kept at low frequency by purifying selection. Testing viability effects of sets of genetic variants that jointly influence 1 of 42 traits, we detected a number of strong signals. In participants of the UK Biobank of British ancestry, we found that variants that delay puberty timing are associated with a longer parental life span ($p \sim 6.2 \cdot 10^{-6}$ for fathers and $p \sim 2 \cdot 10^{-3}$ for mothers), consistent with epidemiological studies. Similarly, variants associated with later age at first birth are associated with a longer maternal life span ($p \sim 1.4 \cdot 10^{-3}$). Signals are also observed for variants influencing cholesterol levels, risk of coronary artery disease (CAD), body mass index, as well as risk of asthma. These signals exhibit consistent effects in the GERA cohort and among participants of the UK Biobank of non-British ancestry. We also found marked differences between males and females, most notably at the CHRNA3 locus, and variants associated with risk of CAD and cholesterol levels. Beyond our findings, the analysis serves as a proof of principle for how upcoming biomedical data sets can be used to learn about selection effects in contemporary humans.

1.2 Introduction

A number of central questions in evolutionary genetics remain open, in particular for humans. Which types of variants affect fitness? Which components of fitness do
they affect? What is the relative importance of directional and balancing selection in shaping genetic variation? Part of the difficulty is that our understanding of selection pressures acting on the human genome is based either on experiments in fairly distantly related species or cell lines or on indirect statistical inferences from patterns of genetic variation \[4, 62, 63\]. The statistical inferences rely on patterns of genetic variation in present-day samples (or, very recently, in ancient samples \[64\]) to identify regions of the genome that appear to carry the footprint of positive selection \[62\]. For example, a commonly used class of methods asks whether rates of nonsynonymous substitutions between humans and other species are higher than expected from putatively neutral sites in order to detect recurrent changes to the same protein \[65\]. Another class instead relies on polymorphism data and looks for various footprints of adaptation involving single changes of large effect \[66\]. These approaches detect adaptation over different timescales and, likely as a result, suggest quite distinct pictures of human adaptation \[4\]. For example, approaches that are sensitive to selective pressures acting over millions of years have identified individual chemosensory and immune-related genes (e.g., \[67\]). In contrast, approaches that are most sensitive to selective pressures active over thousands or tens of thousands of years have revealed strong selective pressures on individual genes that influence human pigmentation (e.g., \[68, 69, 70\]), diet \[3, 71, 72\], as well as sets of variants that shape height \[18, 21, 19\]. Even more recent still, studies of contemporary populations have suggested that natural selection has influenced life-history traits like age at first childbirth as well as educational attainment over the course of the last century \[73, 74, 75, 76, 30, 29, 31\].
Because these approaches are designed (either explicitly or implicitly) to be sensitive to a particular mode of adaptation, they provide a partial and potentially biased picture of what variants in the genome are under selection. In particular, most have much higher power to adaptations that involve strongly beneficial alleles that were rare in the population when first favored and will tend to miss selection on standing variation or adaptation involving many loci with small beneficial effects (e.g., \cite{77, 78, 79, 10}). Moreover, even where these methods identify a beneficial allele, they are not informative about the components of fitness that are affected or about possible fitness trade-offs between sexes or across ages. In line with Lewontin’s proposal to track age-specific mortality and fertility of hundreds of thousands of individuals \cite{80}, we sought a more direct and, in principle, comprehensive way to study adaptation in humans, focusing on current viability selection. Similar to the approach that Allison took in comparing frequencies of the sickle cell allele in newborns and adults living in malarial environments \cite{24}, we aimed to directly observe the effects of genotypes on survival by taking advantage of the recent availability of genotypes from large cohorts of individuals of different ages. Specifically, we tested for differences in the frequency of an allele across individuals of different ages, controlling for changes in ancestry and possible batch effects. This approach resembles a genome-wide association study (GWAS) for longevity yet does not focus on an end point (e.g., survival to an old age) but on any shift in allele frequencies with age. Thus, it allows the identification of possible nonmonotonic effects at different ages or sex differences. Any genetic variant that affects survival by definition has a fitness cost, even if the cost is too small to be effectively selected against (depending on the effective population size, the age
structure of the population, and the age at which the variant exerts its effects. Of course, a genetic variant can influence fitness without influencing survival through effects on reproduction or inclusive fitness. Our approach is therefore considering only one of the components of fitness that are likely important for human adaptation.

As a proof of principle, we applied our approach to two recent data sets: to 57,696 individuals of European ancestry from the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort and, by proxy, to the parents of 117,648 individuals of British ancestry surveyed as part of the UK Biobank interim data. We did so for individual genetic variants then jointly for sets of variants previously found to influence 1 of 42 polygenic traits.

### 1.3 Results

#### 1.3.1 A method for testing for differences in allele frequencies across age bins

If a genetic variant does not influence viability, its frequency should be the same in individuals of all ages. We therefore tested for changes in allele frequency across individuals of different ages, while accounting for systematic differences in the ancestry of individuals of different ages (for example, due to migration patterns over decades) and genotyping batch effects. We used a logistic regression model in which we regressed each individual’s genotype on their age bin, their ancestry as determined by principal component analysis (PCA) (Figure 1.6), and the batch in which they were
genotyped (see Materials and Methods for details). In this model, we treated age bin as a categorical variable; this approach allowed us to test for a relationship between age and the frequency of an allele regardless of the functional form of this relationship. We also tested a model with an interaction between age and sex to assess whether a variant affects survival differently in the two sexes.

We first evaluated the power of this method using simulations. We considered 3 possible trends in allele frequency with age: (i) a constant frequency up to a given age followed by a steady decrease, i.e., a variant that affects survival after a given age (e.g., variants contributing to late-onset disorders), (ii) a steady decrease across all ages for a variant with detrimental effect throughout life, and (iii) a U-shaped pattern in which the allele frequency decreases to a given age but then increases, reflecting trade-offs in the effects at young and old ages, as hypothesized by the antagonistic pleiotropy theory of aging [90] or as may be seen if there are protective alleles that buffer the effect of risk alleles late in life [91] (Figure 1.1). In all simulations, we used sample sizes and age distributions that matched the GERA cohort (Figure 1.7). For simplicity, we also assumed no population structure or batch effects across age bins (see Materials and Methods). For all trends, we set a maximum of 20% change in the allele frequency from the value in the first age bin (Figure 1.1).
Figure 1.1: Power of the model to detect changes in allele frequency with age. (A) Trends in allele frequency with age considered in simulations. The y-axis indicates the allele frequency standardized to the frequency in the first age bin. (B) Power to detect the trends in (A) at $p < 5 \cdot 10^{-8}$, given the sample size per age bin in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort (Figure 1.7) and total sample size of 57,696. Shown are results using models with age treated as a categorical (red) or an ordinal (black) variable, assuming no change in population structure and batch effects across age bins. The curves show simulation results sweeping allele frequency values with an increment value of 0.001 (1,000 simulations for each allele frequency) smoothed using a Savitzky-Golay filter in the SciPy package.

Because of the age distribution of individuals in the GERA cohort (Figure 1.7), our power to detect the trend is greater when most of the change in allele frequency
occurs in middle age (Figure 1.1). For example, for an allele with an initial allele frequency of 0.15 that begins to decrease in frequency among individuals at age 20, age 50, or age 70 years, there is around 20%, 90%, and 60% power, respectively, to detect the trend at $p < 5 \cdot 10^{-8}$, the commonly used criterion for genome-wide significance [92]. We also experimented with a version of the model in which the age bin is treated as an ordinal variable; as expected, this model is more powerful if there is in fact a linear relationship between age and allele frequency. Because we do not know the functional form of the relationship between age and allele frequency a priori in most cases, we used the categorical model for all analyses unless otherwise noted. In the UK Biobank, all individuals were 45–69 years old at enrollment, so the age range of the participants is restricted and our method has low power. However, the UK Biobank participants reported the survival status of their parents: age of the parents if alive or age at which their parents died; following recent studies [82, 83, 84], we therefore used these values (when reported) instead in our model. In this situation, we are testing for correlations between an allele frequency and father’s or mother’s age (if alive) or age at death (if deceased). This approach obviously comes with the caveat that children inherit only half of their genome from each parent and so power is reduced (e.g., [93]). Furthermore, the patterns expected when considering individuals who have died differ subtly from those generated among surviving individuals. Notably, when an allele begins to decline in frequency starting at a given age (Figure 1.1A), there should be an increase in the allele frequency among individuals who died at that age followed by a decline in frequency, rather than the steady decrease expected among surviving individuals (Figure 1.8, see Materials
and Methods for details). In a first analysis, we therefore focused on the majority of participants who reported father’s or mother’s age at death, 88,595 and 71,783 individuals, respectively. We compared the results of this approach with the results of a Cox proportional hazards model \[ 94 \], which allowed us to include individuals who reported their parents to be alive but has the disadvantage of assuming fixed effects across all ages.

We further adapted this model to allow us to test for changes in frequency at sets of genetic variants jointly. Many phenotypes of interest, from complex disease risk to anthropomorphic and life-history traits such as age at menarche, are polygenic \[ 95, 96 \]. If a polygenic trait has an effect on fitness, either directly or indirectly (i.e., through pleiotropic effects), the individual loci that influence the trait may be too subtle in their survival effects to be detectable with current sample sizes. We therefore investigated whether there is a shift across ages in sets of genetic variants that were identified as influencing a trait in GWASs (Table \[1.2\]). Specifically, for a given trait, we calculated a polygenic score for each individual based on trait effect sizes of single variants previously estimated in GWASs and then tested whether the scores vary significantly across 5-year age bins (see Materials and Methods for details). These scores are calculated under an additive model, which appears to provide a good fit to GWAS data \[ 97 \].

If a polygenic trait is under stabilizing selection (e.g., human birth weight \[ 98 \]), i.e., an intermediate polygenic score is optimal, no change in the mean value of polygenic scores across different ages is expected. However, if extreme values of a trait are associated with lower chance of survival, the spread of the polygenic scores
should decrease with age. To consider this possibility, we tested whether the squared difference of the polygenic scores from the mean varies significantly with age (see Materials and Methods for details).

1.3.2 Testing for changes in allele frequency at individual genetic variants

We first applied the method to the GERA cohort using 8,868,517 filtered genotyped and imputed autosomal biallelic single-nucleotide polymorphisms (SNPs) and indels. We focused on a subset of 57,696 filtered individuals who we confirmed to be of European ancestry by PCA (see Materials and Methods, Figure 1.9 and Figure 1.10). The ages of these individuals were reported in bins of 5-year intervals (distribution shown in Figure 1.7). We tested for significant changes in allele frequencies across these bins. For each variant, we obtained a p-value comparing a model in which the allele frequency changes with age to a null model. No inflation was detected in the quantile-quantile plot (Figure 1.11A), indicating that, for common variants at least, our control for population structure (and other potential confounders) is sufficient. To illustrate this point, we looked at the lactose intolerance-linked SNP rs4988235 within the LCT locus, which is among the most differentiated variants across European populations [5]; the trend in the expected allele frequency based on the null model (i.e., accounting for confounding batch effects and changes in ancestry) tracks the observed trend quite well (Figure 1.12).

By our approach, all variants that reached genome-wide significance ($p < 5 \cdot 10^{-8}$)
reside on chromosome 19 near the APOE gene (Figure 1.2A and Figure 1.13). This locus has previously been associated with longevity in multiple studies \cite{99, 100}. The APOE \( \epsilon 4 \) allele is known to increase the risk of late-onset Alzheimer disease (AD) as well as of cardiovascular diseases \cite{101, 102}. We observe a monotonic decrease in the frequency of the \( T \) allele of the \( \epsilon 4 \) tag SNP rs6857 (C, protective allele; T, risk allele) beyond the age of 70 years old (Figure 1.2B). This trend is observed for both the heterozygous and homozygous risk variants (Figure 1.14) and for both males and females (Figure 1.15). No variant reached genome-wide significance testing for age by sex interactions (quantile-quantile plot shown in Figure 1.11B).

We further investigated the trends in frequency with age for the other 2 major APOE alleles defined by rs7412 and rs429358 SNPs: \( \epsilon 2 \) (rs7412-T, rs429358-T) and \( \epsilon 4 \) (rs7412-C, rs429358-T), while \( \epsilon 4 \) is known by rs7412-C and rs429358-C alleles. Unlike the \( \epsilon 4 \) allele, \( \epsilon 2 \) allele carriers are suggested to be at lower risk of AD, cardiovascular disease, and mortality relative to the \( \epsilon 3 \) carriers \cite{99, 103}. We focused on a subset of 38,703 individuals with unambiguous counts of each APOE allele. There is a significant change in the frequency of the \( \epsilon 4 \) allele with age in this subset (\( p \sim 6 \cdot 10^{-12} \)), similar to the trend observed for the tag SNP rs6857 (Figure 1.16). The \( \epsilon 3 \) allele shows the reverse trend, with a significant, monotonic increase in frequency beyond the age of 70 years old (\( p \sim 1.7 \cdot 10^{-8} \)) (Figure 1.16). The enrichment of the \( \epsilon 3 \) allele in elderly individuals can be explained by the corresponding depletion of the \( \epsilon 4 \) allele and does not necessarily imply an independent, protective effect of the \( \epsilon 3 \) allele. The frequency of the \( \epsilon 2 \) allele does not change significantly with age (\( p \sim 0.21 \)), possibly
reflecting low power given its allele frequency of approximately 0.06 (Figure 1.16).

Figure 1.2: Testing for the influence of single genetic variants on age-specific mortality in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. (A) Manhattan plot of p-values for the change in allele frequency with age. The red line marks the $p = 5 \cdot 10^{-8}$ threshold. (B) Allele frequency trajectory of rs6857, a tag SNP for the APOE $\epsilon 4$ allele, with age. Data points are the frequencies of the risk allele within 5-year interval age bins ($\pm 2$ SE), with the center of the bin indicated on the x-axis (except for the first and the last points). Bins with ages below 38 years are merged into 1 bin because of the relatively small sample sizes. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry (see Materials and Methods). In orange are the mean ages at onset of Alzheimer disease for carriers of 0, 1, or 2 copies of the APOE $\epsilon 4$ allele [101].

We considered the possibility that some unobserved confounding variable was driving the strength of this signal at APOE. Since there are 2 genotyped SNPs with signals similar to rs6857 within the locus, genotyping error seems unlikely to be driving the pattern (Figure 1.13). Another concern might be a form of ascertainment bias, in which individuals with AD are underrepresented in the Kaiser Permanente Medical Care Plan. However, there is no correlation in these data between the amount of time that an individual has been enrolled in this plan and the individual’s APOE genotype (Figure 1.17). These observations, along with previously reported associations at
this locus, argue that the allele frequency trends in Figure 1.2B are driven by effects of APOE genotype on mortality (or severe disability). Moreover, the effects that we identified are concordant with epidemiological data on the mean age at onset of AD, given 0 to 2 copies of APOE ε4 allele [101]. This case not only serves as a positive control for our approach, it illustrates the resolution that it provides about age effects of genetic variants.

We estimated that we have about 93% power to detect the trend in allele frequency with age as observed for rs6857 (at a genome-wide significance level, see Materials and Methods). Using both versions of the model treating age bin as a categorical or an ordinal variable, we have similar power to detect other potential trends considered in Figure 1.1 for variants as common as rs6857 and with similar magnitude of effect on survival. Yet across the genome, only APOE variants show a significant change in allele frequency with age for both versions of the model (Figure 1.2 and Figure 1.18). Thus, our finding only APOE ε4 allele indicates that there are few or no other common variants in the genome with an effect on survival as strong as is seen in the APOE region.

We then turned to the UK Biobank data set. We applied our method to individuals of British ancestry whose data passed our filters; of these, 88,595 had death information available for their father and 71,783 for their mother. We analyzed 590,437 genotyped autosomal variants, applying similar quality control (QC) measures as with the GERA data set (see Materials and Methods). We tested for significant changes in allele frequencies with father’s age at death and mother’s age at death stratified in eight 5-year interval bins. As in the GERA data set, no inflation was detected in the
quantile-quantile plots (Figure 1.19).

Consistent with recent studies [82, 83], the variants showing a genome-wide significant change in allele frequency with father’s age at death ($p < 5 \cdot 10^{-8}$) reside within a locus containing the nicotine receptor gene CHRNA3 (Figure 1.3A). The A allele of the CHRNA3 SNP rs1051730 (G, major allele; A, minor allele) has been shown to be associated with increased smoking quantity among individuals who smoke [104]. We observe a linear decrease in the frequency of the A allele of rs1051730 throughout almost all age ranges (Figure 1.3B) $p \sim 1.3 \cdot 10^{-7}$ and $p \sim 2.7 \cdot 10^{-10}$, treating paternal age at death as a categorical or an ordinal variable, respectively). Although it does not reach genome-wide significance, this allele shows a similar trend with age in GERA ($p \sim 8.6 \cdot 10^{-3}$, Figure 1.20). We note that 30,819 of the UK Biobank individuals included in the above analysis were genotyped on the UK BiLEVE Axiom array (see Materials and Methods), selected based on lung function and smoking behavior (while the remaining 57,776 samples were genotyped on the UK Biobank Axiom array) [28]. Expectedly, the frequency of the A allele is significantly higher among UK BiLEVE subjects ($p \sim 2.3 \cdot 10^{-10}$), but the age effects are similar across both arrays ($p \sim 0.72$).

For mother’s age at death, a SNP in a locus containing the MEOX2 gene reached genome-wide significance (Figure 1.3C). The C allele of rs4721453 (T, major allele; C, minor allele) increases in frequency in the age bin centered at 76 years old (Figure 1.21), i.e., there is an enrichment among individuals who died at 74 to 78 years of age, which corresponds to a deleterious effect of the C allele in this period. The trend
Figure 1.3: Testing for the influence of single genetic variants on age-specific mortality in the UK Biobank. (A) Manhattan plot of p-values, obtained from testing for a change in allele frequency with age at death of fathers. (B) Allele frequency trajectory of rs1051730, within CHRNA3 locus, with father’s age at death. (C) Manhattan plot of p-values, obtained from testing for a change in allele frequency with age at death of mothers. (D) Allele frequency trajectory of rs769449, within the APOE locus, with mother’s age at death. Red lines in (A) and (C) mark the $p = 5 \cdot 10^{-8}$ threshold. Data points in (B) and (D) are the frequencies of the risk allele within 5-year interval age bins ($\pm 2$ SE), with the center of the bin indicated on the x-axis (except for the first and the last points). The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry (see Materials and Methods).

is similar and nominally significant for other genotyped common SNPs in moderate linkage disequilibrium with rs4721453 (Figure 1.21). Also, the signal for rs4721453 remains nominally significant when using subsets of individuals genotyped on the
same genotyping array: 44,552 individuals on the UK Biobank Axiom array \((p \sim 6.6 \cdot 10^{-5})\) and 25,231 individuals on the UK BiLEVE Axiom array \((p \sim 1.1 \cdot 10^{-4})\). These observations suggest that the result is not due to genotyping errors, but it is not reproduced in GERA \((p \sim 0.023, \text{Figure 1.22})\) and so it remains to be replicated.

Variants within the APOE locus are among the top nominally significant variants (Figure 1.3C). At the APOE SNP rs769449 (G, major allele; A, minor allele), there is an increase in the frequency of A allele at around 70 years old before subsequent decrease (Figure 1.3D, \(p \sim 1.2 \cdot 10^{-7}\)). This pattern is consistent with our finding in GERA (of a monotonic decrease beyond 70 years of age), considering the difference in patterns expected between allele frequency trends with age among survivors versus individuals who died (Figure 1.8).

We note that by considering parental age at death of the UK Biobank participants—as done also in \[82, 83, 84\]—we introduce a bias towards older participants, who are more likely to have deceased parents (Figure 1.23). We confirmed that our top signals are not significantly affected after adjusting for age of the participants (among other potential confounders, including participants’ sex, year of birth, and socioeconomic status, as measured by the Townsend deprivation index): results remain similar for the SNP rs4721453 near MEOX2 \((p \sim 1.2 \cdot 10^{-9})\), APOE SNP rs769449 \((p \sim 1.5 \cdot 10^{-6})\), and CHRNA3 SNP rs1051730 \((p \sim 1.8 \cdot 10^{-6} \text{ and } p \sim 4.3 \cdot 10^{-9})\), treating paternal age at death as a categorical or an ordinal variable, respectively.

We further tested for trends in allele frequency with parental age at death that differ between fathers and mothers, focusing on 62,719 individuals with age at death information for both parents. No variant reached genome-wide significance level (Figure
1.24 A). The rs4721453 near the MEOX2 gene and APOE variant rs769449 show nominally significant sex effects (\( p \approx 7.2 \cdot 10^{-8} \) and \( p \approx 2.2 \cdot 10^{-3} \), respectively), with stronger effects in females (Figure 1.24B). Variants near the CHRNA3 locus are nominally significant when using the model with parental ages at death treated as ordinal variables (rs11858836, \( p \approx 5.7 \cdot 10^{-4} \)), with stronger effects in males (Figure 1.24B).

1.3.3 Testing for changes in allele frequency at trait-associated variants

We next turned to sets of genetic variants that have been associated with polygenic traits rather than individual genetic variants. We focused on 42 polygenic traits, including disease risk and traits of evolutionary importance such as puberty timing, for which a large number of common variants have been mapped in GWASs (see Table 1.2 for the list of traits and number of loci [86, 87, 88, 89]). For each individual and each trait, we calculated a polygenic score based on the genetic variants that reached genome-wide significance level for association and then tested whether this polygenic score, or its squared difference from the mean in the case of stabilizing selection, is associated with survival (after controlling for covariates, see Materials and Methods).

We first applied the Cox proportional hazards model in the UK Biobank for parental survival, focusing on the participants whose genetic ancestry is British and who reported their father’s or mother’s age or age at death (114,122 and 116,323 individuals,
respectively). We then compared the results with our approach of testing for changes in the polygenic score across parental ages at death. We further analyzed 2 data sets for replication purposes: participants of the UK Biobank of non-British ancestry (29,511 and 30,372 individuals reporting father’s or mother’s age information, respectively) and the GERA cohort.

Using the Cox model, the scores for several traits show significant associations with father’s survival after accounting for multiple testing (Figure 1.4A, Table 1.1): total cholesterol (TC, \( p \sim 4.3 \times 10^{-11} \)), low-density lipoproteins (LDL, \( p \sim 8.1 \times 10^{-9} \)), body mass index (BMI, \( p \sim 1.8 \times 10^{-8} \)), and coronary artery disease (CAD, \( p \sim 9 \times 10^{-6} \)), consistent with 2 recent studies [83, 84]. In addition, we uncovered significant association between the polygenic score for puberty timing (\( p \sim 6.2 \times 10^{-6} \)); in this analysis, we used age at menarche-associated variants in females, motivated by the high genetic correlation between the timing of puberty in males and females [105]). A higher score for puberty timing is associated with longer paternal survival (per year hazard ratio of 0.96) (Table 1.1), indicating that variants delaying puberty timing are associated with a higher chance of survival, consistent with epidemiological studies suggesting early puberty timing to be associated with adverse health outcomes [106].

For all other traits, a higher score is negatively associated with paternal survival: 1 standard deviation (SD) hazard ratio of 1.09 for TC, 1.08 for LDL, and 1.22 for BMI (Table 1.1). With the exception of lipid traits, the effects on survival are not significantly changed after accounting for the effect of the polygenic score of another trait (Figure 1.25). This is especially relevant to BMI and puberty timing, for which there is substantial genetic overlap [57]; the per year hazard ratio is 0.97 for
the puberty timing score \((p \sim 4.8 \cdot 10^{-4})\) after adjusting for the BMI score.

Using our approach instead, i.e., considering the father’s age at death, led to very similar results. Specifically, all traits significantly associated with paternal survival show a significant change in polygenic score with father’s age at death using the model with parental ages at death treated as ordinal variables (Figure 1.26): TC \((p \sim 8.8 \cdot 10^{-9})\), CAD \((p \sim 3.3 \cdot 10^{-8})\), puberty timing \((p \sim 1.6 \cdot 10^{-7})\), LDL \((p \sim 8.6 \cdot 10^{-7})\), and BMI \((p \sim 3.4 \cdot 10^{-6})\). In addition, we uncovered significant changes in polygenic score with father’s age at death for asthma (ATH, \(p \sim 9.4 \cdot 10^{-5}\)) and triglycerides (TG, \(p \sim 4.4 \cdot 10^{-4}\), the effect of which does not seem to be distinct from other lipid traits, Figure 1.25). The score for puberty timing increases monotonically with the father’s age at death (Figure 1.4B), indicative of a protective effect of later predicted puberty timing, whereas all other traits with significant signal show a monotonic decline in score with age (Figures 1.4C-F).

In a Cox survival model, for mothers as with for fathers, scores for TC, CAD, and LDL are significantly associated with survival, with similar hazard ratios (Figure 1.5A and Table 1.1): 1 SD hazard ratio of 1.09 for LDL \((p \sim 5.2 \cdot 10^{-8})\), 1.09 for CAD \((p \sim 5.2 \cdot 10^{-6})\), and 1.07 for TC \((p \sim 7.8 \cdot 10^{-6})\). In addition, the high-density lipoproteins (HDL) score is associated with maternal survival (1 SD hazard ratio of 0.94, \(p \sim 8.9 \cdot 10^{-5}\)). Also, suggestive evidence was detected for protective effects of increased predicted age at first birth (AFB) (per year hazard ratio of 0.94, \(p \sim 1.3 \cdot 10^{-3}\)) as well as predicted puberty timing (per year hazard ratio of 0.97, \(p \sim 2 \cdot 10^{-3}\)) (Figure 1.5A and Table 1.1). Other than the LDL and TC, all signals seem to be distinct (Figure 1.25), including for puberty timing and AFB, despite
the genetic correlation between the 2 phenotypes [88].
Figure 1.4: Testing for the influence of sets of trait-associated variants on survival of the fathers of UK Biobank participants. (A) Quantile-quantile plot for association between the polygenic score of 42 traits (see Table 1.2) with father’s survival, using the Cox model. The red line indicates the distribution of the p-values under the null model. Signs “+” and “−” indicate protective and detrimental effects associated with higher values of polygenic scores, respectively. See Table 1.3 for values and hazard ratios for all traits. (B-F) Trajectory of polygenic score with age at death of fathers for top traits associated with paternal survival (only independent signals are shown, see Figure 1.25): puberty timing (using age at menarche-associated variants) in males (B), total cholesterol (TC) (C), coronary artery disease (CAD) (D), body mass index (BMI) (E), and asthma (ATH) (F). Data points in (B-F) are mean polygenic scores within 5-year interval age bins (± 2 SE), with the center of the bin indicated on the x-axis (except for the first and the last points). The dashed line shows the expected score based on the null model, accounting for confounding batch effects, changes in ancestry, and participant’s age, sex, year of birth, and the Townsend index (a measure of socioeconomic status).

Table 1.1: Associations between sets of trait-associated variants and paternal and maternal survival among the UK Biobank participants of British ancestry, under the Cox model.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Scaling effect</th>
<th>Father</th>
<th>Mother</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Effect size (SE)</td>
<td>HR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puberty timing</td>
<td>1 year</td>
<td>-0.0363 (0.008)</td>
<td>0.96</td>
</tr>
<tr>
<td>AFB</td>
<td>1 year</td>
<td>-0.0398 (0.018)</td>
<td>0.96</td>
</tr>
<tr>
<td>ATH</td>
<td>1 unit log-odds</td>
<td>0.0279 (0.0109)</td>
<td>1.03</td>
</tr>
<tr>
<td>BMI</td>
<td>1 SD</td>
<td>0.1996 (0.0355)</td>
<td>1.22</td>
</tr>
<tr>
<td>CAD</td>
<td>1 unit log-odds</td>
<td>0.0784 (0.0177)</td>
<td>1.08</td>
</tr>
<tr>
<td>HDL</td>
<td>1 SD</td>
<td>-0.0340 (0.0139)</td>
<td>0.97</td>
</tr>
<tr>
<td>LDL</td>
<td>1 SD</td>
<td>0.0806 (0.014)</td>
<td>1.08</td>
</tr>
<tr>
<td>TC</td>
<td>1 SD</td>
<td>0.0901 (0.0137)</td>
<td>1.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait</th>
<th>Scaling effect</th>
<th>Father</th>
<th>Mother</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Effect size (SE)</td>
<td>HR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puberty timing</td>
<td>1 year</td>
<td>-0.0278 (0.009)</td>
<td>0.97</td>
</tr>
<tr>
<td>AFB</td>
<td>1 year</td>
<td>-0.0639 (0.012)</td>
<td>0.94</td>
</tr>
<tr>
<td>ATH</td>
<td>1 unit log-odds</td>
<td>0.0149 (0.0121)</td>
<td>1.02</td>
</tr>
<tr>
<td>BMI</td>
<td>1 SD</td>
<td>0.0823 (0.0055)</td>
<td>1.08</td>
</tr>
<tr>
<td>CAD</td>
<td>1 unit log-odds</td>
<td>0.0892 (0.0196)</td>
<td>1.09</td>
</tr>
<tr>
<td>HDL</td>
<td>1 SD</td>
<td>-0.0605 (0.0154)</td>
<td>0.94</td>
</tr>
<tr>
<td>LDL</td>
<td>1 SD</td>
<td>0.0844 (0.0153)</td>
<td>1.09</td>
</tr>
<tr>
<td>TC</td>
<td>1 SD</td>
<td>0.0679 (0.0152)</td>
<td>1.07</td>
</tr>
</tbody>
</table>

**Abbreviations:** AFB, age at first birth; ATH, asthma; BMI, body mass index; CAD, coronary artery index; HR, hazard ratio; HDL, high-density lipoproteins; TC, total cholesterol

Applying our approach to maternal age at death instead, puberty timing and AFB are the top signals \( p \sim 2.2 \cdot 10⁻⁴ \) and \( p \sim 3.1 \cdot 10⁻³ \), respectively, Figure 1.26. Higher polygenic scores for puberty timing are enriched among longer-lived mothers (Figure 1.5B), as seen for fathers. Similarly, the score for AFB increases with mother’s age at death (Figure 1.5C), indicating an association between variants that delay AFB and
longer life span. Scores for CAD, LDL, and HDL did not show significant monotonic change across mother’s age at death bins ($p \sim 7.7 \cdot 10^{-3}$, $p \sim 0.058$, and $p \sim 0.35$, respectively); however, the trends are suggestive of subtle age-dependent effects, with an effect of CAD score in middle age and late-onset effects of LDL and HDL scores (Figures 1.5D-F). Testing for age by sex interactions, the TC and CAD score trends with parental ages at death are significantly different between fathers and mothers ($p \sim 4 \cdot 10^{-4}$ and $p \sim 7.4 \cdot 10^{-4}$, respectively, Figure 1.27).

To further investigate the age dependency of the effects, we plotted polygenic scores among parents who had survived up to a given age as compared to the trends with parental ages at death (Figure 1.28 and Figure 1.29). All traits associated with paternal survival seemingly show more pronounced effects in middle age (Figure 1.28). Similar patterns were observed for maternal survival-associated traits except for LDL and HDL, which had more pronounced late-age effects (Figure 1.29). We also compared the hazard ratios for ages at death of $\leq 75$ and $> 75$ years (Materials and Methods), similar to a recent study [82]. Consistent with trends in scores with parental age, among the traits associated with paternal survival, almost all traits have seemingly stronger effects among younger fathers, particularly for CAD (Table 1.4): 1 SD hazard ratio of 1.14 for younger fathers ($p \sim 2.6 \cdot 10^{-9}$) and 0.99 for older fathers ($p \sim 0.7$). Unlike in fathers, in mothers, TC, LDL, and HDL scores had more pronounced late-age effects (Table 1.4): for TC, 1 SD hazard ratio of 1.03 for younger mothers ($p \sim 0.15$) and 1.11 for older mothers ($p \sim 1.4 \cdot 10^{-6}$) and for LDL, 1 SD hazard ratio of 1.05 for younger mothers ($p \sim 0.03$) and 1.12 for older mothers.
\( (p \sim 3.3 \cdot 10^{-8}) \).
Figure 1.5: Testing for the influence of sets of trait-associated variants on survival of the mothers of UK Biobank participants. (A) Quantile-quantile plot for association between the polygenic score of 42 traits (see Table 1.2) with mother’s survival, using the Cox model. The red line indicates the distribution of the p-values under the null. Signs “+” and “−” indicate protective and detrimental effects associated with higher values of polygenic scores, respectively. See Table 1.3 for p-values and hazard ratios for all traits. (B-F) Trajectory of polygenic score with age at death of mothers for top traits associated with maternal survival (only independent signals are shown, see Figure 1.25): puberty timing (B), age at first birth (AFB) (C), coronary artery disease (CAD) (D), low-density lipoproteins (LDL) (E), and high-density lipoproteins (HDL) (F). Data points in (B-F) are mean polygenic scores within 5-year interval age bins (± 2 SE), with the center of the bin indicated on the x-axis (except for the first and the last points). The dashed line shows the expected score based on the null model, accounting for confounding batch effects, changes in ancestry, and participant’s age, sex, year of birth, and the Townsend index (a measure of socioeconomic status).

Next, we sought to replicate the top associations observed among the UK Biobank participants of British ancestry (discovery cohort) in 2 other data sets: participants of the UK Biobank of non-British ancestry and the GERA cohort. Applying the Cox model using parental survival for UK Biobank participants of non-British ancestry, the direction of hazard ratios for all traits (as well as the estimated values for most traits) are consistent with the discovery cohort for both fathers and mothers (Table 1.5). The congruence of results in 2 cohorts with different ancestries suggests that our top signals are not false positives caused by poor control for population structure. In the GERA cohort, we tested whether polygenic scores change with the age of the participant, similar to our approach for individual genetic variants in this cohort. All top signals except AFB have directionally consistent effects with the discovery cohort (Table 1.6). Of particular interest, the strongest signal is an increase in the polygenic score for puberty timing with age of the participants ($p \sim 6.7 \cdot 10^{-3}$, Figure
In the discovery cohort, we further investigated if there are significant changes in the squared difference of polygenic scores with parental ages at death, as might be expected if the mean value of the trait leads to the highest chance of survival. No trait shows evidence of such stabilizing selection (Figure 1.31).

1.4 Discussion

We introduced a new approach to identify genetic variants that affect survival to a given age and thus to directly observe viability selection ongoing in humans. Attractive features of the approach include that we do not need to make a decision a priori about which loci or traits matter to viability and focus not on an end point (e.g., survival to an old age) but on any shift in allele frequencies with age, thereby learning about the ages at which effects are manifest and possible differences between sexes.

To illustrate the potential of our approach, we performed a scan for genetic variants that impact age-specific mortality in the GERA and the UK Biobank cohorts. We only found a few individual genetic variants, almost all of which were identified in previous studies. This result is in some ways expected: available data only provide high power to detect effects of common variants (> 0.15–0.2) on survival (Figure 1.1), yet if these variants were under viability selection, we would not expect them to be common, short of strong balancing selection due to trade-offs between sexes, ages, or environments. As sample sizes increase, however, the approach introduced here
should provide a comprehensive picture of viability selection in humans. To illustrate this point, we repeated our power simulation with 500,000 samples and found that we should have high power to detect the trends for alleles at a couple percent frequency in the sample (Figure 1.32).

Already, however, this application raises a number of interesting questions about the nature of viability selection in humans. Notably, we discovered only a few individual variants influencing viability in the 2 cohorts, most of which exert their effect late in life. On first thought, this finding may suggest such variants to be neutrally evolving. We would argue that if anything, our findings of only a few common variants with large effects on survival late in life suggest the opposite: that even variants with late-onset effects have been weeded out by purifying selection. Indeed, unless the number of loci in the genome that could give rise to such variants (i.e., the mutational target size) is tiny, other variants such as the APOE ε4 allele must often arise. That they are not observed when we have very high power to detect them suggests they are kept at lower frequency by purifying selection. Why might they be selected despite affecting survival only at old ages? Possible explanations include that they decrease the direct fitness of males sufficiently to be effectively selected (notably given the large, recent effective population size of humans [107]) or that they impact the inclusive fitness of males or females. If this explanation is correct, it raises the question of why the APOE ε4 allele has not been weeded out. We speculate that the environment has changed recently, making this allele more deleterious. For example, it has been proposed that the evolution of this allele has been influenced by changes in physical activity [108] and parasite burden [109].
Considering 42 traits that have been investigated by GWASs, we found a number of cases in which the mean polygenic score changes with age. Of course, detecting an effect of age on the traits does not imply that these are the phenotypes under viability selection, as the variants that contribute likely have pleiotropic effects on other traits \[86\]. Nonetheless, it is perhaps not surprising that we found detrimental effects of higher genetically predicted TC, LDL, BMI, and risk of CAD on survival, as these phenotypes are studied in GWASs precisely because of their adverse health effects. Intriguingly, however, we also found associations for fertility traits, notably, protective effects of later predicted puberty timing and AFB. If these findings reflect life-history trade-offs (e.g., longer life span at the cost of delayed reproduction), they may help to explain the persistence of extensive variation in such fitness-correlated traits \[110, 111\]. Intriguingly, we see a negative correlation between genetically predicted AFB and number of siblings of the UK Biobank participants, a proxy for the fertility of their parents \(p \sim 4.2 \cdot 10^{-8}, \text{Figure 1.33}\), consistent with previous reports of a genetic correlation between AFB and the number of children ever born \[30, 88\]. These findings underscore that consideration of survival or fertility effects alone does not allow one to infer whether the net effect of a variant or set of variants is beneficial. Instead, to convert effects on viability such as those detected here, or effects on fertility reported elsewhere \[29, 31\], into an understanding of how natural selection acts on an allele requires a characterization of its effects on all components of fitness (including potentially inclusive fitness).

In this regard, it is also worth noting that while our method is designed to detect changes in allele frequencies (and in polygenic scores) caused by genetic effects on
age-specific mortality, such changes could in principle also arise from effects on other components of fitness. For example, if the frequency of a genetic variant in a population decreases over decades due to an effect on fertility, its frequency would increase with the age of surviving individuals sampled at a given time (as in the GERA cohort). This confounding is less of an issue when considering effects on the age at death (what we measured in the UK Biobank). Nonetheless, even in the UK Biobank, fertility effects may manifest as effects on age at death; for example, when sampling a cohort of children, parents with later ages at death are possibly born earlier (Figure 1.34). To this end, in the UK Biobank, we accounted for changes in allele frequencies with year of birth of the participants themselves (ideally, we would want to condition on parents born at similar times, which we cannot do; instead, we used year of birth of the participants as an estimator for year of birth of the parents). Thus, we believe our results in the UK Biobank not to be confounded by fertility effects. Moreover, a number of our findings in this study are consistent with prior knowledge of effects on survival, such as those for disease risk variants like the APOE \( \epsilon 4 \) allele. Nonetheless, some caution is required in interpreting trends with age as strictly reflecting viability effects.

Also of interest are the marked differences between males and females in our analysis of mothers and fathers of individuals in the UK Biobank. The differences between sexes are most notable at the CHRNA3 locus, which shows a strong effect only in fathers, and sets of genetic variants associated with risk of CAD and cholesterol levels, which exhibit different age-dependent effects between fathers and mothers. Results for the CHRNA3 locus, in which variants are associated with the amount of
smoking among smokers, may reflect a gene-by-environment interaction rather than a sex effect per se. Consistent with a more pronounced effect on male than female age at death, smoking prevalence in men has been consistently higher than women over the past few decades in the United Kingdom: from 1970 to 2000, smoking prevalence decreased from around 70% to 36% in middle-aged men, compared to from around 50% to 28% in middle-aged women [112].

Moving forward, the application of approaches such as ours to the millions of samples in the pipeline (such as the UK Biobank [113], the Precision Medicine Initiative program [114], and the BioVU biobank at Vanderbilt University [115]) will allow viability effects of rare as well as common alleles to be examined. These analyses will provide a comprehensive answer to the question of which loci affect survival, helping to address long-standing open questions such as the relative importance of viability selection in shaping genetic variation and the extent to which genetic variation is maintained by fitness trade-offs between sexes or across ages.

1.5 Materials and Methods

Data sets.

GERA cohort. We performed our analyses on the data for 62,318 participants of the GERA cohort (who are members of the Kaiser Permanente Medical Care Plan, Northern California Region and participating in its Research Program on Genes, Environment, and Health), self-reported to be “White-European American”, “South
Asian”, “Middle-Eastern”, or “Ashkenazi” but no other ethnicities among a list of 23 choices on the GERA survey, and genotyped on a custom array at 670,176 SNPs designed for Non-Hispanic white individuals [26, 27]. We considered the age of the participants and the number of years they were enrolled in the Kaiser Permanente Medical Care Plan at the time of the survey (year 2007).

**UK Biobank.** We performed our analyses on the data for 152,729 participants of the UK Biobank study, focusing on 120,286 individuals identified to be “British” by genetic analysis, and all other individuals for replication. They were genotyped on the UK Biobank Axiom or the UK BiLEVE Axiom SNP arrays at a total of 847,441 SNPs in the interim release [85].

**QC.**

**GERA cohort.** We used PLINK v1.9 [116] to remove individuals with missing sex information or with a mismatch between genotype data and sex information, individuals with <96% call rate, and related individuals. We validated self-reported European ancestries using PCA, see below, and removed individuals identified as non-European (Figure 1.9 and Figure 1.10). In the end, 57,696 individuals remained. Using PLINK, we removed SNPs with < 1% minor allele frequency, SNPs with < 95% call rate, and SNPs failing a Hardy-Weinberg equilibrium test with $p < 10^{-8}$ (filtering based on HWE test could potentially exclude true signals of viability selection if selection coefficients were very large [117], but this possibility is much less likely than genotyping error). We additionally tested for a correlation between age (or
sex) and proportion of missing data, which can induce artificial change in the allele frequencies as a function of age (or sex). We thus removed SNPs showing a significant age-missingness or sex-missingness correlation, defined as a chi-squared test with $p < 10^{-7}$. After these steps, 583,357 SNPs remained.

We imputed the genotypes of the filtered GERA individuals using post-QC SNPs and using the 1000 Genomes phase 3 haplotypes as a reference panel \[118\]. We phased observed genotypes using EAGLE v1.0 software \[119\]. The inferred haplotypes were then passed to IMPUTE2 v2.3.2 software for imputation in chunks of 1 Mb using the default parameters of the software \[120\]. To gain computational speed, variants with minor allele frequency of $<0.005$ in the 1000 Genomes European populations were removed from the reference panel. This step should not affect our analysis because our statistical model is not well powered for rare variants, given the GERA data sample size. We called imputed genotypes with posterior probability of $> 0.9$ and then filtered the imputed genotypes, removing variants with IMPUTE2 info score of $< 0.5$ and with minor allele frequency of $<0.01$. We also used imputation with a leave-one-out approach \[121\] to impose a second stage of QC on genotyped SNPs, removing SNPs that were imputed back with high reported certainty (info score $> 0.5$) and with $< 90\%$ concordance between the imputed and the original genotypes. These yielded a total of 8,868,517 imputed and genotyped biallelic SNPs and indels.

For our analysis of the APOE alleles ($\epsilon2$, $\epsilon3$, and $\epsilon4$), which are defined by rs7412 and rs429358 SNPs \[103\], given the lack of tag SNPs for all 3 alleles, we kept a subset of 38,703 individuals with no poorly imputed genotypes for these 2 SNPs, for whom the count of each APOE allele could be determined unambiguously.
UK Biobank. In the UK Biobank, we obtained sets of genotype calls and the output of imputation as performed by the UK Biobank researchers [85]. We first applied QC metrics to the autosomal genotyped SNPs, focusing on the individuals of British genetic ancestry. We used PLINK to remove SNPs with minor allele frequency of < 0.01, SNPs with < 95% call rate, and SNPs failing a Hardy-Weinberg equilibrium test with \( p < 10^{-8} \). These filters were applied separately to SNPs genotyped on the UK Biobank Axiom and the UK BiLEVE Axiom arrays. Then, we divided the genotyped SNPs into 3 sets (SNPs specific to either array and shared SNPs) and then performed additional QC on each set separately: we removed SNPs with significant allele frequency difference between genotyped and imputed calls (chi-squared test \( p < 10^{-5} \)) and SNPs showing a significant correlation between proportion of missing data and age or sex of the participants as well as with participants’ father’s or mother’s age at death (chi-squared test \( p < 10^{-7} \)). We then extracted this list of SNPs from the imputed genotype files available from the UK Biobank (we did not use the full set of imputed genotypes). From this set, we removed SNPs with minor allele frequency of < 0.01, SNPs with < 95% call rate, and SNPs failing a Hardy-Weinberg equilibrium test with \( p < 10^{-8} \), yielding 590,437 SNPs. For variants influencing quantitative traits, we first extracted them from imputed genotype files and then imposed similar QC measures as above. For individuals of non-British ancestry, we first extracted the trait-influencing variants from imputed genotype files and then removed SNPs with minor allele frequency of < 0.01 and SNPs with < 90% call rate.

Each participant was asked to provide the survival status and age of their father and their mother on each assessment visit. For each participant who reported an age at
death of father and/or mother, we averaged over the ages reported at recruitment and any subsequent repeat assessment visits and used PLINK to exclude individuals with > 5-year variation in their answers across visits (around 800 individuals). For those reporting their parents to be alive, the latest assessment visit was considered. We also removed adopted individuals, individuals with a mismatch between genotype data and sex information, and individuals with missing values for the covariates, resulting in 88,595 individuals of British ancestry with age at death information for their father, 71,783 individuals of British ancestry with age at death information for their mother, and 62,719 individuals of British ancestry with age at death information for both parents. For the survival analyses, we further removed individuals with evidently invalid parental survival status, particularly parental ages at death values smaller than their age when still alive, resulting in 114,122 and 116,323 individuals of British ancestry with paternal and maternal survival information, respectively. With similar QC measures, 29,511 and 30,372 individuals of non-British ancestry with paternal and maternal survival information, respectively, were analyzed.

PCA.

We performed PCA using the EIGENSOFT v6.0.1 package with the fastpca algorithm for 2 purposes: (i) as a QC on individuals to validate self-reported European ancestries (only in GERA data set) and (ii) to correct for population structure in our statistical model (for individuals in the UK Biobank of non-British ancestry, we used the PCs provided with the data).
**European ancestry validation.** We used more stringent QC criteria specifically for the PCA compared to the QC steps described above. We filtered a subset of 157,277 SNPs in GERA, retaining SNPs shared between the data sets and the 1000 Genomes phase 3 data, removing nonautosomal SNPs, SNPs with minor allele frequency of $< 0.01$, SNPs with $< 99\%$ call rate, and SNPs failing a Hardy-Weinberg equilibrium test with $p < 10^{-6}$. We then performed LD pruning using PLINK with pairwise $r^2 < 0.2$ in windows of 50 SNPs shifting every 10 SNPs. We used these SNPs to infer principal components for the 1000 Genomes phase 3 data \[118\]. We then projected individuals onto these PCs. We observed that the majority of individuals have European ancestry and marked individuals with PCs deviating from the population mean for any of the first 6 PCs as non-European (Figure 1.9 and Figure 1.10).

**Control for population structure.** After the main QC stage, additional QC steps (as in European ancestry validation) were implemented for PCA. In the UK Biobank, we also removed inversion variants on chromosome 8, which otherwise dominate the PC2 (not shown). A subset of 156,721 SNPs in GERA and 207,657 SNPs in the UK Biobank was then used to infer PCs for individuals passing QC (Figure 1.6). The first 10 PCs were used as covariates in our statistical model.

**Quantitative traits.**

We downloaded the list of variants contributing to 39 traits (all traits but age at menarche, AFB, and age at natural menopause) and their effect sizes recently described in Pickrell et al. \[86\] from: [https://github.com/PickrellLab/](https://github.com/PickrellLab/).
For age at menarche, we used the variants and effect sizes recently identified by Day et al. [87]. We used variants associated with AFB from Barban et al. identified in either sex-specific analyses or analyses of both sexes and used the effect sizes estimated in the combined analysis [88]. We used age at natural menopause-associated variants and their effect sizes from Day et al. [89]. For all traits, we used variants that were genotyped/imputed with high quality in our data (see Table 1.2).

**Statistical model.**

**An individual variant.** Using a logistic regression, we predict the genotype of individual \( j \) (the counts of an arbitrarily selected reference allele, \( G_{ij} = 0, 1, \) or 2) at variant \( i \), using the individual’s ancestry, the batch in which the individual was genotyped, and the individual’s age (as well as sex, see below) as explanatory variables. Specifically, the distribution of \( G_{ij} \) is \( Bin(2, p_{ij}) \), where \( p_{ij} \), the probability of observing the reference allele for individual \( j \) at variant \( i \), is related to explanatory variables as:

\[
\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha + \sum_{l=1}^{10} \beta_l PC_{lj} + \sum_m \gamma_m I_{j \in BATCH_m} + \sum_n \kappa_n J_{j \in BIN_n},
\]

where \( \beta_l \) is the effect of principal component \( l \) (to account for population structure), \( \gamma_m \) is the effect of being in batch \( m \) (to account for potential systematic differences between genotyping packages), \( \kappa_n \) is the effect of being in age bin \( n \), obtained by regression across individuals with nonmissing genotypes at variant \( i \), and \( I \) and \( J \)
are indicator variables for the genotyping batch and age bin, respectively. In the version of the model in which we treat age as an ordinal variable, we replace $J$ age bin variables with 1 age variable. In the GERA data set, age binning is over the age of the participants in 14 categories, from age 19 onwards, in 5-year intervals. For replication purposes, we further binned the ages in 7 categories, in 10-year intervals, to boost our power by increasing the sample size per bin, particularly for younger age bins. In the UK Biobank, we binned the age at death of father or mother over 8 categories, from age 63 onwards, in 5-year intervals. In the UK Biobank, we included all ages at death below 63 in one age bin to minimize the potential noise caused by accidental deaths at young ages.

We tested for an effect of age categories by a likelihood ratio test with a null model using only the covariates (PCs and batch terms) ($H_0 : \kappa_n = 0$, for all $n$) and an alternative also including age terms as predictors ($H_1 : \kappa_n \neq 0$, for at least one $n$):

$$
\begin{align*}
H_0 : \log\left( \frac{p_{ij}}{1-p_{ij}} \right) &= \alpha + \sum_{l=1}^{10} \beta_l PC_{lj} + \sum_m \gamma_m I_{j \in BATCH_m} \\
H_1 : \log\left( \frac{p_{ij}}{1-p_{ij}} \right) &= \alpha + \sum_{l=1}^{10} \beta_l PC_{lj} + \sum_m \gamma_m I_{j \in BATCH_m} + \sum_n \kappa_n J_{j \in BIN_n}
\end{align*}
$$

To test for age by sex effects in GERA, we included 2 sets of additional predictors. The first consists in 2 indicator variables for sex, $K_{\text{male}}$ and $K_{\text{female}}$, which are included to capture possible sex effects induced by potential genotyping errors or mismapping of sex chromosome-linked alleles (we note that because of Hardy-Weinberg equilibrium, mean allele frequency difference between males and females is not expected). The second set of predictors consists in age by sex terms, $J \times K$. We then compare a
model with age and sex terms as predictors to a model also including age by sex terms. To test for sex effects in the UK Biobank, we compared a model with both father and mother age terms separately as predictors to a model with 1 set of age categories for average age at death of both parents, only for individuals reporting the age at death for both parents. In all models, PCs and batch terms were incorporated as covariates. For the top SNPs in the UK Biobank, we additionally tested models also including as covariates the participants’ age, sex, year of birth, and the Townsend index (a measure of socioeconomic status). For rs1051730, we also tested whether allele frequencies or trends in allele frequencies with the father’s age at death vary significantly across the UK Biobank genotyping arrays after adjusting for population structures, using similar models as described above.

**Set of variants.** As for the model described above for an individual variant, we investigated age and age by sex effects on quantitative traits for which a large number of common genetic variants have been identified in GWASs. For a given trait, we used a linear regression with the same covariates and predictors as for the model for an individual variant to predict the polygenic score for individual $j$, $S_j$ (estimated by summing the previously estimated effects of single variants, assuming additivity and that the effect sizes are similar in the GWAS panels and the cohorts considered here):

$$S_j = \alpha + \sum_{l=1}^{10} \beta_l PC_{lj} + \sum_m \gamma_m I_{j \in BATCH_m} + \sum_n \kappa_n J_{j \in BIN_n} + \epsilon_j$$
$S_j$ is calculated as $\sum a_i G_{ij} + \sum 2a_i q_i$ (standardized to mean 0), where the first sum is across variants with nonmissing genotypes, $a_i$ is the effect size for the arbitrary selected reference allele at variant $i$, and the second sum is across the variants with missing genotypes estimating their contribution assuming Hardy-Weinberg equilibrium where $q_i$ is the frequency of the alternate allele. Likelihood ratio tests, as described above, were used to test for age and age by sex effects. In the UK Biobank, we additionally adjusted for participants’ age, sex, year of birth, and the Townsend index.

To evaluate the possibility of stabilizing selection on a trait, we applied the same model, but instead of the polygenic score, we regressed the squared difference of the score from the mean in each bin, $(S_j - \overline{S}_{BIN,j})^2$, on the predictors, where $\overline{S}_{BIN,j}$ is the mean score in the age bin to which individual belongs.

We also used the Cox proportional hazards model [94] to evaluate the association between polygenic scores and parental survival in the UK Biobank. Compared to the model described above, this approach presents the advantage of allowing data from participants with alive parents to be incorporated but has the disadvantage of assuming fixed effects across all ages. Under this model, at a given time $t$ (age in our application):

$$log(\lambda_j(t)) = log(\lambda_0(t)) + \sum_{l=1}^{10} \beta_l P_{C_{ij}} + \sum_m \gamma_m I_{j \in BATCH_m} + \kappa S_j,$$

where $\lambda$ is the hazard rate (probability of death within $t + dt$ conditional on survival to time $t$) given the covariates, and $\lambda_0$ is the baseline hazard rate that describes
the risk for individuals with the value of 0 for all predictors. Not shown in the equation above are covariates to adjust for participants’ age, sex, year of birth, and the Townsend index. Using the R package “Survival” \cite{124}, for a given trait, we tested for a significant effect of polygenic score ($\kappa \neq 0$). In addition, to assess the interdependence of detected effects (Figure 1.25), for each pair of traits $[a, b]$, we tested for the effect of the polygenic score for trait $a$ but also incorporated the polygenic score for trait $b$ as a covariate in the null model (in addition to the covariates mentioned above).

We further investigated the age dependency of the effects in the framework of the survival analysis by comparing hazard ratios in 2 age categories: ages at death of $\leq 75$ and $> 75$ years. For the category of ages at death $\leq 75$ years, all parental ages were included in the analysis, and parents with ages at death beyond 75 years were marked as alive. For the category of ages at death $> 75$ years, only parents who survived beyond 75 years were considered.

All Manhattan and quantile-quantile plots were generated using qqman \cite{125} and GWASTools \cite{126} packages.

**Power simulations.**

We ran simulations to determine the power of our statistical model to detect deviation of allele frequency trends with age across 14 age categories mimicking the GERA cohort’s age structure (57,696 individuals with age distribution as in Figure 1.7) from a null model, which for simplicity was no change in frequency with age, i.e.,
no changes as a result of age-dependent variation in population structure and batch effects. For a given trend in frequency of an allele with age, we generated 1,000 simulated trends in which the distribution of the number of the alleles in age bin $i$ is $Bin(2N_i, f_i)$, where $N_i$ and $f_i$ are the sample size and the sample allele frequency in bin $i$. We then estimated the power to detect the trend as the fraction of cases in which $p < 5 \cdot 10^{-8}$, by a chi-squared test.

**Survival simulations.**

We ran simulations to investigate the relationship between allele frequency with age of the surviving individuals and the age of the individuals who died in a cohort. We simulated $2 \cdot 10^6$ individuals going forward in time in 1-year increments. For each time step forward, we tuned the chance of survival of the individuals based on their count of a risk allele for a given variant such that the number of individuals dying in the increment complies with: (i) a normal distribution of ages at death with mean of 70 years and standard deviation of 13 years, roughly, as is observed for parental ages at death in the UK Biobank and (ii) a given frequency of the risk allele among those who survive. Specifically, we modeled the survival rate of the population, $S$, as the weighted mean for 2-alleles carriers, $S_2$, 1-allele carriers, $S_1$, and noncarriers, $S_0$:

$$S(x) = \sum_{i=0}^{2} f_i S_i(x),$$

where $f$ denotes the frequency of genotypes in the population and $x$ denotes the age. $S_i$ and $S$ are related: $S_i(x) = S(x)f_i(x)/f_i$, where $f_i(x)$ is the genotype frequency
among individuals survived up to age $x$. Given a trend in allele frequency with age, we calculated genotype frequencies with age assuming Hardy-Weinberg equilibrium and then estimated genotype-dependent chance of survival, $S_i(x)$, taking $S(x)$ as the survival function for $N(70, 13^2)$.

**Acknowledgments**

We thank Guy Sella and members of the Pickrell, Przeworski, and Sella labs for helpful discussions and Graham Coop and Jonathan Pritchard for comments on an earlier version of the manuscript. This research was funded in part by Columbia University (a Research Initiative in Science and Engineering grant to MP and JKP) and the National Institutes of Health (grant R01GM121372 to MP and JKP, grant R01MH106842 to JKP and R01GM115889 to Guy Sella). These data analyses were approved by the Columbia University Institutional Review Board, protocols AAAQ2700 and AAAP0478.
1.6 Supplementary Materials

Figure 1.6: Results of principal component analysis (PCA). (A) PCA on 57,696 GERA individuals after quality control removing “non-European” individuals. (B) PCA on 120,286 UK Biobank participants of British ancestry. Result are in agreement with recent studies of these data [123, 127].
Figure 1.7: Age distribution of the GERA individuals at the time of the survey, year 2007. The labels on the x-axis indicate the center of 5-year interval age bins (except the last category).
Figure 1.8: Comparison of trends in allele frequency with age and age at death. (A) Simulated allele frequencies among surviving individuals, reproducing trends as in Figure 1.1A. (B) Trends in allele frequency among individuals who died, corresponding to the trends in (A). Points are allele frequency within 5-year interval age bins (mean ± 2 SE).
Figure 1.9: Validation of European ancestry in GERA. Shown are PCs inferred for all 26 populations in the 1000 Genomes Project phase 3 data. For clarity, in each plot, only a few representative populations are shown. GERA individuals (blue dots) are projected on the inferred PCs. The dashed lines correspond to the dashed lines in Figure 1.10, delimiting the majority of GERA individuals.
Figure 1.10: Distribution of GERA individuals for PCs inferred from 1000 Genome Project phase 3 data. The dashed lines enclose the majority of the data points; beyond, individuals were labeled as “non-Europeans”.
Figure 1.11: Quantile-quantile plots for model results for individual variants in GERA. Quantile-quantile plots for age (A) and age by sex (B) effects. The red lines indicate the distribution of the p-values under the null model (of no age or age by sex effect) and the shaded bands represent the 95% confidence intervals, assuming independent SNPs.
Figure 1.12: Frequency of the G allele of rs4988235 with age of the GERA participants. The data points are the frequencies within 5-year interval age bins (± 2 SE). The x-axis indicates the center of the age bin (except for the first and the last bins). Bins with ages below 38 years are merged into 1 bin because of the relatively small sample sizes. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and, importantly, changes in ancestry.
Figure 1.13: Regional plot for the APOE locus. The y-axis shows p-values obtained from a test of the influence of single genetic variants on age-specific mortality in GERA.
Figure 1.14: Frequency of rs6857 genotypes with age in GERA. Frequency of noncarriers (A), heterozygous (B), and homozygous (C) carriers of the risk allele for rs6857, tagging the ε4 allele of the APOE gene, across GERA age bins. Data points are frequencies within 5-year interval age bins (± 2 SE), with the center of the bin indicated on the x-axis (except for the first and the last bins). Bins with ages below 38 years are merged into 1 bin because of the relatively small sample sizes per bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.15: Frequency of rs6857 genotypes with age among males and females in GERA. Frequency of noncarriers (A), heterozygous (B), and homozygous (C) carriers of the risk allele for rs6857, tagging the ε4 allele of the APOE gene, across GERA age bins. Data points are frequencies within 5-year interval age bins (± 2 SE), with the center of the bin indicated on the x-axis (except for the first and the last bins). Bins with ages below 38 years are merged into 1 bin because of the relatively small sample sizes per bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.16: Frequency of the APOE gene alleles with age in GERA. Frequency of the \( \epsilon_2 \) (A), \( \epsilon_3 \) (B), and \( \epsilon_4 \) (C) alleles across GERA age bins. Data points are frequencies within 5-year interval age bins (± 2 SE), with the center of the bin indicated on the x-axis (except for the first and the last bins). Bins with ages below 38 years are merged into 1 bin because of the relatively small sample sizes per bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.17: Enrollments of individuals in the Kaiser Permanente Medical Care Plan. (A) Years enrolled in the care plan at the time of the survey (mean ± SD) per age bin. The x-axis indicates the center of 5-year interval age bins. (B) Years enrolled in the plan (mean ± SD) for individuals > 70 years old versus the rs6857 (APOE) genotype that they carry.
Figure 1.18: Testing for the influence of single genetic variants on age-specific mortality in the GERA cohort. Manhattan plot of p-values testing for a change in allele frequency with age using the version of the model with age treated as an ordinal variable. The plot only includes the filtered genotyped SNPs in the GERA study. Red line marks the $p = 5 \cdot 10^{-8}$ threshold. The signal for variant on chromosome 18 is presumably caused by genotyping error, as other closely linked variants did not show a similar behavior, and the signal was lost when the variant was imputed using a leave-one-out approach.
Figure 1.19: Quantile-quantile plots for model results for individual variants in the UK Biobank. Quantile-quantile plots for significant change in allele frequency with father’s (A) and mother’s (B) age at death. The red lines indicate distribution of the p-values under the null (no change in frequency) and the shaded bands represent the 95% confidence intervals, assuming independent SNPs.
Figure 1.20: Effect of rs1051730 (CHRNA3) on survival in GERA \( (p \sim 8.6 \cdot 10^{-3}) \). Allele frequency trajectory of rs1051730 with age for males and females together (A) and separately (B). The data points are the frequencies within 10-year interval age bins (± 2 SE). The x-axis indicates the center of the age bin (except for the first and the last bins). The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.21: Allele frequencies of variants in the MEOX2 locus with mother’s age at death in the UK Biobank. Plots are for 4 genotyped SNPs in moderate linkage disequilibrium with $p < 10^{-4}$ for the change in allele frequency with mother’s age at death. Data points are frequencies within 5-year interval age bins ($\pm 2$ SE), with the center of the bin indicated on the x-axis (except for the first and the last bins). The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.22: No significant effect of rs4721453 (near MEOX2) on survival in GERA ($p \sim 0.023$). Allele frequency trajectory of rs4721453 with age for males and females together (A) and separately (B). The data points are the frequencies within 10-year interval age bins (± 2 SE). The x-axis indicates the center of the age bin (except for the first and the last bins). The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.23: Ascertainment bias towards older participants introduced by using parental ages at death in the UK Biobank. Fraction of the participants in each age bin (bin size of 3 years) who reported their father’s or mother’s age at death.
Figure 1.24: Testing for a significant age by sex effect of individual genetic variants in the UK Biobank. (A) Manhattan plot of p-values, testing a difference between fathers and mothers in the change in allele frequency with parental age at death. (B) Allele frequencies as a function of father’s and mother’s age at death for top SNPs with age effects: rs4721453 (near MEOX2), rs11858836 (near CHRNA3), and rs769449 (APOE). The data points are the frequencies within 5-year interval age bins (± 2 SE). The x-axis indicates the center of the age bin (except for the first and the last bins). The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.25: Heat map showing interdependence between the age effects of pairs of trait-associated variants in the UK Biobank. Each square $[i, j]$ shows the effect size (log[hazard ratio]) of the polygenic score for trait $i$ on father’s (left) or mother’s (right) survival in the Cox model, after accounting for the effect of the polygenic score of trait $j$ (i.e., incorporating the polygenic score for trait $j$ as a covariate in the null model, see Materials and Methods). Squares on the diagonal (marked by black rectangles) show the effect size of the polygenic score without accounting for the score for other traits.
Figure 1.26: Testing for age effect of sets of trait-associated variants in the UK Biobank, treating age variables as ordinal. Quantile-quantile plots for changes in polygenic score of 42 traits (see Table 1.2) with father’s (A) or mother’s (B) age at death, after accounting for confounding batch effects, changes in ancestry, and the participant’s age, sex, year of birth, and the Townsend index (a measure of socioeconomic status). The red lines indicate the distribution of the p-values under the null model. Signs “+” and “−” indicate protective and deleterious effects associated with higher values of polygenic scores, respectively.
Figure 1.27: Testing for age by sex effect of sets of trait-associated variants in the UK Biobank. (A) Quantile-quantile plot for changes in polygenic score of 42 traits (see Table 1.2) with parental ages at death that are different between fathers and mothers of the UK Biobank participants. The red lines indicate the distribution of the p-values under the null. (B) The trend in polygenic score with parental ages at death for total cholesterol and coronary artery disease, which show significant age by sex effects. The data points are the mean polygenic scores within 5-year interval age bins (± 2 SE). The x-axis indicates the center of the age bin (except for the first and the last bins). The dashed line shows the expected polygenic score based on the null model, accounting for confounding batch effects, changes in ancestry, and the participant’s age, sex, year of birth, and the Townsend index (a measure of socioeconomic status).
Figure 1.28: Trajectories of polygenic scores with father’s age for traits associated with paternal survival in the UK Biobank. Each plot shows (in blue) the mean polygenic score (± 2 SE) among the fathers who died in a 5-year interval centered around the plotted discs, and (in black) the mean polygenic score among fathers alive up to a given age, i.e., all fathers with age or age at death (if deceased) exceeding a given age. The dashed lines show the expected changes in polygenic scores based on the null model. If there is no effect of the score on survival at a given time (age), then the score among those who died (blue disc) should be the same as the score among those who were alive at the previous time interval. Thus, the divergence between the blue and the black lines in any time interval is an indicator of the effect of the score on survival (and its direction) within that interval. The precise effect, however, also depends on the total hazard rate of the sample, which varies by age.
Figure 1.29: Trajectories of polygenic scores with mother’s age for traits associated with maternal survival in the UK Biobank. Same as Figure 1.28, but plotted for mothers (with red instead of blue).
Figure 1.30: Protective effect of later predicted puberty timing on survival in GERA ($p \sim 6.7 \cdot 10^{-3}$). Polygenic score for puberty timing with age of the participants. The data points are the mean scores within 10-year interval age bins ($\pm$ 2 SE). The x-axis indicates the center of the age bin (except for the first and the last bins). The dashed line shows the expected score based on the null model, accounting for confounding batch effects and changes in ancestry.
Figure 1.31: Testing for stabilizing selection on traits in the UK Biobank. Quantile-quantile plots testing for a change in the squared difference of polygenic score from the mean with fathers’ (A) and mothers’ (B) age at death, treating age variables as ordinal. 42 traits were tested (see Table 1.2). The red line indicates the distribution of the p-values under the null model.
Figure 1.32: Power of the model to detect changes in allele frequency with age. Same as Figure 1.1, but with 500,000 samples evenly distributed among age categories and only showing the results using models with age treated as a categorical variable. As can be seen, there should be substantial power to detect such effects even for relatively rare variants (i.e., at a couple of percent frequency in the population).
Figure 1.33: Association between variants influencing age at first birth and apparent fertility in the UK Biobank ($p \sim 4.2 \cdot 10^{-8}$). Polygenic score versus the number of siblings for 112,130 participants with mother’s age ≥ 50 years. Data points are mean scores (± 2 SE). The polygenic score was regressed on the number of siblings, accounting for the confounding batch effects, changes in ancestry, and the participant’s age, sex, year of birth, and the Townsend index (a measure of socioeconomic status). The dashed line shows the expected score based on the null model.
Figure 1.34: Ascertainment bias towards older participants for older parental age at death categories. Fraction of the participants > 61 years old (last 3 age categories in Figure 1.23) in each parental age bin. Assuming parents of older participants on average belong to earlier generations, older age at death categories will contain parents born earlier.
Table 1.2: List of phenotypes and abbreviations. The numbers of loci passing quality control measures are shown for each data set.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Abbreviation</th>
<th>#loci in UK Biobank&lt;sup&gt;a&lt;/sup&gt;</th>
<th>#loci in UK Biobank&lt;sup&gt;b&lt;/sup&gt;</th>
<th>#loci in GERA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at first birth</td>
<td>AFB</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Age at natural menopause</td>
<td>ANM</td>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at voice drop</td>
<td>AVD</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alzheimer’s disease&lt;sup&gt;c&lt;/sup&gt;</td>
<td>AD</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any allergies</td>
<td>ALL</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asthma</td>
<td>ATH</td>
<td>33</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>Brighton hypermobility</td>
<td>BHM</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index</td>
<td>BMI</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Bone mineral density (femoral neck)</td>
<td>FNBMD</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bone mineral density (lumbar spine)</td>
<td>LSBMD</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast size</td>
<td>CUP</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Childhood ear infections</td>
<td>CEI</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chin dimples</td>
<td>DIMP</td>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>CAD</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>CD</td>
<td>58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>FG</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>HEIGHT</td>
<td>561</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>HB</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-density lipoproteins</td>
<td>HDL</td>
<td>45</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td>Hypothyroidism</td>
<td>HTHY</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low-density lipoproteins</td>
<td>LDL</td>
<td>39</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Male pattern baldness</td>
<td>MPB</td>
<td>44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean cell hemoglobin concentration</td>
<td>MCHC</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean platelet volume</td>
<td>MPV</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean red cell volume</td>
<td>MCV</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migraine</td>
<td>MIGR</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nearsightedness</td>
<td>NST</td>
<td>159</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose size</td>
<td>NOSE</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Packed red cell volume</td>
<td>PCV</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parkinson’s disease</td>
<td>PD</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Photic sneeze reflex</td>
<td>PS</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puberty timing&lt;sup&gt;d&lt;/sup&gt;</td>
<td>PT</td>
<td>359</td>
<td>358</td>
<td>359</td>
</tr>
<tr>
<td>Red blood cell count</td>
<td>RBC</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>RA</td>
<td>68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>SCZ</td>
<td>191</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tonsillectomy</td>
<td>TS</td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>TC</td>
<td>49</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>TG</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>T2D</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unibrow</td>
<td>UB</td>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waist-hip ratio</td>
<td>WHR</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>: Among participants of British genetic ancestry.
<sup>b</sup>: Among participants of non-British genetic ancestry.
<sup>c</sup>: For AD the APOE locus was excluded.
<sup>d</sup>: Age at menarche associated variants were used to proximate puberty timing scores in both males and females because of the strong genetic correlation between the timing of puberty in makes and females.
Table 1.3: Results of the Cox model for association of polygenic scores for 42 traits with survival of parents of the UK Biobank participants.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Father Effect</th>
<th>SE</th>
<th>HR</th>
<th>p-value</th>
<th>Mother Effect</th>
<th>SE</th>
<th>HR</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Puberty timing</td>
<td>-0.0363</td>
<td>0.0080</td>
<td>0.96</td>
<td>6.2·10^{-6}</td>
<td>-0.0278</td>
<td>0.0090</td>
<td>0.97</td>
<td>2·10^{-3}</td>
</tr>
<tr>
<td>AD</td>
<td>-0.0238</td>
<td>0.0205</td>
<td>0.98</td>
<td>0.25</td>
<td>0.0182</td>
<td>0.0228</td>
<td>1.02</td>
<td>0.42</td>
</tr>
<tr>
<td>AFB</td>
<td>-0.0398</td>
<td>0.0180</td>
<td>0.96</td>
<td>0.03</td>
<td>-0.0639</td>
<td>0.0200</td>
<td>0.94</td>
<td>1.4·10^{-3}</td>
</tr>
<tr>
<td>ALL</td>
<td>-0.0096</td>
<td>0.0230</td>
<td>0.99</td>
<td>0.67</td>
<td>0.0245</td>
<td>0.0255</td>
<td>1.02</td>
<td>0.34</td>
</tr>
<tr>
<td>ANM</td>
<td>0.0015</td>
<td>0.0039</td>
<td>1.00</td>
<td>0.70</td>
<td>0.0102</td>
<td>0.0044</td>
<td>1.01</td>
<td>0.02</td>
</tr>
<tr>
<td>ATH</td>
<td>0.0279</td>
<td>0.0109</td>
<td>1.03</td>
<td>0.01</td>
<td>0.0149</td>
<td>0.0121</td>
<td>1.02</td>
<td>0.22</td>
</tr>
<tr>
<td>AVD</td>
<td>0.1032</td>
<td>0.0518</td>
<td>1.11</td>
<td>0.05</td>
<td>0.0370</td>
<td>0.0576</td>
<td>1.04</td>
<td>0.52</td>
</tr>
<tr>
<td>BHM</td>
<td>0.0020</td>
<td>0.0167</td>
<td>1.00</td>
<td>0.90</td>
<td>0.0162</td>
<td>0.0185</td>
<td>1.02</td>
<td>0.38</td>
</tr>
<tr>
<td>BMI</td>
<td>0.1996</td>
<td>0.0355</td>
<td>1.22</td>
<td>1.8·10^{-8}</td>
<td>0.0823</td>
<td>0.0395</td>
<td>1.09</td>
<td>0.04</td>
</tr>
<tr>
<td>CAD</td>
<td>0.0784</td>
<td>0.0177</td>
<td>1.08</td>
<td>9·10^{-6}</td>
<td>0.0892</td>
<td>0.0196</td>
<td>1.09</td>
<td>3.7·10^{-6}</td>
</tr>
<tr>
<td>CD</td>
<td>0.0035</td>
<td>0.0039</td>
<td>1.00</td>
<td>0.37</td>
<td>0.0018</td>
<td>0.0043</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td>CEI</td>
<td>-0.0029</td>
<td>0.0204</td>
<td>1.00</td>
<td>0.89</td>
<td>0.0052</td>
<td>0.0227</td>
<td>1.01</td>
<td>0.82</td>
</tr>
<tr>
<td>CUP</td>
<td>0.0058</td>
<td>0.0149</td>
<td>1.01</td>
<td>0.70</td>
<td>-0.0206</td>
<td>0.0166</td>
<td>0.98</td>
<td>0.21</td>
</tr>
<tr>
<td>DIMP</td>
<td>-0.0103</td>
<td>0.0061</td>
<td>0.99</td>
<td>0.09</td>
<td>0.0041</td>
<td>0.0068</td>
<td>1.00</td>
<td>0.55</td>
</tr>
<tr>
<td>FG</td>
<td>0.0438</td>
<td>0.0194</td>
<td>1.04</td>
<td>0.02</td>
<td>0.0123</td>
<td>0.0216</td>
<td>1.01</td>
<td>0.57</td>
</tr>
<tr>
<td>FNBMD</td>
<td>0.0212</td>
<td>0.0189</td>
<td>1.02</td>
<td>0.26</td>
<td>-0.0147</td>
<td>0.0210</td>
<td>0.99</td>
<td>0.48</td>
</tr>
<tr>
<td>HB</td>
<td>0.0610</td>
<td>0.0278</td>
<td>1.06</td>
<td>0.03</td>
<td>-0.0137</td>
<td>0.0307</td>
<td>0.99</td>
<td>0.66</td>
</tr>
<tr>
<td>HDL</td>
<td>-0.0340</td>
<td>0.0139</td>
<td>0.97</td>
<td>0.01</td>
<td>-0.0605</td>
<td>0.0154</td>
<td>0.94</td>
<td>8.9·10^{-5}</td>
</tr>
<tr>
<td>HEIGHT</td>
<td>-0.0148</td>
<td>0.0083</td>
<td>0.99</td>
<td>0.08</td>
<td>0.0092</td>
<td>0.0092</td>
<td>1.01</td>
<td>0.32</td>
</tr>
<tr>
<td>HTHY</td>
<td>0.0073</td>
<td>0.0083</td>
<td>1.01</td>
<td>0.38</td>
<td>0.0024</td>
<td>0.0093</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>LDL</td>
<td>0.0806</td>
<td>0.0140</td>
<td>1.08</td>
<td>8.1·10^{-9}</td>
<td>0.0844</td>
<td>0.0155</td>
<td>1.09</td>
<td>5.2·10^{-8}</td>
</tr>
<tr>
<td>LSBMD</td>
<td>0.0167</td>
<td>0.0173</td>
<td>1.02</td>
<td>0.34</td>
<td>0.0168</td>
<td>0.0193</td>
<td>1.02</td>
<td>0.38</td>
</tr>
<tr>
<td>MCHC</td>
<td>0.0561</td>
<td>0.0260</td>
<td>1.06</td>
<td>0.03</td>
<td>0.0106</td>
<td>0.0289</td>
<td>1.01</td>
<td>0.71</td>
</tr>
<tr>
<td>MCV</td>
<td>-0.0098</td>
<td>0.0118</td>
<td>0.99</td>
<td>0.40</td>
<td>0.0237</td>
<td>0.0313</td>
<td>1.02</td>
<td>0.07</td>
</tr>
<tr>
<td>MIGR</td>
<td>-0.0176</td>
<td>0.0165</td>
<td>0.98</td>
<td>0.28</td>
<td>-0.0261</td>
<td>0.0182</td>
<td>0.97</td>
<td>0.15</td>
</tr>
<tr>
<td>MPB</td>
<td>0.0045</td>
<td>0.0041</td>
<td>1.00</td>
<td>0.27</td>
<td>0.0017</td>
<td>0.0046</td>
<td>1.00</td>
<td>0.70</td>
</tr>
<tr>
<td>MPV</td>
<td>-0.0104</td>
<td>0.0091</td>
<td>0.99</td>
<td>0.25</td>
<td>0.0072</td>
<td>0.0101</td>
<td>1.01</td>
<td>0.48</td>
</tr>
<tr>
<td>NOSE</td>
<td>0.0070</td>
<td>0.0585</td>
<td>1.01</td>
<td>0.90</td>
<td>0.0240</td>
<td>0.0650</td>
<td>1.02</td>
<td>0.71</td>
</tr>
<tr>
<td>NST</td>
<td>-0.0130</td>
<td>0.0072</td>
<td>0.99</td>
<td>0.07</td>
<td>0.0180</td>
<td>0.0080</td>
<td>1.02</td>
<td>0.02</td>
</tr>
<tr>
<td>PCV</td>
<td>0.0519</td>
<td>0.0284</td>
<td>1.05</td>
<td>0.07</td>
<td>-0.0411</td>
<td>0.0314</td>
<td>0.96</td>
<td>0.19</td>
</tr>
<tr>
<td>PD</td>
<td>-0.0033</td>
<td>0.0094</td>
<td>1.00</td>
<td>0.73</td>
<td>-0.0012</td>
<td>0.0105</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td>PLT</td>
<td>0.0100</td>
<td>0.0143</td>
<td>1.01</td>
<td>0.48</td>
<td>-0.0028</td>
<td>0.0159</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td>PS</td>
<td>0.0147</td>
<td>0.0062</td>
<td>1.01</td>
<td>0.02</td>
<td>-0.0042</td>
<td>0.0069</td>
<td>1.00</td>
<td>0.54</td>
</tr>
<tr>
<td>RA</td>
<td>0.0092</td>
<td>0.0044</td>
<td>1.01</td>
<td>0.03</td>
<td>0.0124</td>
<td>0.0049</td>
<td>1.01</td>
<td>0.01</td>
</tr>
<tr>
<td>RBC</td>
<td>0.0215</td>
<td>0.0193</td>
<td>1.02</td>
<td>0.26</td>
<td>0.0117</td>
<td>0.0213</td>
<td>1.01</td>
<td>0.58</td>
</tr>
<tr>
<td>SCZ</td>
<td>0.0073</td>
<td>0.0054</td>
<td>1.01</td>
<td>0.18</td>
<td>0.0070</td>
<td>0.0061</td>
<td>1.01</td>
<td>0.25</td>
</tr>
<tr>
<td>T2D</td>
<td>0.0162</td>
<td>0.0117</td>
<td>1.02</td>
<td>0.17</td>
<td>0.0120</td>
<td>0.0131</td>
<td>1.01</td>
<td>0.36</td>
</tr>
<tr>
<td>TC</td>
<td>0.0901</td>
<td>0.0137</td>
<td>1.09</td>
<td>4.3·10^{-11}</td>
<td>0.0679</td>
<td>0.0152</td>
<td>1.07</td>
<td>7.8·10^{-6}</td>
</tr>
<tr>
<td>TG</td>
<td>0.0488</td>
<td>0.0160</td>
<td>1.05</td>
<td>2.2·10^{-3}</td>
<td>0.0346</td>
<td>0.0177</td>
<td>1.04</td>
<td>0.05</td>
</tr>
<tr>
<td>TS</td>
<td>0.0005</td>
<td>0.0129</td>
<td>1.00</td>
<td>0.97</td>
<td>0.0278</td>
<td>0.0143</td>
<td>1.03</td>
<td>0.05</td>
</tr>
<tr>
<td>UB</td>
<td>-0.0184</td>
<td>0.0202</td>
<td>0.98</td>
<td>0.36</td>
<td>0.0020</td>
<td>0.0224</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td>WHR</td>
<td>-0.0251</td>
<td>0.0502</td>
<td>0.98</td>
<td>0.62</td>
<td>0.0188</td>
<td>0.0559</td>
<td>1.02</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Table 1.4: Age dependency of hazard ratios for the top associations with parental survival in the UK Biobank under the Cox model.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Age range</th>
<th>Father</th>
<th>Mother</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Effect size (SE)</td>
<td>HR</td>
</tr>
<tr>
<td>Puberty timing</td>
<td>&gt; 75</td>
<td>-0.0167 (0.0129)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>-0.0486 (0.0102)</td>
<td>0.95</td>
</tr>
<tr>
<td>AFB</td>
<td>&gt; 75</td>
<td>-0.0453 (0.0291)</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>-0.0370 (0.0229)</td>
<td>0.96</td>
</tr>
<tr>
<td>ATH</td>
<td>&gt; 75</td>
<td>-0.0122 (0.0176)</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>0.0524 (0.0138)</td>
<td>1.05</td>
</tr>
<tr>
<td>BMI</td>
<td>&gt; 75</td>
<td>0.1658 (0.0575)</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>0.2182 (0.0451)</td>
<td>1.24</td>
</tr>
<tr>
<td>CAD</td>
<td>&gt; 75</td>
<td>-0.0111 (0.0286)</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>0.1337 (0.0224)</td>
<td>1.14</td>
</tr>
<tr>
<td>HDL</td>
<td>&gt; 75</td>
<td>-0.0304 (0.0225)</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>-0.0360 (0.0176)</td>
<td>0.96</td>
</tr>
<tr>
<td>LDL</td>
<td>&gt; 75</td>
<td>0.0431 (0.0227)</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>0.1030 (0.0177)</td>
<td>1.11</td>
</tr>
<tr>
<td>TC</td>
<td>&gt; 75</td>
<td>0.0490 (0.0222)</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>≤ 75</td>
<td>0.1144 (0.0173)</td>
<td>1.12</td>
</tr>
</tbody>
</table>
Table 1.5: Replication of associations in the discovery panel (UK Biobank individuals of British ancestry) in the UK Biobank participants of non-British ancestry.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Father</th>
<th>Mother</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effect size (SE)</td>
<td>HR</td>
</tr>
<tr>
<td>Puberty timing</td>
<td>-0.0497 (0.0164)</td>
<td>0.95</td>
</tr>
<tr>
<td>AFB</td>
<td>-0.0167 (0.0359)</td>
<td>0.98</td>
</tr>
<tr>
<td>ATH</td>
<td>0.0296 (0.0204)</td>
<td>1.03</td>
</tr>
<tr>
<td>BMI</td>
<td>0.2240 (0.0719)</td>
<td>1.25</td>
</tr>
<tr>
<td>CAD</td>
<td>0.0466 (0.0354)</td>
<td>1.05</td>
</tr>
<tr>
<td>HDL</td>
<td>-0.0168 (0.0276)</td>
<td>0.98</td>
</tr>
<tr>
<td>LDL</td>
<td>0.0711 (0.0282)</td>
<td>1.07</td>
</tr>
<tr>
<td>TC</td>
<td>0.0460 (0.0273)</td>
<td>1.05</td>
</tr>
</tbody>
</table>
Table 1.6: Testing for change in polygenic scores with age of the GERA participants.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Beta (SE) ($\times 10^4$)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Puberty timing</td>
<td>3.7 (1.4)</td>
<td>0.0067</td>
</tr>
<tr>
<td>AFB</td>
<td>-0.69 (0.60)</td>
<td>0.25</td>
</tr>
<tr>
<td>ATH</td>
<td>-0.41 (1.0)</td>
<td>0.7</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.26 (0.32)</td>
<td>0.42</td>
</tr>
<tr>
<td>CAD</td>
<td>-1.1 (0.62)</td>
<td>0.088</td>
</tr>
<tr>
<td>HDL</td>
<td>0.042 (0.81)</td>
<td>0.96</td>
</tr>
<tr>
<td>LDL</td>
<td>-0.65 (0.74)</td>
<td>0.38</td>
</tr>
<tr>
<td>TC</td>
<td>-1.0 (0.78)</td>
<td>0.18</td>
</tr>
</tbody>
</table>

*a: Linear regression slope coefficient (polygenic score per year).*
Chapter 2

Variable prediction accuracy of polygenic scores within an ancestry group

Published under: Mostafavi H.*, Harpak A.*, Conley D., Pritchard J.K., Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. BioRxiv (2019); *Contributed equally.

2.1 Abstract

Fields as diverse as human genetics and sociology are increasingly using polygenic scores based on genome-wide association studies (GWAS) for phenotypic prediction. However, recent work has shown that polygenic scores have limited portability across groups of different genetic ancestries, restricting the contexts in which they can be
used reliably and potentially creating serious inequities in future clinical applications. Using the UK Biobank data, we demonstrate that even within a single ancestry group, the prediction accuracy of polygenic scores depends on characteristics such as the age or sex composition of the individuals in which the GWAS and the prediction were conducted, and on the GWAS study design. Our findings highlight both the complexities of interpreting polygenic scores and underappreciated obstacles to their broad use.

### 2.2 Introduction

Genome-wide association studies (GWAS) have now been conducted for thousands of human complex traits, revealing that the genetic architecture is almost always highly polygenic, i.e., that the bulk of the heritable variation is due to thousands of genetic variants, each with tiny marginal effects \[14, 15\]. These findings often make it difficult to interpret the molecular basis for variation in a trait, but they lend themselves more immediately to another use: phenotypic prediction. Under the assumption that alleles act additively, a “polygenic score” (PGS) can be created by summing the effects of the alleles carried by an individual; this score can then be used to predict that individual’s phenotype \[37, 38\]. For highly heritable traits, such scores already provide informative predictions in some contexts: for example, prediction accuracies are 24.4% for height \[17\] and up to 13% for educational attainment \[48\].

This genomic approach to phenotypic prediction has been rapidly adopted in three distinct fields. In human genetics, PGS have been shown to help identify individuals
that are more likely to be at risk of diseases such as breast cancer (e.g., [39, 40, 41, 42]). Based on these findings, a number of papers have advocated that PGS be adopted in designing clinical studies, and by clinicians as additional risk factors to consider in treating patients [39, 128]. In human evolutionary genetics, several lines of evidence suggest that adaptation may often take the form of shifts in the optimum of a polygenic phenotype and hence act jointly on the many variants that influence the phenotype [129, 19, 130]. In this context, PGS are used to examine the evolutionary history of the set of alleles known to impact a complex trait of interest, e.g., height [19, 68, 33, 131, 132, 133]. Finally, in various disciplines of the social sciences, PGS are increasingly used to distinguish environmental from genetic sources of variability [43], as well as to understand how genetic variation among individuals may cause heterogeneous treatment effects when studying how an environmental influence (e.g., a schooling reform) affects an outcome (such as BMI) [44, 45]. In these applications, the premise is that PGS will “port” well across groups—that is that they remain predictive not only in samples very similar to the ones in which the GWAS was conducted, but also in other sets of individuals (henceforth “prediction sets”).

As recent papers have highlighted, however, PGS are not as predictive in individuals whose genetic ancestry differs substantially from the ancestry of individuals in the original GWAS (reviewed in [47]). As one illustration, PGS calculated in the UK Biobank predict phenotypes of individuals sampled in the UK Biobank better than those of individuals sampled in the BioBank Japan Project: for instance, the incremental for height is approximately 11% in the UK versus 3% in Japan [47]. Similarly, using PGS based on Europeans and European-Americans, the largest ed-
ucational attainment GWAS to date (“EA3”) reported an incremental of 10.6% for European-Americans but only 1.6% for African-Americans [48].

To date, such observations have been discussed mainly in terms of population genetic factors that reduce portability [46, 49, 50, 51, 52, 53]. Notably, GWAS does not pinpoint causal variants, but instead implicates a set of possible causal variants that lie in close physical proximity in the genome. The estimated effect of a given SNP depends on the extent of linkage disequilibrium (LD) with the causal sites [134, 15]. Thus, LD differences between populations that arose from their distinct demographic and recombination histories will lead to variation in the prediction accuracy of phenotypes across populations [135]. Because of their distinct demographic histories, populations also differ in the allele frequencies of causal variants. This problem is particularly acute for alleles that are rare in the population in which the GWAS was conducted but common in the population in which the trait is being predicted. Such variants are likely to have noisy effect size estimates in the estimation sample or may not be included in the PGS at all, and yet they contribute substantially to heritability in the target population. Furthermore, causal loci or effect sizes may differ among populations, for instance if the effect of an allele depends on the genetic background on which it arises (e.g., [136]). For all these reasons, we should expect PGS to be less predictive across ancestries.

In practice, given that most individuals (79%) included in current GWAS are of European ancestry [137, 47], PGS are systematically more predictive in European-ancestry individuals than among other people. As a consequence, the clinical applications and scientific understanding to be gained from PGS will predominantly and unfairly ben-
efit a small subset of humanity. A number of papers have therefore highlighted the importance of expanding GWAS efforts to include more diverse ancestries [138, 47, 139, 52, 53].

Importantly, factors other than ancestry could also impact the accuracy and portability of PGS. For example, the educational attainment of an individual depends not only on their own genotype, but on the genotypes of their parents, due to nurturing effects [54], and of their peers, due to social genetic effects [55], as well as of course on non-genetic factors. Also, traits such as height and educational attainment show strong patterns of assortative mating, which can distort estimated effect sizes in GWAS [56, 57, 58]. To what extent these effects remain the same across cultures and environments is unknown, but if they differ, so will the prediction accuracy. More generally, while we still know little about GxE (genotype-environment interactions) in humans, GxE effects are well-documented in other species—notably in experimental settings—and would further reduce the portability of PGS across environments [59, 90, 140]. In addition, environmental variance could differ between groups, which would change the proportion of the variance in the trait explained by a PGS (i.e., the prediction accuracy) even in the absence of genetic differences or GxE effects. Finally, PGS for some traits may include a component of environmental or cultural confounding associated with population structure [33, 34, 141, 142, 143]. This source of confounding can increase or decrease prediction accuracy, depending on the structure in the prediction samples.

Given these considerations, it is important to ask to what extent PGS are portable among groups within the same ancestry. To explore this question, we stratified the
subset of UK Biobank samples designated as “White British” (WB) according to some of the standard sample characteristics of GWAS studies: the ages of the individuals, their sex, and socio-economic status. We chose to focus on these particular characteristics because they vary widely among GWAS samples depending on sample ascertainment procedures. Furthermore, these characteristics have been shown to influence heritability for some traits in a study of a subset of the UK Biobank [144], raising the possibility that these choices also influence prediction accuracy. Indeed, for three example traits, we show that there exist major differences in the prediction accuracy of the PGS among these groups, even though they share highly similar genetic ancestries. For a variety of traits, we further demonstrate that prediction accuracy differs markedly depending on whether the GWAS is conducted in unrelated individuals or in pairs of siblings, even when controlling for the precision of the estimates. This finding is again unexpected under standard GWAS assumptions; it underscores the importance of genetic effects that are included in estimates from some study designs and not others and highlights underappreciated challenges with GWAS-based phenotypic prediction.

At present, it is difficult to fully determine the reasons why we see such variable prediction accuracy across these strata and study designs. Contributing factors probably include indirect genetic effects from relatives, assortative mating, varying levels of environmental variance, GxE interaction effects and perhaps undetected environmental confounding. Nonetheless, our results make clear that the prediction accuracy of PGS can be affected in unpredictable ways by known—and presumably unknown—factors in addition to genetic ancestry.
2.3 Results

2.3.1 Sample characteristics of the GWAS and prediction set can influence prediction accuracy even within a single ancestry

We examined how PGS for a few example traits port across samples that are of similar genetic ancestry but differ in terms of some common study characteristics, e.g., the male:female ratio (henceforth “sex ratio”), age distribution, or socio-economic status (SES). To this end, we limited our analysis to the largest subset of individuals in the UKB with a relatively homogeneous ancestry: 337,536 unrelated individuals that were characterized by the UKB as “White British” (WB). In all analyses, we further adjusted for the first 20 principal components of the genotype data, to account for any population structure within this set of individuals (Materials and Methods). In all analyses, we randomly selected a subset of individuals to be the prediction set; we then conducted GWAS using the remaining individuals and built a PGS model by LD-based clumping of the associations (Materials and Methods). To examine the reliability of the prediction, we considered the incremental, i.e., the increment obtained when adding the PGS to a model with only covariates (referred to as “prediction accuracy” henceforth). Whether this measure is appropriate depends on how PGS are to be used; it is not an obvious choice in human genetics, where the goal is often to identify individuals at high risk of developing a particular disease (i.e., in the tail of the polygenic score distribution). Nonetheless, because it has been
widely reported in discussions of portability across genetic ancestries (e.g., [48, 47]), we also used it here.

As a first case, we considered the prediction accuracy of a PGS for diastolic blood pressure in prediction sets stratified by sex, motivated by reports that variation in this trait may arise for somewhat distinct reasons in the two sexes [145, 146]. We randomly selected males and females as prediction sets (20K individuals each), and used the rest of the individuals for GWAS, matching the numbers of females and males in the GWAS set. Adjusting for mean sex effects and medication use (see Materials and Methods), the prediction accuracy is about 1.31-fold higher for females than for males (Mann-Whitney $p = 1.4 \cdot 10^{-11}$; Figure 2.1A). Thus, despite equal representation of males and females in the GWAS set, the prediction accuracy varies depending on the sex ratio of prediction samples. To examine this further, we repeated the same analysis but performed the GWAS in only one sex. When the GWAS is conducted only in females, the prediction accuracy is about 1.43-fold higher for females than for males; in turn, when GWAS was done in only males, the prediction accuracy in both sexes is similar, as well as somewhat decreased (Figure 2.1B).
Figure 2.1: Variable prediction accuracy even within an ancestry group. Shown are incremental $R^2$ values (i.e., the increment in $R^2$ obtained by adding a polygenic score predictor to a model with covariates alone) in different prediction sets. Each box and whiskers plot is computed based on twenty choices of estimation and prediction sets. Thick horizontal lines denote the medians. (A,C,E) The polygenic scores were estimated in large samples of unrelated WB individuals. Phenotypes were then predicted in distinct samples of unrelated WB individuals, stratified by sex (A), age (C) or Townsend deprivation index, a measure of SES (E). (B,D,F) Same as in A,C,E, but here the polygenic scores are based on a GWAS in a sample limited to one sex, age or SES group. When the GWAS is performed in the group that showed higher prediction accuracy in A,C,E (women, young, low SES), the qualitative trend is the same; but when the GWAS is performed in men, old or high SES, prediction accuracy is diminished and similar across groups.

We then considered two other cases, evaluating prediction accuracy in groups strati-
fied by age for BMI and by adult socio-economic status (SES) for years of schooling, using the Townsend deprivation index as a measure; our choices were motivated by prior evidence suggesting that these characteristics of the GWAS can influence heritability \[147, 148, 149, 144, 150\]. We withheld a random set of 10K individuals in each quartile of age and SES for prediction and performed GWAS using the remaining individuals, matching the sample sizes across quartiles in the GWAS set. Similar to our observation for diastolic blood pressure, the prediction accuracy varies across prediction sets: it is 1.25-fold higher for BMI in the youngest quartile compared to the oldest (Mann-Whitney \( p = 1.7 \cdot 10^{-8}; \) \textbf{Figure 2.1C}), and 1.69-fold higher for years of schooling in the lowest SES quartile compared to the highest (Mann-Whitney \( p = 1.4 \cdot 10^{-11}; \) \textbf{Figure 2.1E}). Furthermore, the differences across groups are again sensitive to the choice of the GWAS set: the differences are marked when GWAS is restricted to the youngest quartile for BMI and the lowest SES quartile for years of schooling, but diminished when the GWAS is performed in the oldest and the highest SES quartiles for BMI and years of schooling, respectively (\textbf{Figure 2.1D,F}). These results remained qualitatively unchanged when we used instead of incremental to measure prediction accuracy (\textbf{Figure 2.4}). In these analyses, we used a p-value threshold of for inclusion of a SNP in the PGS. The choice of how stringent to make the GWAS p-value threshold is important but somewhat arbitrary, with approaches ranging from requiring genome-wide significance to including all SNPs \[151, 152, 153, 154, 155, 25, 133\]. Often, this threshold is chosen to maximize prediction accuracy in an independent validation set. When the goal is to compare prediction performance across different groups, there is no obvious
optimal choice of the p-value threshold. As we show, however, the qualitative trends reported in Figure 2.1 do not depend on the p-value threshold choice (Figure 2.5). These results pertain to three exemplar traits and do not speak to the prevalence of this phenomenon. Nonetheless, they demonstrate that the portability of a polygenic score can vary markedly depending on sample characteristics of both the original GWAS and the prediction set, even within a single ancestry, and that the variation in prediction accuracy across strata can be substantial; in fact, on the same order as reported for different continental ancestries within the UK Biobank [47]. As one example, the prediction accuracy in East Asian samples, averaged across a number of traits, is about half of that in European samples when GWAS was European-based; when the GWAS is done in the lowest SES group for years of schooling, prediction accuracy in the highest SES group is less than half of that in the lowest SES (Figure 2.1F). Moreover, whereas for these traits, we had prior information about which characteristics may be relevant, other aspects that vary across sets of individuals are undoubtedly important as well (e.g., smoking behavior and diet may modify genetic effects on lipid traits; [156, 157]), and for any given trait of interest, much less may be known a priori.

**Possible explanations for the variable prediction accuracy.** Our goal in this paper is to highlight that prediction accuracies can vary across groups of highly similar ancestry, rather than to investigate the likely causes for any particular phenotype. Nonetheless, it is worth noting a couple of possibilities. Perhaps the simplest explanation for our findings is that prediction accuracies vary only because of differences...
in the extent of environmental variance, while the genetic variance is more or less constant. Indeed, the SNP heritabilities vary markedly across strata (see also [144]), and the prediction accuracies track heritability differences (Figure 2.2A,B,C). For all three traits, however, the estimated SNP heritability increases or remains the same with increasing phenotypic variance, in contrast to what would be expected under a model with a fixed genetic variance across strata (Figure 2.2D,E,F).

Another possibility is that there is an interaction between genetic effects and sample characteristics, for instance that different sets of genetic variants contribute to blood pressure levels in males and females or to BMI across different stages of life. This explanation is not supported by bivariate LD-score regression, which indicates that the genetic correlations across strata are close to 1 (Table 2.2; Materials and Methods). Yet when we re-estimate individual SNP effects in the prediction sets for SNPs ascertained in the original GWAS, the estimated effects of trait-increasing alleles are larger in the groups with higher prediction accuracy (Figure 2.6; Materials and Methods). A possible way to reconcile these findings is if effect sizes are highly correlated but systematically larger in the groups with higher prediction accuracy.
Figure 2.2: Differences in environmental variance alone do not explain the variable prediction accuracy. (A,B,C) The x-axes show heritability estimates (± SE) based on LD-score regression in each set. The y-axes show incremental $R^2$ values as in Figure 2.1A,C,E. ‘Q’ denotes quartile of age and SES in (B) and (C), respectively. Throughout, prediction accuracy largely tracks SNP heritability. (D,E,F) The x-axes show phenotypic variance estimates (± SE) across strata after adjusting for covariates (sex, age and 20 PCs). If the heritability differences across strata are due to differences in environmental variance alone, with genetic variance constant, then heritability should be inversely proportional to phenotypic variance. However, the best fitting model for this inverse proportionality (dashed line) provides a poor fit.

Other factors complicate interpretation, however, and may also contribute to our observations. In particular, for the case of educational attainment, conditioning on adult SES induces a form of range restriction, which could contribute to variable prediction accuracy across strata. We note, however, that we see highly variable prediction accuracies across SES strata even when the GWAS is conducted in all individuals (Figure 2.1E); in that regard, our approach mimics what happens in practice when polygenic scores are used to predict phenotypes in a sample with a smaller range of SES (e.g., 158). More generally, although this type of range restriction is artificially amplified in our example, SES differences will often be a problem for GWAS in which the sample is not representative of the population; for instance, the most recent major GWAS of educational attainment 48 included numerous medical data sets and the 23andMe data set, which are not representative of the national population.

Another potentially important factor is that the adjustment for PCs may not be a sufficient control for the different ways in which population structure can confound GWAS results 159, leading to variable prediction accuracy across strata if they differ in their population structure. To examine this possibility, we repeated the analysis
in Figure 2.1B,D,F but using a linear mixed model (LMM) approach (including PCs among other covariates; see Materials and Methods), and obtained qualitatively similar results (Figure 2.7). Although not a perfect fix [160, 161], the fact that we obtain similar results using PCs and LMM suggests that confounding due to population stratification in the UK Biobank alone does not explain the variable prediction accuracies across strata.

### 2.3.2 Potential portability obstacles explored through a comparison of standard and family-based GWAS

Beyond sample characteristics, a number of factors may shape the portability of scores across groups of similar ancestry. Standard GWAS is done in samples of individuals that deliberately exclude close relatives; as implemented, it detects direct effects of the genetic variants, but can also detect any indirect genetic effects of parents, siblings, or peers, effects of assortative mating among parents, and potentially environmental differences associated with fine scale population structure [54, 162, 48, 163, 33, 57, 58].

Given that many of these effects are likely to be culturally mediated (e.g., [164, 165]), it seems plausible that they may vary within as well as across groups of individuals with different ancestries. To the extent that they contribute to GWAS estimates and hence to PGS, they may lead to differences in the prediction accuracy in samples unlike the original GWAS.

To demonstrate that these considerations are not just hypothetical, we compared the prediction accuracy when the PGS is trained on “unrelated” individuals such as
those used in a standard GWAS to one obtained from a sibling-based (or “sib-based”) GWAS (Materials and Methods). In the latter, genotype differences between sibs—a result of random Mendelian segregation in the parents—are tested for association with the phenotypic difference between them. Because the tests depend on phenotypic differences between siblings who, of course, have the same parents, these tests are conditioned on the parental genotypes. Hence, they exclude many of the indirect effects signals that may be picked up in standard GWAS (Supplementary Materials). Differences between standard and sib-based GWAS are thus informative about the relative importance of factors other than direct genetic effects\cite{38, 163, 48, 33, 165}.

A challenge in this comparison is that the UKB contains about 22K sibling pairs, about 19K of which fall in the designation “White British” (WB). The siblings are similar to the unrelated individuals in terms of ages, SES distributions and genetic ancestries (Figures 2.8, 2.9) but include a higher proportion of females; this difference is unlikely to influence our analyses (see below). While a large number, 19K pairs is still too few to have adequate power to discover trait-associated SNPs, when compared to a standard GWAS using the much larger sample of unrelated WB individuals (~340K).

To increase power and enable a direct comparison between the two designs, we split the SNP ascertainment and effect estimation steps as follows (Figure 2.3A): we identified SNPs using a standard GWAS with a large sample size (median ~270K across the traits considered) (see Materials and Methods). We then estimated the effect of each significant SNP using (i) a sib-based association test and (ii) a standard
association test. We chose the size of the estimation set in (ii) such that the median standard error of effect estimates in (i) and (ii) is approximately equal. We then compared the prediction accuracy of the two PGS obtained in this way (“standard PGS” and “sib-based PGS”) in an independent prediction set of unrelated individuals; as we show in the Supplementary Materials, our approach leads to highly similar prediction accuracies of the two approaches under a model with direct effects only (see Materials and Methods for details). A further advantage is that the two scores are compared for the same set of SNPs, such that LD patterns and allele frequencies do not come into play.

We applied the approach to 22 traits, focusing on traits with relatively high heritability estimates as well as social and behavioral traits that have been the focus of recent attention in social sciences. For the majority of the traits, such as diastolic blood pressure, BMI, and hair color, the prediction accuracies of standard and sib-based PGS were similar, as expected under standard GWAS assumptions and as observed for two traits simulated under these assumptions (Figure 2.3B). However, for a range of social and behavioral traits, such as years of schooling completed, pack years of smoking and age at first sexual intercourse, the prediction accuracy of the sib-based PGS was substantially lower than that of the standard PGS (Figure 2.3B). It was also significantly lower for two morphological traits, height and whole body water mass.
Figure 2.3: Comparison of prediction accuracy of standard and sib-based polygenic scores. 

(A) After ascertaining SNPs in a large sample of unrelated individuals, we estimated the effect of these SNPs with a standard regression using unrelated individuals and, independently, using sib-regression. We then used the polygenic scores for prediction in a third sample of unrelated individuals. We chose the sample size of the standard PGS estimation set such that median effect estimate SEs are equal in the two designs, thereby ensuring equal prediction accuracy under a vanilla model with no indirect effects or assortative mating. Numbers in parentheses are median sample size in each set across 22 traits in Table 2.1 (see Table 2.3 for sample sizes for each trait). 

(B) Ratio of prediction accuracy in the two designs across 22 traits. For each trait, we performed 10 resampling iterations of unrelated individuals into three sets for discovery, estimation and prediction (small points). Large points show mean values. 

(C-H) We repeated this procedure with different discovery-set p-value thresholds for including a SNP in the polygenic score. The higher the p-value threshold is, the more SNPs are included. For each p-value threshold, points show 10 iterations as described and lines show mean values. Shown are a subset of traits, with traits appearing in (B) but not shown here presented in Figure 2.13.

A number of factors could contribute to the difference between prediction accuracies for PGS based on sibs versus unrelated individuals, including residual effects of population stratification, indirect genetic effects from parents and assortative mating. The relative importance of each factor will vary across traits \[135, 54, 141, 58, 165\]; for educational attainment, this gap is likely to reflect at least in part the documented contribution of indirect genetic effects to the standard PGS \[48, 54, 162\]. 

We show in the Supplementary Materials that in the presence of indirect genetic effects mediated through parents, standard PGS outperforms sib-based PGS unless direct and indirect effects are strongly anticorrelated (Figure 2.10), which seems unlikely to be the case for years of schooling. The difference in the performance of sib-based and standard PGS observed for other social and behavioral outcomes, such as household income and age at first sexual intercourse (Figure 2.3B), may reflect a similar phenomenon. An additional contribution to divergent prediction accuracies
could come from sibling indirect effects, which contribute differentially to standard and sibling-based PGS.

For height, there may be an important contribution of assortative mating to the difference in prediction accuracies \[36, 57, 48\]. In the Supplementary Materials, we show that under a simple model of positive assortative mating (mating of similar individuals), the prediction accuracy based on a standard PGS is better than that of a sib-based PGS (Figure 2.11). The difference in the performance of sib-based and standard PGS observed for whole body water mass (Figure 2.3B) could possibly reflect the same underlying effects of assortative mating, especially considering the high genetic correlation between the two traits (by bivariate LD score regression, \(\rho_g \approx 0.66, p < 10^{-30}\)). We further confirmed that the difference in the sex ratio of the siblings and unrelated individuals, mentioned earlier, has a negligible effect on these differences (Figure 2.12).

Thus, in comparisons of the prediction accuracies for PGS derived from standard and sib-based association tests, many traits, notably behavioral ones, show substantial differences in performance. We caution that while lower prediction accuracies for PGS based on sib-based GWAS suggest that assortative mating or indirect effects play a substantial role, the magnitude of the ratio also depends on other features of the comparison like the sample sizes used (see Supplementary Materials). By matching the sampling errors of the two approaches (Figure 2.3A), we ensure that prediction accuracies are comparable in the absence of complications such as assortative mating or indirect effects. But in the presence of these complications, the relative prediction accuracies will depend on sample sizes and on the contributions of envi-
ronmental, direct and indirect genetic components to phenotypic variance. Indeed, we show in the Supplementary Materials that in the presence of indirect genetic effects or assortative mating, the difference in prediction accuracies between the two approaches stems in part from the noise-to-signal ratio for sib-based versus standard GWAS. An implication is that the gap between the prediction accuracy of sib-based and standard PGS should depend on the number of SNPs included in the polygenic scores (Figures 2.10, 2.11).

Motivated by these considerations, we examined how the prediction accuracy varies when progressively relaxing the GWAS p-value threshold for inclusion of SNPs, i.e., when including more weakly associated SNPs in the PGS. (In Figure 2.3B, results are shown for the p-value threshold that maximizes the prediction accuracy of the standard PGS, replicating the practice when comparing populations of different ancestry [47].) For hair color and blood pressure, there is little to no difference in prediction accuracy between the two estimation methods, regardless of the number of SNPs included in the score (Figure 2.3C, D). In contrast, for height and whole body water mass, although standard and sib-based PGS perform similarly when based on the most significantly associated SNPs, standard PGS progressively outperforms sib-based PGS when more SNPs are included (Figure 2.3E, F). Similarly, the difference in prediction accuracy between sib-based and standard PGS changes markedly for years of schooling, household income and other social and behavioral traits (Figure 2.3G, H and Figure 2.13). The growing gap in performance with increasing p-value threshold likely reflects a combination of an increasing noise-to-signal ratio in the sib-based PGS (see Supplementary Materials) and changes in the relative importance
of direct effects versus other factors such as indirect parental effects and assortative mating.

In summary, the differences between the prediction accuracies of standard and sib-GWAS seen for a number of traits (Figure 2.3B) demonstrate that standard GWAS estimates often include a substantial contribution of factors other than direct effects. In these cases, even if the power to detect direct effects were comparable, standard GWAS would lead to higher prediction accuracy than sib-GWAS. In some contexts that may be a sufficient reason to rely on PGS derived from standard GWAS. However, that gain stems from the inclusion of factors such as indirect effects and assortative mating that are likely to be modulated by SES, environment and culture (e.g., [164, 165]). Thus, the increased prediction accuracy likely comes at a cost of not always porting well across groups, even of the same ancestry, in ways that may be difficult to anticipate.

2.4 Implications

Although the conversation around the portability of PGS has largely focused on genetic ancestries, our results show that prediction accuracy can also differ, at times to a comparable extent, among groups of similar ancestry—even due to basic study design differences such as age and sex composition. If only due to increased environmental variance, such decreased accuracy would be acceptable, at least for certain applications. But as we have shown, differences in the degree of environmental variance are not the primary explanation for the patterns we report (Figure 2.2), and other fac-
tors, including differences in the magnitude of genetic effects among groups, indirect effects and assortative mating, also lead to differences in the prediction accuracy of PGS, in ways that may make applications of phenotypic prediction problematic, even within a single ancestry group.

Following the discussion of portability across ancestries, we have focused on incremental $R^2$ as a measure of portability, and it remains unknown to what extent the same issues also impact the use of PGS in reliably identifying individuals in the tails of the distribution, i.e., those at elevated risk of developing a disease—the main application of PGS in human genetics, as distinct from social science or evolutionary biology. Nonetheless, the same concerns are likely to apply, especially when the magnitude of genetic effects depends on GWAS characteristics.

In any case, these results make clear that the question of the domain over which a PGS applies is not just about population genetic parameters such as LD patterns and allele frequencies or GxG effects but also the extent of environmental variance, GxE, as well as the contribution of direct effects versus indirect effects, assortative mating and environmental confounding. An important implication is that differences in prediction accuracies among groups with distinct ancestries cannot be interpreted exclusively or even primarily in terms of population genetic parameters when these groups differ dramatically in their SES [166, 167, 168, 169] and other factors that may affect portability—especially when the relative contribution of these factors to GWAS signals remains unknown. Thus, efforts to conduct GWAS in groups that vary in ancestry and geographic locations will need to be accompanied by a careful examination of variation in portability along other dimensions.
In that regard, it is worth noting that while classical twin studies were often constituted to be representative of a reference population (often national in nature) \[147, 170\], the same is not true of most contemporary human genetic datasets, which are skewed towards medical case-control studies, biobanks that are opt-in (and thus tend to be wealthier and better educated than the population average) or direct-to-consumer proprietary genetic databases (which are even more skewed along these dimensions) \[48\]. For instance, individuals in UK Biobank have higher SES than the rest of the British population \[171\] and are presumably self-selected for a certain level of interest in biomedical research. These factors alone raise challenges as to the broad portability of PGS derived from them.

One fruitful way forward may be to study data from related individuals, in which it should be possible to decompose the components of the signals identified in GWAS into direct and indirect effects, the degree of assortative mating and the contribution of residual stratification \[162, 54, 172\]. Not only will this decomposition help us to better interpret the results of GWAS and the resulting PGS, it will make it possible to examine under which circumstances, and for which phenotypes, components port more reliably to other sets of individuals, both unrelated and related. Ultimately, we envisage that in order to be broadly applicable, GWAS-based phenotypic prediction models will need to include not only a PGS but some study characteristics, other social and environmental measures and, perhaps crucially, their interactions.
2.5 Materials and Methods

UK Biobank.

The UK Biobank (UKB) is a large study of about half a million United Kingdom residents, recruited between 2006 to 2010 \cite{28}. In addition to genetic data, hundreds of phenotypes were collected through measurements and questionnaires at assessment centers, and by accessing medical records of the participants.

Inclusion criteria. In this study, we focused on 408,494 participants who passed quality control (QC) measures provided by UKB; specifically, for whom the reported sex (QC parameter “Submitted.Gender”) matched their inferred sex from genotype data (QC parameter “Inferred.Gender”); who were not identified as outliers based on heterozygosity and missing rate (QC parameter “het.missing.outliers”==0); and did not have an excessive number of relatives in the database (QC parameter “excess.relatives”==0). We further restricted ourselves to those individuals identified by UKB to be of “White British” (WB) ancestry (QC parameter “in.white.British.ancestry.subset”==1), which is a label that refers to those who, when given a set of choices, self-reported to be of “White” and “British” ethnic backgrounds and, in addition, were tightly clustered in a principal component analysis of the genotype data, as detailed in \cite{28}. For a given trait, we further conditioned on individuals for which measurement or report of the trait value was available.

Phenotype data.
We focused on 22 traits, including traits with relatively high heritability estimates as well as social and behavioral traits that have been the focus of recent attention in social sciences (see Table 2.1 for a complete list and description of the phenotypes used in this work, and their corresponding numeric field codes in the UKB data showcase). We calculated the phenotype “years of schooling” by converting the maximal educational qualification of the participants to years following Okbay et al. [173] (Table 2.4). For diastolic blood pressure, pulse rate, and forced vital capacity, we took the average of the first two rounds of measurement taken during the same examination at UKB assessment centers. We adjusted the diastolic blood pressure levels for blood pressure lowering medication following Evangelou et al. [174] by shifting the values upward by 10 mmHg for individuals taking medication. For hand grip strength, we took the average of the measurements for the two hands. For categorical phenotypes, we assigned integer values to each category (Table 2.1). For hair color, individuals who reported hair color variable “Other” were excluded from the analyses. For a subset of individuals, multiple measurements of a phenotype were provided, corresponding to multiple visits to UKB assessment centers; in those cases, we used the measurements during the first visit.

Genotype data.

UKB participants were genotyped on either of two similar genotyping arrays, UK Biobank Axiom and UK BiLEVE arrays, at a total of ~850K markers. We focused on autosomal bi-allelic SNPs shared between both arrays, and used plink v. 1.90b5
to filter SNPs with calling rate > 0.95, minor allele frequency > $10^{-3}$, and Hardy-Weinberg equilibrium test p-val > $10^{-10}$ among the WB samples, resulting in 616,323 SNPs.

GWAS and trait prediction methods.

GWAS by sample characteristics. We focused on a set of 337,536 WB samples that were identified by the UKB to be “unrelated” (sample QC parameter “used.in.pca.calculation”==1 as provided by UKB), defined such that no pairs of individuals are inferred to be 3rd degree relatives or closer. We split the sample into non-overlapping sets of individuals by one of the following factors: age at recruitment (in years), sex, and Townsend deprivation index at recruitment (used as a proxy for socioeconomic status or SES). For the Townsend deprivation index and age, we divided into four sets: Q1 [minimum value, first quartile], Q2 (first quartile, second quartile], Q3 (second quartile, third quartile], and Q4 (third quartile, maximum value]. We randomly selected 10K samples in each SES and age group, and 20K of males and 20K of females as held-out prediction sets, and performed GWAS using the remaining samples, matching sample sizes across groups in the GWAS set. We performed nine GWAS: for years of schooling in SES Q1 and SES Q4 (sample size 73,298 for each), and in the pooled sample of all four groups (sample size 293,192); for body mass index (BMI) in Q1 and Q4 (sample size 72,343 for each), and in pooled sample of all four groups (sample size 272,508); and for diastolic blood pressure in males and females (sample size 122,791 for each), and in a pooled sample of males and
females (sample size 245,582). We performed all GWAS using plink v. 2.0 (with flag: –linear), adjusting for sex, age and first 20 PCs as covariates. PCs are principal components of the all genotype data, not just WB, as provided by UKB (calculated using the whole cohort, not just WB). For a subset of cases, (where GWAS was performed in samples restricted by characteristics described above), we additionally performed association tests using a linear mixed model (LMM) as implemented in BOLT-LMM v. 2.3.2 [175], using LD scores computed from 1000 Genomes European-ancestry samples, with sex, age and first 20 PCs as covariates. The GWAS summary statistics were used to construct PGS for the samples in the prediction sets. 

To better understand the performance of PGS across the strata (see “Possible explanations for the variable prediction accuracy”), we estimated the mean effect sizes of significant SNPs in each strata. To avoid overfitting, we first performed an association test in the pooled sample of all strata; then for significantly associated SNPs, we re-estimated the effect sizes in each of the strata. We performed 20 iterations of all above steps (Figure 2.1 and Figures 2.4-2.7).

We also considered two binary phenotypes (i) attained a college degree or not and (ii) attained any degree or not, for the analysis of educational attainment by SES (as described above for years of schooling), confirming that our analysis is robust to how education phenotype is coded (Figure 2.14). For these traits we used a logistic regression model for GWAS (using plink v. 2.0 with flag: –logistic).

**Standard versus sibling-based regression.** We used the genetic relatedness information provided by UKB to infer sibling pairs among the WB samples. Following
Bycroft et al. [28], we marked pairs with and IBS0 > 0.0012 as siblings, where is the estimated kinship coefficient and IBS0 is the fraction of loci at which individuals share no alleles. By this approach, we identified 19,335 sibling pairs including 35,464 individuals across 17,305 families. For a given trait, we included pairs with the property that trait values for both individuals were reported. We then formed two sets of individuals: “Siblings” set, including the sibling pairs randomly sampled to include only one pair per family, and an “Unrelateds” set, including the unrelated individuals identified by the UKB (see section “GWAS by sample characteristics” above), but excluding the Siblings and 7,409 individuals that were related to the Siblings (3rd degree or closer).

We focused on 22 traits (Table 2.1) and two simulated traits (see below). For each trait, we first downsampled the Unrelateds to a sample size such that the median standard error of effect estimates roughly matched the median standard error in the sibling-based regression (see “Estimating $n^*$” below). We then divided the Unrelateds set into three non-overlapping sets: after sampling individuals (Unrelateds-set), we randomly split the rest of the Unrelateds set into an Unrelateds-prediction set (10% of the samples) to be used as a sample for trait prediction (“prediction set”), and an unrelated individuals discovery set (90% of the samples) to be used for the discovery of trait associated variants (see Table 2.3 for sample sizes in each set). For each trait, we performed standard GWAS in the Unrelateds-discovery set, and ascertained SNPs by thresholding on association p-values. We then estimated the effect sizes for these ascertained SNPs in two ways: by a sibling-based association test in the Siblings set (using plink v. 1.90b5’s QFAM procedure; flag: –qfam), and by a standard
association test in the Unrelatesd- set (using plink v. 2.0). Subsequently, for each set of ascertained SNPs in the Unrelatesd-discovery set, two PGS were constructed for the samples in the Unrelatesd-prediction set (see Figure 2.3A for overview of the pipeline). We performed 10 iterations of the above sampling, ascertainment and estimation steps.

Estimating n*. In order to compare the performance of sibling-based and standard GWAS designs, we wanted to match both analyses to have similar prediction accuracy under a vanilla model of no assortative mating, population structure stratification or indirect effects. In the Supplementary Materials, we show that this could be achieved by matching median effect estimate standard errors. For each trait, we therefore calculated \( n^* \), the sample size of a standard GWAS that yields roughly equal standard errors in the standard and sibling-based regressions. Specifically, for each trait, we first performed sibling-based GWAS in the Siblings using plink’s QFAM procedure (using the flag: –qfam mperm=100000 emp-se). We then randomly sampled a range of sample sizes from the set of Unrelatesd, from 5K to 20K in 1K increments. Following Wood et al. [36], for each sample size, we performed a standard GWAS, and investigated the linear relationship between the square root of the sample size and the inverse of the median standard error of the effect size estimates. We then used this linear relationship to estimate the sample size of a standard GWAS that corresponds to the inverse of the median standard error of the effect sizes estimate in the sibling-based GWAS.

All standard association tests were performed using plink v. 2.0 (using the flag: –linear), adjusting for sex, age and first 20 PCs as covariates. For sibling-based
association tests we first residualized the phenotypic values on the same covariates, and then regressed the sibling differences in residuals on sibling genotypic differences using plink’s QFAM procedure as described above.

We also considered a version of the analysis described above, in which we first residualized the phenotypes on covariates in the pooled sample of all WB individuals, and then ran the pipeline on the residuals without further adjustment for covariates in the GWAS or prediction evaluation. As shown in Figure 2.15, this approach produced results that are qualitatively the same to what we present in Figure 2.3.

**Simulated traits.** We wanted to check that given the study design described above, sibling-based and standard GWAS perform similarly with respect to trait prediction, under the vanilla model of no population stratification, assortative mating or indirect genetic effects (Figure 2.3). To this end, we simulated two traits with heritability $h^2 = 0.5$ and (i) $m = 10,000$ causal loci, and (ii) $m = 1,000,000$ causal loci.

We randomly selected the causal SNPs from a set of 10,879,183 imputed SNPs, considering that most causal variants are plausibly not directly genotyped on SNP arrays. We used a set of SNPs that passed quality control procedures by the Neale lab (http://www.nealelab.is/uk-biobank), namely autosomal SNPs, imputed using the haplotype reference consortium (HRC) panel, which have INFO score $> 0.8$ and have minor allele frequency $> 10^{-4}$; we further limited the SNP set to ones that were bi-allelic in the WB sample. As in Martin et al. [46], we randomly assigned effect sizes to these causal SNPs as $\beta \sim N(0, h^2/m)$, and zero for non-causal SNPs. We then calculated genetic component of the trait, $g$, for all WB samples under an additive model by summing the allelic counts weighted by their effect sizes using plink (using
the flag: –score). Allelic counts were determined by converting imputation dosages to genotype calls with no hard calling threshold. We also assigned environmental contributions as $\epsilon \sim N(0, 1 - h^2)$, and then constructed the PGS for each individual,

$$g = \sum_{i}^{m} \beta_i X_i,$$

where $X_i$ is the number of minor alleles at SNP $i$ carried by the individual, and the trait value for the individual is calculated as the sum of genetic and environmental contributions:

$$y = \sqrt{h^2} \frac{g - \bar{g}}{\sigma_g} + \sqrt{1 - h^2} \frac{\epsilon - \bar{\epsilon}}{\sigma_\epsilon},$$

where bars represent averages, $\sigma_g$ is the standard deviation of PGS across individuals and $\sigma_\epsilon$ is the standard deviation of environmental contributions across individuals. These simulated traits were then analyzed using the same pipelines as the other traits (e.g., adjusting for covariates etc.). Importantly, SNP discovery and effect size estimations in GWAS were performed without knowledge of the causal SNPs.

**Polygenic score (PGS) construction and trait prediction.** For all GWAS designs described above, we used p-value thresholding followed by clumping to choose sets of roughly independent SNPs to build PGS. We considered a logarithmically-spaced range of p-values: $10^{-8}$, $10^{-7}$, $10^{-6}$, $10^{-5}$, $10^{-4}$, $10^{-3}$, and $10^{-2}$ (or a subset if no SNP reached that significance level). We then used plink’s clumping procedure (using the flag: –clump) with LD threshold $r^2 < 0.1$ (using 10,000 randomly selected unrelated WB samples as a reference for LD structure) and physical distance threshold
of >1MB. The selected SNPs were then used to calculate PGS for individuals in the prediction sets, by summing the allelic counts weighted by their estimated effect sizes (log of the odds ratios in the case of binary traits) using plink (using the flag: –score). We calculated the incremental $R^2$: we first determined in a regression of the phenotype to the covariates, and then calculated the change in when including the PGS as a predictor. For binary traits, we calculated incremental Nagelkerke’s $R^2$.

**Estimating heritability and genetic correlation.**

We calculated SNP heritability across sex, age and SES groups for diastolic blood pressures, BMI and years of schooling, respectively (as described in the section “GWAS by sample characteristics”) as well as genetic correlations across pairs of groups: we first performed GWAS using all unrelated WB individuals in each group. We then used the GWAS summary statistics to perform LD-score regression with LD scores computed from the 1000 Genomes European-ancestry samples [176]. We also calculated genetic correlation between height and whole body water mass, using all unrelated WB individuals for GWAS.

**Acknowledgements**

We are grateful to Ipsita Agarwal, Daniel Belsky, Jeremy Berg, Graham Coop, Doc Edge, Iain Mathieson, Augustine Kong, Magnus Nordborg, Guy Sella, Alex Young and members of the Przeworski and Sella labs for valuable discussions and Ipsita Agarwal and Doc Edge for comments on a draft of the manuscript. This work was
2.6 Supplementary Materials

Overview of derived results

In the main text, we compare the prediction accuracies of polygenic scores (PGS) based on a standard GWAS of unrelated individuals and a GWAS based on sibling differences, for a number of traits. Here, we describe how this comparison is implemented, and how indirect effects and assortative mating manifest in this comparison.

Matching standard and sib-based prediction accuracies. Current standard GWAS are based on huge sample sizes, leading to less noisy estimates than are afforded by family association studies such as those based on sib differences, which are typically much smaller. This difference in precision needs to be taken into account in making comparisons between the prediction accuracy of scores derived from the two approaches. We show that under a vanilla additive model with no assortative mating, indirect effects, population structure (or other complications), and if the standard GWAS is subsampled to a sample size

\[ n^* \approx \frac{1}{1 + (1 - h^2)(1 - 2\rho_{sibs}) n^{pairs}} \]

where \( n^{pairs} \) is the number of sib pairs, \( h^2 \) is the heritability and \( \rho_{sibs} \) is the correlation
in environmental effects experienced by siblings, the two study designs are expected to have the same (out-of-sample) prediction accuracy (see section “Picking the sample size of the standard GWAS sample to match the prediction accuracy of the score based on the sib-GWAS”). This analytic result is not that useful in practice, however; in particular, it requires prior knowledge about the extent to which environmental effects correlate among siblings. Instead, we took an empirical approach to match the prediction accuracies in the two approaches: following Wood et al. [36], we subsampled the regular GWAS to match the median standard errors of the sib-GWAS. As we show in section “Empirical matching of standard errors”, under our vanilla model, we then expect near identical out-of-sample prediction accuracies for polygenic scores derived from the two study designs.

Indirect parental effects. In the presence of indirect parental effects, the out-of-sample prediction accuracy takes a simple form. For a polygenic score based on a standard GWAS, we obtain

\[ E[R_{ur}^2] = \tau^2 \frac{1}{1 + c}, \]

where \( \tau^2 \) is the ratio of the variance in the trait due to both direct effects and indirect effects of transmitted parental alleles over the total phenotypic variance; and \( c \) is a term representing the noise to signal ratio in a standard GWAS. For the polygenic
score based on sib-GWAS, we obtain

\[ E[R_{sib}^2] = (1 + \rho \frac{\sigma^2}{\sigma_\beta})^2 h_\beta^2 \frac{1}{1 + c\tau^2/h_\beta^2}. \]

where \( \sigma_\beta^2 \) and \( \sigma^2 \) are the variances of random direct and indirect effects, respectively, \( \rho \) is the correlation between direct and indirect effects, and \( h_\beta^2 \) is the proportion of the phenotypic variance explained by direct effects. Our results suggest that under plausible conditions, the presence of indirect effects would lead to higher prediction accuracy in a standard GWAS (after our sampling variance matching procedure). This result holds whether direct and indirect effects are positively correlated, uncorrelated or even somewhat negatively correlated (Figure 2.10).

**Assortative mating.** We investigated several models of assortative mating by simulation. Standard GWAS-based polygenic scores have greater prediction accuracies than those based on sib-GWAS when the parental phenotypes are positively correlated, and the reverse is true if they are negatively correlated (Figure 2.11A,B). The relative difference in prediction accuracies of the two study designs grows with the inclusion of more SNPs in the polygenic score model (Figure 2.11D,F).

In our analytic considerations, we ignored the ascertainment step of our study design, in which it is decided which SNPs to include in the polygenic score. We assumed that SNPs are pre-ascertained and that the set of ascertained SNPs includes all causal
ones. In a subset of simulations, we implemented the ascertainment step based on an independent simulated GWAS (see below). In both settings, we refer (somewhat loosely therefore) to the regression on ascertained SNPs in a sample of unrelated individuals as “standard GWAS” and the regression of the difference in phenotypes on the difference in sib genotypes as “sib-GWAS.”

**Picking the sample size of the standard GWAS sample to match the prediction accuracy of the score based on the sib-GWAS**

We look for the sample size $n^*$ of a standard GWAS performed on sample of unrelated individuals such that, under our vanilla model, the resulting polygenic score has the same (out-of-sample) prediction accuracy as the polygenic score obtained from a sib-GWAS with sample size $n_{pairs}$. We begin by assuming that all causal sites $i$ are known; that they are unlinked; that they have only additive, direct effects on the phenotype; and that there is no population stratification or assortative mating. We first find the sampling variance of the effect size estimate for a single site obtained from each of the two study designs. We will then examine (and ultimately match) the prediction accuracy of the polygenic scores obtained from effect sizes estimated in the estimation sets, $\hat{\beta}_{ur}, \hat{\beta}_{sit}$, on a new, independent prediction sample of unrelated individuals $\{(x', y')\}$. 
Sampling error of the estimated effect size at a single site

Our model for the phenotypic value $y$ is

$$y = g + e$$

where $e$ is a Normally distributed environmental effect (which includes all sources of random noise) and

$$g = \beta_0^{ar} + \sum_i \beta_i x_i$$

where $x_i \in \{0, 1, 2\}$ are random genotypes. The genotype is coded as the the number of alleles with effect $\beta_i$ carried by the individual at site $i$. Effect sizes $\beta = \{\beta_i\}$ are treated as fixed parameters throughout (except when noted otherwise in the very last step leading to eq. 2.22). We can rewrite our model to focus on the effect size at a single site $i$:

$$y = \beta_0 + \beta_i x_i + \epsilon_i,$$  \hspace{1cm} (2.1)

where

$$\epsilon_i = g - \beta_i x_i + e,$$

with variance

$$Var[\epsilon_i] = Var[g - \beta_i x_i] + Var[e] = Var[y] - \beta_i^2 Var[x_i]$$
In an OLS regression, the standard error for the effect of an allele at site $i$ is

$$Var[\hat{\beta}_i] = \frac{Var[\epsilon_i]}{(n - 1)Var[x_i]} = \frac{Var[y] - \beta_i^2Var[x_i]}{(n - 1)Var[x_i]},$$

(2.2)

where $n$ is the sample size. In sib-GWAS, our model for site $i$ is

$$\Delta y = \beta_0^s + \beta_i \Delta x_i + \Delta \epsilon_i,$$

with variance

$$Var[\Delta \epsilon_i] = Var[\Delta g - \beta_i \Delta x_i] + Var[\Delta e] =

Var[\Delta g] + \beta_i^2 Var[\Delta x_i] - 2 \beta_i^2 Var[\Delta x_i] + Var[\Delta e].$$

Recall that for siblings (denoted with subscripts $A$ and $B$), we expect

$$Cov[x_{i,A}, x_{i,B}] = \frac{1}{2} Var[x_i],$$

$$Cov[g_A, g_B] = \frac{1}{2} Var[g].$$

Plugging these back in, we obtain

$$Var[\Delta \epsilon_i] = Var[g] - \beta_i^2 Var[x_i] + 2 Var[e](1 - \rho_{sibs})$$

where $\rho_{sibs} = Cor[e_A, e_B]$ is the correlation in environmental effects between sibs.
The variance of the estimated effect size in sib-GWAS is therefore

$$Var[\hat{\beta}_i] = \frac{Var[\Delta \epsilon_i]}{(n_{pairs} - 1)Var[\Delta x_i]} = \frac{Var[y] - \beta_i^2Var[x_i] + Var[e](1 - 2\rho_{sibs})}{(n_{pairs} - 1)Var[x_i]}.$$  \hspace{1cm} (2.3)

Sample size required for matched prediction accuracy

We measure prediction accuracy as the expected squared correlation between polygenic scores $\hat{g}$ and phenotypic values in an independent prediction set of unrelated individuals, denoted $\{(x', y')\}$,

$$E[R^2] = \frac{Cov[\hat{g}(x'), y']}{Var[y']Var[\hat{g}(x')]},$$

by requiring

$$E[R^2(\text{standard GWAS with sample size } n^*)] = E[R^2(\text{sib regression with sample size } n_{pairs})]$$

or equivalently

$$\frac{Cov[\hat{g}_{sib}(x'), y']}{Var[\hat{g}_{sib}(x')]} = \frac{Cov[\hat{g}_{ur}(x'), y']}{Var[\hat{g}_{ur}(x')]}.$$ \hspace{1cm} (2.4)

We solve eq. 2.4 for a sample size $n^*$ to be used for estimation of the polygenic score in a standard GWAS that satisfies eq. 2.4. We note that if the vector of estimates $\hat{\beta}$ is given, then

$$Cov_{\{(x', y')\}}[y', \hat{g}(x')] = Cov_{\{(x', y')\}}[g(x'), \hat{g}(x')] = Cov_{\{(x', y')\}}[g(x'), g(x')] + \sum_{i=1}^{m} x'_i(\hat{\beta}_i - \beta_i)|\hat{\beta}| =$$
\[ \text{Var}_{\{(x', y')\}}[g(x')|\hat{\beta}] + \sum_{i}^{m} \text{Cov}_{\{(x', y')\}}[\beta_i x'_i, (\hat{\beta}_i - \beta_i)x'_i|\hat{\beta}] = \sum_{i}^{m} \text{Var}[x'_i] \beta_i \hat{\beta}_i, \quad (2.5) \]

To incorporate randomness both in the estimation set (summarized by the Multivariate Normal distribution of \( \hat{\beta} \)) and the prediction set \( \{(x', y')\} \), we apply the law of total expectation on eq. 2.4.

\[
\frac{E_{\hat{\beta}}[\text{Cov}_{\{(x', y')\}}[\hat{g}_{sib}(x'), y']^2]}{E_{\hat{\beta}}[\text{Var}_{\{(x', y')\}}[\hat{g}_{sib}(x')]]} = \frac{E_{\hat{\beta}}[\text{Cov}_{\{(x', y')\}}[\hat{g}_{ur}(x'), y']^2]}{E_{\hat{\beta}}[\text{Var}_{\{(x', y')\}}[\hat{g}_{ur}(x')]]}. \quad (2.6)
\]

Since for every \( i \), we have
\[ E[\hat{\beta}_i^{ur}] = E[\hat{\beta}_i^s] = \beta_i, \]
we obtain
\[ E_{\hat{\beta}}[\text{Cov}[y', \hat{g}^*(x')|\hat{\beta}^s]] = \sum_{i}^{m} \text{Var}[x'_i] \beta_i^2 = E_{\hat{\beta}}[\text{Cov}[y', \hat{g}^{ur}(x')|\hat{\beta}^{ur}]], \]
which turns the requirement of eq. 2.6 into
\[ E_{\hat{\beta}}[\text{Var}_{\{(x', y')\}}[\hat{g}_{sib}(x')]] = E_{\hat{\beta}}[\text{Var}_{\{(x', y')\}}[\hat{g}_{ur}(x')]], \]

or simply
\[ \sum_{i}^{m} \text{Var}[X_i] \text{Var}[\hat{\beta}_i^{ur}] = \sum_{i}^{m} \text{Var}[X_i] \text{Var}[\hat{\beta}_i^s]. \quad (2.7) \]

Plugging the sampling variance results from eq. 2.2 and eq. 2.3 into eq. 2.7 and...
reordering, we obtain

\[
\frac{n^* - 1}{n_{pairs} - 1} = \frac{\sum_i^m \text{Var}[y] - \beta_i^2 \text{Var}[x_i]}{\sum_i^m \text{Var}[y] - \beta_i^2 \text{Var}[x_i] + \text{Var}[e](1 - 2\rho_{sibs})},
\]

or, assuming that the trait is polygenic such that \(m \gg 1\),

\[
\frac{n^*}{n_{pairs}} \approx \frac{1}{1 + (1 - h^2)(1 - 2\rho_{sibs})}.
\] (2.8)

Eq. 2.8 can in principle be applied to the estimation of \(\rho_{sibs}\) for a given trait, under our model assumptions, and given an independent estimate of \(h^2\).

**Empirical matching of standard errors**

The result of eq. 2.8 is the same as we would obtain if we required

\[
\forall i \text{ Var}[^{\hat{\beta}_i^{sib}}(x_i)] \overset{\perp}{=} \text{ Var}[^{\hat{\beta}_i^{ur}}(x_i^{sib})]
\] (2.9)

without taking into account randomness in the prediction set. In practice (and in the results shown in the main text), we have no prior knowledge about \(\rho_{sibs}\) and instead we find a sample size \(n^*\) for the standard GWAS such that

\[
\text{median}_i(\text{Var}[^{\hat{\beta}_i^{sib}}(x)]) \overset{\perp}{=} \text{median}_i(\text{Var}[^{\hat{\beta}_i^{ur}}(x)])
\] (2.10)
We note that the condition in eq. 2.9 is approximately met because, if we assume that $y$ is a highly polygenic trait where

$$ \forall i \ \beta_i^2 Var[x_i] \ll Var[y], $$

then

$$ \forall i \ Var[\hat{\beta}_i^{\text{IB}}(x)] = Var[\hat{\beta}_i^\text{ur}(x)] = \frac{D}{Var[x_i]} $$

where $D$ is the same for sib-GWAS and standard GWAS estimates (given that a sample of $n^*$ is used), and approximately independent of $\beta_i$. Eq. 2.10 can be thought of as a weighted-median of $D$. In conclusion, the requirement of eq. 2.10 leads to equal prediction accuracy of standard and sib-GWAS under the vanilla model assumptions.

We note further that in the main text (Figure 2.3), to follow common practice, we use incremental $R^2$ throughout rather than $R^2$; However, as we show in Figure 2.15, the use of $R^2$ instead gives highly similar qualitative results.

**Indirect parental effects**

**Distribution of the effect size estimate at a single site**

We consider an additive model with direct effects as well as indirect parental effects, assuming no interaction between the parents and the polygenic score of the children and ignoring possible indirect effects of siblings on each other. We start by considering the model

$$ y = \beta_0 + g + n + e $$
where $g$ is the sum of direct effects in an individual with genotype (effect-allele count) $x_i$ at each site $i$,

$$g = \sum_{i}^{m} \beta_i x_i,$$

and

$$n = \sum_{i}^{m} \eta_i (x_i + \tilde{x}_i^m + \tilde{x}_i^p)$$

is the sum of parental indirect effects, with parental allele counts $x_i + \tilde{x}_i^p + \tilde{x}_i^m$ at each site, where $\tilde{x}_i^m$ is the untransmitted maternal effect allele count, and $\tilde{x}_i^p$ the untransmitted paternal effect allele count, with $\tilde{x}_i^m, \tilde{x}_i^p \in \{0,1\}$. As we show, when we choose the standard GWAS sample size $n^*$ such that the effect size estimate sampling errors match those of the sib-GWAS, the prediction accuracies of the two polygenic scores differ in an independent sample: unless there is a large negative correlation between indirect and direct effects, the polygenic score from standard GWAS is expected to outperform the one based on sib-GWAS.

We first examine the distribution of an estimated effect size of $x_i$ on the phenotype. The OLS regression for a single site in a standard GWAS follows eq. 2.1 and can be rewritten as

$$y = \beta_0 + (\beta_i + \eta_i) x_i + \eta_i (\tilde{x}_i^p + \tilde{x}_i^m) + \epsilon_i \tag{2.11}$$

with

$$\epsilon_i = g + n + e - (\beta_i + \eta_i) x_i - \eta_i (\tilde{x}_i^p + \tilde{x}_i^m).$$
By the assumption of no assortative mating or other population structure,

\[ \text{Cov}[\tilde{x}_i^p, \tilde{x}_i^m] = \text{Cov}[x_i, \tilde{x}_i^m] = \text{Cov}[x_i, \tilde{x}_i^p] = 0. \quad (2.12) \]

It directly follows that under the generative model specified by eq. 2.11, the OLS regression of \( y \) to \( x_i \) and \( \tilde{x}_i^p + \tilde{x}_i^m \) is a regression involving two independent variables. Therefore, \( \hat{\beta}^{ur}_i \) is Normally distributed with expectation

\[ E[\hat{\beta}^{ur}_i] = \beta_i + \eta_i. \]

We next calculate the variance of \( \hat{\beta}^{ur}_i \). From assumption 2.12 and

\[ \text{Var}[\tilde{x}_i^m + \tilde{x}_i^p] = \text{Var}[x_i], \]

we obtain

\[ \text{Var}[\epsilon_i] = \text{Var}[y] + (\beta_i + \eta_i)^2 \text{Var}[x_i] + \eta_i^2 \text{Var}[x_i] - 2 \text{Cov}[g+n, (\beta_i + \eta_i)x_i] - 2 \text{Cov}[n, \eta_i(\tilde{x}_i^m + \tilde{x}_i^p)] = \]

\[ = \text{Var}[y] - \text{Var}[x_i][(\beta_i^2 + 2\beta_i\eta_i + 2\eta_i^2)]. \]

Finally,

\[ \text{Var}[\hat{\beta}^{ur}_i] = \frac{\text{Var}[\epsilon_i]}{(n-1)\text{Var}[x_i]} = \frac{\text{Var}[y] - \text{Var}[x_i][(\beta_i^2 + 2\beta_i\eta_i + 2\eta_i^2)]}{(n-1)\text{Var}[x_i]}. \]
In sib regression, we have

\[ \Delta y = \Delta g + \Delta e \]

since indirect parental effects cancel out when taking the difference between sibs (as siblings have an equal parental effect allele count). Thus, the expected estimate is the same as it was in the absence of indirect effects. Using the same considerations as in section “Picking the sample size of the standard GWAS sample to match the prediction accuracy of the score based on the sib-GWAS” for the variance in sib differences, we obtain

\[ \hat{\beta}_{sib} \sim N(\beta_i, \frac{Var[g] - \beta_i^2 Var[x_i] + Var[e](1 - 2\rho_{sibs})}{(n_{pairs} - 1)Var[x_i]}), \]

where \( \rho_{sibs} \) is again the correlation in environmental effects between siblings.

**Polygenic score prediction accuracy**

We now examine the difference in prediction accuracy of \( \hat{g}^{ur} \) and \( \hat{g}^{sib} \) after matching

\[ Var[\hat{\beta}_{ur}] \overset{1}{=} Var[\hat{\beta}_{sib}^{sib}] \] (2.13)

by choosing a standard GWAS sample size \( n^* \) that empirically satisfies the condition, as we do in the main text (see also section “Empirical matching of standard errors”). We can derive the expected prediction accuracy by averaging over both the estimation set (which we again shorthand as the distribution of \( \hat{\beta} \)) and the prediction set
\{(x', y')\}. By the law of total expectation,

\[
E[R^2] = E_{\hat{\beta}}[E_{(x', y')}(R^2)] \approx \frac{E_{\hat{\beta}}[\text{Cov}_{(x', y')}(\hat{g}(x'), y' | \hat{\beta})]^2}{E_{\hat{\beta}}[\text{Var}_{(x', y')}(y')E_{\hat{\beta}}[\text{Var}_{(x', y')}(\hat{g}(x') | \hat{\beta})]},
\]

(2.14)

where we approximate the expected prediction accuracy by its first-order Taylor expansion. The numerator of eq. 2.14 is

\[
E_{\hat{\beta}}[\text{Cov}_{(x', y')}(\hat{g}(x'), y' | \hat{\beta})] = \sum_{i} (\beta_i + \eta_i)E_{\hat{\beta}}[\text{Cov}_{(x', y')}(\hat{\beta}_i, x'_i, x'_i | \hat{\beta})] = \sum_{i} \text{Var}[x_i](\beta_i + \eta_i)E[\hat{\beta}_i],
\]

(2.15)

and the denominator terms are

\[
E_{\hat{\beta}}[\text{Var}_{(x', y')}(y' | \hat{\beta})] = \text{Var}[y]
\]

(2.16)

and

\[
E_{\hat{\beta}}[\text{Var}_{(x', y')}(\hat{g}(x') | \hat{\beta})] = E_{\hat{\beta}}[\sum_{i} \text{Var}[x_i](\hat{\beta}_i^2] = \sum_{i} \text{Var}[x_i](E[\hat{\beta}_i]^2 + \text{Var}[\hat{\beta}_i]).
\]

(2.17)

Plugging eq. 2.15, 2.16, 2.17 back into eq. 2.14, we obtain

\[
E[R^2] \approx \frac{(\sum_{i} \text{Var}[x_i](\beta_i + \eta_i)E[\hat{\beta}_i]^2)}{\text{Var}[y](\sum_{i} \text{Var}[x_i]\text{Var}[\hat{\beta}_i] + \sum_{i} \text{Var}[x_i]E[\hat{\beta}_i]^2)}.
\]

(2.18)

We note that

\[
\tilde{C} := \text{Var}[y] \sum_{i} \text{Var}[x_i]\text{Var}[\hat{\beta}_i]
\]

is the same for sib-GWAS and standard GWAS under the requirement of eq. 2.13.
We therefore have

\[
E[R_{ur}^2] \approx \frac{\left(\sum_i^m Var[x_i](\beta_i + \eta_i)^2\right)^2}{C + Var[y]\sum_i^m Var[x_i](\beta_i + \eta_i)^2},
\]

(2.19)

and

\[
E[R_{sib}^2] \approx \frac{\left(\sum_i^m Var[x_i](\beta_i + \eta_i)\beta_i\right)^2}{C + Var[y]\sum_i^m Var[x_i]\beta_i^2}.
\]

(2.20)

If we denote the proportion of the phenotypic variance explained by direct effects by

\[ h_{\beta}^2 := \frac{\sum_i^m Var[x_i]\beta_i^2}{Var[y]}, \]

the proportion of the phenotypic variance explained by indirect effects of transmitted parental alleles by

\[ r_{\eta}^2 := \frac{\sum_i^m Var[x_i]\eta_i^2}{Var[y]}, \]

and the proportion of phenotypic variance explained by both direct and indirect effects of transmitted alleles by

\[ r^2 := \frac{\sum_i^m Var[x_i](\beta_i + \eta_i)^2}{Var[y]} \]

then the square of eq. 2.19 can be written as

\[
E[R_{ur}^2] \approx r^2 \frac{1}{1 + c},
\]

(2.21)
where we defined
\[ c := \frac{\sum_i^m \text{Var}[x_i] \text{Var}[\hat{\beta}_i]}{\sum_i^m \text{Var}[x_i](\beta_i + \eta_i)^2}. \]

Here, \( c \) can be thought of as a summary of the noise-to-signal ratio—with respect to the signal coming from both direct and indirect effects of transmitted alleles. If we treat effects \( \beta \) and \( \eta \) as random, treating results obtained thus far as conditional on \( \beta \) and \( \eta \), and further assume that effects are i.i.d. across sites (implying, in particular, that effect sizes and allele frequencies are independent),

\[
\begin{pmatrix}
\beta_i \\
\eta_i
\end{pmatrix}
\sim
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\sigma_\beta^2 & \rho \sigma_\beta \sigma_\eta \\
\rho \sigma_\beta \sigma_\eta & \sigma_\eta^2
\end{pmatrix},
\]

the expectation of the numerator of eq. 2.20 is

\[ E_{\beta,\eta}\sum_i^m \text{Var}[x_i] \beta_i(\beta_i + \eta_i)|\beta, \eta \] = \[ \sum_i^m \text{Var}[x_i] E_{\beta,\eta}[\beta_i^2 + \beta_i \eta_i] = \sum_i^m \text{Var}[x_i](\sigma_\beta^2 + \rho \sigma_\beta \sigma_\eta) \]

and thus eq. 2.20, in expectation, is:

\[ E[R_{\text{sib}}^2] \approx E_{\beta,\eta}[E[R_{\text{sib}}^2|\beta, \eta]] = (1 + \rho \frac{\sigma_\eta^2}{\sigma_\beta^2})h_\beta^2 \frac{1}{1 + c/\alpha}. \quad (2.22) \]

where

\[ \alpha := h_\beta^2 / \tau^2 = \frac{\sum_i^m \text{Var}[x_i] \beta_i^2}{\sum_i^m \text{Var}[x_i](\beta_i + \eta_i)^2}. \]

We examined the fit of this prediction to simulated data. Specifically, we ran simulations to estimate effect sizes in a sib-GWAS and in a standard GWAS, after choosing \( n^* \) to match their sampling variances. Finally, we used the polygenic scores to predict
phenotypic values in a sample of unrelated individuals (see section “Simulations of indirect effects” for further detail).

Figure 2.10A,C,D show the analytic result alongside simulation results, for different correlation coefficients between indirect and direct effect sizes. Even in the absence of a correlation between indirect and direct effect sizes, the polygenic score based on standard GWAS outperforms the polygenic score based on sib-GWAS.

To understand this behavior and dependency of the \( \frac{R^2_{sib}}{R^2_{ur}} \) ratio on other parameters, we divide eq. 2.22 by eq. 2.21 and obtain

\[
E\left[ \frac{R^2_{sib}}{R^2_{ur}} \right] \approx \left( 1 + \rho \frac{\sigma_\eta}{\sigma_\beta} \right)^2 \alpha \frac{1 + c}{1 + c/\alpha}.
\]

Noting further that

\[
\left( 1 + \rho \frac{\sigma_\eta}{\sigma_\beta} \right)^2 = \left( \frac{\sigma_\beta + \rho \sigma_\eta}{\sigma_\beta} \right)^2 \frac{\sigma_\beta^2}{\sigma_\beta^2 + 2 \rho \sigma_\beta \sigma_\eta + \sigma_\eta^2} = 1 - (1 - \rho^2) \frac{\tau_\eta^2}{\tau^2},
\]

we obtain

\[
E\left[ \frac{R^2_{sib}}{R^2_{ur}} \right] \approx \left[ 1 - (1 - \rho^2) \frac{\tau_\eta^2}{\tau^2} \right] \frac{1 + c}{1 + c/\alpha}. \tag{2.23}
\]

A few conclusions emerge from eq. 2.23 and from the accompanying simulations. First, the sib-GWAS based polygenic score will outperform the standard GWAS-based polygenic score only if direct and indirect effects are strongly negatively correlated
(see Figure 2.10A-D for illustration). Second, the term

\[
\frac{1 + c}{1 + c \frac{\sigma^2}{h^2}} = 1 + \frac{\sum m \text{Var}[\hat{\beta}] \text{Var}[x_i]}{\sigma^2} / \left( 1 + \frac{\sum m \text{Var}[\hat{\beta}] \text{Var}[x_i]}{h^2} \right)
\]  

(2.24)

can be interpreted as the dependence on the noise-to-signal ratio (where the signals are the proportions of phenotypic variance explained by direct and indirect effects of transmitted alleles). For a given sampling variance (matched across the two study designs), the extent of the signal will differ between standard GWAS and sib-GWAS.

Importantly, the sampling variance influences the ratio of prediction accuracies. If indirect effects do not exist or make negligible contributions to the trait in question, then the ratio of prediction accuracies is expected to be close to one. In the presence of indirect effects, however, the magnitude of the deviation from 1 depends on the relationship between direct and indirect effects (and their covariance) as well as on the (matched) sampling variance. Simulations of several parameter combinations suggest that the overall effect of this dependence on the noise-to-signal ratio is a decrease in \( \frac{R^2_{sibs}}{R^2_{ur}} \) as noise increases; as more SNPs are included in the polygenic scores, the advantage of the standard GWAS-based polygenic score over the sib-GWAS based one grows larger (Figure 2.10E-H). These considerations provide a possible interpretation for the patterns observed in Figure 2.3C-H of the main text.
Simulations of indirect effects

For each set of simulated individuals (discovery, estimation and prediction sets), we first simulated mother-father pairs, assigning parental alleles as $Bernoulli(p_i)$, where $p_i$ denotes the effect allele frequency at site $i$. We then sampled the parental alleles at random to generate offspring (one offspring per each mother-father pair to simulate a sample of unrelated individuals and two offspring to generate sibling pairs). Phenotypes of the offspring were assigned under an additive model, sampling from a Normal distribution with mean

$$\sum_{i}^{m} \beta_i x_i + \eta_i (x_i^p + x_i^m)$$

(where $x_i^m$ and $x_i^p$ are the maternal and paternal effect allele counts, respectively) and variance $\sigma_e^2$, representing the total variance of environmental effects. When there is no correlation between direct and indirect effects, $\sigma_e^2 = 1 - h^2 + 2\tau^2$. Using this approach, we generated a set of sibling pairs and estimated SNP effect sizes from these simulated data using a sib-GWAS. We calculated $n^*$ as follows: we simulated sets of unrelated individuals with a range of sample sizes. In each set, we performed a simple linear regression of the phenotypic values on the genotypes. We then estimated a linear relationship between the inverse of the median standard error of effect size estimates (as a dependent variable) and the square root of the sample size. Using this linear relationship, we predicted the sample size for the unrelated set that gives a median standard error equal to the median standard error of sib-GWAS effect size estimates ($n^*$). Finally, we simulated a set of unrelated individuals with sample
size $n^*$ and compared the prediction accuracy ($R^2$) of the polygenic score based on
standard GWAS on this sample with the one obtained from sib-GWAS.

We additionally investigated the effect of the number of SNPs included in the
polygenic scores. For this analysis, we sorted the SNPs based on the association
p-value obtained in an independent simulated set of unrelated individuals.

In these simulations, we used the following parameter values:

- The ratio of the phenotypic variance accounted for by direct effects versus by indirect
effects ($h^2_β/\tau^2_η$): 5
- The phenotypic variance explained by offspring and parental alleles, given no cor-
relation between direct and indirect effects ($h^2_β + 2\tau^2_η$): 0.25 or 0.5
- The ratio of the variance of direct effects to the variance of indirect effects ($\sigma^2_β/\sigma^2_η$):
5
- Allele frequencies are drawn from a truncated exponential distribution, truncated
on the left such that the minimum allele frequency is 1%.
- The number of loci, assumed independent (i.e., in linkage equilibrium): 100 (all
causal), or 10,000 (all causal) or 10,000 (20% causal)
- SNP effect sizes drawn as

$$
\begin{pmatrix}
\beta_i \\
\eta_i
\end{pmatrix}
\sim N
\left(
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\sigma^2_β & \rho\sigma_β\sigma_η \\
\rho\sigma_β\sigma_η & \sigma^2_η
\end{pmatrix}
\right),
$$

where $\rho$ is the correlation between direct and indirect effect sizes. Effects sizes were
then re-scaled to satisfy $\sum^m_i 2\beta^2_ip_i(1-p_i) = h^2_β$ and $\sum^m_i 2\eta^2_ip_i(1-p_i) = \tau^2_η$. Effects
were set to 0 for non-causal loci.

- The number of sibling pairs for sib GWAS: 10,000
- The number of unrelated individuals for prediction: 10,000
- The number of unrelated individuals for discovery GWAS (i.e., to decide which SNPs to include): 20,000
- Number of iterations used to estimate $n^*$ and $R^2$ for a given set of parameters: 10

**Assortative mating**

We consider assortative mating with regard to a phenotype, whereby the parents of individuals were more likely to mate if they were similar with respect to that phenotype. This process generates a correlation between genetic variants that contribute to the phenotype (i.e., linkage disequilibrium). Consequently, in a standard GWAS, the effect sizes of causal SNPs will partially capture the effect of other causal SNPs as well. Estimated effect sizes are thus expected to be inflated under positive assortative mating (mating of similar individuals) and deflated under negative assortative mating (mating of dissimilar individuals). In turn, in a sib-GWAS, the estimates are in expectation unaffected by assortative mating, because genetic differences between siblings arise from random Mendelian segregation in the parents.

**Simulations of assortative mating**

We used simulations to examine the phenotypic prediction accuracies of polygenic scores based on sib- and standard GWAS under a model with assortative mating (assuming no indirect effects or population stratification beyond assortative mating);
this end, we considered a sample of unrelated individuals, varying the degree of
correlation between parental phenotypes $\rho_a$. Similar to our simulations for indirect
effects (section “Simulations of indirect effects”), we first simulated the estimation
procedure in a sibling-based and in a standard GWAS (with sample size $n^*$). We
then computed the prediction accuracy $R^2$ in an independent sample of unrelated
individuals (see “Further simulation details” below).

We first considered the simple case of a single generation of assortative mating.
In the presence of positive assortative mating ($\rho_a > 0$), polygenic scores based on
standard GWAS outperform those based on sib-GWAS, whereas the opposite is true
in the case of negative assortative mating ($\rho_a < 0$) (Figure 2.11A). In simulations
of two generations of assortative mating, the gap between the prediction accuracies
of scores based on standard and sib-GWAS (Figure 2.11B) widens, suggesting
that our qualitative findings apply to scenarios of sustained assortative mating as well.

We further investigated prediction accuracy as a function of the number of SNPs
included in the polygenic scores, by progressively increasing the p-value threshold,
using p-values obtained from an independent GWAS in unrelated samples (similar
to our analysis in Figure 2.3). We considered two genetic architectures scenarios:
(i) in which all SNPs are causal and (ii) the case in which 20% of of SNPs are
causal (leading polygenic scores to include non-causal SNPs). Under both scenarios,
the gap in prediction accuracies between standard and sib-GWAS grows with the
number of SNPs (Figure 2.11C-F).
Further simulation details. We simulated parental and offspring alleles as described for indirect effects in section “Simulations of indirect effects”. To mimic assortative mating between parents, we first simulated i.i.d. genotypes (with effect allele counts $x_i$ at each SNP $i$) and randomly assigned “mother” and “father” labels to each individual. We then generated corresponding parental phenotypes under an additive model as

$$N\left(\sum_{i} \beta_i x_i, \sqrt{1-h^2}\right)$$

where $\beta_i$ is the effect size of SNP $i$, and $h^2$ is the heritability. The same model was used to generate offspring phenotypes.

To induce a given correlation between parental phenotypes, $\rho_a$ (mimicking the consequence of the assortative mating process), we paired mothers and fathers as follows:

First, we generated a random matrix

$$
\begin{pmatrix}
  u_{m,i} \\
  u_{p,i}
\end{pmatrix} \sim N\left(
\begin{pmatrix}
  \overline{y}_m \\
  \overline{y}_p
\end{pmatrix},
\begin{pmatrix}
  \sigma_{y_m}^2 & \rho_a \sigma_{y_m} \sigma_{y_p} \\
  \rho_a \sigma_{y_m} \sigma_{y_p} & \sigma_{y_p}^2
\end{pmatrix}
\right),
$$

where $\overline{y}_m$ and $\overline{y}_p$ are the average phenotypes of mothers and fathers, respectively, $\sigma_{y_m}$ and $\sigma_{y_p}$ are the standard deviation of the phenotypes of mothers and fathers, respectively. We then sorted the mothers and fathers sets such that the ranks of values in $y_m$ and $y_p$ match the ranks of values in $u_m$ and $u_p$, respectively, to obtain

$$cor(y_m, y_p) \approx cor(u_m, u_p) = \rho_a.$$
we simulated the generation of the grandparents similarly. We compared the performance of polygenic scores based on standard and sib-GWAS as described in section “Simulations of indirect effects”. In the simulations, we used the following parameter values:

- Heritability under random mating: 0.5

- The number of loci, assumed independent (i.e., in linkage equilibrium) under random mating: 10,000 (all causal) or 10,000 (20% causal)

- Allele frequencies, \( p \), drawn from a truncated exponential distribution, truncated on the left such that the minimum allele frequency is 1%.

- SNP effect sizes set to 0 for non-causal loci and drawn as \( \beta_i \sim N(0, \sigma^2) \), choosing \( \sigma^2 \) to satisfy \( \sum_i \beta_i^2 p_i (1 - p_i) = h^2 \) for causal loci.

- The number of sibling pairs for sib GWAS: 10,000

- The number of unrelated individuals for prediction: 10,000

- The number of unrelated individuals for discovery GWAS (i.e., to decide which SNPs to include in the polygenic score): 20,000

- The number of iterations used to estimate \( n^* \) and \( R^2 \) for a given set of parameters: 10
Table 2.1: Traits analyzed and their corresponding data fields in UK Biobank. For diastolic blood pressure, in addition to the raw measurements, we used medication data to adjust the blood pressure levels. For hand grip strength, we used the measurements for both hands. In parentheses are units of measurements.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Description</th>
<th>UKB data field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at first sex</td>
<td>Self-reported age at first sexual intercourse (years)</td>
<td>2139</td>
</tr>
<tr>
<td>Alcohol intake frequency</td>
<td>Self-reported category, encoded as an integer: 1, “Daily or almost daily”; 2, “Three or four times a week”; 3, “Once or twice a week”; 4, “One to three times a month”; 5, “Special occasions only”; 6, “Never”</td>
<td>1558</td>
</tr>
<tr>
<td>Basal metabolic rate</td>
<td>Estimated from body composition impedance measurements (KJ)</td>
<td>23105</td>
</tr>
<tr>
<td>Birth weight</td>
<td>Self-reported birth-weight (Kg)</td>
<td>20022</td>
</tr>
<tr>
<td>Body mass index</td>
<td>Constructed from height and weight measurements (Kg/m^2)</td>
<td>21001</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>Measured using automated devices (mmHg); values are adjusted for medicine use (see Materials and Methods)</td>
<td>4079, 6153, 6177</td>
</tr>
<tr>
<td>Fluid intelligence</td>
<td>Unweighted sum of the number of correct answers given to 13 fluid intelligence questions</td>
<td>20016</td>
</tr>
<tr>
<td>Forced vital capacity</td>
<td>Calculated from breath spirometry (liters)</td>
<td>3062</td>
</tr>
<tr>
<td>Hand grip strength</td>
<td>Measured right and left hand isometric grip strength (Kg)</td>
<td>46, 47</td>
</tr>
<tr>
<td>Height</td>
<td>Measured standing height (cm)</td>
<td>50</td>
</tr>
<tr>
<td>Hip circumference</td>
<td>Measured hip circumference (cm)</td>
<td>49</td>
</tr>
<tr>
<td>Household income</td>
<td>Self-reported average total annual household income before tax category, encoded as an integer: 1, “Less than $18,000”; 2, “$18,000 to $30,999”; 3, “$31,000 to $51,999”; 4, “$52,000 to $100,000”; 5, “Greater than $100,000”</td>
<td>738</td>
</tr>
<tr>
<td>Neuroticism score</td>
<td>Derived summary score, based on participants' responses to 12 neurotic behaviour-related questions</td>
<td>20127</td>
</tr>
<tr>
<td>Overall health rating</td>
<td>Self-reported category, encoded as an integer: 1, “Excellent”; 2, “Good”; 3, “Fair”; 4, “Poor”</td>
<td>2178</td>
</tr>
<tr>
<td>Pack years of smoking</td>
<td>Calculated for individuals who have ever smoked as the number of cigarettes smoked per day, divided by twenty, multiplied by the number of years of smoking (years)</td>
<td>20161</td>
</tr>
<tr>
<td>Pulse rate</td>
<td>Measured during the automated blood pressure readings (bpm)</td>
<td>102</td>
</tr>
<tr>
<td>Waist circumference</td>
<td>Measured waist circumference (cm)</td>
<td>48</td>
</tr>
<tr>
<td>Whole body fat mass</td>
<td>Estimated from body composition impedance measurements (Kg)</td>
<td>23100</td>
</tr>
<tr>
<td>Whole body water mass</td>
<td>Estimated from body composition impedance measurements (Kg)</td>
<td>23102</td>
</tr>
<tr>
<td>Years of schooling</td>
<td>Education qualifications converted to years (see Table 2.4)</td>
<td>6138</td>
</tr>
</tbody>
</table>
Table 2.2: Genetic correlations across samples that vary by a study characteristic. Numbers are genetic correlations estimated using LD score regression for BMI, years of schooling and diastolic blood pressure, across samples stratified by age, Townsend deprivation index (a measure of socioeconomic status, SES), and sex, respectively. ‘Q’ denotes quartile of age or SES.

<table>
<thead>
<tr>
<th>Trait/characteristic</th>
<th>Pair of strata</th>
<th>Genetic correlation (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BMI/Age</strong></td>
<td>(Q1,Q2)</td>
<td>0.93 (0.036)</td>
</tr>
<tr>
<td></td>
<td>(Q1,Q3)</td>
<td>0.95 (0.035)</td>
</tr>
<tr>
<td></td>
<td>(Q1,Q4)</td>
<td>0.95 (0.039)</td>
</tr>
<tr>
<td></td>
<td>(Q2,Q3)</td>
<td>0.89 (0.032)</td>
</tr>
<tr>
<td></td>
<td>(Q2,Q4)</td>
<td>0.91 (0.036)</td>
</tr>
<tr>
<td></td>
<td>(Q3,Q4)</td>
<td>1.00 (0.040)</td>
</tr>
<tr>
<td><strong>Years of schooling/SES</strong></td>
<td>(Q1,Q2)</td>
<td>0.98 (0.054)</td>
</tr>
<tr>
<td></td>
<td>(Q1,Q3)</td>
<td>0.99 (0.067)</td>
</tr>
<tr>
<td></td>
<td>(Q1,Q4)</td>
<td>0.93 (0.068)</td>
</tr>
<tr>
<td></td>
<td>(Q2,Q3)</td>
<td>0.97 (0.063)</td>
</tr>
<tr>
<td></td>
<td>(Q2,Q4)</td>
<td>1.09 (0.074)</td>
</tr>
<tr>
<td></td>
<td>(Q3,Q4)</td>
<td>1.04 (0.074)</td>
</tr>
<tr>
<td><strong>Diastolic blood pressure/Sex</strong></td>
<td>(male,female)</td>
<td>0.93 (0.031)</td>
</tr>
</tbody>
</table>
Table 2.3: Sample sizes for siblings and unrelated sets. As described in Figure 2.3A, for the comparison of prediction accuracies of polygenic scores based on standard and sib-GWAS, we first ascertain SNPs in a large sample of unrelated individuals (“Unrelated-discovery”) and then estimate the effect of these SNPs with a standard regression using unrelated individuals (“Unrelated-n*”) and, independently, using sib-regression (in the “Siblings” set). Finally, we used the polygenic scores for prediction in a third sample of unrelated individuals (“Unrelated-prediction”). This table shows sample sizes used for each set across the traits analyzed.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Siblings (pairs)</th>
<th>Unrelateds-discovery</th>
<th>Unrelateds-n*</th>
<th>Unrelateds-prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at first sex</td>
<td>13677</td>
<td>244929</td>
<td>8843</td>
<td>27214</td>
</tr>
<tr>
<td>Alcohol intake frequency</td>
<td>17288</td>
<td>276872</td>
<td>10977</td>
<td>30763</td>
</tr>
<tr>
<td>Basal metabolic rate</td>
<td>16802</td>
<td>269728</td>
<td>13490</td>
<td>29969</td>
</tr>
<tr>
<td>Birth weight</td>
<td>6753</td>
<td>159237</td>
<td>5608</td>
<td>17693</td>
</tr>
<tr>
<td>Body mass index</td>
<td>17223</td>
<td>274887</td>
<td>12377</td>
<td>30543</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>14795</td>
<td>253343</td>
<td>9424</td>
<td>28149</td>
</tr>
<tr>
<td>Fluid intelligence</td>
<td>3889</td>
<td>101070</td>
<td>2928</td>
<td>11229</td>
</tr>
<tr>
<td>Forced vital capacity</td>
<td>14611</td>
<td>252859</td>
<td>9731</td>
<td>28095</td>
</tr>
<tr>
<td>Hair color</td>
<td>16859</td>
<td>272151</td>
<td>11825</td>
<td>30238</td>
</tr>
<tr>
<td>Hand grip strength</td>
<td>17070</td>
<td>275067</td>
<td>10884</td>
<td>30563</td>
</tr>
<tr>
<td>Height</td>
<td>17248</td>
<td>270065</td>
<td>18085</td>
<td>30007</td>
</tr>
<tr>
<td>Hip circumference</td>
<td>17254</td>
<td>275957</td>
<td>11615</td>
<td>30661</td>
</tr>
<tr>
<td>Household income</td>
<td>13244</td>
<td>239274</td>
<td>8787</td>
<td>26585</td>
</tr>
<tr>
<td>Neuroticism score</td>
<td>11759</td>
<td>227111</td>
<td>6825</td>
<td>25234</td>
</tr>
<tr>
<td>Overall health rating</td>
<td>17195</td>
<td>276628</td>
<td>10365</td>
<td>30736</td>
</tr>
<tr>
<td>Pack years of smoking</td>
<td>2307</td>
<td>85626</td>
<td>1604</td>
<td>9513</td>
</tr>
<tr>
<td>Pulse rate</td>
<td>14791</td>
<td>253877</td>
<td>8790</td>
<td>28208</td>
</tr>
<tr>
<td>Skin color</td>
<td>16903</td>
<td>274150</td>
<td>10342</td>
<td>30461</td>
</tr>
<tr>
<td>Waist circumference</td>
<td>17257</td>
<td>275863</td>
<td>11757</td>
<td>30651</td>
</tr>
<tr>
<td>Whole body fat mass</td>
<td>16750</td>
<td>270569</td>
<td>12046</td>
<td>30063</td>
</tr>
<tr>
<td>Whole body water mass</td>
<td>16804</td>
<td>269694</td>
<td>13535</td>
<td>29965</td>
</tr>
<tr>
<td>Years of schooling</td>
<td>17041</td>
<td>273599</td>
<td>11873</td>
<td>30399</td>
</tr>
<tr>
<td>Simulated trait 1</td>
<td>17305</td>
<td>276331</td>
<td>11362</td>
<td>30747</td>
</tr>
<tr>
<td>Simulated trait 2</td>
<td>17305</td>
<td>276382</td>
<td>11749</td>
<td>30709</td>
</tr>
</tbody>
</table>
Table 2.4: Academic degree to years of schooling conversion table. The qualifications were converted to years of schooling following Okbay et al. [173]

<table>
<thead>
<tr>
<th>Qualifications (UKB data field 6138)</th>
<th>Years of schooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>College or University degree</td>
<td>20</td>
</tr>
<tr>
<td>NVQ or HND or HNC or equivalent</td>
<td>19</td>
</tr>
<tr>
<td>Other professional qualifications eg: nursing, teaching</td>
<td>15</td>
</tr>
<tr>
<td>A levels/AS levels or equivalent</td>
<td>13</td>
</tr>
<tr>
<td>O levels/GCSEs or equivalent</td>
<td>10</td>
</tr>
<tr>
<td>CSEs or equivalent</td>
<td>10</td>
</tr>
<tr>
<td>None of the above</td>
<td>7</td>
</tr>
</tbody>
</table>
Figure 2.4: Variable prediction accuracy (measured as $R^2$) even within an ancestry group. This figure mirrors Figure 2.1 of the main text, except for the y-axis showing are $R^2$ values, rather than incremental $R^2$. Each box and whiskers plot was computed based on twenty choices of estimation and prediction sets. Thick horizontal lines denote the medians. (A,C,E) The polygenic scores were estimated in large samples of unrelated “White British” (WB) individuals. Phenotypes were then predicted in distinct sets of unrelated WB individuals, stratified by sex (A), age (C) or Townsend deprivation index, a measure of SES (E). (B,D,F) Same as in A,C,E, but here the polygenic scores are based on a GWAS in a sample limited to one sex or to one age or SES stratum. When the GWAS is performed in the sex or stratum that shows higher prediction accuracy in A,C,E (female, young, low SES), the qualitative trend is the same, but when the GWAS is performed in men, old or high SES groups, prediction accuracy is diminished and similar across prediction sets.
Figure 2.5: Dependence on the number of SNPs included in the polygenic score. This figure extends Figure 2.1 of the main text, showing the prediction accuracies as a function of the p-value threshold for inclusion of a SNP in the polygenic score. The higher the p-value threshold is, the more SNPs are included. Shown are incremental $R^2$ values in different prediction sets. Points and error bars are mean and central 80% range computed based on twenty choices of estimation and prediction sets. (A,D,G). The polygenic scores were estimated in large samples of unrelated “White British” (WB) individuals. Phenotypes were then predicted in distinct samples of unrelated WB individuals, stratified by sex (A), age (D) or Townsend deprivation index, a measure of SES (G). (B-I) Same as in A,D,G, but here the polygenic scores are based on a GWAS in a sample limited to one sex, age or SES group. The trends shown in Figure 2.1 of the main text are for p-value threshold of $10^{-4}$, and are qualitatively similar to the trends at other p-value thresholds.
Figure 2.6: Estimating mean effect size across strata. SNPs were ascertained in large samples of unrelated “White British” (WB) individuals. The effects of trait-increasing alleles were then re-estimated in an independent set of unrelated WB individuals (that were excluded from the original GWAS) stratified by sex for diastolic blood pressure (A), by age for BMI (B) and by Townsend deprivation index, a measure of SES for years of schooling (C). Points and error bars are mean and central 80% range computed based on twenty choices of ascertainment and estimation sets, plotted as a function of the p-value threshold (for p-values obtained in the discovery GWAS).
Figure 2.7: Variable prediction accuracy within an ancestry: sensitivity to controls for population structure. This figure mirrors the panels in Figure 2.5, except that here, the standard GWAS estimates were obtained from a linear mixed model (LMM). Shown are the prediction accuracies, measured as incremental $R^2$, as a function of the p-value threshold for inclusion of a SNP in the polygenic score. Points and error bars are mean and central 80% range computed based on twenty choices of estimation and prediction sets. The polygenic scores are based on a GWAS in a sample limited to one sex, age or SES group. Phenotypes are then predicted in distinct samples of unrelated individuals, stratified by sex (A), age (B) or Townsend deprivation index, a measure of SES (C). The qualitative trends are similar to those in Figure 2.5, which uses a standard linear regression with PCs as a control for population structure when testing for an association between the phenotypes and genotypes. The similarity suggests that the observed differences in prediction accuracies across strata are not driven to a large degree by population structure confounding.
Figure 2.8: Comparison of siblings and unrelated individuals in the UK Biobank with respect to age, SES, and sex ratio. Panels show the distribution of Townsend deprivation index, a measure of SES (A), the age distribution (B), and the proportion of males (C) for the siblings and unrelated sets used in the analysis described for Figure 2.3 of the main text. For each sibling pair, one sibling was randomly selected for these comparisons. The asterisk symbol marks a significant difference at the 1% level between siblings and unrelated individuals, as assessed by a Mann-Whitney test. SES and age distributions are quite similar in siblings and unrelated sets, whereas the proportion of males is significantly smaller in the siblings.
Figure 2.9: Comparison of siblings and unrelated individuals in the UK Biobank with respect to population structure. Panels show the distribution of PCs (principal components of the genotype data) for the siblings and unrelated sets used in the analysis described for Figure 2.3 of the main text. For each sibling pair, one sibling was randomly selected for these comparisons. The asterisk symbol marks a significant difference at the 1% level between siblings and unrelated individuals, as assessed by a Mann-Whitney test. Despite slight but significant differences, siblings and unrelated sets are broadly similar with respect to their genetic ancestries.
Figure 2.10: Simulation results for polygenic scores based on standard GWAS and sib-GWAS in the presence of indirect effects. (A,B) Simulation results as a function of the correlation between direct and indirect effects, $\rho$. Simulations were performed with $h_{\beta}^2 = 0.5$, $\tau_{\beta}^2 = 0.1$, and $\sigma_{\beta}^2/\sigma_{\gamma}^2 = 5$. The size of the estimation set in the sib-GWAS is 10,000, and the size of the estimation set in the standard GWAS is chosen to match sampling variances between the two study designs. The polygenic scores is based on 10,000 causal loci; its performance was evaluated in an independent set of 10,000 unrelated individuals. As long as direct and indirect effects are not strongly negatively correlated, out-of-sample prediction accuracy is higher for the polygenic scores based on standard GWAS. (C) Same as (A) but with three-fold greater environmental noise. (D) Same as (A) but with 100 causal loci. In (A-D) points are mean ± 2 SD in 10 simulation iterations. Solid lines are values based on analytic expressions derived in section “Polygenic score prediction accuracy”. (E-H) Simulations results, with the same parameters as in (A) but $\rho = 0.5$, as a function of the number of SNPs included in the polygenic scores, with all loci being causal (E,F), or with 20% of loci being causal (G,H). SNPs are added in an increasing order of their association p-value in an independent set of 20,000 unrelated individuals. In both cases, the ratio of prediction accuracies of polygenic scores based on sib- versus standard GWAS becomes smaller with the inclusion of more weakly associated SNPs, a behavior qualitatively similar to observations in Figure 2.3 in the main text. Points are mean ± 2 SD in 10 simulations.
Figure 2.11: Simulation results for polygenic scores based on standard GWAS and sib-GWAS in the presence of assortative mating. (A) Simulation results as a function of the approximate correlation between parental phenotypes, $\rho_a$. Simulations were performed with $\tau^2 = 0.5$ under random mating. The size of the estimation set in the sib-GWAS is 10,000, and the size of the estimation set in the standard GWAS is chosen to match sampling variances between the two study designs. The polygenic scores is based on 10,000 causal loci; its performance was evaluated in an independent set of 10,000 unrelated individuals. Standard-GWAS based polygenic scores outperforms (underperforms) sib-GWAS based polygenic scores under positive (negative) assortative mating. (B) Ratio of prediction accuracies of the polygenic scores based on sib- versus standard GWAS, as a function of $\rho_a$, for two sets of simulations with one or two generations of assortative mating, with same parameters as in (A). (C-F) Simulations results, with the same parameters as in (A) but $\rho_a = 0.5$, as a function of the number of SNPs included in the polygenic score, with all loci being causal (C,D), or with 20% of loci being causal (E,F). SNPs are added in the order of their association p-value in an independent set of 20,000 unrelated individuals. In both cases, the ratio of prediction accuracies for scores based on sib-GWAS versus standard GWAS becomes smaller with the inclusion of more weakly associated SNPs, a behavior that is qualitatively similar to observations in Figure 2.3 in the main text. Points are mean ± 2 SD in 10 simulation iterations.
Figure 2.12: Comparison of prediction accuracies of polygenic scores based on standard and sib-GWAS matched for sex ratio. This figure mirrors Figure 2.3B of the main text, but here the samples of siblings and unrelated individuals used in the analysis are matched for sex ratio. Results are shown for diastolic blood pressure, as the prediction accuracy differed between sexes (Figure 2.1), the related phenotype of pulse rate, and a subset of the traits for which the prediction accuracy varied by GWAS design (Figure 2.3B). Small points show the ratio of the prediction accuracies in the two designs across 10 iterations; in each iteration, we resample sets of unrelated individuals to constitute three sets: for discovery, estimation and prediction. Large points show mean values.
Figure 2.13: Prediction accuracy of polygenic scores based on sib-and standard GWAS, for a range of traits. This figure complements Figure 2.3C-H of the main text, showing the results of the study design depicted in Figure 2.3A for all traits presented in Figure 2.3. As described for Figure 2.3, we randomly divided unrelated individuals to constitute three non-overlapping sets for discovery, estimation and prediction. Small points correspond to 10 iterations of resampling these three sets. The prediction accuracy is plotted as a function of the p-value threshold, where p-values come from the discovery GWAS. Lines show mean values.
Figure 2.14: Variable prediction accuracy within an ancestry group for binary education phenotypes. Shown are incremental Nagelkerke's $R^2$ for two binary education phenotypes, attained a college degree or not (A), and attained any degree or not (B). The polygenic scores are estimated in large samples of unrelated individuals. Phenotypes are then predicted in distinct samples of unrelated individuals, stratified by Townsend deprivation index, a measure of SES. Points and error bars are mean and central 80% range computed based on twenty choices of estimation and prediction sets, plotted as a function of the p-value threshold (with p-values obtained in the original GWAS).
Figure 2.15: Comparison of prediction accuracies of polygenic scores (measured as $R^2$) based on standard and sib-GWAS. This figure mirrors Figure 2.3B of the main text, but here we first residualized the phenotypes on covariates, and then ran the same pipeline described as that used to generate Figure 2.3B on the residuals without further adjustment for covariates in the GWAS or prediction evaluation. Thus, this figure relates more directly to the analytical derivation in section “Picking the sample size of the standard GWAS sample to match the prediction accuracy of the score based on the sib-GWAS”. However, the results in Figure 2.3B and here are qualitatively similar. Small points show the ratio of the prediction accuracies in the two designs across 10 iterations; in each iteration, we resample sets of unrelated individuals to constitute three sets: for discovery, estimation and prediction. Large points show mean values.
Future directions

In Chapter 1, I presented results on viability selection using the two large data sets, the GERA cohort and the UK Biobank, testing whether the frequency of an allele (or set of alleles identified in GWAS to influence variation in diseases and anthropomorphic traits) differs across individuals born at different times accounting for variation in ancestry and batch effects. Although these cohorts include a wide range of ages, they often lack younger than middle-aged individuals, which are of particular importance to detect variants influencing survival early in life. Incorporating young individuals recruited for studies of childhood diseases or longitudinal studies, such as MSSNG \([177]\) and ALSPAC \([178]\) data sets, could close this gap.

Another complementary approach is to analyze nuclear families to investigate the viability effect of variants during early development and embryogenesis. These can be identified, for instance, by looking at transmission distortion in parent-offspring trios: for a bi-allelic SNP, an advantageous allele will appear over-transmitted (compared to the other allele) from the parents to the offspring. To this end, a transmission distortion test (TDT) could be applied to human pedigrees \([179]\), such as Framingham Heart Study \([180]\) (e.g., Meyer et al. \([117]\)). This approach will become more powerful
with the collection of more family-based data in the future; current sample sizes (e.g., \( \sim 10,000 \) trios in MSSNG) are much smaller compared to GWAS which are reaching sample sizes on the scale of a million.

This work is one of the first to utilize the genetic data that are now available through biomedical studies, in order to directly observe the ongoing evolutionary change in contemporary humans. The application of approaches such as ours to the millions of samples in the pipeline (such as the Precision Medicine Initiative program [114]) will provide a more comprehensive picture of viability selection. Moving forward, studies should consider other components of fitness as well, fertility and inclusive fitness, ideally with large scale multi-generational longitudinal designs to tract some of the long-standing questions in human evolutionary genetics, such as: How, and to what extent, does natural selection shape common genetic variation? Such a goal is reminiscent of Lewontin’s suggestion in 1968 for investigating natural selection on ABO blood groups [80]: “What we need to do is a brute force, unsophisticated, simple study in which you take 300,000 or 400,000 … offspring at birth, blood-type them for ABO, … keep track of them until they get to be about 35 and record when each of them dies, if he does, and when each of them has an offspring.”

One thing Lewontin did not foresee is the importance of subtle population structure, among other challenges involved with the analyses enabled by large-scale data sets.

In Chapter 2, I discussed underappreciated factors influencing the portability of polygenic scores (PGS), which are increasingly being used in diverse disciplines such as human genetics, social sciences and evolutionary biology. Previous studies have mainly focused on the lack of portability across genetic ancestry groups, largely at-
tributing this limitation to distinct demographic histories. However, we show that in addition to these considerations, factors such as sample characteristics (e.g., age, sex, socioeconomic status) and GWAS study design contribute to the prediction accuracy of PGS, and consequently to their portability between as well as within ancestry groups. Given these increasing uses, there is a pressing need to quantify the relative contribution of these factors to PGS prediction accuracy and determine the axes along which PGS are more or less portable.

One direction is to investigate whether, and by how much, portability across ancestries is improved if PGS are based on fine-mapped loci (i.e., loci determined as causal within a genomic region detected in GWAS, e.g., \[181\]) compared to when based on tag SNPs alone. This effort will be informative about the contribution of linkage disequilibrium differences to portability across human populations. Another direction is to study portability across groups stratified by both ancestry and sample characteristics. Consider the example of years of schooling: in Chapter 2 we demonstrated that PGS prediction accuracy is variable within White British (WB) samples in the UK Biobank depending on their socioeconomic status (SES). This analysis can be extended to include prediction samples from other ancestries to evaluate by how much the drop in portability across ancestries is mediated by SES, e.g., by performing GWAS in low and high SES WB individuals and making prediction in low and high SES WB as well as low and high SES East Asian ancestry individuals. Lastly, portability can be evaluated when PGS use direct genetic effects only, dissected using family-based approaches (such as sibling-GWAS presented here).

In summary, for GWAS results to be reliably used for the study of polygenic adapta-
tion and phenotypic prediction will require novel approaches to characterize different sources of bias in GWAS that contribute to these analyses (e.g., [33, 162]). Expanding efforts to collect large-scale data bases from related individuals would be particularly useful in this regard.
Bibliography


[97] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. “Prioritizing GWAS results: a review of statistical methods and recommendations for their appli-


[160] Jennifer Listgarten, Christoph Lippert, and David Heckerman. “FaST-LMM-Select for addressing confounding from spatial structure and rare variants.” In: Nature Genetics 45.5 (2013), pp. 470–471. ISSN: 1061-4036. DOI: 10.1038/ng.2620 URL: http://dx.doi.org/10.1038/ng.2620.


Appendix

In this section I list the works that I contributed as a co-author and/or were not directly relevant to the topic of this thesis to be presented as a chapter.

- **Entropic forces drive self-organization and membrane fusion by SNARE proteins**

  (Published under Mostafavi et al. PNAS 2017)

  **Abstract**

  SNARE proteins are the core of the cell’s fusion machinery and mediate virtually all known intracellular membrane fusion reactions on which exocytosis and trafficking depend. Fusion is catalyzed when vesicle-associated v-SNAREs form trans-SNARE complexes (“SNAREpins”) with target membrane-associated t-SNAREs, a zippering-like process releasing $\sim 65$ kT per SNAREpin. Fusion requires several SNAREpins, but how they cooperate is unknown and reports of the number required vary widely. To capture the collective behavior on the long timescales of fusion, we developed a highly coarse-grained model that retains key biophysical SNARE properties such as the zippering energy landscape.
and the surface charge distribution. In simulations the \(\sim 65\text{-}kT\) zipper energy was almost entirely dissipated, with fully assembled SNARE motifs but uncomplexed linker domains. The SNAREpins self-organized into a circular cluster at the fusion site, driven by entropic forces that originate in steric–electrostatic interactions among SNAREpins and membranes. Cooperative entropic forces expanded the cluster and pulled the membranes together at the center point with high force. We find that there is no critical number of SNAREs required for fusion, but instead the fusion rate increases rapidly with the number of SNAREpins due to increasing entropic forces. We hypothesize that this principle finds physiological use to boost fusion rates to meet the demanding timescales of neurotransmission, exploiting the large number of v-SNAREs available in synaptic vesicles. Once in an unfettered cluster, we estimate \(\geq 15\) SNAREpins are required for fusion within the \(\sim 1\text{-}ms\) timescale of neurotransmitter release.

This is my work at Columbia’s Chemical Engineering department before joining the Przeworski lab.

- **Reduced signal for polygenic adaptation of height in UK Biobank**

  (Published under Berg et al. eLife 2019)

  Abstract

  Several recent papers have reported strong signals of selection on European polygenic height scores. These analyses used height effect estimates from the
GIANT consortium and replication studies. Here, we describe a new analysis based on the the UK Biobank (UKB), a large, independent dataset. We find that the signals of selection using UKB effect estimates are strongly attenuated or absent. We also provide evidence that previous analyses were confounded by population stratification. Therefore, the conclusion of strong polygenic adaptation now lacks support. Moreover, these discrepancies highlight (1) that methods for correcting for population stratification in GWAS may not always be sufficient for polygenic trait analyses, and (2) that claims of differences in polygenic scores between populations should be treated with caution until these issues are better understood.

My main contribution to this work was to perform a sibling-based GWAS for height, as well as to curate sets of independent height-associated SNPs for polygenic score analyses.

- Measuring intolerance to mutation in human genetics

(Published under Fuller et al. Nat. Genet. 2019)

Abstract

In numerous applications, from working with animal models to mapping the genetic basis of human disease susceptibility, knowing whether a single disrupting mutation in a gene is likely to be deleterious is useful. With this goal in mind, a number of measures have been developed to identify genes in which
protein-truncating variants (PTVs), or other types of mutations, are absent or kept at very low frequency in large population samples—genes that appear ‘intolerant’ to mutation. One measure in particular, the probability of being loss-of-function intolerant (pLI), has been widely adopted. This measure was designed to classify genes into three categories, null, recessive and haploinsufficient, on the basis of the contrast between observed and expected numbers of PTVs. Such population-genetic approaches can be useful in many applications. As we clarify, however, they reflect the strength of selection acting on heterozygotes and not dominance or haploinsufficiency.

In this perspective paper, I contributed to the structure and writing of the piece, as well as to the study design applying the pLI measure to simulated data.