



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Inbreeding depression on human complex traits

David W. Clark



PhD Thesis
The University of Edinburgh
2021



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g., PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Declaration

I hereby declare that:

- I composed this thesis entirely on my own.
- The work presented in the thesis is my own, and for the elements of the thesis that were a result of collaboration, I declare that I have made a substantial contribution to the work, and I have clearly indicated this contribution below.
- This work has not been submitted for any other degree or professional qualification except as specified.

Edinburgh, August 2021.

David W. Clark

Collaboration

Some of the results in this thesis are drawn from data collected by consortium of collaborating researchers (called ROHgen). With assistance from Professor Jim Wilson and Dr Peter Joshi, I designed analyses to be executed by members of the ROHgen consortium. I personally implemented these analyses in software which was distributed to members of ROHgen. I played no part in gathering individual sample data, nor in executing the analysis software (except for the cohorts specified in Supplementary Table 1). Summary statistics from each ROHgen cohort were returned to me in Edinburgh, and I meta-analysed these cohort contributions. Many of the results from ROHgen have been previously published in *Nature Communications* as “Associations of autozygosity with a broad range of human phenotypes”¹ by Clark et al. A few paragraphs in this thesis are reproduced verbatim from that article, but I alone composed those words and figures.

Acknowledgements

I would like to thank:

- All those who contributed their personal data to the consented studies which are the foundation of this thesis.
- All researchers involved in the creation of UK Biobank and the many ROHgen cohorts. I would particularly like to thank their generosity in making this data available to an unpublished PhD student.
- The many analysts who ran the ROHgen software, and patiently worked with me when things went wrong. I would particularly like to thank Yuki Okada at Biobank Japan, Dan Mason at Born in Bradford, and Kristján Helgi Swerford Moore at deCODE who ran several iterations of the software.
- All those who generously shared their knowledge, experience and wisdom when discussing this research. To name but a few: Nicola Pirastu, Andrew Bretherick, Tõnu Esko, Ruth McQuillan, Vera Kaiser, Brian Charlesworth, Kristján Helgi Swerford Moore, and Agnar Helgason.
- Peter Joshi, whose own PhD formed the springboard for this thesis, and whose intelligence and enthusiasm for knowledge are an inspiration to his colleagues.
- The Medical Research Council for funding my scholarship.
- Philippa Saunders, Susan Farrington, Martin Taylor, Kate Hardman and everyone involved in Precision Medicine Doctoral Training Program scheme.
- Martin Taylor and Harry Campbell for generously giving their time to guide my thesis committee.
- My supervisors Chris Haley and Jim Wilson for their support, advice and encouragement throughout this PhD. In particular, I thank Jim Wilson for patiently keeping me on track, despite my many distractions and diversions.
- All my friends and family, but especially my wonderfully supportive parents and beloved partner, Vera.

Table of Contents

Declaration	iii
Acknowledgements	iv
Table of Contents	v
Abbreviations	ix
Abstract	x
Lay summary	xi
1 Introduction	12
1.1 Key concepts	12
1.1.1 Autozygosity	12
1.1.2 Consanguinity, endogamy and small effective population size	13
1.1.3 Inbreeding coefficients (F)	13
1.1.4 Inbreeding load	14
1.2 Theoretical framework of inbreeding depression	15
1.2.1 Overdominance versus partial dominance	15
1.2.2 Partial dominance theory	16
1.3 Empirical evidence of inbreeding depression	19
1.3.1 Model organisms	19
1.3.2 Wild populations	21
1.3.3 Domesticated plants, pets, and livestock	22
1.3.4 Humans	23
1.4 Study aims	24
2 Source data description	25
2.1 Introduction	25
2.1.1 Estimating Inbreeding coefficients	25
2.2 Genetic data	28
2.2.1 ROHgen	28
2.2.2 UK Biobank	30
2.3 Methods	31
2.3.1 ROH calling	31
2.3.2 Calculating estimators of the Inbreeding Coefficient (F)	34
2.3.3 Population structure and plotting F_{ROH} against F_{IS}	36
2.3.4 Statistical power contributed by individual samples	37
2.4 ROH Simulation methods	37
2.4.1 Simple ROH simulation method	37
2.4.2 Full haplotype inheritance ROH simulation method	40
2.5 Results	42

2.5.1	Plotting F_{ROH} against F_{IS} for ROHgen cohorts	42
2.5.2	Mean F_{ROH} by country of birth in UK Biobank.....	44
2.5.3	Distribution of F_{ROH} in selected ROHgen cohorts.....	44
2.5.4	Statistical power of samples as a function of F_{ROH} in UK Biobank	46
2.5.5	Genomic distribution of ROH in UK Biobank	48
2.6	ROH simulation results	55
2.6.1	Estimating ROH age from ROH length	55
2.6.2	Simulated F_{ROH} for different relative types	58
2.6.3	Simulating multiple generations of cousin marriage.....	60
2.6.4	Correlation between siblings of first cousin parents.....	62
3	Associations between inbreeding coefficients (F) and traits	64
3.1	Introduction.....	64
3.2	Method	67
3.2.1	Phenotypes.....	67
3.2.2	Overlap between ROHgen and UKB.....	72
3.2.3	Estimating $\beta_{F_{ROH}}$ for quantitative traits	73
3.2.4	Binary phenotypes	76
3.2.5	Covariates.....	77
3.2.6	Reducing bias in statistical tests	79
3.2.7	Multivariate trait.....	83
3.2.8	Sex-specific estimates of $\beta_{F_{ROH}}$	84
3.3	Results	85
3.3.1	ROHgen traits associated with F_{ROH}	85
3.3.2	Survey of UK Biobank traits associated with F_{ROH}	86
3.3.3	ICD10 codes associated with F_{ROH} in UK Biobank	87
3.3.4	Self-reported disease codes associated with F_{ROH} in UK Biobank.....	89
3.3.5	Combined UKB results.....	89
3.3.6	Sex-specific $\beta_{F_{ROH}}$	95
3.3.7	Comparison of different estimators of F	97
3.3.8	Effect of different ROH lengths.....	101
3.4	Discussion	104
3.4.2	Comparing estimates of F	106
3.4.3	Sex-specific effects	108
3.4.4	Effects of different ROH lengths	109
4	Non-genetic confounding of F_{ROH}	111
4.1	Introduction.....	111
4.1.1	Confounding.....	112
4.1.2	Assortative mating	116

4.1.3	Introduction summary	117
4.2	Methods	117
4.2.1	Relationship between F_{ROH} and early-life SES	117
4.2.2	Married couples in UKB.....	118
4.2.3	Combining the effects of multiple confounders	119
4.2.4	F_{ROH} -based methods to quantify the role of confounding	121
4.2.5	Comparing $\beta_{F_{ROH}}$ to the effect of hemizygous deletions	132
4.2.6	Comparing effect estimates from different methods.....	132
4.2.7	Testing confounding-free methods	134
4.3	Results	136
4.3.1	Relationships between F_{ROH} and confounders in UKB	136
4.3.2	Accounting for confounders in $\beta_{F_{ROH}}$	144
4.3.3	Comparing $\beta_{F_{ROH}}$ to effect of hemizygous deletions	157
4.4	Discussion	162
4.4.1	Introduction	162
4.4.2	Relationships between F_{ROH} and possible confounders.....	162
4.4.3	Effects of F_{ROH} resemble effects of suspected confounders	164
4.4.4	Constancy of F_{ROH} effects between different populations.....	165
4.4.5	Linearity of effects with F_{ROH}	166
4.4.6	Methods to reduce confounding in $\beta_{F_{ROH}}$	166
4.4.7	Effects of F_{ROH} resemble effects of hemizygous deletions.....	170
4.4.8	Discussion summary.....	171
5	Effects of ROH in specific genomic locations.....	172
5.1	Introduction.....	172
5.2	Methods	174
5.2.1	Estimating local effects of ROH in 3 Mb wide windows	174
5.2.2	Summarising genomic annotations in 3 Mb-wide windows.....	176
5.2.3	Effect of ROH on each chromosome.....	178
5.2.4	Calculating F_{ROH} at specific genomic annotations	178
5.2.5	Simulation of inbreeding depression at known causal loci	182
5.3	Results	184
5.3.1	ROH effects in specific 3 Mb wide windows	184
5.3.2	Associations between window effects and genomic features	195
5.3.3	Distribution of ROH effects in UK Biobank.....	199
5.3.4	F_{ROH} calculated at specific genomic annotations	204
5.4	Discussion	212
5.4.1	ROH effects at specific genomic locations	212
5.4.2	Genome wide distribution of inbreeding depression	213

6	Conclusions	217
6.1	Main findings	217
6.1.1	Cause or Confounding?	217
6.1.2	Complex traits affected by inbreeding.....	218
6.1.3	Genetic architecture of inbreeding depression	220
6.2	Future directions	221
6.2.1	Sibling studies.....	221
6.2.2	ROH calling methods.....	222
6.2.3	Whole genome sequence	223
6.2.4	Trait variance.....	224
6.2.5	Unexplained results from this study	224
6.2.6	Recommendations for future studies	225
6.3	Social context	226
6.4	Final words	226
	References.....	227
Appendix A	Supplementary Figures.....	239
Appendix B	Supplementary Table legends	243
Appendix C	Comparison of inbreeding coefficient estimates	248
Appendix D	Interpretation of $Trait \sim F_{ROH} + F_{GRM}$ models.....	256
Appendix E	Proofs.....	260
Appendix F	Traits affected by F_{ROH}	265
Appendix G	Simulation R script.....	268
Appendix H	Other publications.....	270

Abbreviations

CI	Confidence interval
cn-LoH	Copy-neutral Loss of Heterozygosity
CNV	Copy Number Variant
DNA	Deoxyribonucleic Acid
EA	Educational Attainment
FEV1	Forced Expiratory Volume in one second
GRM	Genomic Relationship Matrix
GWAS	Genome-Wide Association Study
HDL	High-Density Lipoprotein
HWE	Hardy Weinberg Equilibrium
ID	Inbreeding Depression
IMD	Index of Multiple Deprivation
LD	Linkage Disequilibrium
LDL	Low-Density Lipoprotein
MAF	Minor Allele Frequency
pLI	predicted Loss of function Intolerant
pLoF	Putative Loss of Function
PTV	Protein Truncating Variant
PV	Polycythaemia Vera
ROH	Run Of Homozygosity
SD	Standard Deviation
SE	Standard Error
SES	Socio-Economic Status
SNP	Single Nucleotide Polymorphism
UPD	Uniparental Disomy
UKB	UK Biobank
WGS	Whole Genome Sequence

Abstract

In many species the offspring of related parents suffer reduced survival, vigour and reproductive success – a phenomenon known in evolutionary genetics as inbreeding depression. Historically, inbreeding depression has been experimentally studied in model organisms and wild populations, but the huge quantities of human genetic and phenotype data being currently gathered provides new opportunities. In humans, it is well known that reproduction between close relatives (consanguinity) increases the risk of recessive Mendelian disorders, but the extent of inbreeding depression on complex traits remains less well-established – partly because effects are small, and consanguinity is often confounded by social and cultural factors. This thesis explores the characteristics of human inbreeding, identified in genetic microarray data as continuous Runs of Homozygous genotypes (ROH), in a worldwide consortium (ROHgen) with 1,246,000 genotyped samples, and in the richly phenotyped UK Biobank (UKB). In ROHgen, an estimated inbreeding coefficient (F_{ROH}) is associated with significant changes in 18 out of 49 complex traits investigated, and in UKB F_{ROH} is associated with at least 155 complex traits. Effect sizes are small, but almost exclusively deleterious to health and reproductive success, and appear to be caused by the increased homozygosity of rare, recessive variants within ROH segments. Importantly, a novel, within-siblings method is introduced which, with larger sample sizes, will eliminate any biases caused by social confounding. For now, sibling sample sizes are sufficient to determine that most complex traits associated with F_{ROH} have a substantial genetic component. The distribution of inbreeding depression effects throughout the human genome was investigated and found to be remarkably even with, intriguingly, no enrichment at protein-coding genes. This is consistent with the theory of mutation-selection balance where deleterious variants in less constrained regions reach higher allele frequencies. In summary, inbreeding depression occurs in modern humans and manifests as small deleterious effects on a multitude of human complex traits. These effects appear to be caused by numerous rare, recessive genetic variants evenly distributed throughout the genome.

Lay summary

Humans, like almost all animals, have two copies of their genome – the DNA instruction manual from which our bodies are built and operated. Each of these copies contains millions of different variations, which are vital to allow our species to adapt to changing environments, but can also cause problems to us as individuals. It is these variations which explain many of the differences between us, in visible features such as height, and intelligence, but also how likely we are to suffer from common diseases, such as diabetes and cancer. Fortunately, the most harmful variations are uncommon, and rarely found on both copies of our DNA. In most cases, we escape their worst effects, by relying on the one good copy. However, in the children of related parents, the copy may also be impaired if inherited from the same ancestor. This thesis investigates the effects of having identical variations on both copies of our DNA. We found that identical sections cause small changes in many commonly measured traits, and that these changes are almost always harmful to health, or to the chances of having children. While the effects are largest in people whose parents are closely related, they are relevant to everyone, since all parents are related to some degree, and we all carry small identical sections of DNA. Intriguingly, the effects are not concentrated in genes already identified by other types of genetic study, indicating that our understanding of the consequences of genetic variation remains incomplete.

1 Introduction

In many species the offspring of related parents suffer reduced vigour, survival and reproductive success - a phenomenon known as inbreeding depression. Since it was first described by Charles Darwin², inbreeding depression has remained an important concept in evolutionary genetics, as it is presumed to explain the preference for outcrossing mating systems in almost all species³. However, despite insights from both theory and empirical studies, many questions about the genetic causes of inbreeding depression remain unanswered. Furthermore, in humans, the importance (or even the existence) of inbreeding depression has been questioned, partly because reproduction between close relatives is both rare in many cultures and frequently associated with confounding social factors⁴. This thesis aims to determine whether inbreeding depression currently affects human complex traits, despite the enormous social and demographic changes our species has experienced since its evolutionary origins. Subsequently, the genetic architecture of inbreeding depression is investigated using the large quantities of human genetic and phenotypic data which are becoming available.

1.1 Key concepts

1.1.1 Autozygosity

Inbreeding increases the homozygosity of genetic variants because identical copies of an ancestral variant may be inherited on both the maternal and paternal allele. When a variant is made homozygous by inbreeding, it is described as autozygous, or identical-by-decent (ibd), to distinguish from the chance homozygosity of alleles inherited from unrelated parents, described as identity-by-state (ibs). This is an arbitrary, and occasionally confusing distinction, since all parents are related to some degree, and sufficiently small homologous segments will always trace back to a common ancestor some distant number of generations ago (chromosomal rearrangements notwithstanding). In practice, identity-by-decent is usually defined relative to an arbitrary (and often unspecified) base generation in which all alleles are assumed to be distinct.

1.1.2 Consanguinity, endogamy and small effective population size

Consanguinity refers to reproduction between close relatives. The definition of close relatives is arbitrary (and may differ by author) although in clinical genetics, consanguinity usually refers to reproduction between second cousins, or closer⁵. However, autozygosity may also be caused by more distant common ancestry, and this becomes more likely when the number of distinct ancestors is much smaller than the theoretical maximum (2^N , where N is the number of generations). This situation can be summarised as a small effective population size, and may be caused by either physical or social barriers. For example, the anabaptist religions of the USA (e.g., Hutterites, Amish) have high average autozygosity traceable to a small number of founder individuals⁶. Similarly, despite geographic dispersal, Ashkenazi Jews have high average rates of autozygosity attributed to endogamy, the practice of marrying within a social group⁷. Finally, autozygosity increases from west to east in the Polynesian islands due to successive bottlenecks attributed to repeated founder events⁸.

1.1.3 Inbreeding coefficients (F)

Whatever the demographic cause, the excess homozygosity caused by inbreeding can be conveniently summarised in a single genomic metric, the inbreeding coefficient (F). Although originally defined as the correlation between genotypes^{9,10}, the inbreeding coefficient of a diploid organism is today most often defined “*the probability that two alleles at any locus are identical by descent from a common ancestor of the two parents*”.

However, the term inbreeding is used to describe several related facets of reproduction between relatives, which can cause some confusion¹¹. Inbreeding is always a relative measure, with different usages of the word corresponding to different reference populations. The definition given above may be described as “pedigree inbreeding” where the reference population is a base generation, or the start of a genealogy. In other contexts, inbreeding may refer to a situation where reproducing couples are more closely related than two random individuals selected

from the population. In this usage of the word (inbreeding as non-random mating), the current generation is the implicit reference population. Hence, with a small effective population size, it is possible to have significant “pedigree inbreeding” despite random mating (see also sections 2.3.3 and 2.5.1).

1.1.4 Inbreeding load

Theoretical models predict that average fitness will decline with increasing inbreeding coefficient (see section 1.2), and the magnitude of inbreeding depression can be described by the gradient of this decline¹². Most often, in wild populations or model organisms, the traits studied are components of fitness (e.g., survival or fertility), where a value of zero implies zero fitness, and where effects are assumed to be multiplicative. In this situation, the logarithm of the trait is expected to decline linearly with F , and the gradient of this regression, called the inbreeding load (B), can be used to compare the magnitude of inbreeding depression across different traits and species (Figure 1).

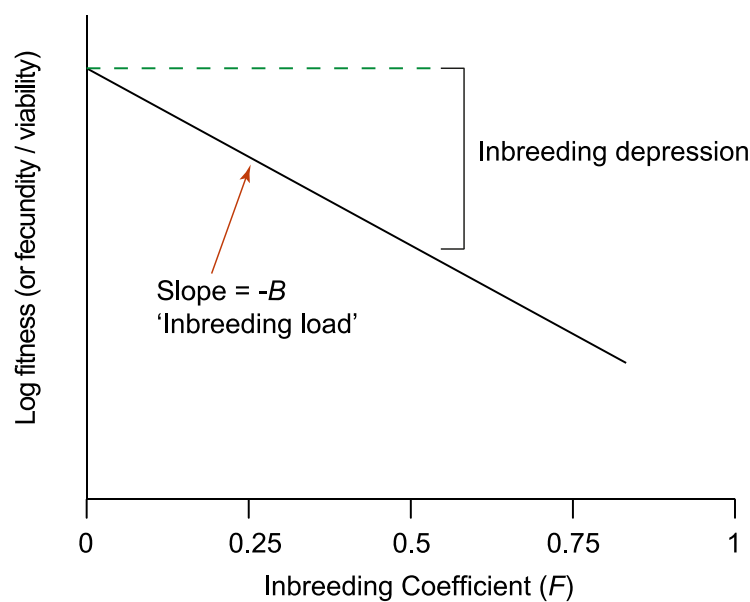


Figure 1: **Estimating inbreeding load (B) by regression.** Figure reproduced from “Inbreeding effects in wild populations” by Keller and Waller¹¹.

However, when the studied traits are not components of fitness (say, height or cognition, in humans) the inbreeding load has little meaning, and effects are as likely to be additive as multiplicative. In this situation it is more customary to report

the linear effects of F , sometimes normalised by the standard deviation of the trait¹³. This latter approach is the one adopted in this thesis.

1.2 Theoretical framework of inbreeding depression

1.2.1 Overdominance versus partial dominance

Two main theories have been proposed to explain how increased homozygosity causes inbreeding depression: the overdominance theory and the partial dominance theory¹⁴. The overdominance theory postulates that, at most causal loci, heterozygotes have higher evolutionary fitness than either homozygote, while the partial dominance theory suggests that most causal loci are partially recessive, and heterozygotes have intermediate fitness (Figure 2).

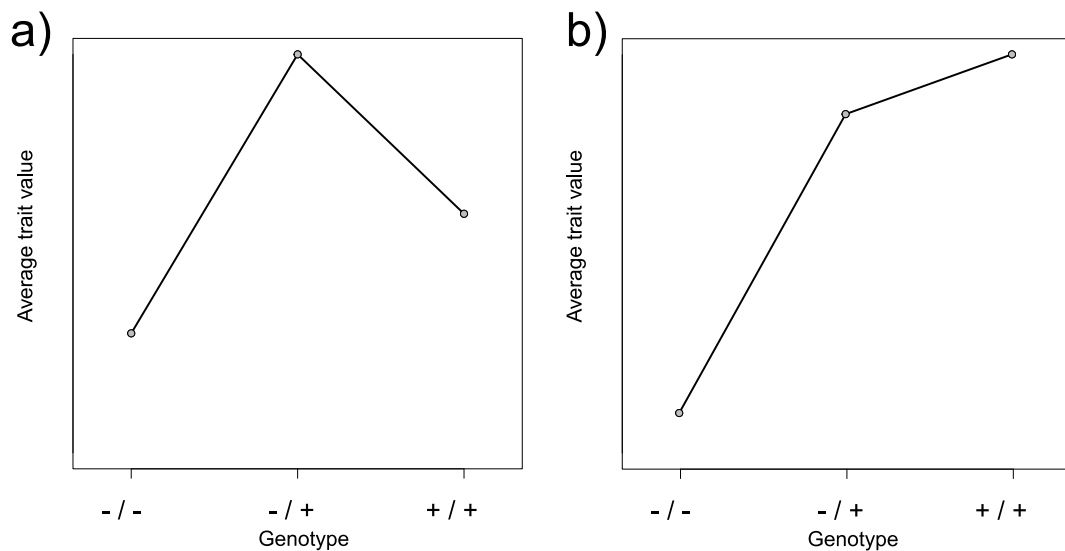


Figure 2: **Comparison of overdominance and partial dominance.** a) At overdominant loci heterozygotes have highest trait values. b) At partially dominant loci the wild-type homozygote (+/+) has the highest trait value.

In both cases, increasing homozygosity reduces average fitness, but the two theories make different predictions about the maintenance of genetic variance of fitness, and the allele frequencies of causal variants. In overdominance, the highest population fitness is achieved if heterozygotes are common, which requires intermediate frequencies of both the wild-type and alternative alleles. In contrast, with partial dominance, the highest population fitness is achieved when the less fit allele is purged from the population. In practice, both types of loci may contribute

to inbreeding depression, although the majority of empirical evidence appears to support the partial dominance theory³.

1.2.1.1 Associative overdominance

It is also worth noting that even if all causal loci are partially recessive, heterozygous markers in linkage disequilibrium with the causal loci may have an apparent selective advantage, through a phenomenon called associative overdominance¹⁵. For example, if two haplotypes segregating in a population carry different (potentially numerous) recessive variants, individuals heterozygous for the haplotype may be fitter than those who are homozygous for either haplotype, despite all the causal loci being recessive.

1.2.2 Partial dominance theory

Dominance (or partial dominance) at many recessive loci is believed explain a large fraction of inbreeding depression³, and mutation-selection balance provides a theoretical framework for understanding this hypothesis¹².

1.2.2.1 Mutation-selection balance

The simplest model of mutation-selection balance concerns mutations at a single autosomal site in an infinitely large, randomly mating, diploid population¹². In this model, mutation from a common wild-type allele to a deleterious variant is assumed to occur at a mutation rate u . The variant allele is assumed to be deleterious and have fitness $1 - hs$ in a heterozygous state, and $1 - s$ when homozygous. The wild-type homozygote has fitness 1, s is the selection coefficient, and h is the coefficient of dominance.

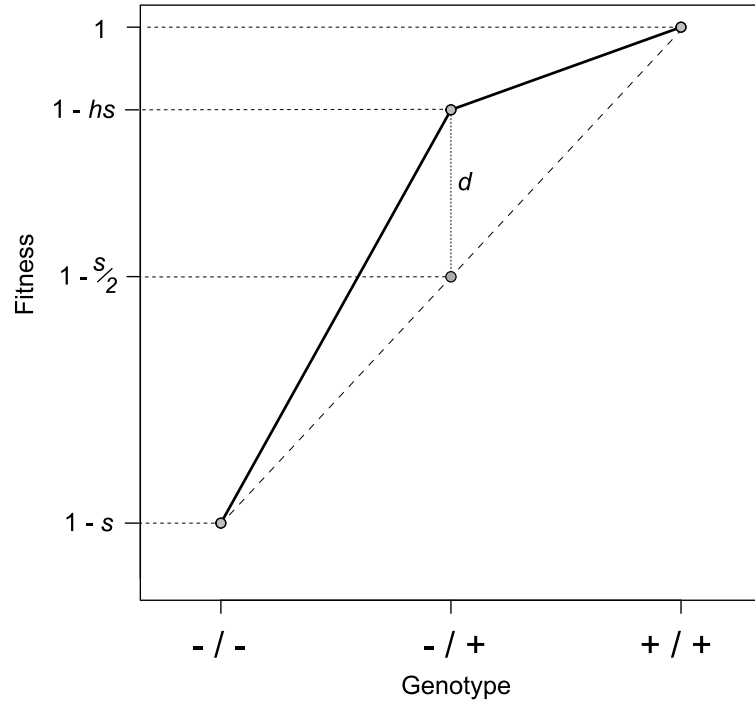


Figure 3: **Genotype fitness in mutation-selection balance model.** The wild-type allele is represented by a plus (+) and the mutant allele by a minus (-).

With a few simplifying assumptions it can be shown that the variant allele is expected to reach an equilibrium frequency (q^*) equal to

$$q^* \approx \frac{u}{hs} \quad (1)$$

Hence, for a given mutation rate (u) and selection coefficient (s), recessive variants (with $h < 0.5$) are expected to reach higher equilibrium frequencies than additive ($h = 0.5$) or dominant ($h > 0.5$) variants.

Furthermore, the magnitude of inbreeding depression at a biallelic locus can be expressed as

$$ID = 2pqdF \quad (2)$$

Where p and q are the minor and major allele frequencies, d is the heterozygote value (relative the mid-point of the two homozygote values), and F is the inbreeding coefficient¹⁶.

Substituting equation (1) into (2), assuming q^* is small, and expressing d in terms of h and s , gives

$$ID = \frac{u(1 - 2h)F}{h} \quad (3)$$

I.e., inbreeding depression at a locus in mutation-selection balance depends on the mutation rate (u), the inbreeding coefficient (F), and the inverse of the dominance coefficient (h). As expected, additive variants ($h = 0.5$) do not contribute any inbreeding depression, but increasingly recessive variants contribute increasingly important amounts of inbreeding depression.

It is interesting to note that the magnitude of inbreeding depression in equation (3) does not depend on the strength of selection (s). This can be understood because inbreeding depression is the product of allele frequency and effect size, and highly deleterious variants reach lower equilibrium frequencies which exactly balance their larger effects.

1.2.2.2 *Evolution of dominance*

The two of the most influential geneticists of the 20th century, Ronald Fisher and Sewall Wright, both tried to explain why a large majority of mutations are either partially or completely recessive to their wild-type counterparts¹⁷. Fisher suggested that most deleterious mutations are codominant (additive) when they first arise, but that selection on modifier mutations causes the heterozygote value to gradually approach the fitter, wild-type value¹⁸. Wright strongly objected to this explanation, arguing that selection of dominance modifiers would not be strong enough to overcome genetic drift, nor the pleiotropic effects of the putative dominance modifier¹⁹. Instead, Wright suggested that dominance was an inherent property of metabolic pathways²⁰ – a view which has subsequently gained favour¹⁷. In support of Wright's theory, Kacser and Burns demonstrated that the connected nature of enzymatic pathways is sufficient to explain the recessive effects of individual enzyme activity on higher-level phenotypes²¹. Furthermore, Fisher's theory suffered blows from two observations by which cannot be explained by selection of dominance modifiers. Specifically, Charlesworth and Charlesworth noted that,

in *Drosophila*, dominance coefficients are inversely correlated with selection coefficients²² – a finding which cannot be explained by Fisher’s theory.

Furthermore, Orr observed that mutations which occur in the haploid organism, *Chlamydomonas reinhardtii*, are almost entirely recessive in artificially produced diploids, despite no opportunity for selection on heterozygotes in the natural state²³.

In summary, the modern consensus is that dominance is an inherent property of diploid organisms. While this is demonstrable true, it seems an unsatisfactory explanation for evolutionary genetics, as it seems likely that this inherent property has arisen by selection.

1.3 Empirical evidence of inbreeding depression

1.3.1 Model organisms

Because inbreeding is rare in most wild populations, a large majority of important results have been established in model organisms, which can be easily mated with close relatives in laboratory experiments. In particular, the fruit fly, *Drosophila melanogaster*, offers many advantages to the experimental geneticist²⁴ (e.g., short reproductive cycle, inexpensive to maintain, known balancer chromosomes) and has been extensively studied over the past 120 years²⁵. The existence of inbreeding depression in *Drosophila* is well established²⁶, and most recent research has focused on understanding the genetic architecture of causal variants. In contrast, relatively little attention had been paid to which *Drosophila* traits are affected, beyond bristle number²⁷, egg-to-adult viability²⁸ and adult survivorship²⁹. A much smaller number of studies exploring how inbreeding manifests in observable traits have been performed in other model organisms, notably the guppy³⁰ (*Poecilia reticulata*) and the zebra finch³¹ (*Taeniopygia guttata*).

1.3.1.1 *Drosophila melanogaster*

The discovery of *Drosophila* balancer chromosomes was particularly important for the experimental study of inbreeding depression, and much of what is currently known about the genetic architecture of inbreeding depression can be attributed to

these experiments³². Balancer chromosomes have numerous inversions and rearrangements which suppress meiotic recombination with their homologs, and can be used to maintain variants that would be highly deleterious in a homozygous state. For example, experiments using balancer chromosomes to make entire chromosomes homozygous, have shown that as many as 30% of naturally occurring *Drosophila* chromosomes are lethal when made homozygous. The homozygous lethal mutations map to many different loci and are almost always viable in a heterozygous state, consistent with the predictions of partial dominance and mutation-selection balance. In addition, even viable homozygous chromosomes have been shown to substantially reduce whole lifetime fitness, indicating that relatively minor deleterious mutations are recessive and contribute substantially to inbreeding depression. In contrast, X chromosomes from wild *Drosophila* populations rarely carry homozygous lethal mutations, consistent with the prediction that such mutations would be purged from the population in the hemizygous state which occurs in males³³.

Further studies in *Drosophila* have demonstrated that homozygous lethal mutations typically have very small viability effects as heterozygotes, implying dominance coefficients of the order of a few percent^{26,34}. In contrast, mildly deleterious *Drosophila* mutations are estimated to have much larger dominance coefficients³⁵, perhaps between 0.2 and 0.35^{26,34}. This relationship between h and s is not predicted by Fisher's theory of selection for dominance, and implies that mutations of large effect are generally expected to be more recessive than those of small effect.

1.3.1.2 *Guppies*

A number of inbreeding studies have been conducted on the guppy fish, *Poecilia reticulata*. The guppy is a popular model organism in all manner of biological research, especially studies of sexual selection since female mate preference depends on the unique pattern of colourful spots and stripes carried by sexually mature males³⁶. The guppy is both genetically and phenotypically highly variable, and severe inbreeding depression has been reported in both wild³⁷ and laboratory

populations^{30,38–41}. Many guppy traits have been reported to be affected by inbreeding, but amongst the most interesting, van Oosterhout et al. reported that sib-mating produced guppies with fewer, smaller colour spots, which are consequently less attractive to females³⁰. Sib-mated guppies also show a strong decline in male sexual motivation and mating success^{38,41}. Other guppy traits affected by inbreeding include: sperm competitiveness and number⁴², juvenile survival³⁹, saline tolerance³⁹, male development rate⁴⁰, female size⁴³, and parasite loads⁴⁴. In summary, in the guppy, there is substantial evidence that inbreeding is harmful to both survival and reproductive success. However, inbreeding in these studies is often substantially higher ($F = 0.25$, or even $F = 0.5$) than common in natural populations, and it is not established whether these effects would be proportional at lower inbreeding coefficients.

1.3.2 Wild populations

Many species have elaborate mechanisms of inbreeding avoidance (ostensibly to avoid inbreeding depression) and consequently inbreeding is typically rare, and difficult to observe, in wild populations⁴⁵. Indeed, as recently as the 1990s, some authors questioned the importance of inbreeding depression in the wild, arguing that frequent bottlenecks would purge deleterious recessive mutations, and that inbreeding avoidance mechanisms would nullify the effect of inbreeding depression^{46,47}. However, several important reviews of inbreeding in wild populations make it clear that inbreeding depression does occur in essentially all well-studied, outbreeding plants and animals^{11,48}. In self-pollinating plants, and the few animal species which routinely mate with close relatives (e.g., naked mole rat, social insects), inbreeding depression appears to be lower due to purging, although deleterious effects are occasionally reported⁴⁹.

Neaves et al. recently summarized the relevant scientific literature, and provide a searchable database of 703 distinct studies of the fitness consequences of inbreeding in natural populations⁵⁰. A large majority of these studies (67%) were conducted in plants, while only 5% are studies of mammals. When inbreeding depression is reported in wild-populations, it is frequently on life-history traits with

clear connections to evolutionary fitness. For example, inbreeding adversely affects: germination rates in many plant species, hatch rate in many bird species, winter survival in sheep and deer, etc.⁵⁰ It is not clear whether the relative scarcity of associations reported with non-life-history traits is due to their absence, or the absence of investigation. Examples of inbreeding depression on less-clearly fitness-related traits include: body size in several mammals⁵¹, and resistance to disease or parasites in many species^{52,53}.

1.3.2.1 Conservation genetics

Inbreeding depression in wild populations is particularly relevant in the field of conservation. Almost by definition, endangered species suffer small effective population sizes, and the consequent increase in autozygosity which that entails. Furthermore, this problem may be particularly acute when habitat loss fragments remaining populations, and an increased expression of recessive deleterious mutations may lead to a gradual decline in population fitness and eventual extinction. A better understanding of inbreeding depression may find practical application in the management of small remaining populations.

1.3.3 Domesticated plants, pets, and livestock

A very large number of studies have considered the effects of inbreeding in domesticated plants and livestock, but these are somewhat difficult to relate to natural populations, since the genetic architecture of these species has been heavily shaped by artificial selection⁵⁴. This artificial selection frequently results in a small effective population size and consequently elevated levels of autozygosity, which may have a deleterious effect on traits, including those which were not necessarily the target of selection. Domesticated species which have been the subject of inbreeding research include: cattle, pigs, sheep, horse, chickens, dogs, cats and various forestry trees.

A recent meta-analysis of 57 livestock studies concluded that inbreeding depression is significant on many livestock traits – for example, birth weight, body weight, fecundity, offspring survival, litter weight, milk yield, fat yield and milk quality.

Interestingly, inbreeding depression is, on average, greater for production traits (which may have been targets of artificial selection) than for fitness traits. This unexpected finding emphasises that results from domesticated species may be difficult to relate to natural populations, but also highlights the relationship between selection (whether artificial or natural) and inbreeding depression.

More than in wild populations, or model organisms, studies of domesticated livestock have employed the use of genomics to estimate inbreeding coefficients, and to attempt to map causal regions. It is not the intention of this thesis to review this large body of research, other than to note that the effects of inbreeding are almost universally deleterious, and mapping has identified candidate genes for both production and fertility traits^{55,56}.

1.3.4 Humans

In humans, it is well known that autozygosity caused by consanguinity causes an increased risk of congenital birth defects and other Mendelian disorders. Since individuals with these conditions are, on average, less likely to survive and reproduce, these effects can be understood as a manifestation of inbreeding depression. However, in studies of natural populations and model organisms, only a small fraction of inbreeding depression is typically attributed to variants of large effect, and numerous, smaller, deleterious changes are thought to make a substantial contribution to inbreeding load²⁶. In humans, it is less well established to what extent inbreeding affects the numerous quantitative traits which vary continuously in healthy populations, nor to what extent these changes affect evolutionary fitness. As described in section 3.1, inbreeding effects on many human complex traits have already been reported, but this study aims to survey the full scope of affected traits, understand their connection to evolutionary fitness, and establish the genetic nature of the observed effects.

1.4 Study aims

Almost all species, including humans, avoid reproducing with close relatives - ostensibly to avoid the harmful effects of inbreeding depression. However, in humans, it is not fully established how inbreeding depression is expressed in observable traits, and whether genetic effects are detectable through the noise introduced by social confounders. This study aims to:

- 1) Estimate the effect of inbreeding on a broad catalogue of human complex traits.
- 2) Determine whether any significant effects are manifestations of inbreeding depression, or artifacts of social confounding.
- 3) Investigate the genetic architecture of inbreeding depression in humans.
- 4) Make recommendations relevant to future studies.

2 Source data description

2.1 Introduction

This chapter describes the genetic data used in this thesis to identify autozygosity, and then estimate inbreeding coefficients. The distribution of autozygosity, both genome-wide and between populations, is described here and then associations between inbreeding coefficients and various phenotypes are explored in Chapters 3, 4 and 5.

2.1.1 Estimating Inbreeding coefficients

Although it is the excess homozygosity of causal variants which gives rise to inbreeding depression⁵⁷, the location of these variants is unknown, and estimates of the inbreeding coefficient (F) are presumed to correlate well with the causal homozygosity. Many different methods have been proposed for estimating F , as methods and data availability have evolved^{58–62}. In the pre-genomic era, inbreeding coefficients were calculated from known parental relationships (F_{PED}), but gathering large family trees is time-consuming and rarely possible over more than a few generations. As it became possible to directly measure the homozygosity of genetic markers (first allozymes, then microsatellites, and more recently single-nucleotide-polymorphisms (SNPs) and whole-genome-sequence (WGS)), methods were proposed to estimate the realised autozygosity from genetic markers^{60–63}. In principle, this should be an improvement over F_{PED} , since realised autozygosity may differ from the expectation, either because of the stochastic nature of meiotic segregation, or because of incomplete, or mis-specified, pedigree data⁶⁴. In practice, a sufficient number of markers are required to ensure that the sampling error is not greater than the potential benefit^{65,66}. Genomic estimates of F are discussed in more detail in Chapter 3, but can be summarised as either marker-by-marker methods which estimate excess homozygosity, or multipoint methods which require a genetic map, and attempt to identify autozygous segments. Primarily the latter approach is used in this thesis, and is described below.

2.1.1.1 Identifying autozygous segments by Runs of Homozygosity (ROH)

Autozygous segments occur when identical copies of the same ancestral haplotype are inherited on both maternal and paternal chromosomes. This can only happen if a subject's mother and father share at least one common ancestor, thus forming a pedigree loop (Figure 4).

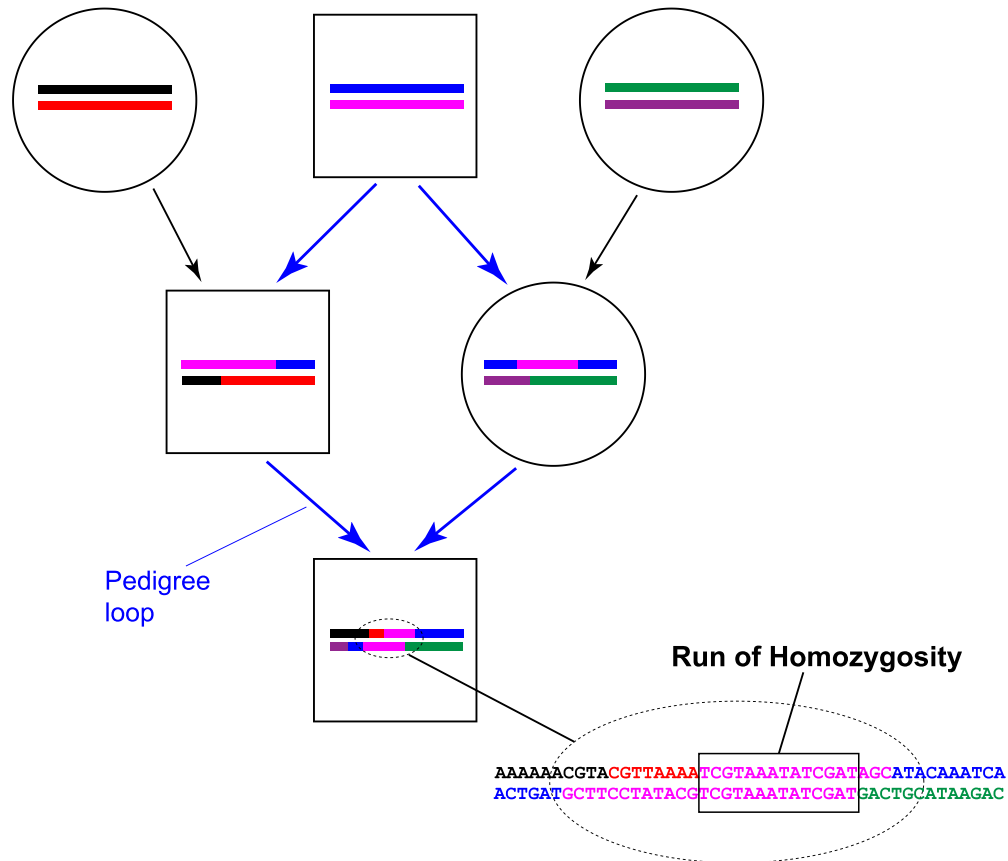


Figure 4: **Identifying autozygous segments by ROH.** Example pedigree showing possible haplotype inheritance for a subject whose parents are half-siblings. Males are represented by squares and females by circles. Haplotypes in the base generation are shown as solid colour lines, where each colour represents a unique sequence of genetic markers. Because of the pedigree loop, the pink and blue haplotypes carried by the subject's grandfather may be inherited on both the subject's maternal and paternal chromosomes. When a haplotype is inherited at the same locus on both chromosomes (the pink haplotype in this example) this can be detected as a Run of Homozygosity.

With the exception of mutations which have occurred within the pedigree loop, autozygous segments are expected to make all markers homozygous. If a sufficient number of independent markers within a homozygous segment are polymorphic,

then it should be possible to determine whether a segment is autozygous, rather than a chance alignment of homozygous genotypes. Consecutive homozygous genotypes which are assumed to correspond to autozygous segments are often called “Runs of Homozygosity” or ROH.

Identifying ROH requires a sufficient marker density, such that an adequate number of independent polymorphic markers are expected within each autozygous segment, as well as a genetic map defining the physical order of markers on each chromosome⁶. For these reasons, calling ROH only really became practical with the advent of dense SNP microarrays (e.g., Gibson 2006 in HapMap phase 1)⁶¹. Since then, several methods have been proposed for identifying ROH, and these are briefly discussed below.

The most commonly used algorithm for identifying ROH is implemented in the popular genetic analysis software PLINK^{67,68}. This method implements a sliding window (of default size 50 SNPs) which advances, one SNP at a time, across the genome. At each window position, a call is made whether that window is entirely homozygous – allowing for a maximum number of missing genotypes (default 5) and heterozygous calls (default 1). SNPs covered by a significant fraction (default 0.05) of entirely homozygous windows are called as ROH, subject to occurring in sufficiently large consecutive tracts (default 100 SNPs and 1 Mb).

The simple method implemented in PLINK takes no account of population parameters (e.g., allele frequencies and linkage disequilibrium (LD)), and thus requires appropriate parameterisation, according to the quality and density of the SNP genotypes to which it is applied. Furthermore, since PLINK takes no account of LD between markers, some studies LD prune before ROH calling^{69–71}, while others do not^{13,72,73}. These choices have been thoroughly investigated by Howrigan et al, who conclude that PLINK performs well compared to two other algorithms (GERMLINE and BEAGLE), and make recommendations for optimum settings⁷⁴.

ROH calling in PLINK takes no account of allele frequencies, so the homozygosity of a very rare allele is considered no more informative than the homozygosity of a

common allele. In addition, PLINK treats all SNPs equally, so homozygous SNPs in strong LD are given the same weight as uncorrelated SNPs. Various, more sophisticated, ROH calling algorithms have been proposed to account for the potential shortcomings of the simple PLINK method^{6,62,75–78}, and GARLIC⁷⁸, a Hidden-Markov-Model (HMM) method has been used in several recent publications^{79,80}. HMM methods have the potential to identify shorter autozygous segments more reliably than PLINK, but do not seem to significantly improve the calling of the longer ROH (> 1.5 Mb), which account for most of the variation in F_{ROH} (Figure 13). However, accounting for allele frequency might be particularly advantageous in whole genome sequence data, where very low frequency variants could strongly support a ROH call (note that ROH is usually pronounced like “row” /rəʊ/, hence “a ROH” is used here, rather than the alternative “an ROH”).

2.2 Genetic data

The data used in this thesis are mainly drawn from two sources:

1. ROHgen¹³ - a consortium of worldwide genetic studies.
2. The UK Biobank⁸¹ - a large, deeply-phenotyped UK population cohort.

The two datasets became available at different points in the project and have complementary characteristics, as described below.

2.2.1 ROHgen

2.2.1.1 Overview

Inbreeding coefficients are low in most human populations^{82,83}, and consequently very large samples are required to accurately estimate the genetic effects of inbreeding⁸⁴. To maximize sample size, a collaborative consortium (ROHgen¹³) was (re)established in early 2017, and research groups administering cohorts with SNP chip genotyping were invited to participate. To ensure that all participants performed uniform and repeatable analyses, an automated software pipeline was developed and executed locally by each research group. This software pipeline required cohorts to provide only quality-controlled genotypes (in plink binary format) and standardised phenotypes (in plain-text) and used familiar software (R,

PLINK^{68,85}, KING⁸⁶) to perform the analyses described below. Results from each cohort were returned to Edinburgh for meta-analysis. Initial findings from ROHgen were published as “Associations of autozygosity with a broad range of human phenotypes”¹ in January 2019 and are expanded upon in this thesis.

2.2.1.2 Cohort recruitment

In total, 117 independent, genetic epidemiological study cohorts contributed to ROHgen. Of these, 116 were studies of adults and contributed multiple phenotypes, while 1 was a study of children and contributed only birth weight. To minimize any potential confounding or bias caused by within-study heterogeneity, studies were split into single-ethnicity sub-cohorts wherever possible. Each sub-cohort was required to use only one genotyping array and be of uniform ancestry and case-status. For example, if a study contained multiple distinct ethnicities, sub-cohorts of each ancestry were created and analysed separately. At minimum, ancestry was defined on a sub-continental scale (i.e., European, African, East Asian, South Asian, West Asian, Japanese, and Hispanic were always analysed separately) but more precise separation was used when deemed necessary, for example in cohorts with a large representation of Ashkenazi Jews. In case-control studies of disease, separate sub-cohorts were created for cases and controls and phenotypes associated with disease status were not analysed in the case cohort: for example, fasting plasma glucose was not analysed in Type 2 diabetes case cohorts. Occasionally, cohorts had been genotyped on different SNP genotyping microarrays and these were also separated into sub-cohorts. There was one exception to the single microarray rule (deCODE-Iceland), where the intersection between two arrays used exceeded 150,000 SNPs. In this cohort the genotype data from two arrays was merged since the correspondence between F_{ROH} for the individual arrays and F_{ROH} the intersection dataset was found to be very strong ($\beta_{merged,hap}=0.98$, $r^2=0.98$, $\beta_{merged,omni}=0.97$, $r^2=0.97$). Dividing studies using these criteria yielded 234 sub-cohorts, details of which are provided in Supplementary Table 1.

2.2.1.3 Genotyping

All ROHgen samples were genotyped on high-density (minimum 250,000 markers), genome-wide SNP microarrays supplied by Illumina or Affymetrix. Genotyping arrays with highly variable genomic coverage (such as Exome chip, Metabochip or Immunochip) were judged unsuitable for the ROH calling algorithm and were not permitted. Imputed genotypes were also not permitted since higher error rates are known to downwardly bias estimates of F_{ROH} (see Appendix C.4). Only called genotypes in PLINK binary format were accepted. Each study applied their own GWAS quality controls before additional checks were made in the common analysis pipeline: SNPs with > 3% missingness or MAF < 5% were removed, as were individuals with > 3% missing data. Where available X-chromosome genotypes were included, although not used in any autosome-wide analyses, such as the calculation of F_{ROH} . Additional, cohort-specific, genotyping information is available in Supplementary Table 2.

2.2.2 UK Biobank

2.2.2.1 Overview

The UK Biobank (UKB) is a deeply-phenotyped prospective study of approximately 502,000 UK residents. Although there are known participation biases⁸⁷, UKB was sampled from the resident UK population and therefore includes a broad range of ethnic backgrounds.

For traits measured in many cohorts (for example height), the statistical power to estimate $\beta_{F_{ROH}}$ is much lower in UKB than in ROHgen, however the availability of individual-level data and deep-phenotyping permit additional analyses which complement those performed in ROHgen.

2.2.2.2 Genotyping

Genotyping in UK Biobank was performed using two very similar versions the UK Biobank Axiom Array⁸¹. Genotypes were therefore available at 803,113 SNP markers for 488,288 UKB samples which passed quality control.

2.3 Methods

2.3.1 ROH calling

Throughout this thesis, ROH were called by a relatively simple, but well-established method, implemented in PLINK^{68,85}. More sophisticated methods, based on Hidden-Markov-Models (HMM) are now available⁷⁸, but these were not considered necessary in this setting. Even in outbred populations, the power to estimate $\beta_{F_{ROH}}$ is largely provided by high F_{ROH} samples with long average ROH length (see section 2.5.4). HMM methods improve the calling of short ROH, but have very similar performance to PLINK above 1.5Mb.⁷⁴ Especially in ROHgen, where analyses were run by many different research groups, it was deemed preferable to use familiar software already installed on most servers. Howrigan et al. recommended PLINK over other algorithms, but also suggested light LD pruning to optimise the trade-off between identifying short ROH and erroneously calling non-autozygous ROH. However, throughout this thesis, ROH are called without LD pruning, primarily for consistency with the preceding work in the same group (i.e., Joshi et al¹³, and McQuillan et al.^{82,84}) Also, it is not known whether autozygosity from very distant ancestors (> 50 generations) causes inbreeding depression, and LD pruning potentially risks removing older autozygous segments. The subject of ROH calling is revisited in the discussion of this thesis, Chapter 6.

The ROH calling in this thesis is therefore identical to that recommended by Joshi et al¹³. In summary, SNPs with minor allele frequencies below 5% were removed, before continuous runs of homozygous SNPs (ROH) were identified using PLINK with the following parameters: homozyg-window-snp 50; homozyg-snp 50; homozyg-kb 1500; homozyg-gap 1000; homozyg-density 50; homozyg-window-missing 5; homozyg-window-het 1. No linkage disequilibrium pruning was performed. These parameters have been previously shown to call ROH that correspond well to autozygous segments in which all SNPs (including those not present on the chip) are homozygous-by-descent, not chance arrangements of independent homozygous SNPs, and inbreeding coefficient estimates calculated by this method (F_{ROH})

correlate well with pedigree-based estimates (F_{PED})⁸². Moreover, they have also been shown to be reasonably robust to array choice¹³.

2.3.1.1 *Excluding hemizygous deletions*

Hemizygous deletions up to several Megabases in length are known to occur in apparently healthy individuals⁸⁸. A large majority of CNVs (many of which are presumed to be hemizygous deletions) occur well below the ROH-calling threshold of 1.5 Mb used here, but a small number of phenotypically important deletions (e.g., del 22q11.2 DiGeorges syndrome) are longer than this threshold and have been previously observed in population cohorts such as UKB^{88,89}. These hemizygous deletions are potentially relevant to this study since they are likely to be called as ROH unless specifically removed by QC procedures. McQuillan et al. showed that the variation in F_{ROH} caused by deletions is likely to be small compared to the variation caused by parental relatedness⁸² and that removing these deletions had negligible impact on $\hat{\beta}_{F_{ROH}}$ in their study cohort from the Orkney islands of Scotland. This is consistent with the observation that, even in outbred populations, the statistical power to estimate $\beta_{F_{ROH}}$ is primarily due to high F_{ROH} samples (Figure 13) – with average ROH lengths well above even the longest deletions. For this reason, and since identifying deletions requires access to SNP marker intensities, deletions were not specifically removed by the ROHgen pipeline.

However, the log intensity ratio (log2R) for all genotype calls is provided by UKB and was used to further investigate the role of hemizygous deletions. Although sophisticated methods have been developed to identify CNVs⁹⁰ these were not required to find the long hemizygous deletions of interest here. Indeed, long hemizygous deletions are readily identifiable by a reduction in log2R coincident with a ROH (Figure 5).

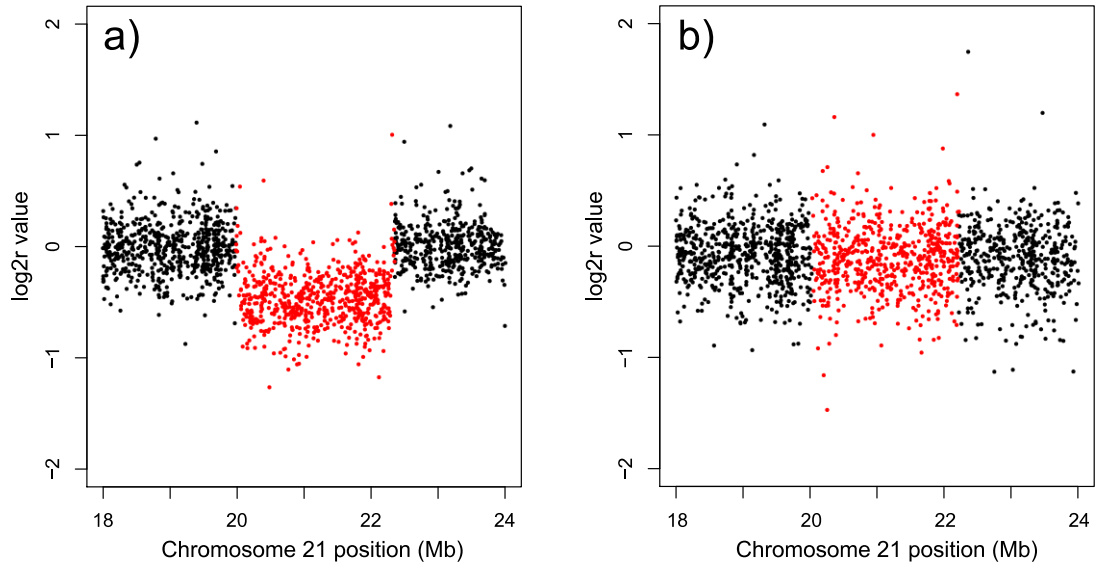


Figure 5: **Log2R values plotted against chromosome position for two UKB subjects with ROH identified at a locus on chromosome 21.** Values of log2R within each ROH are shown in red. (a) ROH identified as a hemizygous deletion. (b) ROH identified as homozygous segment.

The average log2R value was therefore calculated within each previously called ROH and, based on Figure 6, any with a mean value < -0.3 was excluded from the subsequent calculation of F_{ROH} .

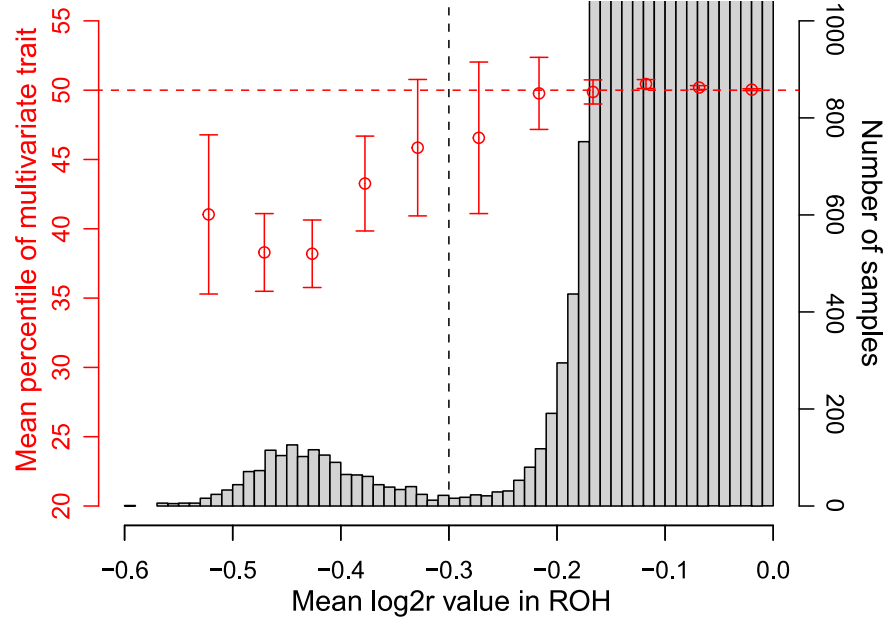


Figure 6: **Histogram of mean log2R values in all called ROH.** Both x and y axes of the histogram are truncated to emphasise a small peak at -0.45 corresponding to hemizygous deletions. Also shown, in red, is the average *multivariate* trait percentile (see section 3.2.7) of all subjects who carry a ROH in 11 log2r ranges – i.e., samples carrying a deletion have less advantageous trait values.

Applying these criteria, 1448 out of 2,450,349 autosomal ROH in UKB were called as deletions. These deletions occur in 1382 unique samples, on all chromosomes, and average 2.4Mb in length (min = 1.5Mb, max=22.2Mb).

Consistent with other studies of CNVs in UKB⁸⁹, only a small number of deletions occur at locations associated with well described microdeletion syndromes (e.g., DiGeorge's syndrome (N=13), Williams (N=0), Prader-Willi/Angelman (N=2) , Wolf-Hirschhorn (N=1)) while many others, such as the 21q11.2 deletion shown in Figure 5, appear to be either sufficiently rare, or phenotypically unobtrusive, as to be less well described in the literature.

2.3.2 Calculating estimators of the Inbreeding Coefficient (F)

For all subjects, three separate estimates of Inbreeding Coefficient (F) were calculated. Two of these are single-SNP methods (F_{SNP} and F_{GRM}) while the primary coefficient used in this thesis (F_{ROH}) requires a genomic map. Other partial measures of homozygosity were also calculated (e.g., $F_{ROH<3Mb}$ and $F_{ROH>5Mb}$), although these are not strictly estimates of F .

2.3.2.1 F_{ROH}

F_{ROH} is the fraction of each genome in ROH > 1.5 Mb. For example, in a sample for which PLINK had identified n ROH of length l_i (in Mb), $i \in \{1...n\}$, then F_{ROH} was then calculated as:

$$F_{ROH} = \frac{\sum_{i=1}^n l_i}{3.0Gb} \quad (4)$$

The choice of denominator in equation (4) is not self-evident. Many previous authors have used the SNP-mappable length of the autosomes (approximately 2.77Gb)^{69,70,84} while others have used the total length (approximately 3.0Gb)¹³. If the objective is to give unbiased estimates of $\beta_{F_{ROH}}$ then neither choice is optimum. Using 2.77Gb implicitly assumes the un-genotyped fraction of the genome has exactly the same autozygosity as the genotyped fraction. In contrast, using 3.0Gb gives unbiased $\hat{\beta}_{F_{ROH}}$ when there is no correlation between ROH at different loci, which is only approximately true in a highly outbred population. In practice, the

correlation between the autozygous fraction in the genotyped and un-genotyped genome is likely to be variable and cohort specific, as it is between different chromosomes. Since the choice on denominator affects only the magnitude of $\hat{\beta}_{F_{ROH}}$, 3.0Gb was used here to follow the precedent set by Joshi et al.

To compare the relative effects of different ROH lengths, the genomic fractions were calculated in three Mb length categories ($F_{ROH<3}$, $F_{3<ROH<5}$, and $F_{ROH>5}$). It follows that

$$F_{ROH} = F_{ROH<3} + F_{3<ROH<5} + F_{ROH>5} \quad (5)$$

2.3.2.2 F_{SNP}

Single-point inbreeding coefficients can also be estimated from individual SNP homozygosity without any reference to a genetic map. For comparison with F_{ROH} , a method of moments estimate of inbreeding coefficient was calculated⁹¹, referred to here as F_{SNP} , and implemented in PLINK by the command --het.

$$F_{SNP} = \frac{O(HOM) - E(HOM)}{N - E(HOM)} \quad (6)$$

Where $O(HOM)$ is the observed number of homozygous SNPs, $E(HOM)$ is the expected number of homozygous SNPs, i.e., $\sum_{i=1}^N (1 - 2p_i q_i)$, and N is the total number of non-missing genotyped SNPs.

2.3.2.3 F_{GRM}

A further single point estimator of the inbreeding coefficient, described by Yang *et al.*⁶⁰ as $\hat{F}^{III}$, is implemented in PLINK by the parameter -ibc (Fhat3) and was also calculated for all samples.

$$F_{GRM} = \hat{F}^{III} = \frac{1}{N} \sum_{i=1}^N \frac{(x_i^2 - (1 + 2p_i)x_i + 2p_i^2)}{2p_i(1 - p_i)} \quad (7)$$

Where N is the number of SNPs, p_i is the reference allele frequency of the i^{th} SNP in the sample population and x_i is the number of copies of the reference allele.

2.3.3 Population structure and plotting F_{ROH} against F_{IS}

The F -statistics considered so far are sometimes described as the correlation between uniting gametes - that is to say, the excess of homozygotes within an individual compared to the expectation for two alleles sampled randomly from a reference population (often the current generation of the whole population). An excess of heterozygotes may be caused by a tendency to marry within families (consanguinity), or within social groups (endogamy). Inbreeding refers to reproduction with relatives, which causes increased homozygosity due to identify-by-decent from a relatively recent common ancestor, but when endogamy occurs within large social groups (say ethnicities) then increased homozygosity (compared to the total population) may be caused by very distant ancestry. Thus, SNP-based F statistics capture both recent inbreeding (of interest) and coarse-scale population structure (of less interest here). For this reason, if the purpose is to measure inbreeding, F statistics (particularly F_{SNP} and F_{GRM}) should only be calculated in populations with homogenous distant ancestry (as required in ROHgen). It is interesting to note that F_{ROH} is less subject to this constraint, since the minimum ROH length naturally excludes homozygosity caused by very distant common ancestry.

Sewall Wright considered the problem of F statistics in structured populations, and introduced three homozygosity measures (F_{IS} , F_{IT} , and F_{ST}), where the subscripts refer to Individual, Sub population and Total population⁹². For example, F_{IS} is the average homozygosity of alleles sampled from an individual, divided by the average homozygosity of alleles sampled from a homogeneous sub-population (for example, a ROHgen cohort). F_{IS} is thus a measure of inbreeding defined as the deviation from random mating in the most recent generation, and can be calculated from the mean individual departure from Hardy-Weinberg equilibrium (i.e., $F_{IS} = \bar{F}_{SNP}$)¹¹.

In a small effective population ROH can occur even with successive generations of random mating (i.e., when $F_{IS}=0$.) and cohorts can therefore be categorised by plotting mean F_{ROH} against F_{IS} (Figure 9). When ROH are caused entirely by non-random mating in the current generation (i.e., consanguinity) then $\bar{F}_{ROH} = F_{IS}$ and

consanguineous cohorts lie along the diagonal unity line. In contrast, when ROH are caused entirely a small effective population, cohorts may have substantial $\bar{F}_{ROH}$ even when $F_{IS} = 0$.

2.3.4 Statistical power contributed by individual samples

It is intuitive that in outbred populations the highest F_{ROH} samples will have most influence on estimates of $\beta_{F_{ROH}}$, but it may be informative to quantify the relative importance of this effect.

For the univariate regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (8)$$

The standard error of the regression gradient estimate ($\hat{\beta}_1$) is defined as:

$$se_{\hat{\beta}_1} = \sqrt{\frac{\sum_i \hat{\varepsilon}_i^2}{(n-2) \sum_i (x_i - \bar{x})^2}} \quad (9)$$

Hence, the relative contribution of a sample (i) is determined by the square of its deviation from the cohort mean F_{ROH} .

$$relative_power_i = \frac{(F_{ROH}^i - \bar{F}_{ROH})^2}{\sum_i (F_{ROH}^i - \bar{F}_{ROH})^2} \quad (10)$$

This is exploited in section 2.5.4 to quantify the relative importance of very high F_{ROH} samples in UK Biobank.

2.4 ROH Simulation methods

To understand some of the ROH features observed in real data, two simulation models were constructed. The first is simple, and allows large samples to be simulated quickly and approximately, while the second is more detailed allowing complex pedigrees to be realistically simulated, albeit at the expense of speed.

2.4.1 Simple ROH simulation method

A method was required to quickly estimate the distribution of ROH lengths arising in the offspring of different relative types. For simplicity, only the autosomes were

simulated in this method, but both chromosome lengths, and the variation of recombination rate throughout the genome were accounted for.

Each generation, meiotic recombination introduces breakpoints which define transitions from one ancestral haplotype to another. Therefore, in an inbred individual, where (parts of) the paternal and maternal chromosomes are potentially inherited from one common ancestor, key parameters are the number of meiotic divisions experienced down the paternal and maternal legs of the pedigree loop ($n_meiosis_paternal$ and $n_meiosis_maternal$). The total genetic length of the human autosomes ($total_genetic_length$) is approximately 35.5 Morgans, and therefore, at each meiosis, the number of recombination breakpoints is Poisson distributed with rate (Lambda) equal to 35.5. These breakpoints occur approximately independently of each other, and are therefore distributed randomly (in genetic map distance) throughout the genome.

Potential ROH segments were therefore simulated by a Poisson distributed number of haplotype transitions:

$$\lambda = 35.5 * (n_meiosis_paternal + n_meiosis_maternal) \quad (11)$$

randomly distributed throughout the genome. Since ROH are necessarily terminated at chromosome ends, 23 further fixed breakpoints were added at the genetic map positions corresponding to all chromosome ends. The simulation of these potential ROH segments is implemented by the R code provided in Appendix G.

Each potential ROH segment has an equal probability of being autozygous (equal to $2^{-(n_meiosis_paternal+n_meiosis_maternal-1)}$) and therefore all segments can be used to build a ROH length distribution for the simulated relative type. Although the potential ROH segments are simulated on the genetic map, they can be mapped back to physical distance via the genetic map, if required.

2.4.1.1 Estimating ROH ages from genetic map lengths (cM)

The simulation method described above (section 2.4.1) can be used to estimate the probability of observing a ROH of length (l_i), given that the ROH originates from an ancestor N generations prior, i.e., $Pr(l_i|N)$, as shown in Figure 21. However, in real data it will often be more useful to estimate the probable ‘age’ (N) of a ROH from its length i.e., $Pr(N|l_i)$. These two probabilities can be connected by Bayes’ theorem which states that:

$$Pr(A|B) = \frac{Pr(B|A) * Pr(A)}{Pr(B)} \quad (12)$$

Substituting l_i and N in equation (12) gives:

$$Pr(N|l_i) = \frac{Pr(l_i|N) * Pr(N)}{Pr(l_i)} \quad (13)$$

$Pr(l_i) = 1$, since a ROH of length l_i has been observed, so only the prior probability of common ancestry at generation N , $Pr(N)$, remains to be estimated. In a randomly mating population, the probability of coalescence at generation N is approximated by:

$$P_c(N) = \frac{1}{2N_e} e^{-\frac{N-1}{2N_e}} \quad (14)$$

And if N_e is sufficiently large, compared to N , equation (14) can be treated as approximately constant. Hence:

$$Pr(N|l_i) \propto Pr(l_i|N) \quad (15)$$

Or the probability of a ROH of length l_i originating from an ancestor N generations ago, is proportional to the probability of observing a ROH of length l_i given a common ancestor N generations ago. This can be visualised in Figure 21, by considering that the age distribution of a ROH of length l_i can be determined from the relative probabilities plotted at $x = l_i$.

2.4.2 Full haplotype inheritance ROH simulation method

To corroborate the results of the simple simulation method described in Section 2.4.1, and to understand the more complex inheritance patterns of the X chromosome, an explicit pedigree simulation model was developed. In this method, a complete pedigree is modelled with simulated haplotype inheritance for all related and inbred individuals. For example, to simulate the offspring of first cousins, a pedigree could be defined as shown in Figure 7.

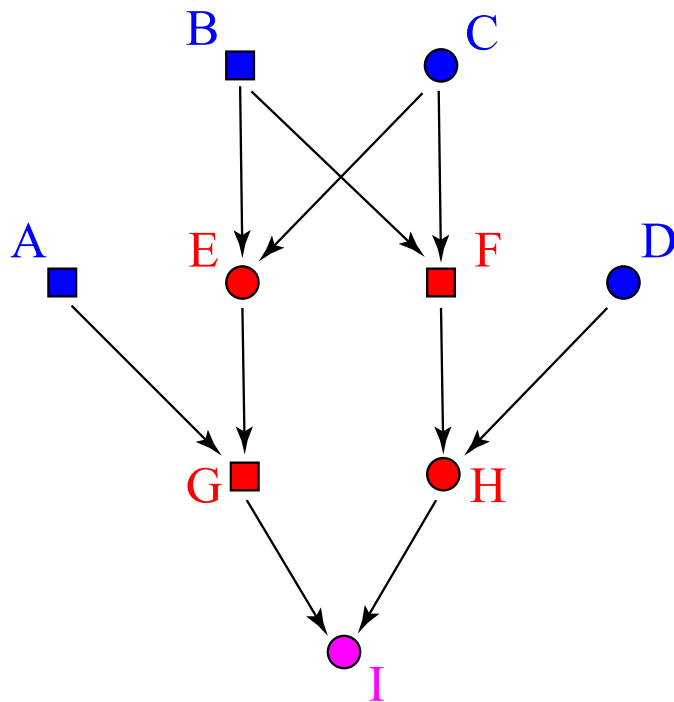


Figure 7: **Simulation pedigree for the offspring of first cousins.** Unrelated base generation individuals (A, B, C & D) are shown in blue, related individuals (E, F, G & H) in red and the focal inbred individual (I) is shown in magenta. Males are represented by squares and females by circles.

Base generation ancestors (A, B, C & D) would be then assigned unbroken, genome-wide, haplotypes (1 & 2 and 3 & 4 and 5 & 6 and 7 & 8, respectively). Note, there is no need for these unrelated individuals to occupy the same generation, so long as they are unrelated to all other “base generation” individuals. Each related individual would then be generated by simulating meiosis – i.e., by randomly choosing recombination breakpoints in both parents, and assigning their offspring the resulting, randomly assorted, paternal and maternal alleles.

Recombination breakpoints were randomly determined from the hap3 recombination rate map, and take account of the absence of X chromosome recombination in males. This simulated haplotype inheritance was completed in all related individuals, finishing with the focal inbred individual, who, in this case, has a mosaic of base generation haplotypes on both the paternal and maternal chromosomes. The haplotypes of one such simulated offspring of first cousins are shown in Figure 8.

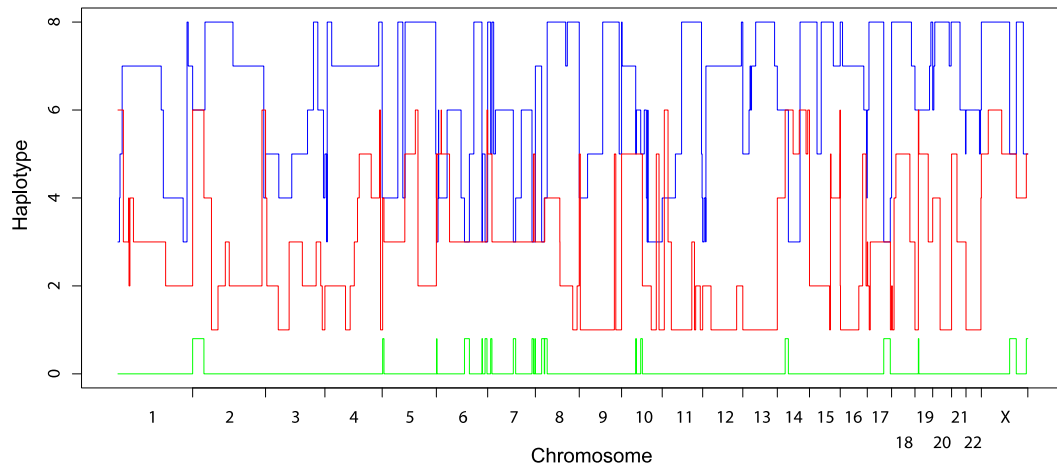


Figure 8: Simulated base-generation haplotypes in an offspring of first cousin parents. The unrelated base-generation of this simulated individual has four members, and therefore eight haplotypes, which are denoted 1, 2, 3, 4, 5, 6, 7 & 8 (see Figure 7). The paternal and maternal haplotypes inherited by the simulated offspring are shown in red and blue, respectively. When the maternal and paternal haplotypes are equal an autozygous segment results, and this is shown by a green track where 1 indicates autozygosity.

This full simulation method is very flexible since any pedigree which can be drawn may be resimulated as many times as required. For example, simulating 10,000 offspring off full siblings takes about 2 hours on a single high-performance-computing (HPC) core.

2.5 Results

2.5.1 Plotting F_{ROH} against F_{IS} for ROHgen cohorts

As described previously (section 2.3.3), the ROH origins of a population can be examined by plotting $\bar{F}_{ROH}$ against F_{IS} – this is shown for all 234 ROHgen cohorts in Figure 9.

Several features are apparent in Figure 9. Firstly, many of the South and West Asian cohorts lie along the $\bar{F}_{ROH} = F_{IS}$ line, indicating ROH caused exclusively by consanguinity. The most consanguineous populations are Pakistanis from the Born in Bradford study and UK Biobank, as well as the Saudi Arabian and Iranian cohorts. However, the cohort with the highest average F_{ROH} (Polynesia east) has very little consanguinity, but rather ROH caused by a small effective population. This is consistent with a known history of successive founder populations as the Polynesian islands were settled from west to east⁹³. Other small effective population cohorts, such as the Amish and Hutterites, descend from a small number of founder individuals, and actually display moderate inbreeding avoidance.

Equally, the single largest cohort (UKB British, 402,981 samples) exhibits a small amount of inbreeding avoidance which, coupled with a large effective population size, results in relatively modest statistical power (5.9% of the ROHgen total). Biobank Japan (BBJ) is the single most powerful cohort (37.7% of the ROHgen total) due to its substantial size (158,284 samples) and relatively high mean F_{ROH} (0.012). Although BBJ shows some evidence a small effective population (presumably caused by social or geographical barriers), the raised mean F_{ROH} is also caused by a substantial fraction of subjects who are most likely the offspring of first cousins (1 in 20 with $0.04 < F_{ROH} < 0.10$).

Finally, a small number of cohorts have $F_{IS} > 0$ without substantial F_{ROH} indicating population structure caused by heterogeneity in sample ancestry. This is evident in the admixed Hispanic cohorts, and is particularly pronounced in a catch-all cohort for all UKB samples with mixed ancestry. In these cohorts F_{SNP} and F_{GRM} capture

population structure, as well as inbreeding, while F_{ROH} is thought to remain a valid measure of inbreeding (see Section 2.3.3).

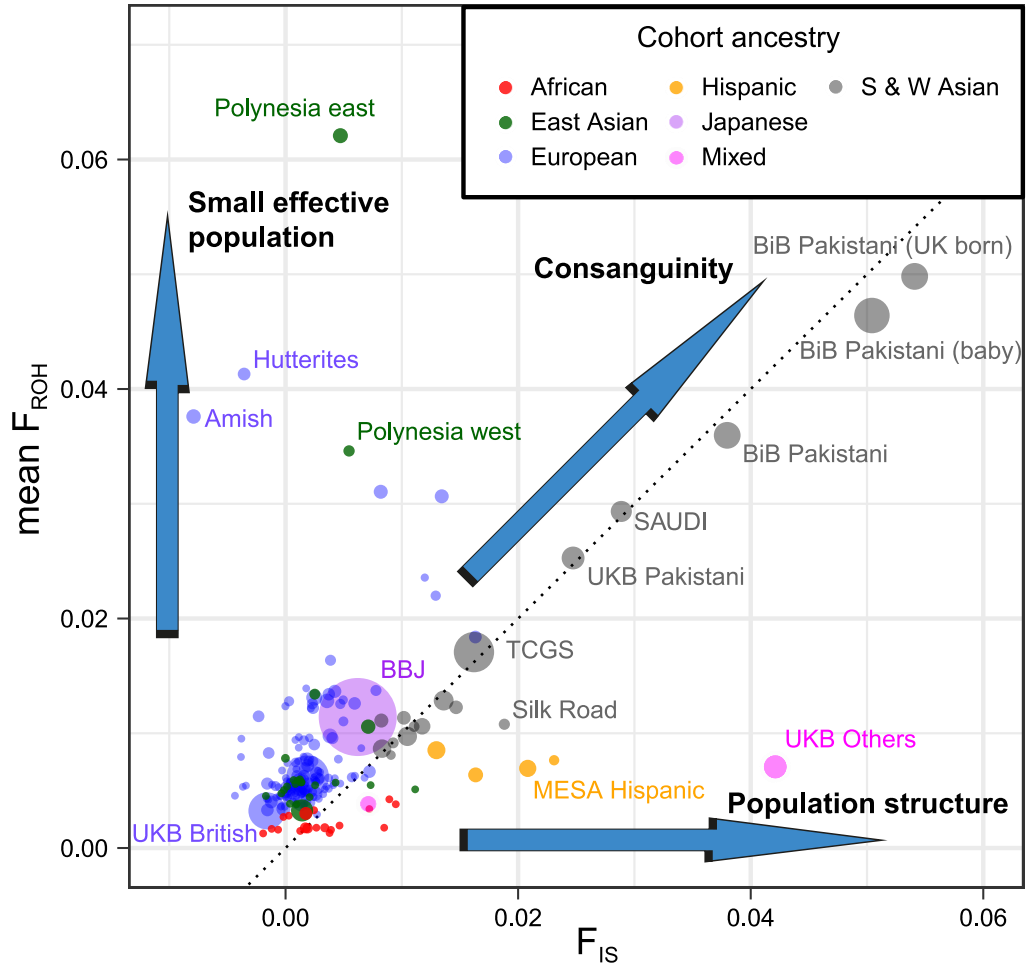


Figure 9: **Mean F_{ROH} plotted against F_{IS} for 234 ROHgen sub-cohorts.** Each cohort is represented by a circle whose area is proportional to the approximate statistical power ($N\sigma_{F_{ROH}}^2$) contributed to estimates of $\beta_{F_{ROH}}$. Mean F_{ROH} can be considered as an estimate of total inbreeding relative to an unknown base generation, approximately tens of generations past. F_{IS} measures inbreeding in the current generation, with $F_{IS} = 0$ indicating random mating, $F_{IS} > 0$ indicating consanguinity, and $F_{IS} < 0$ inbreeding avoidance¹¹. In cohorts along the y-axis, such as the Polynesians, Hutterites and Amish, autozygosity is primarily caused by small effective population size rather than preferential consanguineous unions. In contrast, in cohorts along the dotted unity line, all excess SNP homozygosity is accounted for by ROH, as expected of consanguinity within a large effective population. A small number of cohorts along the x-axis, such as Hispanic and mixed-race groups, show excess SNP homozygosity without elevated mean F_{ROH} , indicating population genetic structuring, caused for instance by admixture. A few notable cohorts are labelled. BBJ, Biobank Japan; BiB, Born in Bradford; UKB, UK Biobank; MESA, Multi-ethnic Study of Atherosclerosis; TCGS, Tehran Cardiometabolic Genetic Study.

2.5.2 Mean F_{ROH} by country of birth in UK Biobank

Country of birth is recorded for almost all participants in UK Biobank, and the mean F_{ROH} of subjects born in a country may give an indication of levels of autozygosity present in the indigenous population. Of course, a subject could be born in a country coincidental to their ancestry, and individuals who emigrate to the UK may not be representative of the whole population, but none-the-less the broad range of countries represented (46,178 samples in 198 non-UK countries) gives a unique snapshot of worldwide autozygosity (Supplementary Table 21).

The highest autozygosity levels are often recorded in countries where Islam is the dominant religion – notably: Pakistan, Tunisia, Afghanistan, Iraq, Turkey, and many East African nations. The average F_{ROH} in these countries is around 2%, implying about 30% of marriages are between first cousins. Non-Islamic nations with high average F_{ROH} include Nepal and Tibet, Sri Lanka, the Philippines, Israel, and Japan.

2.5.3 Distribution of F_{ROH} in selected ROHgen cohorts

The mean F_{ROH} of all ROHgen cohorts is reported in Figure 9 and Supplementary Table 3, but it is also informative to examine the distribution of F_{ROH} for a few selected cohorts (Figures 10, 11 and 12).

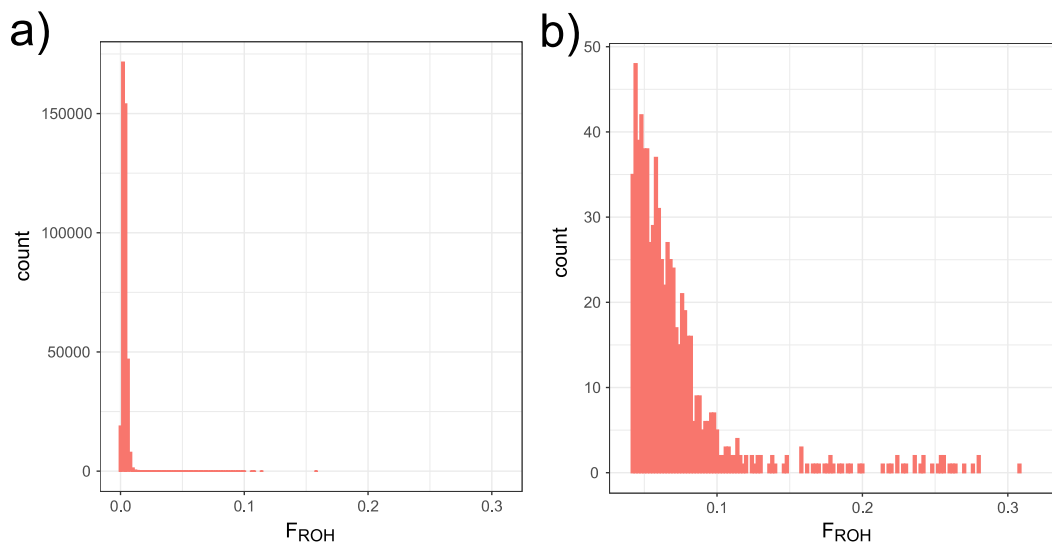


Figure 10: Histogram of F_{ROH} distribution in genetically British UK Biobank subjects. (a) All 402,981 genetically British samples. (b) 732 genetically British samples with $F_{ROH} > 0.04$.

In the UK Biobank British cohort 99.1% of samples have F_{ROH} less than 0.01 (Figure 10a). In this relatively outbred cohort, the parental relationship of the few higher autozygosity samples can be inferred with some confidence (Figure 10b). For example, the 646 samples with F_{ROH} between 0.04 and 0.1, are most likely the offspring of first cousins, and the 26 samples with F_{ROH} greater than 0.2 are approximately normally distributed around 0.25 - as expected of offspring of first-degree relatives⁷².

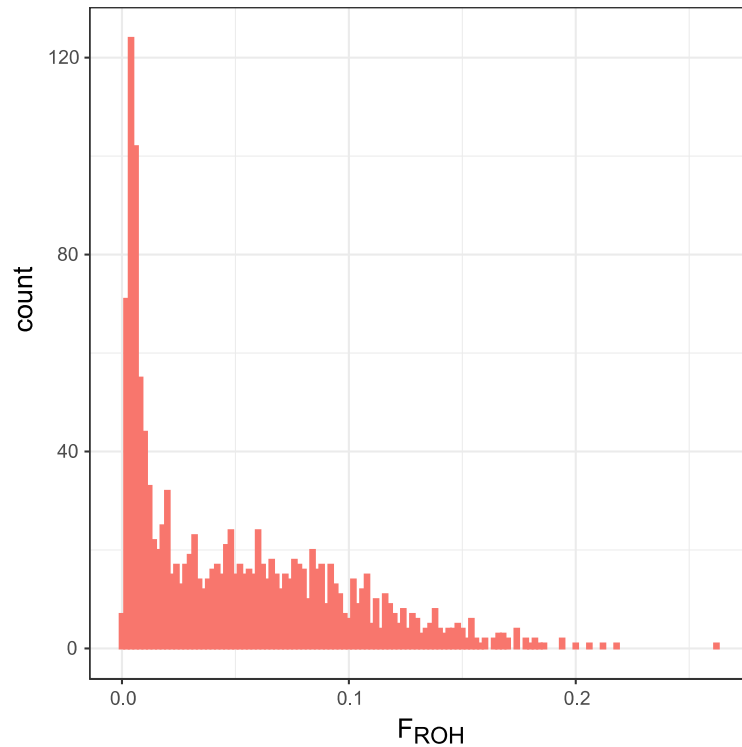


Figure 11: Histogram of F_{ROH} distribution in the Born in Bradford (UK born Pakistani) cohort.

In the Born in Bradford (Figure 11) cohort (drawn from a highly consanguineous, UK born, Pakistani population), the rate of first cousin marriages (1 in 3 with $0.04 < F_{ROH} < 0.1$) is approximately 200 times higher than in the genetically British. In contrast to Figure 10b, the density of individuals with $F_{ROH} > 0.10$ appears to decrease continuously, making the parental relationship unclear for these samples. In particular, the samples with $F_{ROH} > 0.2$ are not centred on 0.25 (as they are in UKB) making them less likely to be the offspring of first-degree relatives. Perhaps a more

plausible explanation for this long tail of high F_{ROH} samples is successive generations of cousin marriage, as simulated in section 2.6.3.

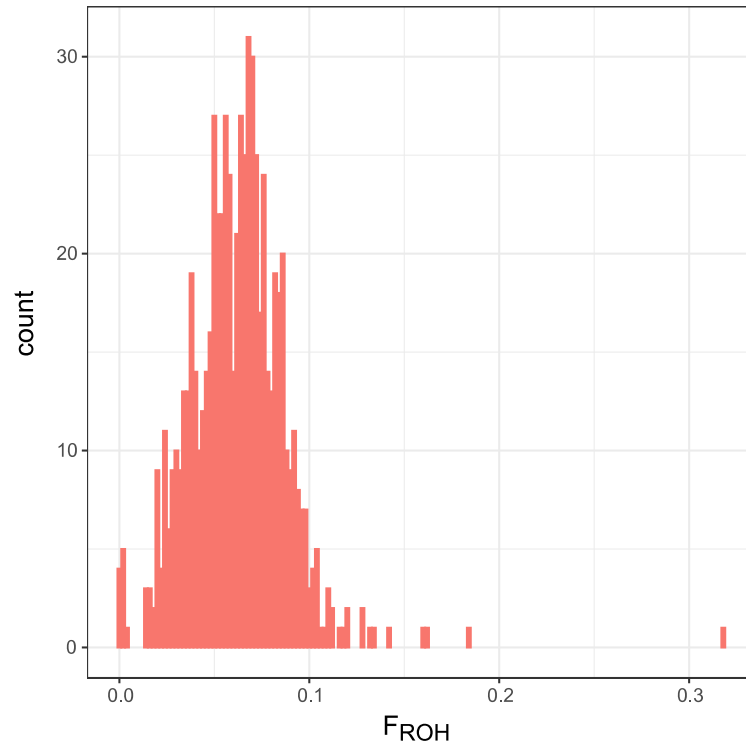


Figure 12: **Histogram of F_{ROH} distribution in the East Polynesian cohort.**

The distribution of F_{ROH} in the East Polynesian ROHgen cohort (Figure 12) is radically different from the distribution in Born in Bradford (Figure 11), although the mean F_{ROH} are comparable (0.062 and 0.050, respectively). In the East Polynesian cohort, there are very few outbred samples, and the distribution of F_{ROH} is approximately normal and centred on 0.06. This normal distribution would be consistent with the random inheritance of haplotypes from a very small founder population. Similarly, the average ROH length in East Polynesia is 3.1 Mb compared to 7.7 Mb in Born in Bradford, implying common ancestry many generations prior (see section 2.6.1).

2.5.4 Statistical power of samples as a function of F_{ROH} in UK Biobank

Many of the results presented in subsequent chapters derive from linear regression of phenotypes on F_{ROH} . Although very large samples sizes are used in these regressions, the relatively few high F_{ROH} samples will certainly have most importance. This was explored in the full UK Biobank cohort, which was used alone

for many results, and has a mean F_{ROH} (0.0038) only slightly lower than the full ROHgen consortium (0.0063). The relative statistical power of a sample is given in equation (10) and was used to calculate the cumulative power of the UKB cohort as a function of F_{ROH} (Figure 13).

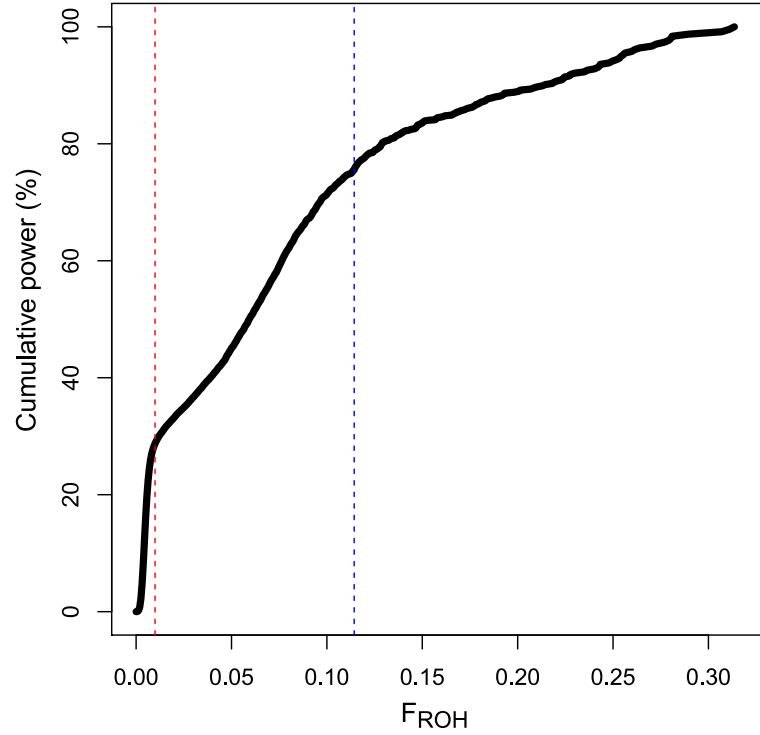


Figure 13: **Cumulative statistical power of all UKB samples ordered by F_{ROH} .** The 98th percentile of F_{ROH} (≈ 0.01) is marked in dashed red and the 99.96th percentile of F_{ROH} (≈ 0.12) is marked in dashed blue.

In UKB, only 29% of power to estimate $\beta_{F_{ROH}}$ is provided by the 98% of samples with F_{ROH} less than 0.01. Furthermore, 25% of the total statistical power is derived from the 0.04% of samples ($N=160$) with F_{ROH} greater than 0.12 (Figure 13). This clearly emphasises the leverage of high F_{ROH} samples, even in this relatively outbred sample.

Never-the-less, 29% of the cohort power is provided by the large number ($N=478,673$) of samples with $F_{ROH} < 0.01$. The relevance of these low F_{ROH} samples is even more pronounced in the genetically British UKB cohort, due to the relative scarcity of higher F_{ROH} samples (Figure 14).

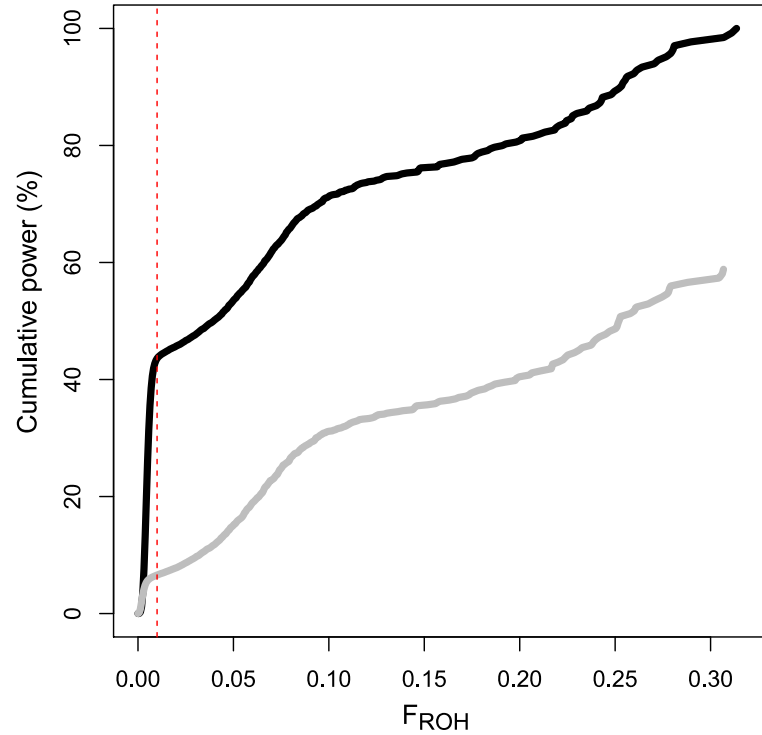


Figure 14: **Cumulative statistical power of genetically British UKB samples ordered by F_{ROH} .** The cumulative power including all ROH is shown in black, and the cumulative power when ROH less than 1cM are removed is shown in grey. $F_{ROH} = 0.01$ is marked as dashed red line.

In the genetically British UKB cohort, 43.6% of the total statistical power is provided by samples with $F_{ROH} < 0.01$. Interestingly, almost all of the variation of F_{ROH} in these low F_{ROH} (< 0.01) samples occurs in ROH islands (see section 2.5.5): when ROH of less than 1cM (almost all of which occur in ROH islands) are removed, the total statistical power is reduced by 41.2%. That is to say, approximately 41.2% of the total statistical power of the genetically British UKB cohort is contained in ROH of less than 1cM. It is though that the variation of these short ROH is mainly stochastic, as they are most likely formed by the unlikely reunion of two alleles both copied from the same haplotype in a distant common ancestor (see section 2.5.5.2).

2.5.5 Genomic distribution of ROH in UK Biobank

The distribution of ROH throughout the genome was investigated in the UK Biobank British cohort. Initially, the fraction of samples in ROH was calculated at every 100 kb throughout the genome, both on the autosomes and X-chromosome.

As described by previous studies, the fraction of samples with ROH at a locus is highly variable, with many loci autozygous in a small fraction of samples (around 0.03% in UK Biobank British), but the remainder autozygous in much greater numbers⁹⁴. These loci where ROH occur more frequently have been previously described as ‘ROH islands’⁹⁴, as they appear to rise rapidly out of a background of low ROH (Figures 15 and 16). It is thought that these ROH islands are caused by the meeting of short haplotypes which have become relatively common in the population, either by genetic drift, or by selection.

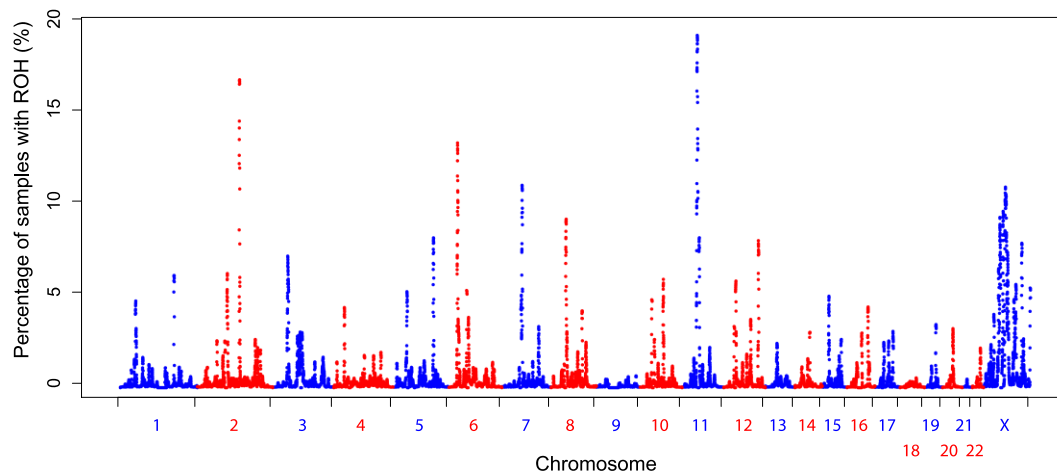


Figure 15: **Genome-wide variation in the percentage of British UK Biobank samples with a ROH at a locus.** The fraction of samples in ROH was calculated every 100 kb and is plotted against physical location. Odd-numbered chromosomes are coloured blue, and even chromosomes coloured red.

ROH islands occur on all chromosomes although the maximum height is highly variable: the highest ROH island (at p11.2 on chromosome 11) occurs in 19.3% of samples, while three chromosomes (9, 18 and 21) have ROH islands which peak just over 0.5% (which is still 15 times higher than the baseline rate). The frequency of ROH islands on the X chromosome is higher than the autosomes, as is the background level between islands, which is consistent with the smaller effective population size of the X chromosome. In general, ROH islands appear to be less frequent on the shorter chromosomes, and towards chromosomes ends, and since meiotic recombination follows a similar pattern, this gives a clue to the apparent clustering of ROH.

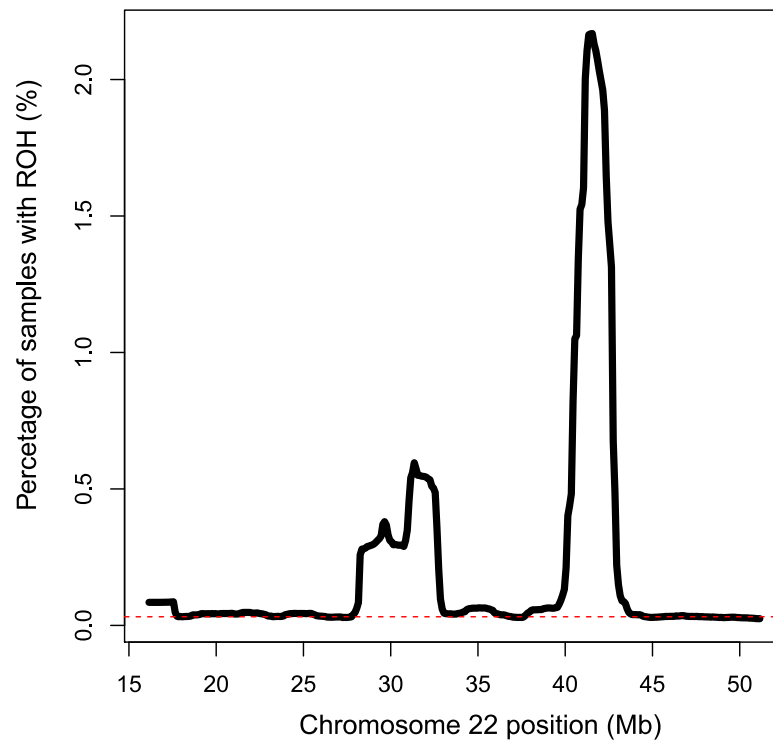


Figure 16: **Percentage of British UK Biobank samples with ROH at loci on chromosome 22.** The fraction of samples in ROH was calculated every 100 kb and is plotted against physical location (Mb). Other chromosomes show a similar overall pattern, with ROH islands rising out of a baseline of low ROH occurrence. The approximate level of this baseline ROH occurrence (0.03%) is shown by a dashed red line.

To investigate the role of recombination rate, the autosomes were divided into 2000 windows of 1.5 Mb length. Within each window the average fraction of samples in ROH was calculated, as was the average recombination rate within the window (in cM / Mb), from the published hap3 recombination rate map⁹⁵. Figure 17 shows that local recombination rate has a strong influence on the fraction of samples with ROH at the loci. Indeed, a 6th order polynomial fit explains 78% of the variance in ROH fraction between windows. Some of the residual variation is likely to be explained by short-comings of the method: all windows contain a range of recombination rates and given the highly non-linear effect of recombination rate on ROH fraction, averaging in windows will obscure the true effect. It is also possible that recent positive selection may have raised the allele frequency of certain advantageous haplotypes above the frequency expected by genetic drift. For example, the lactase persistence locus (*LCT*) is believed to have experienced recent

positive selection in European populations, and almost 10% of UKB British samples are autozygous at this locus, despite an average local recombination rate (0.89 cM/Mb) which more frequently results in ROH in approximately 0.2% of samples.

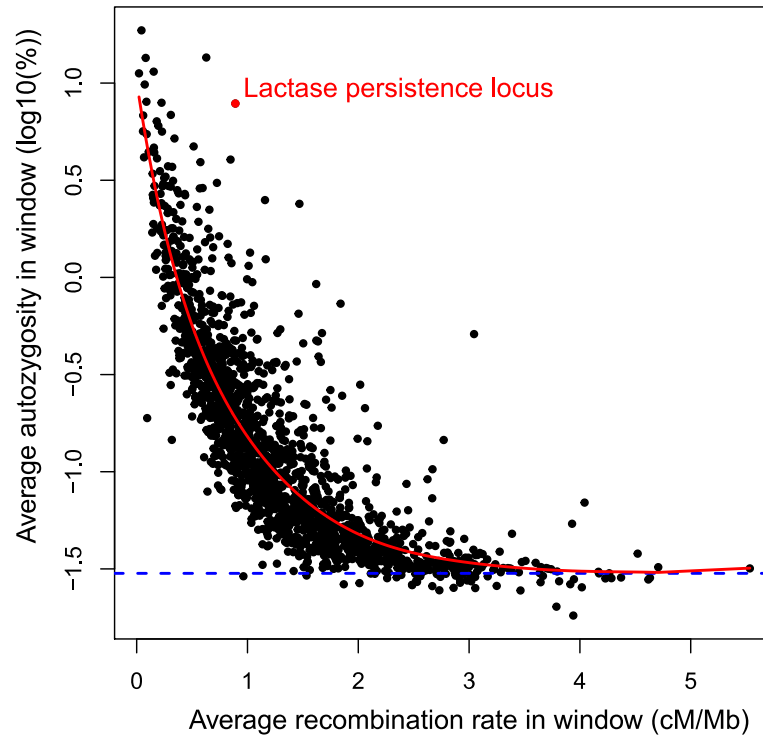


Figure 17: Average fraction of a window in ROH plotted against average recombination rate in the same window. Each autosome was divided into non-overlapping windows 1.5 Mb in length. The total length of ROH in each window was determined and divided by the number of samples to give the average window autozygosity, plotted on the y-axis. The average recombination rate in each window was obtained from published genetic maps. A 6th order polynomial fit is shown in red and the approximate level of background ROH (0.03%) is shown in dashed blue. Only windows containing more than 50 SNPs are plotted, since ROH calling is unreliable in sparsely genotyped regions.

Figure 17 shows that most genomic regions with a local recombination rate greater than 2cM/Mb are autozygous in approximately 0.03% ($10^{-1.5}$) of samples. Since most ROH in the offspring of close relatives will be longer than 1.5 Mb, even in these higher recombination rate regions, this 0.03% level captures the rate of consanguinity in this UKB British population. However, in the lower recombination rate regions, haplotypes inherited from more distant common ancestors are much more likely to remain above 1.5 Mb, and as local recombination rate reduces, ROH

may represent autozygosity from progressively more distant base generations. Hence, ROH islands are believed to represent autozygosity, albeit of haplotypes inherited from an ancestor perhaps many 10s, or even 100s, of generations prior (see section 2.6.1).

2.5.5.1 Filtering out ROH islands

There may be occasions when it is desirable to filter out ROH islands, leaving only autozygosity caused by consanguinity between close relatives. Since ROH islands are formed almost exclusively of short haplotypes, they can be effectively removed by retaining only ROH above a suitable genetic length (cM) threshold. This will also remove some short ROH caused by consanguinity, but in the offspring of close relatives the average ROH length is well above the thresholds considered here (see section 2.6.2).

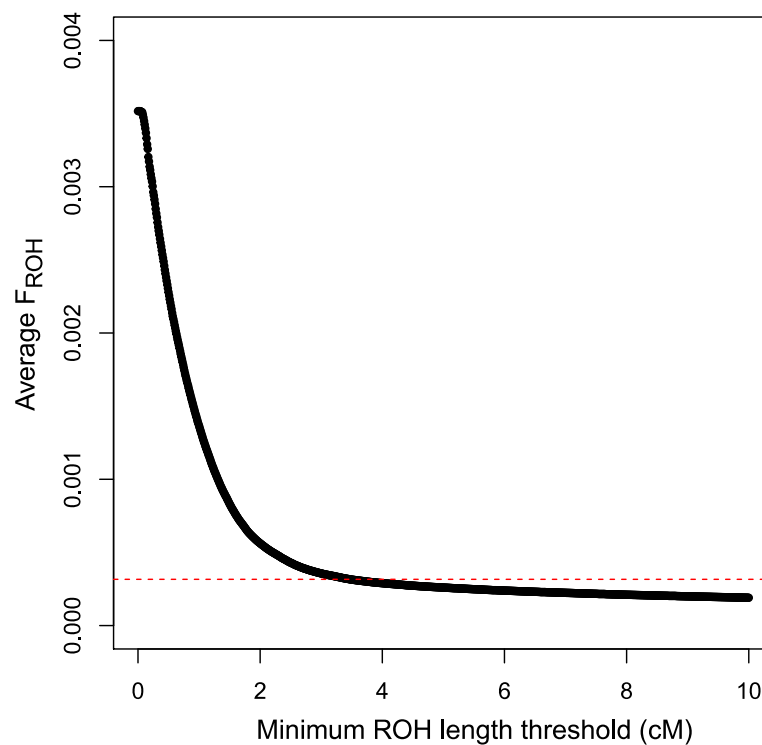


Figure 18: **Average F_{ROH} of all UKB British samples when including only ROH above the threshold plotted on the x-axis.** The summed length of all ROH above each length threshold was calculated and divided by the number of samples to give the average F_{ROH} plotted on the y-axis. The approximate level of baseline ROH (0.03%) is shown in dashed red.

In UK Biobank British, 0.35% of loci are in a ROH of some length, but only 0.03% are in ROH greater than 3 cM. In other words, 90% of the total ROH length in UK Biobank British is in ROH of less than 3 cM length. These short ROH occur predominantly in ROH islands, and when all ROH less than 3 cM are removed, the genome wide distribution of ROH becomes relatively uniform around 0.03% (Figure 19). The relatively uniform distribution of Figure 19 is much closer to theoretical models of inbreeding where each locus has an equal probability (F) of being autozygous - where F is determined only by the sample's pedigree.

It is interesting to note that the average autozygosity of the X-chromosome is approximately double that of the autosomes. Simulation shows that the expected ratio is a complex function of both parental relatedness, and the sexes of their related ancestors (see section 2.6.2). However, increased X chromosome autozygosity is consistent with both the smaller effective population size of the X chromosome, and the average ratios simulated in section 2.6.2.

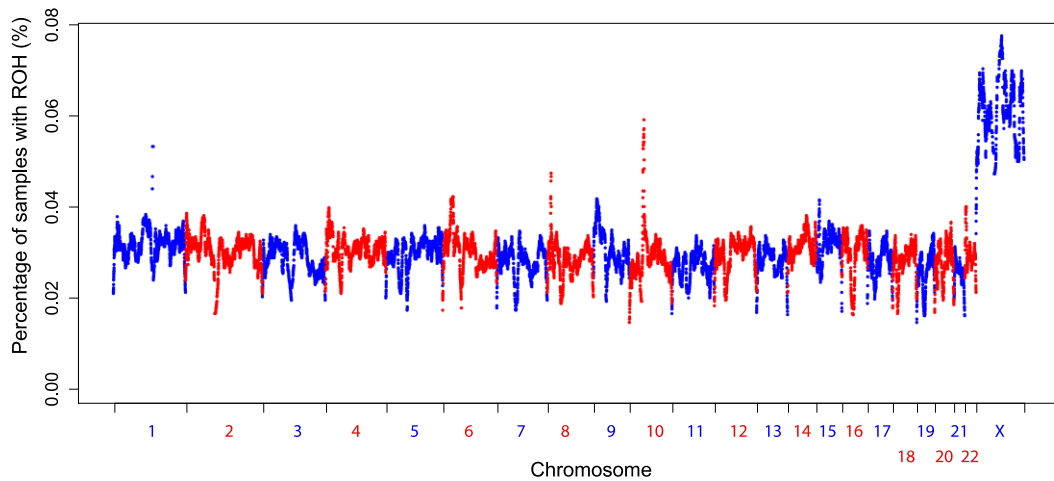


Figure 19: **Genome wide variation in the percentage of British UK Biobank samples with a ROH longer than 3 cM at a locus.** The fraction of samples in ROH (longer than 3cM) was calculated every 100 kb and is plotted against physical location. Odd-numbered chromosomes are coloured blue, and even chromosomes coloured red.

Despite a large fraction of total ROH burden occurring in ROH less than 3 cM, removing these shorter ROH has only a modest impact on the statistical power to estimate $\beta_{F_{ROH}}$ – since most of the variation of F_{ROH} is caused by long ROH associated with consanguinity. For example, in UK Biobank British removing ROH

below 3 cM reduces total ROH burden by 90% but only reduced the variance of F_{ROH} by 14%. In the full, multi-ethnicity, UKB cohort the corresponding figures are 83% and 12%.

2.5.5.2 Distribution of short (< 1 cM) ROH in UK Biobank British

It was suggested earlier that the occurrence of short ROH in a sample is a chance event caused by the unlikely reunion of two alleles both copied from the same haplotype in a distant common ancestor. The simulation results of section 2.6.1 suggest that ROH less than 1 cM are most likely formed by haplotypes copied from a common ancestor perhaps 50 generations ago. In an infinite population size, the probability of both copies descending from the same ancestral haplotype 50 generations ago are vanishingly small (2^{-100}), but 50 generations of recombination, and a finite population size, effectively create a vast number of opportunities (or trials). By this model, the number of short ROH occurring in a sample should follow a binomial distribution with a minute probability of success, but a large number of trials. The limiting case of a binomial distribution, with a large number of trials and small probability of success, is the Poisson distribution, defined by a single parameter, λ , which is equal to both the mean and variance of the distribution. Hence, if short ROH are caused by the stochastic inheritance of distant haplotypes, the number of short ROH in a sample should be Poisson-distributed with λ equal to the mean number of short ROH.

Figure 20, below, shows a histogram of the number of ROH less than 1 cM in all UK Biobank British samples. The analysis is restricted to genetically British samples to ensure homogeneous distant ancestry, and hence effective population size. Overlaying the histogram is the Poisson distribution with $\lambda = 3.608$ (the mean number of short ROH in this cohort).

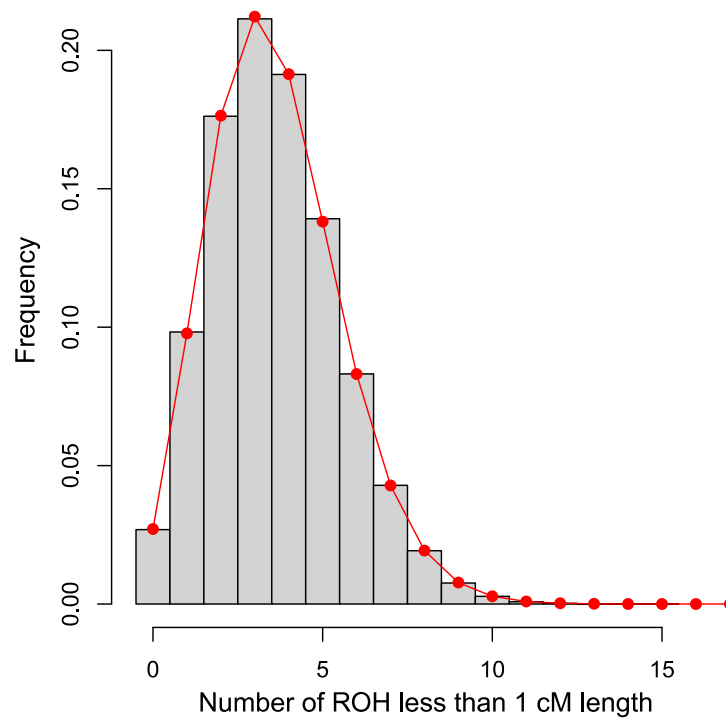


Figure 20: **Histogram of number of ROH less than 1 cM length in UK Biobank British samples.** Overlaying the histogram, in red, is a Poisson distribution with $\lambda = 3.608$ (the mean number of short ROH in this cohort).

The agreement between the theoretical and observed distributions is remarkable (the observed number of short ROH is within 1% of the theoretical prediction up to $N = 9$), making a strong case that short ROH are indeed caused by the random inheritance of distant haplotypes. This stochastic distribution implies that the presence one ROH in a sample is completely uninformative about the chances of there being others, and is therefore also unrelated to all possible confounders.

2.6 ROH simulation results

2.6.1 Estimating ROH age from ROH length

The ROH simulation method described in section 2.4.1 was used to simulate the distribution of ROH lengths (in cM) for haplotypes inherited from a common ancestor between $N=2$ (offspring of full siblings) and $N=100$ generations ago (Figure 21). As expected, for all relative types, the distribution of ROH lengths is approximately exponential and ROH originating from more distant ancestors are

more likely to be shorter (since meiotic recombination has had more opportunities to shorten the ancestral haplotypes).

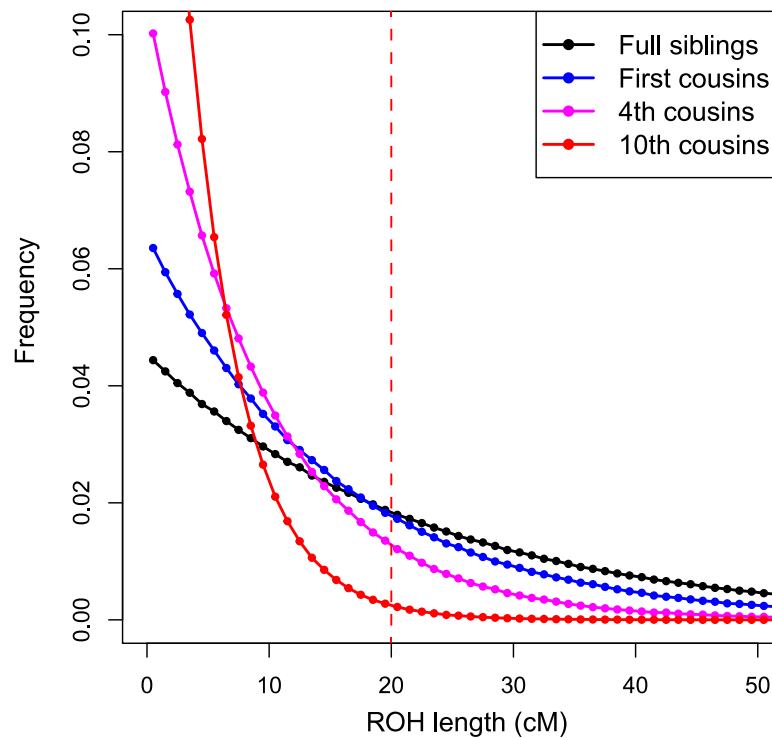


Figure 21: **Distribution of ROH lengths in the offspring of four types of relative.** A dashed red line at 20 cM highlights the relative probabilities for a ROH of 20 cM length, and illustrates how these translate into the ROH age distribution in Figure 22d. Note that a 20 cM ROH is approximately equally likely to occur in the offspring of full siblings and first cousins, corresponding to the similar y-coordinate of the first two points in Figure 22d.

In practice, it is perhaps more useful to estimate how many generations ago a common ancestor might have lived to give rise to a ROH of an observed length (henceforth referred to as the ROH “age”). Methods section 2.6.1 shows that the probable age distribution of a ROH can be estimated from its genetic length, and the probability of observing that ROH length from common ancestry at different numbers of generations ago. This can be visualised as vertical slices through Figure 21 at different ROH lengths. Estimated age distributions are shown at four ROH genetic lengths in Figure 22 below.

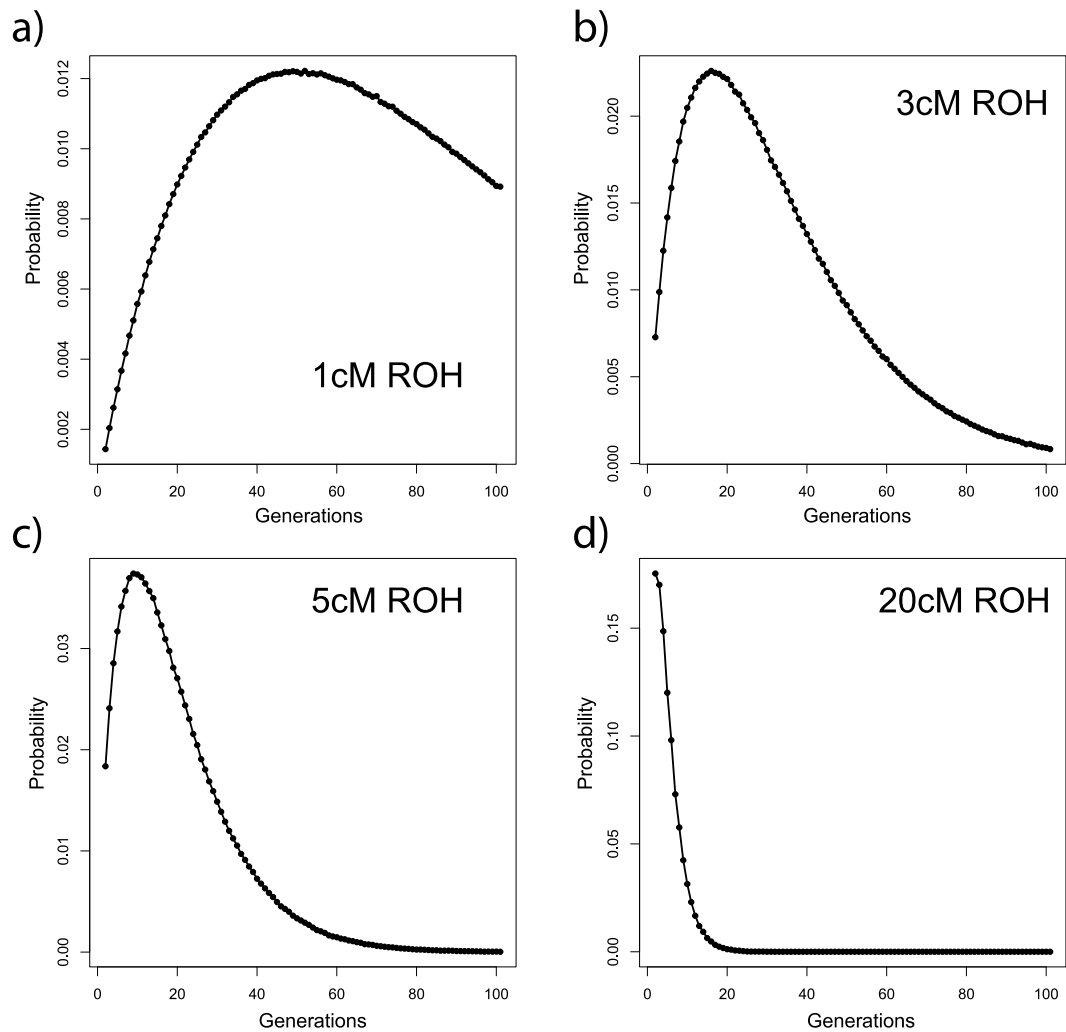


Figure 22: Estimated age distributions for four ROH lengths. The probability of an observed ROH having arisen from a common ancestor N generations ago, is plotted on the y-axis, and against N on the x-axis. Distributions are shown for four ROH genetic lengths (parts a-d), but note that the genome-wide the average recombination rate is approximately 1 cM per Mb, so these distributions are very similar for ROH lengths measured on a physical distance scale (Mb).

ROH of 1 cM in length are most likely to be caused by a common ancestor approximately 55 generations ago, although the distribution is relatively flat and much longer timescales remain quite likely. In contrast, ROH of 20 cM length are most likely caused by an ancestor two or three generations ago (i.e., offspring of full siblings or first cousins) and are very unlikely to be attributable to an ancestor more than 20 generations ago. It should be noted that these estimates assume random mating and that consanguinity is likely to increase the relative probability of ROH originating from a small number of generations ago, particularly for longer ROH.

2.6.2 Simulated F_{ROH} for different relative types

The full haplotype inheritance simulation model (described in section 2.4.2) was used to investigate expected ROH metrics in the offspring of various types of relative. 50,000 replicates of each relative type were simulated and, to correspond with real data, only ROH longer than 1.5 Mb were included in the reported summary statistics (Table 1). An extended version of Table 1, with more columns, is given in Supplementary Table 30.

Table 1: Simulated ROH metrics in the offspring of various types of relative. All columns are calculated only from ROH longer than 1.5 Mb, as this is the minimum length called in real data. Additional columns are reported in Supplementary Table 30.

Relative type	Mean F_{ROH}	S.d. F_{ROH}	Mean N ROH	S.d. N ROH	Mean F_{ROH_X}	S.d. F_{ROH_X}	Ratio F_{ROH_X} / F_{ROH}
Full siblings	0.249	0.047	36.53	5.03	0.250	0.242	1.00
Father – daughter	0.249	0.054	29.35	4.81	0.500	0.336	2.01
Mother – son	0.250	0.054	29.39	4.83	0.499	0.263	2.00
Half siblings (same mother)	0.124	0.040	18.25	4.36	0.250	0.243	2.01
Half siblings (same father)	0.125	0.040	18.30	4.39	0.000	0.000	0.00
Uncle – niece (brother’s daughter)	0.124	0.036	21.67	4.70	0.248	0.241	2.00
Uncle – niece (sister’s daughter)	0.124	0.036	21.67	4.71	0.124	0.177	1.00
Double first cousins (same sex siblings)	0.124	0.033	24.89	4.79	0.187	0.211	1.51
Double first cousins (opposite sex siblings)	0.124	0.033	24.92	4.76	0.124	0.177	1.00
First cousins (Male siblings)	0.062	0.026	12.46	3.87	0	0	0
First cousins (Female siblings)	0.062	0.026	12.43	3.85	0.187	0.212	3.02
First cousins	0.062	0.026	12.42	3.86	0.125	0.179	2.02

(opposite sex siblings)							
Second cousins (random sexes)	0.015	0.012	3.85	2.22	0.020	0.076	1.32
Third cousins (random sexes)	0.004	0.005	1.12	1.18	0.005	0.034	1.36
Sixth cousins (random sexes)	6×10^{-5}	5×10^{-4}	0.025	$\frac{0.16}{4}$	9×10^{-5}	3×10^{-3}	1.47
Tenth cousins (random sexes)	3×10^{-7}	3×10^{-5}	1×10^{-4}	$\frac{0.01}{2}$	0	0	0

For all relative types, the simulated mean F_{ROH} is very close to the theoretical expectation (F_{PED}). For close relatives, only a small fraction of ROH fall below 1.5 Mb, although this rises to 50% in the offspring of 10th cousins (Supplementary Table 30). However, although the mean ROH length drops as the number of meioses to the common ancestors increases, the average ROH length remains as high as 5.2 Mb in the offspring of 10th cousins. Also, in the offspring of distant relatives, the standard deviation of F_{ROH} is large compared to mean F_{ROH} , indicating that F_{ROH} will be a poor metric for differentiating between different distant relative types, and indeed between outbred and slightly inbred individuals.

The ratio of F_{ROH_X} to F_{ROH} varies between 0 and 3 and depends both on the relationship between the simulated parents, and the sexes of their related ancestors. Although calculating the expectation of F_{ROH_X} is necessarily more complex than the autosomal equivalent, the method of loops⁵⁸ can be modified to account for the unique inheritance pattern of the X chromosome. In general, male to female inheritance within a pedigree loop results in higher average F_{ROH_X} , since, if a male carries X chromosome haplotypes from a common ancestor, these are necessarily transmitted to his daughter. In contrast, male to male inheritance within the pedigree loop reduces F_{ROH_X} to zero, since any potentially autozygous X-linked haplotypes are lost. For the offspring of close relatives, relatively few permutations of sex exist, and these were simulated separately in Table 1. For example, first cousins may be related through their fathers (who are brothers), mothers (who are

sisters) or one of each (opposite sex siblings). However, for the offspring of more distant relatives, the permutations become too numerous to simulate separately, so sexes were chosen randomly. Averaging across all relative types, the ratio of F_{ROH_X} to F_{ROH} is greater than one, which is consistent with the theoretical expectation, since the effect population size of the X chromosome is smaller than that of the autosomes. The simulation also confirms that for all relative types, the mean ROH length is longer on the X chromosome than the autosomes, due to a reduced opportunity for recombination.

2.6.3 Simulating multiple generations of cousin marriage

In several highly consanguineous cohorts the distribution of F_{ROH} extends to values well above the expectation for the offspring of first (or even double first) cousins (for example, see Figure 11). It seems possible that this could be caused by multiple generations of cousin marriage, but the expected Inbreeding Coefficient (F) in this scenario was not found in the scientific literature. Therefore, 20 generations of continuous first cousin marriage were simulated, using the full haplotype inheritance model (section 2.4.2), in a constant population size of 1000 individuals, and where every individual has two opposite-sex children with their first cousin. This mating structure is represented (for a population size of 8) in Figure 23.

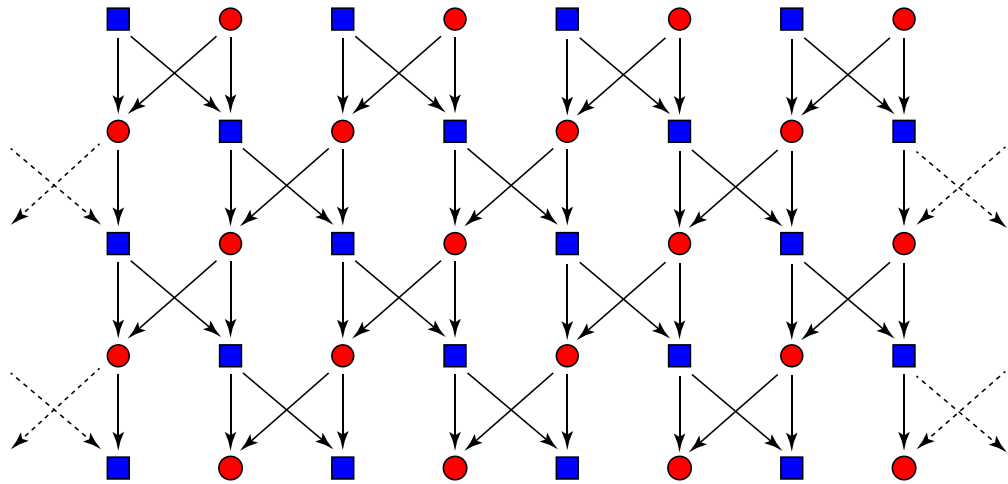


Figure 23: **Mating structure used to simulate multiple generations of first cousin marriage.** Males are shown as blue squares and females as red circles. All individuals from generation 3 onwards have two opposite sex children with a first cousin. For clarity, only 5 generations are shown in a population size of 8, although 20 generations were simulated in a population size of 1000.

The average F_{ROH} across all 1000 individuals was calculated after each generation and is plotted in Figure 24.

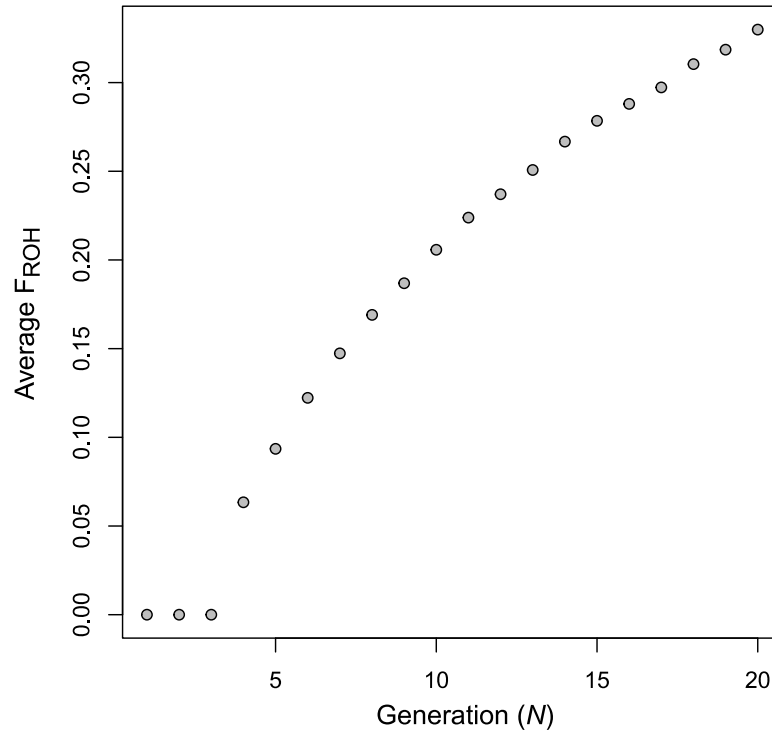


Figure 24: **Average F_{ROH} of 1000 simulated individuals after subsequent generations of first cousin marriage.** The simulated pedigree is shown in Figure 23, and, with the base generation indexed as generation 1, the first inbred individuals (offspring of first cousins) appear in generation 4.

The third generation is the first with reproduction between relatives, and therefore the first inbred individuals (offspring of first cousins) appear in generation four. The mean F_{ROH} in generation 4 is 0.0633, close to the expectation for the offspring of first cousins ($E(F_{ROH}) = 0.0625$). After two generations of cousin marriage (generation 5) the average F_{ROH} is 0.0935 and this continues to rise (at a reducing rate) with subsequent generations. After 10 generations of cousin marriage (generation 13) the average F_{ROH} is 0.251, and hence it seems likely that multiple generations of cousin marriage may explain the long tail of F_{ROH} seen in highly consanguineous cohorts (e.g., Figure 11).

2.6.4 Correlation between siblings of first cousin parents

In Section 4.2.4.3, a method is proposed for estimating the genetic effect of F_{ROH} ($\beta_{F_{ROH}}$) in full sibling families. Unfortunately, the number of sibling genotypes currently available is insufficient to determine whether $\beta_{F_{ROH}}$ differs from zero for most traits. However, it is believed that sufficiently large sibling sample sizes will soon become available and, to estimate the required size of these future studies, it may be useful to simulate the variation of F_{ROH} between siblings in the most common type of marriage between close relatives (i.e., first cousin marriages). Using the full haplotype inheritance model (section 2.4.2), 10,000 first cousin marriages were simulated, each with two offspring (as shown in Figure 25).

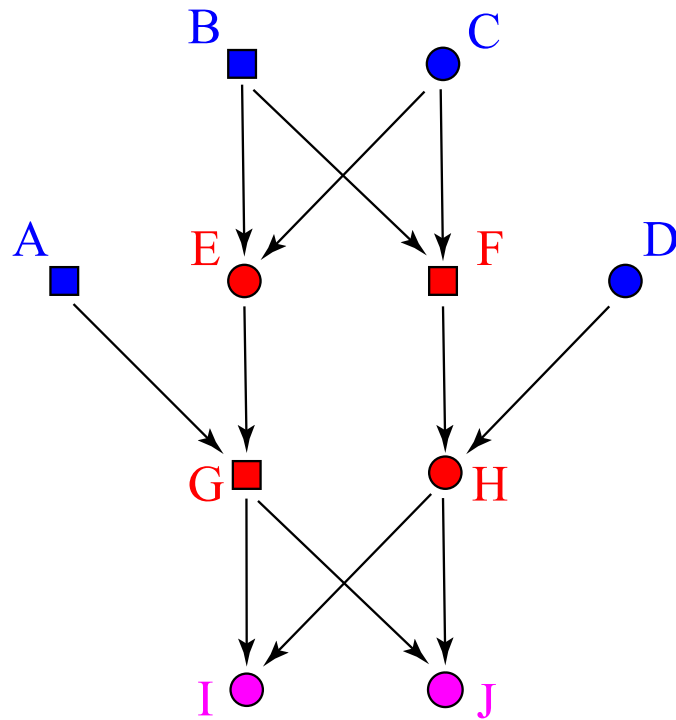


Figure 25: **Simulation pedigree for full siblings (I & J) who are the offspring of first cousins.** Unrelated base generation individuals (A, B, C & D) are shown in blue, related individuals (E, F, G & H) in red and the focal inbred siblings (I & J) are shown in magenta. Males are represented by squares and females by circles.

Although full siblings share the same two parents, they may inherit different haplotypes and therefore have different amounts of autozygosity. This can be seen in Figure 26a, where the F_{ROH} of each sibling pair is plotted for 10,000 simulated first cousin marriages.

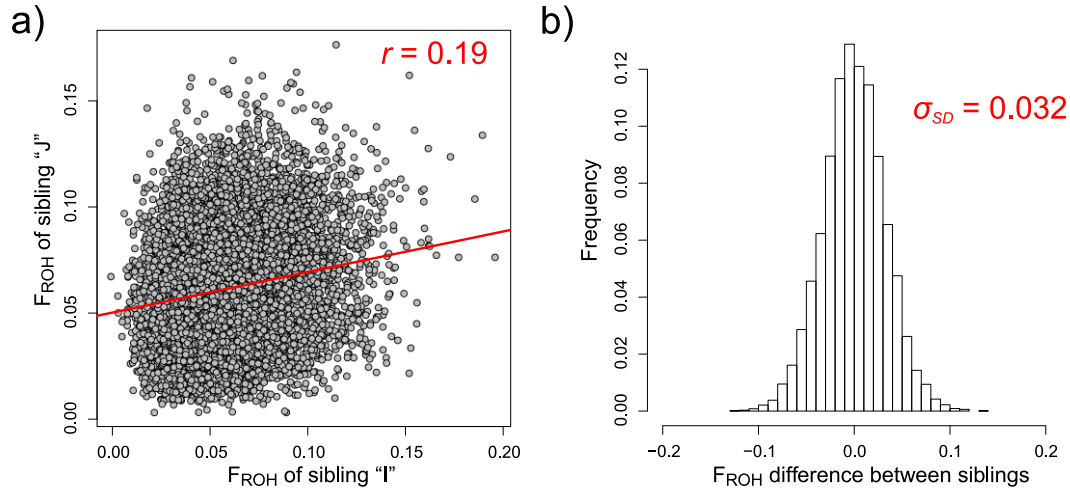


Figure 26: Simulated F_{ROH} of two siblings from 10,000 first cousin marriages. (a) The F_{ROH} of one sibling (J) is plotted against F_{ROH} of the other (I). There is only a weak correlation ($r = 0.19$) between the two values. (b) Histogram of the difference in F_{ROH} between the two simulated siblings. Delta F_{ROH} is normally distributed with standard deviation equal to 0.032.

There is only a weak correlation ($r = 0.19$) of F_{ROH} between siblings, i.e., most of the variation of F_{ROH} is within families, which is beneficial for the statistical efficiency of this method. The difference in F_{ROH} between siblings from first cousin marriages is approximately normally distributed with standard deviation equal to 0.032 (Figure 26b). This value is used in subsequent power calculations (section 6.2.1 and Appendix E.4) to estimate the number of first cousin marriages needed to precisely estimate the genetic effects of F_{ROH} . However, for now, note that the variation of F_{ROH} within first cousin families is substantial, and indeed comparable to the variation in the most powerful ROHgen cohorts (for example, the standard deviation of F_{ROH} in Born in Bradford (UK Pakistani) is 0.046).

3 Associations between inbreeding coefficients (F) and traits

3.1 Introduction

The initial aim of this chapter is to explore which human complex traits are affected by inbreeding, as measured by F_{ROH} . Many previous studies^{13,84,96–99} have addressed the same question but, the sample size of ROHgen, and number of phenotypes available in UK Biobank, are larger than previously available and therefore merit an up-to-date assessment.

Until the discovery of genetic markers, studying the effect of human inbreeding typically required genealogical records, which limited both sample sizes and the range of traits studied. Never-the-less, some inbreeding effects reported from genealogical data (cognition^{100,101}, fertility^{102,103}) appear to have stood the test of time, while others (late onset diseases¹⁰⁴, heart disease¹⁰⁵, hypertension¹⁰⁶, osteoporosis¹⁰⁷, cancer¹⁰⁸, psychiatric disorders^{109,110}) now appear inconsistent with better powered genomic estimates. The abundance of reported associations that now appear inconsistent with current estimates, highlights both the need for large sample sizes, and the potential for confounding by non-genetic factors (which is discussed in Chapter 4).

Detecting the effects of inbreeding has gathered pace with the ever-larger sample sizes available in the genomic era. The rate of this increase has been remarkable: the earliest studies using human genomic data (around 2007) estimated heterozygosity at short tandem repeats (STRs) in hundreds of samples, while today (2021) microarray data at millions of single nucleotide polymorphisms (SNPs) is routinely gathered in millions of subjects. Listing only trait associations which have not subsequently been refuted, the human complex traits so far associated with genomic inbreeding coefficients include: total cholesterol, LDL cholesterol and forced expiratory flow in 2007⁹⁶, height in 2012⁸⁴, household income and educational attainment in 2014⁹⁷, cognition in 2015¹³, number of children fathered, handgrip strength, waist/hip ratio, visual and auditory acuity in 2017⁹⁸, age at first sex in 2018⁷¹, alcohol consumption, smoking, risk-taking behaviour, likelihood of

having children, overall health, heart rate, facial ageing, lymphocyte percentage and haemoglobin levels in 2019¹. More recently (2021) an association has been reported between F_{ROH} and Alzheimer's disease⁹⁹, although replication is suggested since case-control study designs are especially prone to environmental confounding¹¹¹.

In summary, a large number of human complex traits, both binary and quantitative, have been previously associated with F_{ROH} , or other measures of inbreeding. While the genetic nature of these associations remains unproven for many researchers^{112,113}, it has been frequently noted that the effects are often in a direction assumed to be deleterious - as implied by the theory of inbreeding depression on fitness. The traits affected by inbreeding are hence occasionally referred to as fitness-related⁷¹, or subject to directional selection^{13,97,98,114}, although any theory connecting evolutionary fitness to quantitative traits appears to be lacking. Inbreeding depression on fitness can be theoretically explained by mutational-selection balance on partially recessive variants, but why this reduction in fitness should translate to observable changes in some traits, but not in others, is not well explained. Physical traits which change with inbreeding are occasionally described as subject to directional selection, but it seems unlikely that traits such as cognition (which has reliably been shown to reduce with increased F_{ROH} ^{13,70}) are subject to directional selection. Directional selection on a heritable trait should result in a response to selection, but it seems unlikely that genetic scores for cognition are increasing substantially, despite abundant genetic variation¹¹⁵. It seems more plausible to this author that the majority of genetic variation in quantitative traits is neutral for fitness, consistent with the theory that components of fitness should have low heritability. A heuristic explanation for the traits affected by inbreeding depression is discussed in Appendix F.

The second part of this chapter (section 3.3.7), compares three different estimators of inbreeding coefficients (F_{ROH} , F_{SNP} , and F_{GRM}), and attempts to draw conclusions about the allele frequencies of causal variants from that comparison.

There are currently differing opinions about whether F_{ROH} or F_{GRM} (also called F_{UNI} , or $\hat{F}_{III}$) is most suitable for estimating inbreeding depression in human genomic data^{98,114} - and no doubt each has its own merits in different situations. Early genomic studies of inbreeding used genome-wide heterozygosity (h)¹¹⁶, although this was quickly improved by accounting for allele frequencies, and hence replaced by excess homozygosity (F_{SNP})⁹⁶. More recently, Yang et al. proposed a measure based on the correlation between uniting gametes, which they called $\hat{F}_{III}$, and which gives greater weight to rare homozygotes⁶⁰. This measure has become the preferred marker-by-marker metric since it has been shown to predict inbreeding depression better than F_{SNP} in real data⁹⁸. Alternatively, inbreeding coefficients may be estimated by identifying autozygous genomic segments, as in F_{ROH} . This has great appeal since it accounts for what is known about the discrete nature of DNA inheritance, but has the disadvantage of requiring parameterisation to correctly identify autozygous segments.

Several previous studies have compared the performance of F_{ROH} and SNP-based measures of genome-wide homozygosity – with occasionally conflicting conclusions^{98,114,117,118}. Keller et al. conclude that “ F_{ROH} is preferable to F_{PED} and to marker-by-marker estimates of F (F_{alt} [F_{GRM}] and F_h [F_{SNP}]) for detecting both recent and ancient inbreeding, even in cases where the level of inbreeding is likely to be high”¹¹⁴. In contrast, Yengo et al. state that “Our study illustrated some of the shortcomings of the most commonly used ROH-based estimates of ID, which not only are biased but also have large SE (approximately three times larger compared with F_{UNI} [F_{GRM}])”⁹⁸. Appendix C.4 shows that this latter conclusion is most likely caused by a mis-parameterisation of the ROH calling algorithm, however even if there is a building consensus that F_{ROH} is typically the best available estimate of F , there may be scenarios in which F_{GRM} is a better, or at least complementary, predictor of inbreeding depression. Yengo et al. show that variation in waist-to-hip ratio is more associated with F_{GRM} than F_{ROH} (replicated here in section 3.3.7), suggesting that variants causing this effect may be predominantly common. Since the variation of F_{GRM} which is independent of F_{ROH} is not caused by recent

inbreeding (but rather chance, in a homogeneous population) it may be more appropriate to call this effect directional dominance instead of inbreeding depression.

The final part of this chapter (section 3.3.8) estimates the effects of different ROH lengths, as also done previously by other studies^{70,71}. However, as noted by Johnson et al., dividing F_{ROH} into correlated parts increases the standard errors, which has made it difficult to draw firm conclusions from the sample sizes available to date. In addition, results can be interpreted in different ways¹¹¹. For example, significant differences between ROH lengths can be interpreted as evidence of differential confounding, or explained by arguing that the variants made homozygous in shorter ROH have potentially been exposed to purifying selection for longer than those in long ROH.

3.2 Method

3.2.1 Phenotypes

3.2.1.1 In ROHgen

The initial ROHgen analysis plan specified the analysis of 42 quantitative traits. These traits were chosen to represent different domains of health and reproductive success, with consideration given to presumed data availability. Many of these traits had been the subject of existing genome-wide association meta-analyses (GWAMA), and phenotype modelling, such as inclusion of relevant covariates, was copied from the relevant consortia (GIANT for anthropometry, EGG for birth weight, ICBP for blood pressures, MAGIC for glycaemic traits, CHARGE-Cognitive, -Inflammation & -Haemostasis working groups for cognitive function, CRP, fibrinogen, CHARGE-CKDgen for eGFR, CHARGE-ReproGen for ages at menarche and menopause, Blood Cell & HaemGen for haematology, GUGC for urate, RRgen, PRIMA, QRS & QT-IGC for electrocardiography, GLGC for classical lipids, CREAM for spherical equivalent refraction, Spirometa for lung function traits, and SSGAC for educational attainment and number of children ever born). Most traits were analysed with both sexes

combined, although two traits (age at first birth (afb) and number of children ever born (neb)) were also analysed as separate traits for men and women.

Shortly after the initial ROHgen analysis plan was distributed to cohorts (in late 2016) the first phase of UK Biobank data was released. UK Biobank is a large, deeply-phenotyped, population cohort and numerous traits were found to be associated with F_{ROH} (Section 3.3.2). Therefore, seeking independent replication, 3 further traits (ever had children, ever smoked, and alcohol units per week) were added to the ROHgen analysis plan from April 2017.

Altogether, the initial 42 traits, the 4 four sex-specific traits, and the 3 late additions make a total of 49 traits reported here from ROHgen. Further information about individual phenotypes is available in Supplementary Tables 6, 7, and 8.

3.2.1.2 In UK Biobank

3.2.1.2.1 Preparing UK Biobank phenotypes

UKB Phenotyping was done by at least one, and up to three, assessment centre visits. During these visits, participants completed an extensive touchscreen questionnaire, biometric assessment, and blood draw. Subsequent linkage to UK electronic health records also provides ongoing phenotypes such as hospital in-patient ICD10 codes and death registration.

The UKB phenotype file available to this study contains 13496 columns and 502,000 rows - representing almost all of the traits collected on 502,000 UKB participants. Each phenotype column is labelled in the format f.FIELD_ID.x.y, where FIELD_ID is a unique trait identifier, and x and y are indices indicating multiple measurements (y) made at up to three separate assessment centre visits (x).

The 13496 columns of the phenotype file hence represent multiple measurements of 2383 unique FIELD_IDs – each corresponding to a trait described in detail in the UKB data showcase (<https://biobank.ndph.ox.ac.uk/ukb/>) and summarized in Supplementary Table 40. These traits were initially prepared for association testing as described below:

- For all numerical traits: integer values commonly used in UKB to convey special meaning (-999, -818, -121, -10, -3, -2, -1) were removed.
- For all Date traits: date values were converted to an integer number of days passed since 1st January 1900.
- For categorical traits: values of “Do not know” and “Prefer not to answer” were removed and the remaining categories placed in order of the mean F_{ROH} for participants who had given that response. Manually assigning values to each response was also considered, but this missed associations where the ordering of responses was not self-evident. However, ordering categorical responses by mean F_{ROH} was also found to bias $\hat{\beta}_{F_{ROH}}$ for traits containing a large number of categories, since a participant’s F_{ROH} contributes to the mean F_{ROH} of that category. Therefore, for categorical phenotypes with greater than 50 categories, the category order was recalculated for each participant with their own F_{ROH} excluded.

This processing results in 2383 numerical columns which can be tested for association with F_{ROH} . However, not all of these columns are meaningful phenotypes of the participant. For example, some columns describe the device IDs used to measure traits (e.g., blood pressure, or grip strength) and others describe phenotypes which cannot be causally affected by F_{ROH} , such as sex, or ethnicity. Each trait was therefore manually assigned to one of three groups:

- 1) Traits which might reasonably be affected by F_{ROH} (1786 traits)
- 2) Traits which cannot be affected by F_{ROH} , but might be considered as covariates (77 traits)
- 3) Other variables of no interest (514 traits)

$\beta_{F_{ROH}}$ was first estimated for all traits in group 2 (results in Supplementary Table 22) conditioning on an initial set of 64 a priori covariates (sex, age, age², year of birth, assessment centre, country of birth, ethnicity, Ashkenazi, easting and northing of home address, easting and northing of birth address, Townsend deprivation at home address, Index of Multiple Deprivation at home and birth address, breastfed

as a baby, mother still alive, mothers age at birth, mothers age at death, father still alive, fathers age at birth, fathers age at death, genotyping array, binary variable indicating membership of white British ancestry subset and 40 principal components). Although many of these a priori covariates are strongly associated with F_{ROH} (Supplementary Table 28), some are not (for example genotyping array, and several principal components). These F_{ROH} non-associated covariates were nonetheless retained, since the cost of including covariates is small when there are many degrees of freedom, and they may explain variation in other phenotypes.

12 “group 2” traits (out of 77) are significantly associated with F_{ROH} at a 5% false-discovery rate when conditioning on these 64 a priori covariates. These 12 traits were each manually assessed (and in some cases combined) yielding only five F_{ROH} -associated covariates which had not been previously considered: maternal smoking, adopted as a baby, number of full siblings, missingness, and attendance at a follow-up appointment. Several of the most strongly associated group 2 traits are alternative measures of homozygosity (e.g., f.22003 Heterozygosity) and were therefore not included as covariates. Separately, the results in Chapter 4 identified five individual traits (i.e., group 1 traits) that are likely to be markers of social confounding, namely: Educational attainment, religious group attendance, pork consumption, lamb consumption and alcohol drinking status. To help further control confounding these five traits were also included as covariates. Therefore, a total of 74 covariates (64 + 5 + 5) were used in all subsequent UKB phenotype models.

$\beta_{F_{ROH}}$ was then estimated for all group 1 traits conditioning on these 74 covariates (results in Supplementary Table 23). Of the 1786 potential phenotypes, 255 are significantly associated with F_{ROH} at a 5% false discovery rate. As done previously for the potential covariates, these 255 significant traits were manually assessed, and in some cases combined, resulting in a list of 133 UKB traits significantly associated with F_{ROH} . For example, highly correlated traits were either averaged (e.g., when measured on right and left sides) or the most strongly associated trait

chosen (e.g., trunk fat free mass was chosen as representative of 10 highly correlated impedance-derived phenotypes).

3.2.1.2.2 UK Biobank hospital admission ICD10 codes

The UK Biobank phenotype data includes two fields (41202 and 41204) reporting all main and secondary ICD10 codes recorded across all subjects' hospital in-patient records. These ICD10 codes have a hierarchical structure where an initial single letter (A-Z) describes a category of diagnosis and up to four subsequent digits (0-9) give increasing levels of detail. Due to the unique structure of this phenotype, with over 11,657 unique ICD10 codes in 398,472 subjects, this data was analysed separately from the other phenotypes described previously.

For each ICD10 code, a binary variable was created indicating whether a subject had that code (or matching sub-code) recorded anywhere in their main or secondary diagnoses. Because of the hierarchical structure of ICD10 codes, all 13646 'nodes' in the tree were tested, not only the 11,657 used codes. For example, the ICD10 code "E" (Endocrine, nutritional and metabolic diseases) is not itself used, but captures any more precise code starting with "E" (e.g., E119 Non-insulin-dependent diabetes mellitus without complications).

Approximately 50% of ICD10 codes are recorded in fewer than 20 subjects and 90% are recorded in fewer than 1000 subjects, so correction for multiple testing is necessarily stringent, and potentially real associations in common ICD10 codes may be overlooked by correcting for less interesting, rare ICD10 codes. Therefore, the false-discovery-rate Q values reported in Supplementary Table 24 were calculated by including only ICD10 codes with an equal or greater number of cases than the test being corrected.

3.2.1.2.3 UK Biobank self-reported illnesses

UK Biobank field 20002 reports all non-cancer self-reported illness for UKB participants across 66 phenotype columns (i.e., subjects have up to 66 self-reported illnesses). Unlike ICD10 codes these self-reported disease codes are not hierarchical, so only the 446 unique codes themselves were tested for association with F_{ROH} . In

all other respects the self-reported illness analysis was identical to the analysis of ICD10 codes.

3.2.2 Overlap between ROHgen and UKB

A number of analyses described in this thesis were conducted in both ROHgen and UKB. Since UKB contributed to ROHgen, the results of these analyses are not fully independent. Of the 155 traits which are significantly associated with F_{ROH} in UKB, 17 were also included in ROHgen. For these 17 traits UKB contributes a variable fraction of the total ROHgen meta-analysis power, ranging from approximately 10% for commonly measured traits, such as height and weight, to 58% for cognitive g and 84% for alcohol consumption. However, none of the 18 significantly associated ROHgen traits drop below experiment-wise significance when UKB is excluded from the meta-analysis and the agreement between estimates from UKB alone and ROHgen (excluding UKB) is generally good (Figure 27).

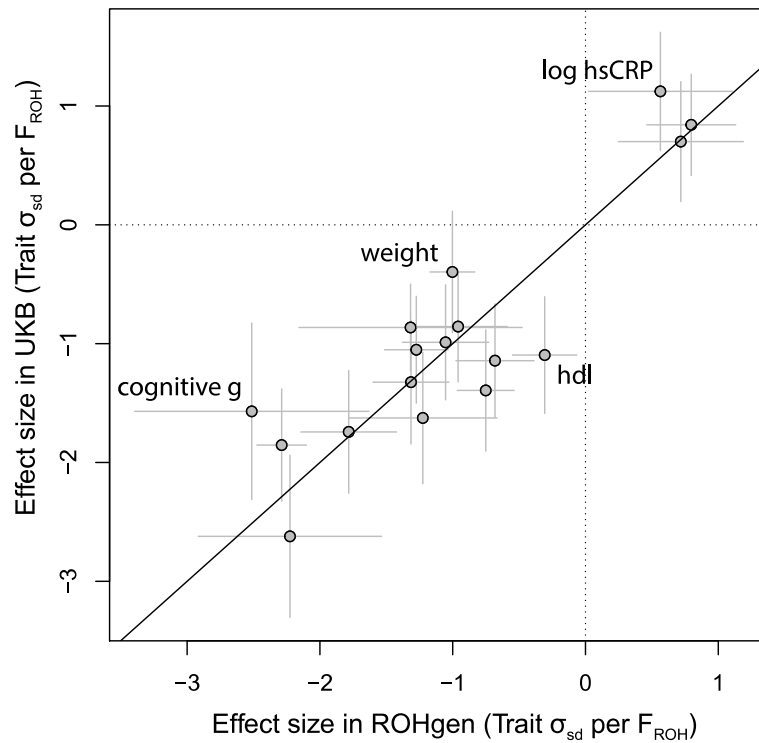


Figure 27: **Comparison of $\hat{\beta}_{F_{ROH}}$ in UK Biobank and ROHgen (excluding UKB) for 17 significant traits.** All traits are scaled by the intra-sex standard deviation and all error bars represent 95% confidence intervals.

It might also be noted that Clark et al. reported on 100 traits: the 49 ROHgen traits, and 51 traits measured only in UKB. Since this thesis reports the UKB results separately, the 51 traits unique to UKB have been removed from the ROHgen meta-analyses reported in this thesis.

3.2.3 Estimating $\beta_{F_{ROH}}$ for quantitative traits

3.2.3.1 In ROHgen

In each cohort of n samples, and for each of the quantitative traits measured in that cohort, trait values were modelled as

$$\mathbf{y} = \beta_{F_{ROH}} * \mathbf{F}_{ROH} + \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \quad (16)$$

where $\mathbf{y}$ is a vector ($n \times 1$) of measured trait values, $\beta_{F_{ROH}}$ is the unknown scalar effect of F_{ROH} on the trait, $\mathbf{F}_{ROH}$ is a known vector ($n \times 1$) of individual F_{ROH} , $\mathbf{b}$ is a vector ($m \times 1$) of unknown fixed covariate effects (including a mean, μ), $\mathbf{X}$ is a known design matrix ($n \times m$) for the fixed effects, and $\boldsymbol{\varepsilon}$ is an unknown vector ($n \times 1$) of residuals.

For reasons of computational efficiency, it was occasionally necessary to solve equation (16) in two steps. In the first step, the trait ($\mathbf{y}$) was regressed on all fixed covariates to obtain the maximum likelihood solution of the model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}' \quad (17)$$

In the ROHgen pipeline all subsequent analyses were performed using the vector of trait residuals $\boldsymbol{\varepsilon}'$, which may be considered as the trait values corrected for all known covariates.

In cohorts with a high degree of relatedness, mixed-modelling was used to correct for family structure, although, because ROH are not narrow-sense heritable, this is believed to be less essential than in Genome-Wide Association Studies. When mixed modelling is used equation (16) becomes

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u} + \boldsymbol{\varepsilon}' \quad (18)$$

Where $\mathbf{u}$ is an unknown vector ($n \times 1$) of polygenic effects with multivariate normal distribution of mean 0 and covariance matrix $\sigma_g^2 \mathbf{A}$, where $\mathbf{A}$ is the genomic relationship matrix (GRM). In these related cohorts, a GRM was calculated using PLINK v1.9 and Grammar+ residuals of equation (18) were estimated using GenABEL¹¹⁹. These Grammar+ residuals ($\boldsymbol{\varepsilon}'$) were then used in subsequent analyses. To estimate $\beta_{F_{ROH}}$ for each trait, trait residuals were regressed on F_{ROH} to obtain the maximum likelihood (ML) solution of the model:

$$\boldsymbol{\varepsilon}' = \mu + \beta_{F_{ROH}} * \mathbf{F}_{ROH} + \boldsymbol{\varepsilon} \quad (19)$$

The sex-specific estimates of $\beta_{F_{ROH}}$ (Supplementary Table 11) were obtained from equation (19) applied to the relevant sex.

For all traits, corresponding estimates of $\beta_{F_{SNP}}$ and $\beta_{F_{GRM}}$ were obtained from the models:

$$\boldsymbol{\varepsilon}' = \mu + \beta_{F_{SNP}} * \mathbf{F}_{SNP} + \boldsymbol{\varepsilon} \quad (20)$$

$$\boldsymbol{\varepsilon}' = \mu + \beta_{F_{GRM}} * \mathbf{F}_{GRM} + \boldsymbol{\varepsilon} \quad (21)$$

and the effects of different ROH lengths, (Figure 43), were estimated from the model:

$$\begin{aligned} \boldsymbol{\varepsilon}' = & \mu + (\beta_1 * \mathbf{F}_{ROH < 3Mb}) + (\beta_2 * \mathbf{F}_{3Mb < ROH < 5Mb}) \\ & + (\beta_3 * \mathbf{F}_{ROH > 5Mb}) + \boldsymbol{\varepsilon} \end{aligned} \quad (22)$$

3.2.3.2 In UK Biobank

Where UKB contributed to the ROHgen meta-analyses it was processed in the same manner as any other cohort. I.e., the full cohort was split into 10 separate ethnicities (based on self-declared ethnicity and principal components) and analysed by the ROHgen software implementing the analyses described in sections 3.2.3.1 and 3.2.4.1. In subsequent analyses, performed only in UKB, a few changes were made with the benefit of hindsight and without the constraints imposed by distributing software to 117 different analysts.

In the standalone UKB analyses, all estimates of $\beta_{F_{ROH}}$ were obtained from full phenotype models where the effects of F_{ROH} and all 74 covariates were estimated simultaneously. In ROHgen all quantitative trait effects had been estimated in two steps: first fitting covariates and then F_{ROH} . In practice, estimates of $\beta_{F_{ROH}}$ from the two-step approximation are almost identical to the full model (Figure 28) because F_{ROH} has minimal multicollinearity with the other covariates (in UKB the variance inflation factor of F_{ROH} in the full model is 1.06). Never-the-less the full model was used in UKB since it is the more accurate method and access to high-performance-computing made this feasible.

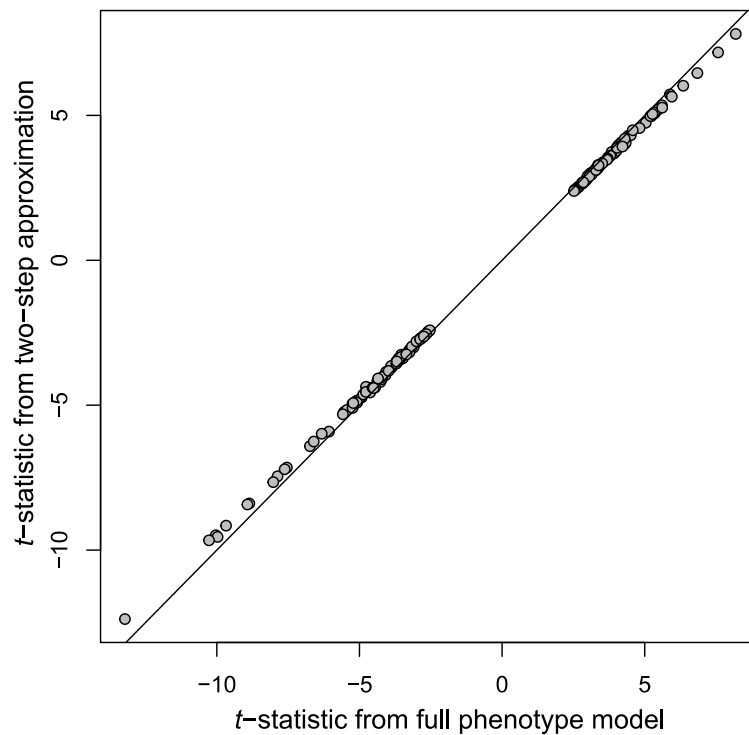


Figure 28: **Comparison of the t -statistics of $\hat{\beta}_{F_{ROH}}$ for 155 significant UKB traits, estimated by two different methods.** $\beta_{F_{ROH}}$ is estimated simultaneously with 74 covariates on the x-axis and from a two-step approximation on the y-axis.

In ROHgen, the UKB cohort was split into 10 separate ethnicity cohorts to follow the guidelines specified in the ROHgen analysis plan. However, analysing 10 separate cohorts complicates the standalone UKB analyses and can be equally well accounted for by fitting ethnicity as a covariate. Splitting the cohort into 10 distinct ethnicity cohorts removes any between ethnicity effects, but meta-analysing the

results assumes the same effect in each cohort – fitting ethnicity as a covariate is much simpler and achieves the same objectives. Splitting the cohort does have the advantage that the effects in different ethnicities can be compared before meta-analysis, but this can also be achieved by fitting an interaction between ethnicity and F_{ROH} (as done in Sections 4.2.4.1.2 and 4.3.2.1).

3.2.4 Binary phenotypes

3.2.4.1 In ROHgen

The 46 traits in the first ROHgen analysis plan were all quantitative traits, but it subsequently became clear that it would also be useful to analyse binary traits. This was initially motivated by the realisation that the association of F_{ROH} with whether a subject had any children (*ever_had_children*) was substantially stronger than the association with the number of children they had (*neb*). Therefore, two binary traits (*ever_had_children* and *ever_smoked*) were added to the ROHgen analysis plan and analysed by two different methods - logistic and linear regression. The primary estimates of $\beta_{F_{ROH}}$ reported in Clark et al. (Fig. 3b and Supplementary Data Table 10), were obtained from full logistic models:

$$g(E[\mathbf{y}]) = \mathbf{X}\mathbf{b} \quad (23)$$

Where $g()$ is the link function (logit), and where F_{ROH} and all applicable covariates were fitted simultaneously. Mixed modelling for family structure was not attempted in the logistic models since an accepted method was not found.

For all other results reported in Clark et al., binary trait values ($\mathbf{y}$) were scaled by $1/\sigma_y^2$ and analysed by linear models, as for quantitative traits, including mixed modelling where appropriate for family studies. For reasonably balanced case-control ratios, this method of estimating binary traits with simple linear models gives asymptotically unbiased estimates of $\beta_{F_{ROH}}$ and $se(\beta_{F_{ROH}})$ on the $\ln(\text{Odds-Ratio})$ scale¹²⁰ (See also Clark et al. Supplementary Figure 8).

The results from ROHgen showed that the response of the two binary phenotypes to F_{ROH} could be well described by a linear model (where probability of being a case

is proportional to F_{ROH}) – see Figure 63b. In practice, for the small effects observed, a logistic model predicts an approximately linear response over the entire range of F_{ROH} , so the choice of logistic or linear modelling is of little consequence. Therefore, for simplicity, all results reported in this thesis are from simple linear regression.

3.2.4.2 In UKB

There are a large number of binary traits in UKB, both in the ordinary phenotypes and, particularly, in the analysis of ICD10 codes. As for the binary ROHgen results reported in this thesis, the UKB binary traits were analysed by linear regression – in the same way as a quantitative trait. This both describes the behaviour of the response well (Figure 63b) and solves the computational challenges of fitting tens of thousands of logistic models with many covariates.

However, linear models of binary variables potentially violate two of the assumptions of linear regression: normality, and homoscedasticity of the residuals. This is particularly problematic for traits with extreme case-control ratios when F_{ROH} has a relatively large effect. In this situation the mean probability of being a case may vary substantially between high and low F_{ROH} and, since for binomial distributions the variance is equal to the mean, there may be substantial heteroskedasticity across the range of F_{ROH} . Although $\hat{\beta}_{F_{ROH}}$ is unbiased in this scenario, $se(\hat{\beta}_{F_{ROH}})$ may be downwardly biased causing erroneously significant p -values. To address the issues of heteroskedasticity, all standard errors of $\hat{\beta}_{F_{ROH}}$ in UKB, for both binary and quantitative traits, were estimated by resampling - as described in section 3.2.6.1.

3.2.5 Covariates

To estimate $\beta_{F_{ROH}}$ with minimum bias and maximum efficiency it is generally desirable to explain as much trait variation as possible by fitting covariates in the phenotype models. Since the number of samples available is typically very large, particularly in UKB, the reduction of degrees-of-freedom caused by additional covariates is not of great concern. However, as the aim is to estimate the causal effect of F_{ROH} on a trait ($\beta_{F_{ROH}}$) some care must be taken not to fit covariates which

are themselves causally affected by F_{ROH} (unless the conditional effect is desired). In general, any phenotype which correlates with F_{ROH} and/or a trait of interest, and is not itself causally affected by F_{ROH} is a suitable covariate. For example, age and sex are unlikely to be affected by F_{ROH} and were used as covariates in all models. The situation is slightly less obvious for ancestry-informative principal components (PCs) which often correlate with F_{ROH} , and might conceivably be affected by autozygosity. However, PCs are typically calculated from additive Genomic Relationship Matrix (GRM) and are probably not able to capture non-additive variation. This understanding is supported by Figures 86 and 87 which shows that F_{ROH} is not significantly associated with any PCs when ancestry is accounted for. In practice, all available PCs were fit as covariates in all models.

3.2.5.1 In ROHgen

In ROHgen, the covariates included in each model were chosen with reference to the leading GWAMA consortium for that trait and are detailed in Supplementary Table 7. For all traits, these covariates included: age (and/or year of birth), sex, and at least the first 10 principal components of the genomic relatedness matrix (GRM). Where necessary, additional adjustments were made for study site, medications, and other relevant covariates (Supplementary Table 5).

3.2.5.2 In UK Biobank

In UKB, deep phenotyping and the availability of individual-level data permitted a large number of variables to be considered as covariates. All available UKB traits were tested for association with F_{ROH} and all those which associated significantly, and were thought not to be causally affected by F_{ROH} , were included as covariates in future models.

It was less clear cut whether to include traits which might be confounders, but might also be directly affected by F_{ROH} – such as educational attainment. However, since previous studies have identified education, socio-economic status and religiosity as potential confounders of F_{ROH} ^{71,112}, it was felt better to also include all potentially relevant traits as covariates.

UK Biobank does contain several traits which are obvious markers of education and socioeconomic status (e.g., Education attained, Townsend deprivation Index of home address, etc.) but did not directly ask any questions about religious beliefs. Nevertheless, it did ask about social activities, and F_{ROH} is strongly associated with taking part in a religious group as a social activity. It seems unlikely this is a causal effect, but more plausible that increased F_{ROH} is confounded by religious beliefs. Equally, F_{ROH} associations with pork, lamb, and alcohol consumption are thought to be markers of religious or cultural confounding (see Chapter 4) and these traits were fitted as covariates in all models.

The full list of 74 covariates fitted to all UKB phenotype models is available in Supplementary Table 28.

3.2.6 Reducing bias in statistical tests

All regression results reported in this thesis are obtained by linear regression and therefore subject to its assumptions, namely:

- 1) Linear relationship between dependent and predictor variables.
- 2) Independence between residuals.
- 3) Normality of the residuals.
- 4) The residuals have equal variance at all levels of the predictor variables, i.e., homoscedasticity.

As in almost all real-world applications of linear regression, all of these assumptions are violated to some extent in the models employed here. However, various precautions were employed to reduce any bias caused by these violations:

- 1) Linear relationship. It seems reasonable to assume that the effects of autozygosity would be proportional to F_{ROH} , and even for binary traits this does appear to be the case (Figure 63). Some traits with very heavily skewed distributions were Log or Inverse normal transformed, primarily to address the assumption of normality, although these transformations are also likely to improve the linearity of the dependent-predictor relationship.

- 2) Independence. Any unfitted covariate is likely to introduce correlation between the model residuals, and generally all reasonable covariates were encouraged. In ROHgen cohorts with a high degree of relatedness, linear mixed models were used to reduce residual correlation between family members. In UKB, mixed modelling was not attempted due to the large sample size, although smaller cohorts confirm this is unlikely to be important since F_{ROH} itself is not heritable.
- 3) Normality. The central limit theorem implies that for large samples sizes the regression coefficient estimates will be approximately normally distributed even when the residuals are not. Nevertheless, in ROHgen some traits with heavily skewed distributions were Log or Inverse normal transformed. In UKB non-normality was addressed by resampling as described in section 3.2.6.1.
- 4) Heteroskedasticity of the residuals does not bias $\hat{\beta}$, but may downwardly bias estimates of the standard error of $\hat{\beta}$ resulting in incorrect p -values and potentially false inferences. In UKB several traits are severely skewed and the problem of heteroskedasticity is exacerbated by the highly skewed distribution of F_{ROH} itself – the variance of the trait residuals is largely informed by the large number of low F_{ROH} samples while much of the statistical power comes from relatively few high F_{ROH} samples. Therefore, for all the main UKB results, null t -statistic distributions were obtained by resampling.

3.2.6.1 *Resampling test statistics in UKB*

It was not computationally practical to resample every full phenotype model in UKB (each model takes about 1 minute), so residuals were first calculated by fitting a single model, with all 74 covariates, for each trait (equation (17)). Because there is only weak multicollinearity between F_{ROH} and the 74 covariates, taking this two-step approach substantially reduces computational time with minimal compromise to accuracy (see Figure 28).

For each trait, the null distribution of t -statistics was determined by permuting trait residuals (ϵ') and fitting the model $trait_residuals \sim F_{ROH}$ 250,000 times. These null distributions were then used to transform observed test statistics into equivalent Gaussian t -statistics by an inverse rank-normal transformation as depicted in Figure 29. For real traits, test statistics outside the resampled range are occasionally observed (when the Null hypothesis is false) and this part of the distribution was estimated by extrapolating based on the 50 most extreme resampled test statistics at either end of the distribution.

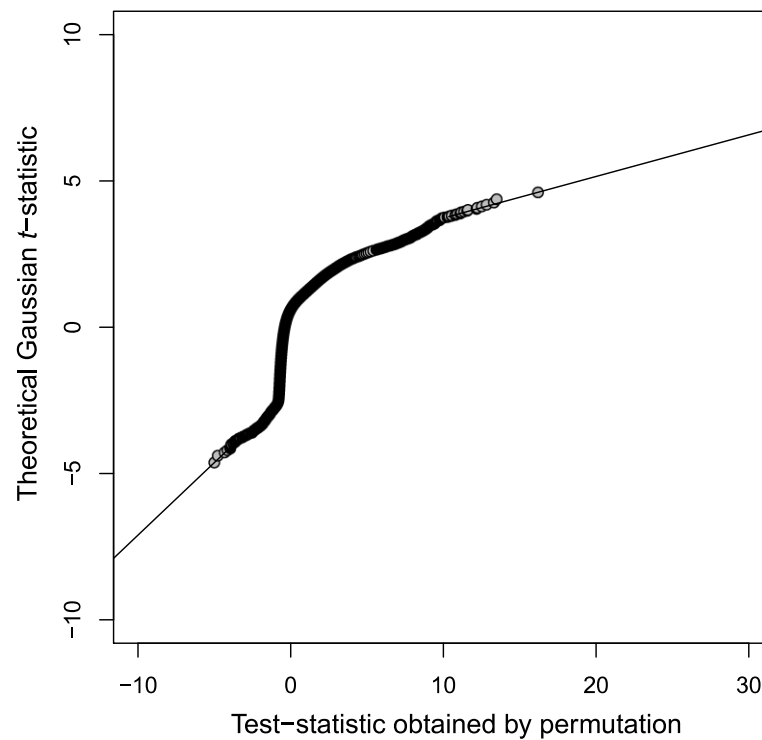


Figure 29: Gaussian equivalent t -statistics plotted against resampled test statistics obtained by permutation for the trait `icd10_male_infertility`. Only 36 subjects have an `icd10` code for male infertility and linear regression on this extremely skewed trait causes upwardly biased test statistics. When the phenotype is randomly shuffled between all 488,288 UKB participants, 250,000 times, the highest test-statistic recorded was 16.2 compared to 4.6 expected from a Gaussian distribution. For some traits, the observed linear regression test statistic lies outside the resampled range, and these were mapped onto equivalent theoretical Gaussian t -statistics by extrapolating based on the 50 most extreme test statistics at either end of the distribution – shown as solid black lines.

Since the null distribution of test statistics differs for different models, this resampling of test statistics was repeated for all 155 traits for:

- UPD model, F_{ROH_wUPD} - equation (35)
- Genetically British model – equation (32)
- Adopted model, $F_{ROH_wAdopted}$ - equation (38)
- Within siblings model, F_{ROH_wSibs} - equation (34)
- Centimorgan model, F_{ROH_cM} - equation (37)
- X Chromosome model, F_{ROH_X} – equation (45)
- Deletions model, F_{DEL} – equation (39)
- Marker homozygosity models, F_{SNP} and F_{GRM} – equation (26)
- Gene density model, F_{GENES} – equation (47)
- Each sex separately – section 3.2.8,

For the many binary traits associated with ICD10 codes and self-reported illnesses it was not computationally practical to resample every trait. Instead, 16 example traits with a range of case numbers (1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000 and 100000) were each resampled 250,000 times. For traits with cases numbers between these 14 resampled values, the resampled t -statistic was estimated by interpolation between the bounding values. For traits with more than 100,000 cases, the distribution for 100,000 was used as this is already approximately Gaussian.

All t -statistics quoted in this thesis are thus expected to be normally distributed with mean=0 and standard deviation=1 under the null hypothesis. The one exception is Supplementary Table 23 where the very large number of traits precluded the correction described here.

3.2.7 Multivariate trait

F_{ROH} was found to have relatively small effects on many traits (see section 3.3.5) and it was occasionally useful to summarise these multiple effects into a single, well-powered, “compound” trait. Since the effects of F_{ROH} are approximately linear (Figure 63), a suitable compound trait would be: the linear combination of individual traits which has the strongest association with F_{ROH} . This can be estimated by regressing F_{ROH} on the traits’ residuals (ϵ') from equation (17):

$$F_{ROH} \sim trait1 + trait2 + trait3 \dots + traitN \quad (24)$$

A predicted value of F_{ROH} can then be calculated for each sample from trait residuals (ϵ') using the R function predict. So that the F_{ROH} of a sample does not contribute to the estimates used to predict its own value, this prediction was done by a 50- part out of sample method. This weighted combination of phenotypes is referred to in this thesis as the “*multivariate*” trait, and since it summarises the effects of F_{ROH} on a large number of phenotypes it might be considered somewhat analogous to evolutionary fitness.

$$multivariate_i = \hat{\beta}_1 trait1_i + \hat{\beta}_2 trait2_i \dots + \hat{\beta}_N traitN_i \quad (25)$$

The *multivariate* trait was calculated in two large cohorts: UKB and deCODE-Iceland. Since different phenotypes were measured in each cohort, the two traits are not directly comparable, so each was normalised to have a regression coefficient of -1 with F_{ROH} .

In UKB, *multivariate* was calculated from 145 individual traits – the 155 significant traits identified in Section 3.3.5 minus 10 traits either suspected to be associated with confounders (*alcohol_drinker*, *lamb*, *pork*, *religious_group*, *edu* and *townsend_deprivation*) or where the observed associations are not clearly caused by germline ROH (*icd10_d45_polycythaemia_vera*, *erythrocyte_distribution_width*, *immature_reticulocyte_fraction*, and *high_light_scatter_reticulocyte_percentage* – see section 5.3.1.3.1).

In deCODE-Iceland 14 individually significant traits were used to calculate the *multivariate* trait: height, whr, fev1, edu, hb, log_lymphoc, hr, neb, afb_men, ever_smoked, ever_had_children, alcohol_units, weight and tot_chol.

3.2.8 Sex-specific estimates of $\beta_{F_{ROH}}$

For all traits, in both ROHgen and UKB, $\beta_{F_{ROH}}$ was estimated separately in both men and women. In ROHgen this done on the trait residuals from equation (17) and in UKB the full phenotype model with F_{ROH} and 74 covariates was fitted separately in men and women.

In hindsight it is now appreciated that the full model approach is preferable in this situation. When covariates are fitted with both sexes combined (as in equation (17)) their estimated effects are necessarily the same for both sexes. If, however, a covariate (for example, ethnicity) actually has a bigger effect in one sex, then subsequent estimates of $\beta_{F_{ROH}}$ in that sex will be confounded by the sex-specific effect of ethnicity. A similar pitfall of the two-step approach was recently highlighted in an article advising against the common practice of Inverse-Rank normalisation of model residuals¹²¹ – a practice which is not uncommon in statistical genetics⁷³. In general, any manipulation of model residuals, whether it be sub-setting or normalisation, is liable to reintroduce some of the covariate effects.

The ROHgen sex-specific effects are reported never-the-less since this imperfect accounting for covariates is not thought to be the cause of the observed sex-specific effects. Furthermore, the trait with the most substantial sex-specific effect (*age_at_first_birth*) happens to be one of two analysed entirely separately by sex.

3.3 Results

3.3.1 ROHgen traits associated with F_{ROH}

Of the 49 complex traits analysed in ROHgen, 18 are significantly associated with F_{ROH} after Bonferroni correction (Figure 30). Effect sizes are relatively small, most often between one and two trait standard deviations per complete genome in ROH.



Figure 30: **Effect of F_{ROH} on 18 complex traits in ROHgen.** To facilitate comparison between traits, effect estimates are presented in units of intra-sex standard deviations. Sample sizes, intra-sex standard deviations, and effect estimates in measurement units are shown in Supplementary Table 9. Traits are grouped into five broad categories: body size, cognition, reproductive success, addictive behaviours, and overall health.

Height is the most significantly associated ROHgen trait (p -value 4×10^{-135}); however, the largest effect (4.0 trait sd $\pm$ 0.90) is on the less commonly measured trait of age of first birth in men (74,182 samples in 25 cohorts). Surprisingly, the corresponding trait in women has a non-significant effect (0.36 trait sd $\pm$ 0.4), despite being more commonly measured (306,867 samples in 50 cohorts).

The significantly associated ROHgen traits can be grouped in to five broad categories: body size, cognition, reproductive success, addictive behaviours, and overall health. For the majority of traits, the direction of effect is consistent with reduced evolutionary fitness – i.e., increased F_{ROH} causes trait changes which are likely to reduce fitness. The apparent exceptions to this pattern include both addictive behaviour traits: i.e., reduced smoking and alcohol consumption seem more likely to increase the modern notion of fitness. Equally, reduced LDL and total cholesterol might be expected to increase evolutionary fitness due to their cardio-protective effects. The explanation for these apparent discrepancies is not clear, although the current selection pressures on these traits are probably radically different from those experienced for most of human evolution. All traits are of course correlated with many others, and a deleterious effect on one trait may be offset by correlated advantageous changes to others. For example, in UKB there are positive correlations between reproductive success and both smoking and alcohol consumption which may help explain the observed effect directions. Equally, in UKB there are positive correlations between self-reported overall health and all measures of cholesterol (HDL, LDL and total).

3.3.2 Survey of UK Biobank traits associated with F_{ROH}

$\hat{\beta}_{F_{ROH}}$ are reported for 1786 UKB traits in Supplementary Table 23. Of these 1786 traits, 255 are significantly associated with F_{ROH} at a false discovery rate of 5% and a QQ plot shows that many more would be significantly associated in a larger sample (Figure 31).

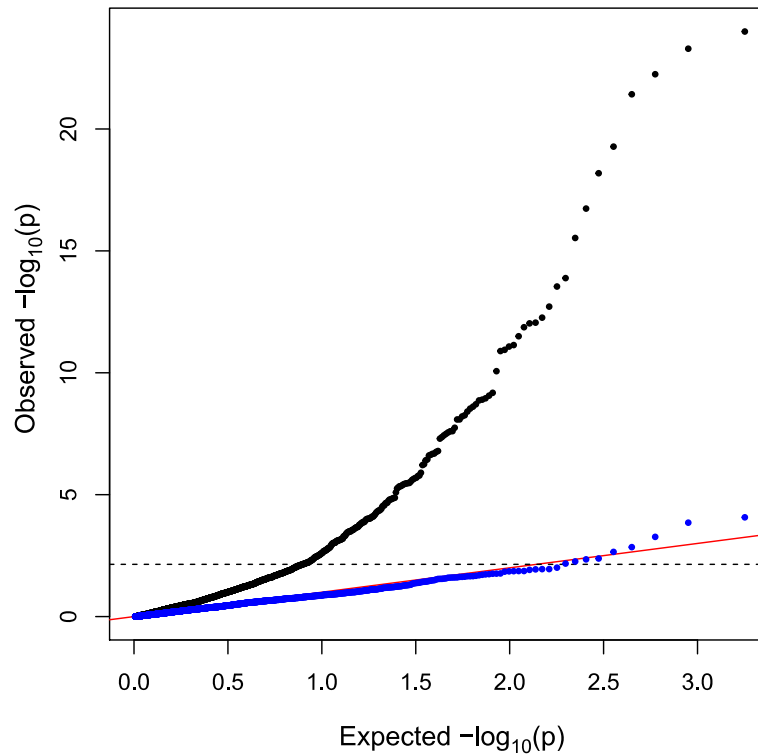


Figure 31: **QQ plot of p -values for the association between F_{ROH} and 1786 UKB traits.** Observed p -values (shown in black) reveal inflation above the expectation (shown as a solid red line). This inflation starts well below the 5% FDR significance threshold (shown as a horizontal dashed black line). When phenotypes are randomly shuffled between samples, p -values (shown in blue) follow the null expectation demonstrating that the observed inflation is not due to statistical bias caused, for example, by trait distributions which are not Gaussian.

The 255 traits which reach 5% FDR significance were each manually assessed and consolidated into 155 traits (see methods section 3.2.1.2.1), which are discussed further in section 3.3.5.

3.3.3 ICD10 codes associated with F_{ROH} in UK Biobank

$\beta_{F_{ROH}}$ was estimated for 13,646 binary traits indicating the presence of corresponding ICD10 codes in a subject's electronic health record (Supplementary Table 24). Of these 13,646 ICD10 codes, 73 are significantly associated with F_{ROH} at a false-discovery-rate of 5% - although due to the hierarchical structure of ICD10 codes these tests are not all independent.

Increased F_{ROH} increases ICD10 code prevalence for 64 of the 73 significant traits and reduces prevalence for only 9. In accord with the complex trait results (Section

3.3.5), these 9 reduced codes include several related to childbirth (e.g., Z30-Z39 Persons encountering health services in circumstances related to reproduction and O70 Perineal laceration during delivery) and others possibly linked to risk-taking behaviour (Z89.7 Acquired absence of both lower limbs and W26 Contact with knife, sword or dagger).

As for the complex traits results (Figure 31), a QQ plot for all commonly used ICD10 codes shows substantial inflation at all p -values (Figure 32). These p -values are calculated by resampling so this inflation cannot be caused by statistical bias. Indeed, when the binary traits indicating ICD10 codes are randomly assigned to samples, the inflation disappears. In conclusion, F_{ROH} appears to affect the prevalence of many, or all, commonly used ICD10 codes to some degree.

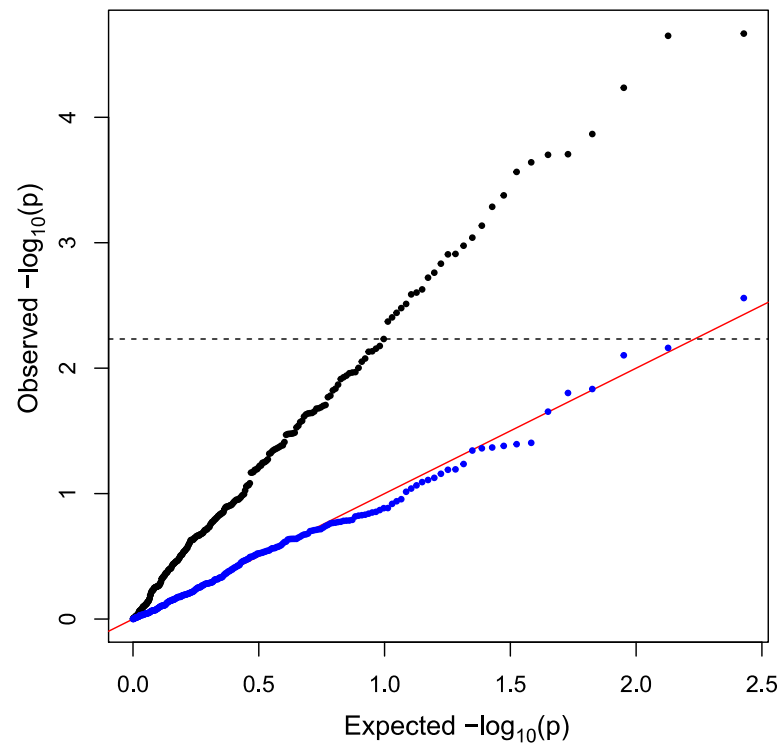


Figure 32: **QQ plot of p -values for the association between F_{ROH} and 268 commonly used ICD10 codes.** Results are shown for all icd10 codes recorded in more than 10,000 subjects. Observed p -values (shown in black) reveal inflation above the expectation (shown as a solid red line). This inflation starts well below the 5% FDR significance threshold (shown as a horizontal dashed black line). When ICD10 codes are randomly shuffled between samples, p -values (shown in blue) follow the null expectation demonstrating that the observed inflation is not due to statistical bias caused, for example, by trait distributions which are not Gaussian.

3.3.4 Self-reported disease codes associated with F_{ROH} in UK Biobank

$\beta_{F_{ROH}}$ was estimated for 446 binary traits indicating the presence of a corresponding self-reported illness in a subject's touchscreen questionnaire (Supplementary Table 25). Only four self-reported disease codes reach 5% FDR significance: polycythaemia vera (p -value 6×10^{-06}), cataract (p -value 0.0001), glaucoma (p -value 0.0002) and male infertility (p -value 0.0001). The first three of these self-reported illnesses also have corresponding hospital ICD10 codes (D45, H25 and H4, respectively) which are significantly associated with F_{ROH} . The ICD10 code for male infertility (N46) is not significantly associated with F_{ROH} (p -value 0.24) although this maybe be partly explained by the age distribution of UKB subjects – ICD10 codes record incident disease in a cohort between 37 and 74 years old at recruitment, while self-reported illness may capture earlier diagnoses.

3.3.5 Combined UKB results

Traits significantly associated with F_{ROH} were found in three separate UKB analyses: regular complex traits (section 3.3.2), hospital in-patient ICD10 codes (section 3.3.3) and self-reported illnesses (section 3.3.4). These three sources were consolidated into 155 significantly associated traits (Supplementary Table 26) which are summarised in Figures 33 and 34 below.

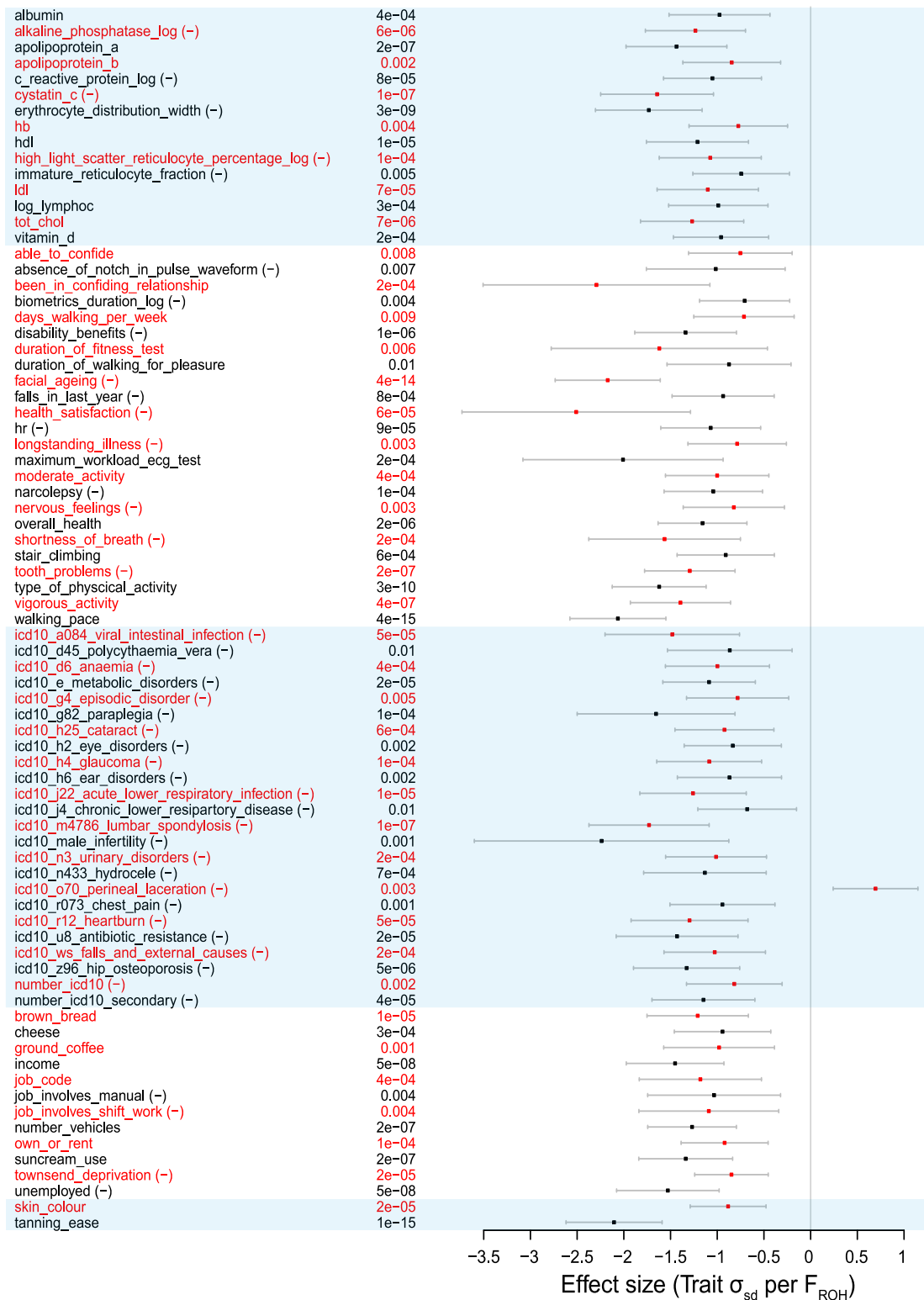


Figure 33: Effect of F_{ROH} on 77 significant traits in UKB. Traits are grouped into five categories (biochemistry, wellbeing, diseases, SES and skin colour). Traits marked (-) have been reversed so that increased trait values correlate with increased self-reported overall health. All disease traits were reversed, regardless of correlation with self-reported overall health, since all ICD10 codes are presumed to be undesired.



Figure 34: **Effect of F_{ROH} on 78 significant traits in UKB.** Traits are grouped into six categories (body size and strength, reproductive success, cognition, eyesight, hearing, and conservatism). Traits marked (-) have been reversed so that increased trait values correspond to increases in the first principal component of all traits in the same category. Where necessary this first principal component was reversed to be consistent with the direction implied by the category name.

As in ROHgen, effect sizes for the 155 significant UKB traits are small - most often below two trait standard deviations per complete genome in ROH. The lowest p -values are associated with the following traits: alcohol drinker status (1×10^{-36}), peak expiratory flow (3×10^{-22}), grip strength (3×10^{-21}), and number of sexual partners (6×10^{-21}). As in ROHgen, the traits can be grouped into broader categories, although the wider range of phenotypes implies a greater number of categories. The division of these categories is somewhat subjective, and a few traits could equally well be assigned to more than one, but broadly the UKB traits associated with F_{ROH} fit into 11 categories: biochemistry, wellbeing, disease, SES, skin colour, body size/strength, reproduction, cognition, eyesight, hearing, and conservatism.

Some traits are negatively correlated with the direction implied by their broader category, and for these traits the sign of $\hat{\beta}_{F_{ROH}}$ has been reversed in Figures 33 and 34 (for example, reaction time is a measure of cognitive function¹²², but is negatively correlated with fluid intelligence). For the biochemistry category there is no obvious direction implied, so biochemistry traits were reversed if they had a negative correlation with self-reported overall health. In summary, almost without exception, increased F_{ROH} causes trait changes associated with:

- 1) Reduced well-being.
- 2) Biochemical changes associated with reduced overall health.
- 3) Increased disease prevalence.
- 4) Reduced socio-economic status.
- 5) Lighter skin colour.
- 6) Smaller body size and strength.
- 7) Reduced reproductive success.
- 8) Reduced cognitive ability.
- 9) Poorer eyesight.
- 10) Poorer hearing.
- 11) Reduced risk-taking and increased conservatism.

Only three traits are inconsistent with this summary: Although increased F_{ROH} increases all other ICD10 codes it reduces the likelihood of having a perineal laceration - this could be explained by the negative effect of F_{ROH} on reproductive success. Equally, F_{ROH} reduces almost all measures of body size, but increases waist circumference and waist-to-hip ratio. Again, this could be explained if these two traits were assigned to well-being category, rather than body size, since both traits are negatively correlated with overall health.

3.3.5.1 *Correlations among the 155 significant UKB traits*

The UKB traits significantly associated with F_{ROH} are used in many subsequent analyses, and although the most strongly correlated traits were removed or combined by consolidation, the final set of 155 is not completely independent. This non-independence has the potential to downwardly bias p -values in subsequent analyses. Indeed, applying the method described by Li et al.¹²³, these 155 correlated traits are equivalent to only 78 independent traits. Therefore, all t -statistics, standard errors, and p -values referring to the ensemble of these traits (e.g., Figure 66, Figure 67, etc.) were obtained by permutation.

3.3.5.2 *UKB multivariate trait*

F_{ROH} has small effects on many traits, but these can be combined into a single, well-powered compound trait. The calculation of this “*multivariate*” trait is described in methods section 3.2.7, and takes account of the correlations described in section 3.3.5.1.

The UKB *multivariate* trait is very strongly associated with F_{ROH} (p -value = 2×10^{-148}) and has a substantially larger effect size (-6.6 trait sd [95% CI -6.1 to -7.1]) than any of the 145 traits from which it is composed. In fact, $\beta_{F_{ROH}}$ is sufficiently large that a trend is visible on a simple scatter plot of *multivariate* versus F_{ROH} (Figure 35).

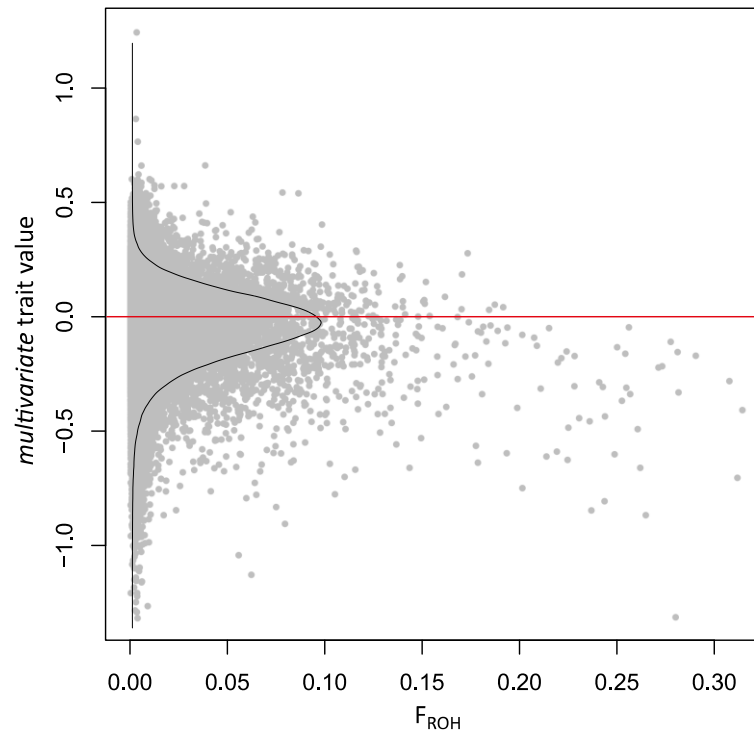


Figure 35: **Scatter plot of *multivariate* versus F_{ROH} for all UKB subjects.** The average *multivariate* value (= 0) is marked as a red line. A density plot of *multivariate* is shown, in black, on the y-axis.

Confounding of F_{ROH} by environmental factors is discussed extensively in Chapter 4, but note here that of the 42 samples with $F_{ROH} > 0.2$ (thought to be the offspring of first-degree relatives) not one has an above average value of *multivariate*, and 16 out of 42 (38%) are in the very lowest percentile. This is consistent with a genetic effect since, even in the unlikely scenario that *multivariate* is a precise measure of some unmeasured confounder, it is unlikely that the offspring of first-degree relatives would experience such extreme environments quite so consistently.

Socioeconomic status (SES) is an obvious potential confounder, but there is little evidence that these very high F_{ROH} subjects experienced substantial socioeconomic deprivation. Their average Index of Multiple Deprivation (see section 4.2.1.1) at birth address (29.7 ± 2.90) is in keeping with that of the whole cohort (30.9 ± 0.03). 14 out of 42 report having been adopted as a child (in the UK, adoptive parents tend to be above-average SES), and of the 28 who were not adopted, 12 report having 3 or more full siblings – perhaps indicating they are unaware of their parentage.

3.3.5.3 Effect of removing hemizygous deletions from F_{ROH}

Hemizygous deletions were removed from F_{ROH} for all UKB results presented in this thesis (as described in Section 2.3.1.1). This precaution was not taken in ROHgen, since the results in McQuillan et al.⁸² (and here) show that this is of little importance. In UKB, removing the identified hemizygous deletions has a negligible effect on F_{ROH} and, despite the relatively large phenotypic effects of the deletions themselves (see Section 4.2.5), also has a negligible effect on $\hat{\beta}_{F_{ROH}}$ for all 155 significant traits (Figure 36). While this is true for the autosomes, a very different conclusion is reached for the X-Chromosome (section 5.3.3.2.1).

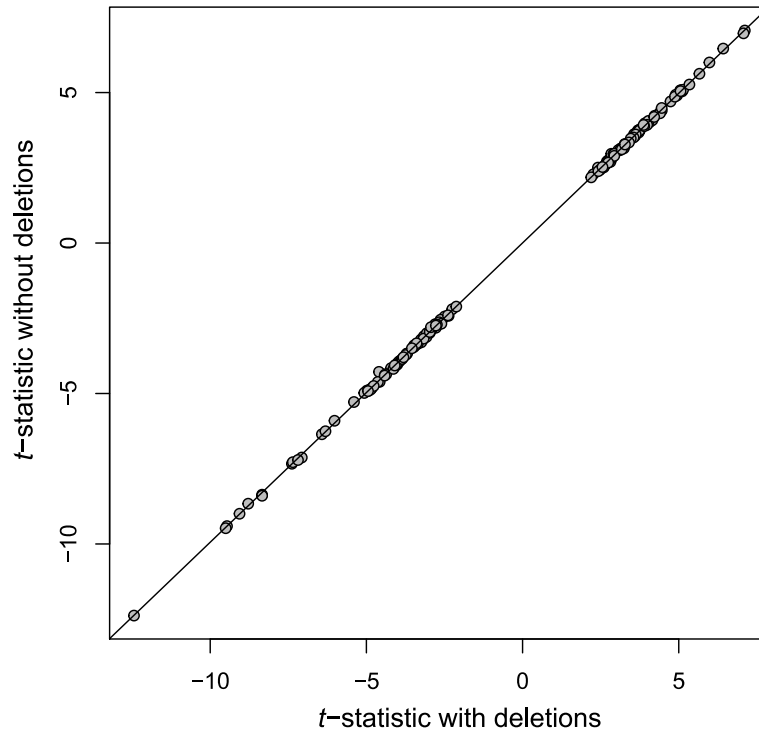


Figure 36: **Comparison of the t -statistics of $\hat{\beta}_{F_{ROH}}$ for 155 significant UKB traits, with and without hemizygous deletions removed.** Although preferable, removing hemizygous deletions from F_{ROH} has a negligible effect on $\hat{\beta}_{F_{ROH}}$.

3.3.6 Sex-specific $\beta_{F_{ROH}}$

For the 43 ROHgen traits measured in both sexes, sex-specific $\beta_{F_{ROH}}$ were estimated as described in Section 3.2.8. Although there is no systematic difference between $\hat{\beta}_{F_{ROH}}$ in men and women, the sex-specific estimates do differ significantly for 13 out of 43 traits (Supplementary Table 11 and Figure 37).

The effect of F_{ROH} on body weight (which was measured in 1,245,371 samples in 203 cohorts) is the most significantly sex-specific effect (p -value = 1×10^{-26}). In men, increased F_{ROH} is associated with reduced weight ($\beta_{F_{ROH}} = -1.84$ trait sd $\pm$ 0.23) while in women there is no effect ($\beta_{F_{ROH}} = 0.01$ trait sd $\pm$ 0.23). As mentioned previously, the effect of F_{ROH} on age at first birth (afb) is substantial in men ($\beta_{F_{ROH}} = 4.0$ trait sd $\pm$ 0.90), but non-significant in women ($\beta_{F_{ROH}} = 0.36$ trait sd $\pm$ 0.40). Unfortunately, afb_men was not recorded in UKB which limits the sample size to 74,182 subjects in 25 cohorts.

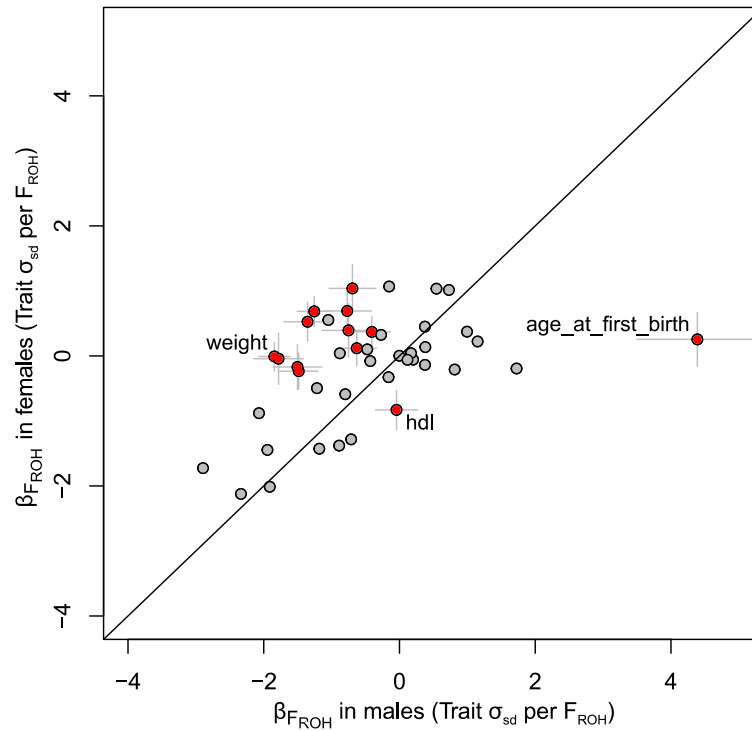


Figure 37: **Sex-specific estimates of $\beta_{F_{ROH}}$ for 43 ROHgen traits.** All estimates are plotted in units of intra-sex standard deviation. Error bars representing 95% confidence intervals are shown for 13 traits (plotted in red) where estimates differ significantly between men and women.

In UKB, sex-specific $\beta_{F_{ROH}}$ were estimated for all 147 significant traits measured in both sexes (Supplementary Table 37 and Figure 38). With a smaller sample size and greater multiple correction (compared to ROHgen), only one UKB trait (*alcohol_drinker*) differs significantly between men and women – and this effect is believed to be largely caused by confounding. As in ROHgen, there is no systematic difference between the effects in men and women – meta-analysing across all traits

(excluding *alcohol_drinker*) the average effect in women is 94% [95% CI = 86% to 102%] of the effect in men.

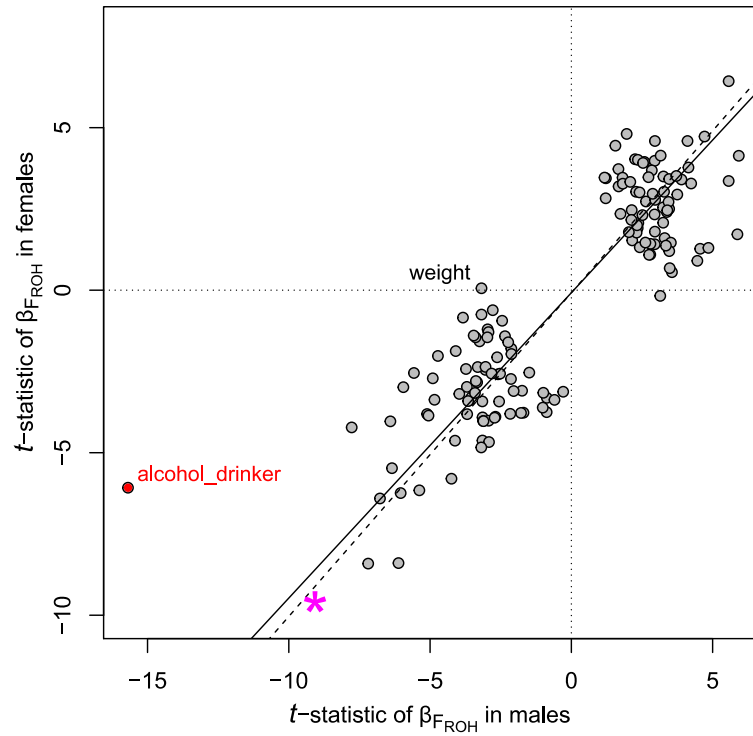


Figure 38: **Sex-specific t -statistics of $\hat{\beta}_{F_{ROH}}$ for 147 ROHgen traits.** One trait (*alcohol_drinker*) which differs significantly between men and women is shown in red. The expected gradient assuming equal effect sizes is shown in dashed black, and a Deming regression best-fit line is shown in solid black. The best-fit gradient is 94% [95% CI = 86% to 102%] of the expectation for equal effect sizes. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

3.3.7 Comparison of different estimators of F

Three estimators of F (F_{ROH} , F_{SNP} , and F_{GRM}) were calculated for all samples (see section 2.3.2). These different estimators were compared by fitting bivariate models, e.g.,

$$trait \sim F_{SNP} + F_{ROH} + covariates \quad (26)$$

for all ROHgen traits and for the 155 significant UKB traits (Supplementary Tables 13, 14, 38, and 39).

Since these estimators are highly correlated, the difference between $\beta_{F_{ROH}}$ and $\beta_{F_{SNP}}$ (or $\beta_{F_{GRM}}$) is often non-significant. However, for well-powered traits (such as height in ROHgen and *multivariate* in UKB), F_{ROH} is a better predictor of inbreeding depression than both F_{SNP} (Figure 39) and F_{GRM} (Figure 40). Since ROH make all variants homozygous (both common and rare), while F_{SNP} and F_{GRM} are calculated only from common SNPs, this implies that the observed effects are mainly caused by rare variants. This intuitive explanation is confirmed in Appendix D, by demonstrating that the homozygosity of common SNPs is better predicted by F_{GRM} , while the homozygosity of rare SNPs is better predicted by F_{ROH} .

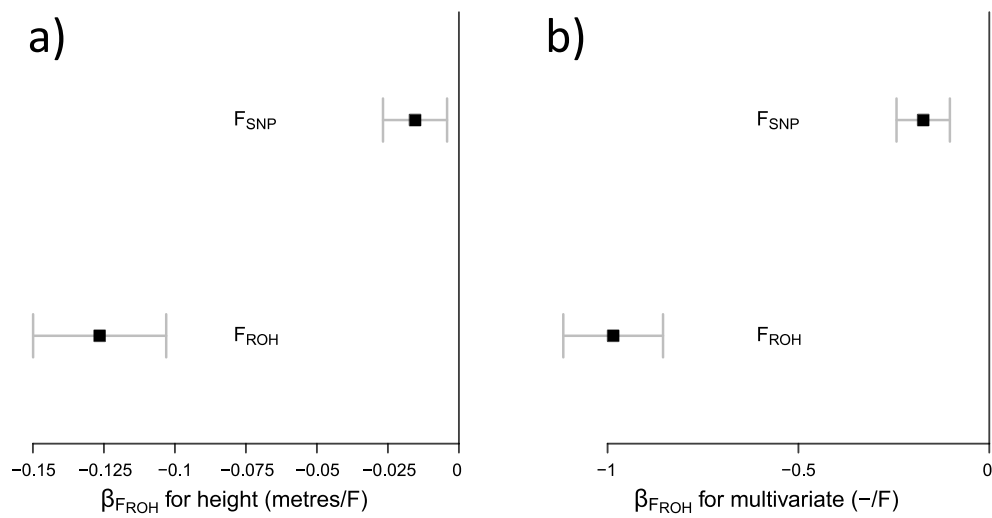


Figure 39: **Comparison of $\hat{\beta}_{F_{SNP}}$ and $\hat{\beta}_{F_{ROH}}$ in bivariate models.** (a) Bivariate model of height in ROHgen. (b) Bivariate model of *multivariate* in UKB. For both traits, F_{SNP} has a small, significant effect, but the conditional effect of F_{ROH} is consistent with the univariate estimate, and significantly greater than $\hat{\beta}_{F_{SNP}}$.

Figures 39 and 40 suggest that F_{GRM} is a slightly better predictor of inbreeding depression than F_{SNP} (although both are less predictive than F_{ROH}). This is confirmed by the bivariate model *multivariate* $\sim F_{GRM} + F_{SNP}$ in UKB British (Figure 41). This suggests that although F_{ROH} is the preferred metric of inbreeding, F_{GRM} should be used in preference to F_{SNP} when ROH cannot be called (for example, in species where no genetic map available).

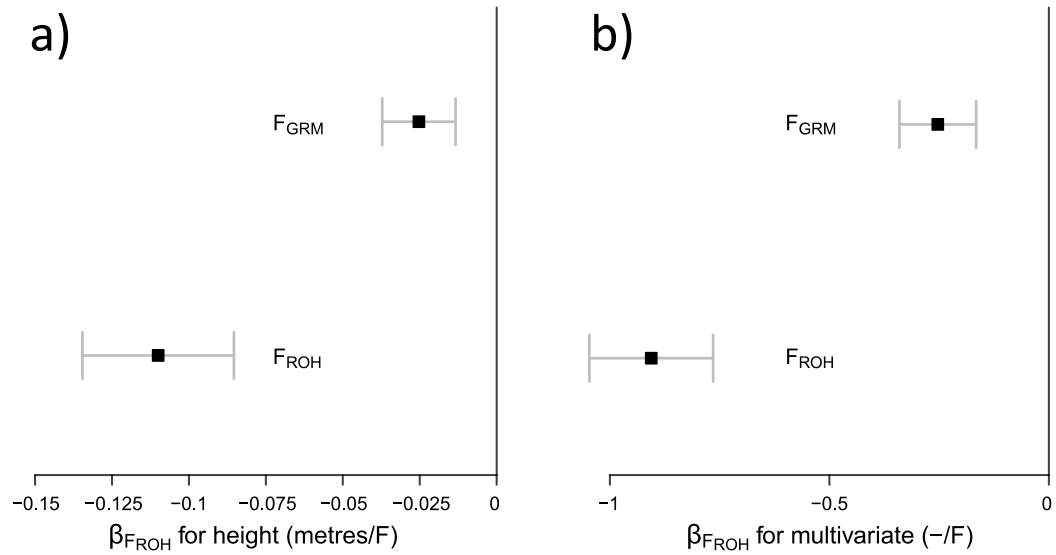


Figure 40: **Comparison of $\hat{\beta}_{F_{GRM}}$ and $\hat{\beta}_{F_{ROH}}$ in bivariate models.** (a) Bivariate model of height in ROHgen. (b) Bivariate model of *multivariate* in UKB British. For both traits, F_{GRM} has a small, significant effect, but the conditional effect of F_{ROH} is significantly greater than $\hat{\beta}_{F_{GRM}}$.

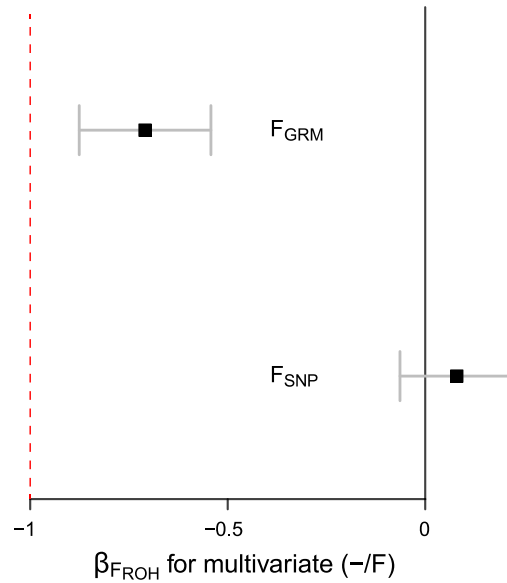


Figure 41: **Comparison of $\hat{\beta}_{F_{GRM}}$ and $\hat{\beta}_{F_{SNP}}$ in a bivariate model of *multivariate* in UKB.** The conditional effect of F_{GRM} is highly significant, while the conditional effect of F_{SNP} is consistent with zero. The univariate effect of F_{ROH} ($\beta_{F_{ROH}} = -1$) is shown as a dashed red line.

The conditional effects of F_{ROH} and F_{GRM} on all 155 significant UKB traits are compared in Figure 42. Averaging across all traits, the conditional effect of F_{GRM} is 25% [95% CI = 16 to 34, p -value = 2×10^{-8}] of the conditional effect of F_{ROH} . As

expected, this fraction is consistent with the effect on F_{GRM} on the *multivariate* trait (Figure 40b). For 16 traits, $\hat{\beta}_{F_{ROH}}$ remains significant despite conditioning on F_{GRM} , while one trait (whr), $\hat{\beta}_{F_{GRM}}$ is significant when conditioned on F_{ROH} .

Overall, these results suggest that, for most traits, F_{ROH} is a better predictor of inbreeding depression than F_{GRM} or F_{SNP} . Never-the-less, despite conditioning on F_{ROH} , F_{GRM} remains informative for at least some traits, suggesting that some excess homozygosity outside of the called ROH contributes to inbreeding depression. It is not known whether this is caused by the chance excess homozygosity of common SNPs, or shorter autozygous segments that have not been identified by the ROH calling algorithm used (for example, because they may be below 1.5 Mb).

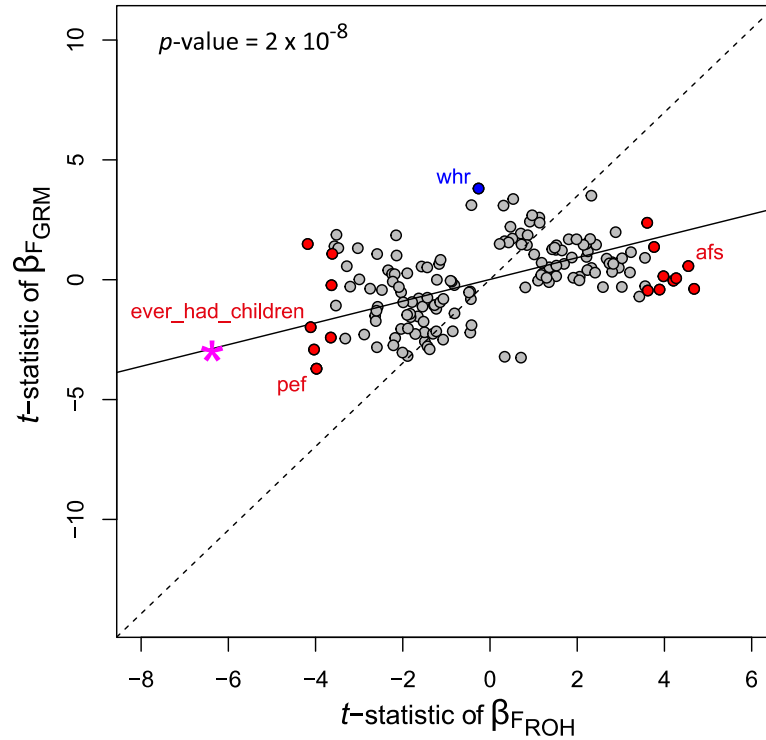


Figure 42: **Conditional effects of F_{GRM} and F_{ROH} for 155 significant UK Biobank traits.** A Deming regression best-fit line is shown in solid black and has a gradient which differs significantly from zero (p -value 2×10^{-8}). The expected regression gradient for equal effect sizes is shown in dashed black. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area. Traits where $\hat{\beta}_{F_{ROH}}$ is significant are shown in red. One trait (whr) where $\hat{\beta}_{F_{GRM}}$ is significant is shown in blue.

3.3.8 Effect of different ROH lengths

To evaluate the relative effects of ROH of different physical (Mb) lengths, F_{ROH} was split into three sub-measures ($F_{ROH<3}$, $F_{3<ROH<5}$, $F_{ROH>5}$), representing the genomic fractions in ROH less than 3 Mb, between 3 and 5 Mb, and greater than 5 Mb, respectively. Multivariate models with these three fractions ($trait \sim F_{ROH<3} + F_{3<ROH<5} + F_{ROH>5}$) were then fitted for all ROHgen traits, and for the 155 significant UKB traits (Supplementary Tables 15 and 36)

The variance of the shorter ROH measures ($F_{ROH<3}$ and $F_{3<ROH<5}$) are very much lower than the variance of $F_{ROH>5}$ and, consequently, the standard errors of $\hat{\beta}_{F_{ROH<3}}$ and $\hat{\beta}_{F_{3<ROH<5}}$ are substantially greater than the standard errors of $\hat{\beta}_{F_{ROH>5}}$. In ROHgen, the effect of ROH less than 3 Mb is only significant (p -value < 0.001) for two traits (height and interleukin 6) and the effect of ROH between 3 and 5 Mb is also only significant for two traits (height and Educational Attainment). These large standard errors also mean there is very little inconsistency between the estimated effects of short and long ROH. Indeed, only one effect, the effect of 3 to 5 Mb ROH on Educational Attainment ($\hat{\beta}_{F_{3<ROH<5}} = -21.4$ years; 95% CI = -29.5 to 13.4), is significantly different (p -value = 0.0003) from its longer-than-5-Mb counterpart ($\hat{\beta}_{F_{ROH>5}} = -6.3$ years; 95% CI = -7.8 to -4.7). This discrepancy for educational attainment may reflect confounding of 3 to 5 Mb ROH by socioeconomic factors in some cohorts.

In UKB, with a smaller sample size and more multiple testing, none of the effects of shorter ROH differ significantly from the effect of >5 Mb ROH. However, despite this lack of power for individual UKB traits, the ‘average’ effect of short ROH can be examined via the *multivariate* trait (Figure 43b).

For the *multivariate* trait, ROH of all lengths have a significant effect, and there is no significant difference between the three length groupings. Never-the-less, there is a suggestion (p -value = 0.14) that the shortest ROH (1.5 to 3 Mb) have a slightly reduced effect ($\hat{\beta}_{F_{ROH}<3} = -0.79$; 95% CI = -1.08 to -0.49) compared ROH longer than 5 Mb ($(\hat{\beta}_{F_{ROH}>5} = -1.02$; 95% CI = -1.11 to -0.93)).

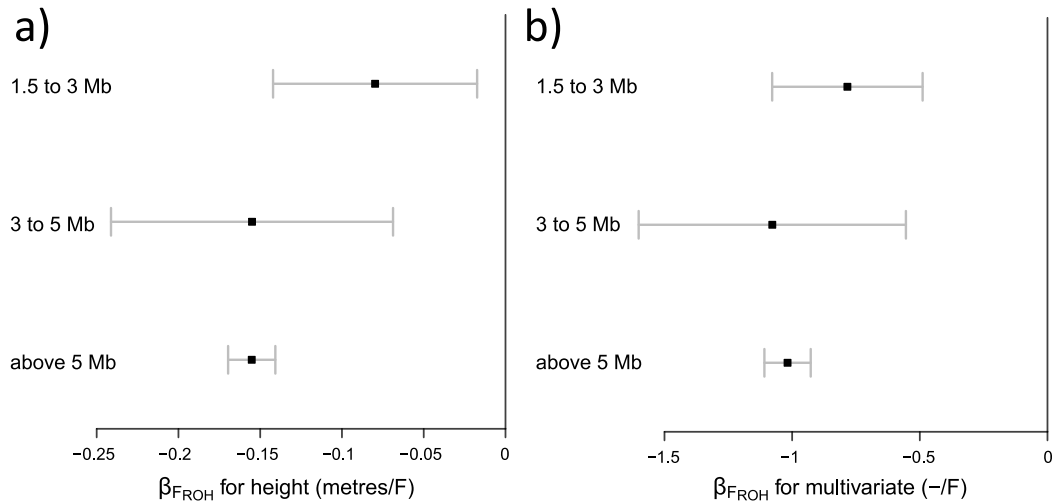


Figure 43: **Estimated effect of ROH of different physical lengths.** (a) Effects on height in ROHgen. (b) Effects on *multivariate* in UKB British. Error bars represent 95% confidence intervals.

Similarly, for the best powered ROHgen trait, height (Figure 43a), all three length groupings differ significantly from zero and the effect of the shortest ROH ($\hat{\beta}_{F_{ROH}<3} = -0.08$ metres; 95% CI = -0.142 to -0.017) is smaller (p -value = 0.02) than the effect of longer ROH ($\hat{\beta}_{F_{ROH}>5} = -0.155$ metres; 95% CI = -0.170 to -0.141).

It was shown previously than many ROH shorter than 3 Mb occur in ROH islands, and that ROH islands are more likely to appear in low recombination regions of the genome (section 2.5.5). ROH which occur in these islands may be subject to increased purging, and are perhaps less likely to represent truly autozygous segments. Therefore, to help quantify the effect of ROH islands, all UKB ROH were grouped by genetic map length (centimorgans, cM) and the genomic fractions in four genetic length groupings (below 1 cM, 1 to 2 cM, 2 to 5 cM, and above 5 cM) were calculated for all UKB samples. In UKB, 99.6% of ROH less than 1 cM occur in the 32.0% of the genome where ROH islands occur (defined as > 0.2% of samples

autozygous). The effects of ROH in these four genetic length groupings were then determined for the *multivariate* trait by the model:

$$\begin{aligned} \text{multivariate} \sim & F_{ROH < 1cM} + F_{1cM < ROH < 2cM} + F_{2cM < ROH < 5cM} \\ & + F_{ROH > 5cM} \end{aligned} \quad (27)$$

Effect estimates from equation (27) are shown in Figure 44 below.

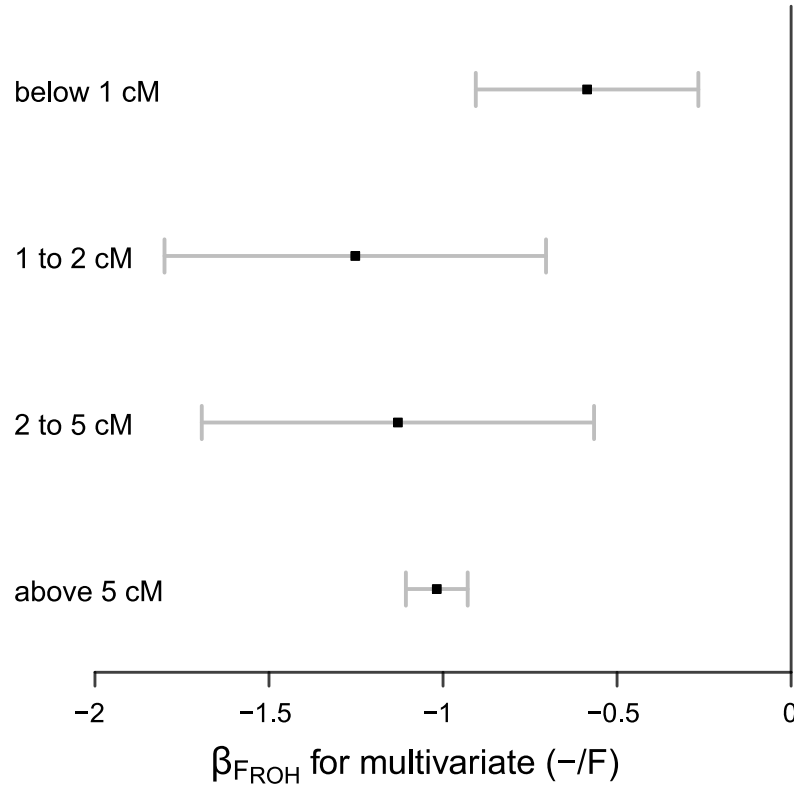


Figure 44: **Estimated effect of ROH of different genetic lengths on the *multivariate* trait in UKB British.** Error bars represent 95% confidence intervals.

The effect of ROH less than 1 cM (-0.59 95% CI = -0.91 to -0.27) is significantly less than (p -value = 0.01) the effect of ROH longer than 5 cM (-1.02 95% CI = -1.10 to -0.93), but is also significantly greater than zero (p -value = 0.0003). I.e., even the shortest called ROH, occurring almost exclusively in ROH islands, have a significant effect on the *multivariate* trait. Equally, other short ROH between 2 and 5 cM genetic length, have significant effects which are consistent with the effect of ROH longer than 5cM. 54.0% of ROH between 1 and 2 cM occur in the previously defined ROH islands, compared to 18.0% of ROH between 2 and 5 cM.

3.4 Discussion

3.4.1 Survey of complex traits affected by inbreeding

As described in the introduction to this chapter (section 3.1), many human complex traits have previously been associated with F_{ROH} , or other estimates of inbreeding. By performing the largest meta-analysis of F_{ROH} to date, and by examining thousands of traits in UK Biobank, this chapter reproduces many of the previously reported associations, but also illuminates the broad scope of human complex traits affected by inbreeding depression.

Many of the traits described in Figures 33 and 34 can be categorised with previously reported associations, but the total number of traits affected is much wider than any previously reported. In addition, correlating traits with self-reported overall health, highlights that the observed effects of inbreeding are almost exclusively deleterious to wellbeing. Interestingly, this extends to the biochemical changes caused by increased F_{ROH} , many of which have not been previously reported, and which are uniformly associated with reduced overall health. Of particular note, total cholesterol is reduced by inbreeding, despite being a well-known cardiovascular disease risk factor. Intriguingly, in UK Biobank increased total cholesterol is correlated with increased overall health (and a reduced number of hospital ICD10 codes), which, together with the inbreeding effect, suggests that increased total cholesterol is, in some sense, beneficial to evolutionary fitness.

Although many of the reported effects of F_{ROH} correlate with reduced overall health, some (in the domains of body size, reproductive success and risk/conservatism) do not. For this reason, traits in these groups were arranged to correlate with their grouping, rather than overall health. Inbreeding is known to reduce body size and reproductive success in other mammals¹¹, so such effects in human might be considered unsurprising.

However, the traits on which reduced reproductive success manifests are potentially informative. Notably, the effect of F_{ROH} on whether a subject has children is substantially stronger than the effect on the number of children they

have. Indeed, in both ROHgen and UK Biobank, F_{ROH} has a non-significant effect on number of children amongst those who have at least one. The reasons for this are unknown, although it seems to imply that mate-choice and/or fertility are more affected than fecundity. Interestingly, a recent study reported (qualitatively) identical effects associated with an increased burden of heterozygous protein-truncating variants (PTVs), and deduced that the reduction in the likelihood of having children is attributable to cognitive or behaviour traits¹²⁴. As well as suggesting which traits which could mediate the effect of F_{ROH} on the likelihood of having children, the almost indisputably genetic effect of PTVs supports a genetic hypothesis for the effect of F_{ROH} (as opposed to confounding).

Interestingly, amongst the largest $\hat{\beta}_{F_{ROH}}$, in both ROHgen and UK Biobank, are several behavioural traits strongly correlated with reproductive success: age at first birth (in men only), age at first sex, and number of sexual partners. These traits might equally have been assigned to domain of risk/conservatism, and indeed two of the three (age at first sex and number of sexual partners) were recently included in a large GWAS of risk tolerance¹²⁵. Linnér et al. show that risk-taking behaviour has a genetic basis, and report genetic correlations between general risk tolerance and several of the conservatism/risk traits associated with F_{ROH} (including: cannabis use, alcohol consumption, smoking, motorway speeding, and contraceptive use). While there is clearly increased potential for social confounding within this grouping, the associations between risk-taking and reproductive success support a hypothesis that certain risky behaviours are related to evolutionary fitness, and consequently affected by inbreeding depression. Although the unique context of human society cannot be ignored, changes to reproductive behaviour in other inbred animals are not uncommon in the scientific literature^{43,126–128}.

Although, a large number of complex traits are now significantly associated with F_{ROH} , it is intriguing to note that QQ plots for both quantitative traits (Figure 31) and common ICD10 codes (Figure 32) show substantial inflation of p -values for almost all traits tested. Statistical bias was excluded as an explanation, so it seems likely that almost all traits are affected by F_{ROH} to some degree. This might be explained

by an “omnigenic” model of genetic variation which proposes that “*gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways*”¹²⁹. If all genes have effects on all traits, then it is perhaps not implausible that the homozygosity of genes also has an effect on all traits. One further implication of the observed inflation is that when larger, more powerful cohorts become available, substantially more traits are likely to become significantly associated with F_{ROH} .

3.4.1.1 Combining individual traits

As already described, UK Biobank, and other large cohorts, are sufficiently well-powered to identify many traits substantially affected by inbreeding (i.e., effects greater than approximately 1 trait SD / genome). However, interest is now turning towards dissecting the effects of F_{ROH} , for example, the effects of different ROH lengths, or ROH in specific genomic locations, which will require even larger datasets. Until these are available, an alternative approach may be to leverage the independent effects of ROH on multiple traits. For example, Yengo et al. recently combined test statistics from 9 individual traits to help identify genomic patterns of ROH distribution⁷³. In this chapter, an alternative method is proposed which summarises many correlated traits into a well-powered “*multivariate*” trait. This compound trait can easily be used in subsequent analyses and, in UKB, increases expected test-statistics by a factor of approximately 3.5 over the best powered individual trait (height). However, what is gained in statistical power is lost slightly in interpretability, since of course *multivariate* is not a conventional, observable trait.

3.4.2 Comparing estimates of F

An important early consideration in any study of inbreeding is how to estimate inbreeding coefficients from genomic data. Some authors advocate a marker-by-marker approach⁹⁸, while others argue¹¹⁴ that F_{ROH} is a lower variance estimate of F . Since F is a theoretical construct, defined in terms of an arbitrary base generation, it

seems pragmatic to suggest that the best estimate of F is the one that best predicts the effects of inbreeding. Hence, the best available estimate might depend on many factors, including: the genetic architecture of the focal trait, the identifiability of autozygous segments, knowledge of causal variants, etc.

For example, Appendix E.3 shows that F_{GRM} calculated at causal variant locations, and weighted by their effect sizes, is a perfect predictor of inbreeding depression. Unfortunately, the genetic variants causing inbreeding depression are far from known, and this is very much a theoretical suggestion. However, if many of the causal variants are genotyped (or in strong LD with markers that are) then F_{GRM} can be a better predictor of inbreeding depression than F_{ROH} (Figure 95). Conversely, if causal variants are rare (or otherwise not in strong LD with genotyped markers) then F_{ROH} should be a better predictor of ID than F_{GRM} .

In the real data conditions of this thesis, F_{ROH} is a better predictor of inbreeding effects than F_{GRM} (calculated from all SNPs with MAF > 5%) for most traits (Figure 42), implying that the variants causing ID are predominantly rare (< 5%). Never-the-less, averaging across all traits, F_{GRM} does predict some ID not captured by F_{ROH} , demonstrating that F_{ROH} is not the optimum estimate. This could be explained in at least two ways. Firstly, some of the ID causal variants may be in strong LD with the genotyped markers (i.e., common). Or secondly, some causal ROH segments may be missing from F_{ROH} - for example because they fall below the calling threshold of 1.5 Mb. In this latter case, careful parameterization of more sophisticated ROH calling algorithms (e.g., GARLIC⁷⁸) may improve estimates of ID.

An additional, practical, advantage of F_{ROH} is its relative insensitivity to population structure¹³. This is especially true for the highest F_{ROH} samples which make the greatest contribution to statistical power¹³. In a multi-ethnicity cohort such as UK Biobank, the possibility of using every sample can make a substantial difference to statistical power. In contrast, marker-by-marker estimates of genome-wide heterozygosity, such as F_{GRM} and F_{SNP} , are extremely sensitive to expected allele frequencies and should only be calculated in populations with homogeneous ancestry.

If, for whatever reason, identifying ROH is not possible the results presented here demonstrate that F_{GRM} is a better predictor of ID than the canonical estimate of F based on excess homozygosity (F_{SNP}). Considering the increased weighting given to rare homozygotes by F_{GRM} , this further highlights the importance of rare variants.

3.4.3 Sex-specific effects

Theoretical and empirical evidence suggests that sex-specific inbreeding depression may develop when the strength of selection differs between males and females^{128,130–132}. In the theory of sexual selection, for example, one sex (usually males) competes for access to mates, and a deleterious phenotype may be more damaging to the reproductive success of the competing sex. In humans, the importance of sexual selection remains contested¹³³, but both anthropological and genetic evidence tend to support a slightly smaller effective population size in males¹³⁴.

Only a few published studies have estimated human inbreeding depression separately in men and women, and it is not easy to draw conclusions from the results^{1,135}. In ROHgen 13 traits differ significantly between men and women, but 7 of these are traits not associated with F_{ROH} in the combined-sex cohort, which makes it difficult to say whether the effect is more deleterious in one sex or the other. In the remaining six traits, the effect (assumed to be deleterious) is greater for men five times out of six. Notably, F_{ROH} substantially increases age at first birth for men, but not for women, consistent with stronger selection in men. Equally, body size is classically reported as subject to sexual selection in other species^{136,137} and, in both ROHgen and UKB, F_{ROH} decreases body weight in men, but not in women. However, despite these few sex-specific effects, there is little evidence that the effects of inbreeding are systematically greater in men. Specifically, averaging across 147 traits in UK Biobank, the effect in men is almost identical to that in women (Figure 38). This equality also holds true for the traits most closely related to reproductive success (*age_at_first_sex*, *number_sexual_partners*, *neb* and *ever_had_children*), all of which have sex-specific estimates within 25% of their combined-sex average. Interestingly, the lack of sex-specificity in these traits,

contrasts with the findings of Gardner et al. who report that the ***heterozygous*** burden of protein-truncating variants (PTVs) has a sex-specific effect on *ever_had_children*¹²⁴.

In summary, the results reported in this thesis paint a slightly confusing picture of sex-specific inbreeding depression in humans. For most traits F_{ROH} has a similar effect in males and females, but for a small number of important traits the effect is significantly larger in males. It might be worth noting that confounders such as socio-economic status (see *weight* in Figure 56) and religiosity (see *alcohol_drinker* in Figure 38), may have sex-specific effects, so future studies will be needed to replicate the sex-specific results reported here, and to establish their genetic causation.

3.4.4 Effects of different ROH lengths

The large sample of ROHgen, and *multivariate* method in UKB, allow the effects of different ROH lengths to be estimated with greater precision than possible in previous studies^{70,71}. However, partitioning the variation of F_{ROH} into separate (potentially correlated) variables necessarily increases the uncertainty around effect estimates and currently restricts meaningful interpretation to only the best powered individual traits (such as height in ROHgen) or averages of many traits (such as *multivariate* in UKB).

For both traits, the effects of longer ROH are consistent with $\hat{\beta}_{F_{ROH}}$ (see Figures 43 and 44), perhaps unsurprisingly since, even in a relatively outbred cohort like UK Biobank, ROH > 3 cM contribute 88% of the statistical power.

However, more interestingly, even the shortest (< 1 cM), and therefore ‘oldest’, ROH also have a significant effect on the *multivariate* trait (Figure 44) – which can be thought of as a sort of average of all significant UKB traits. The simulations in section 2.6.1 imply that ROH < 1 cM derive from common ancestors typically more than 50 generations ago, and their distribution in UKB British (section 2.5.5.2) is consistent with random inheritance. This has several interesting implications: firstly, the effect of ROH < 1 cM should be unconfounded by the environmental biases

discussed in Chapter 4, and might therefore be considered an improved estimate of the true genetic effect of F_{ROH} . Secondly, any variants made homozygous in $ROH < 1$ cM are likely to have arisen as mutations many generations ago (and at least two copies have survived to the current generation), and are therefore unlikely to be ultra-rare in the population.

The vast majority of $ROH < 1$ cM occur in ROH islands (see section 2.5.5), which some authors have doubted to represent autozygous segments⁹⁴. However, Figure 17 shows that recombination rate is a major determinant of ROH density, supporting an interpretation that ROH islands do represent autozygosity – albeit from a distant ‘base’ generation. Also, the observed inbreeding depression effect of $ROH < 1$ cM implies that ROH islands are associated with the homozygosity of ID causal variants, as expected if they reflect autozygous haplotypes. If the causal effect of short ROH is confirmed by future studies, it may be worth considering that, in terms of inbreeding depression, the relevant inbreeding coefficient appears to be identity-by-descent from a distant base generation. ROH calling methods such as LD-pruning and HMM models seem to implicitly estimate autozygosity from more recent base generations by discounting the LD induced by distant common ancestors.

In summary, the analyses of sections 3.3.7 and 3.3.8 are consistent with the interpretation that inbreeding depression is mainly caused by the homozygosity of rare, but not ultra-rare, genetic variants (approximately 0.01% to 5%). Hopefully, larger future studies will be sufficiently well powered to confirm this for more individual traits, and perhaps explore the effect of ROH even shorter than those called here.

4 Non-genetic confounding of F_{ROH}

4.1 Introduction

Potential confounding of the genetic effects of autozygosity by unknown, or unobserved, non-genetic factors is a serious concern which should be addressed in any study of the inbreeding in human populations.^{111–113}.

Estimating the effect of F_{ROH} on a trait is different to estimating the effect of a variant (e.g., a SNP) on a trait (as done in GWAS). In this latter type of study, as RA Fisher wrote, *“Genetics is indeed in a peculiarly favoured condition in that Providence has shielded the geneticist from many of the difficulties of a reliably controlled comparison”*¹³⁸; This is because Mendelian segregation causes genetic variants to be inherited randomly and essentially invisibly. At least for common variants, in the absence of population structure, this brings about a situation where SNPs can be assumed to be distributed randomly among samples. This randomness means that the presence of a SNP is unlikely to be associated with any non-genetic factors affecting the trait. Unfortunately, this is far from the case for F_{ROH} . Cousin marriage is well known to be associated with socio-economic and cultural conditions⁴, and inbreeding coefficients are therefore, unlike SNPs, not randomly distributed in a population. Instead, inbreeding coefficients are inextricably linked to both endogamy and consanguinity and therefore to potentially confounding factors.

4.1.1 Confounding

The problem of confounding is illustrated in Figure 45: this study aims to estimate the causal genetic effect of F_{ROH} on human complex traits ($\beta_{F_{ROH}}$), and the observed associations have, so far, been assumed to be due to this effect.

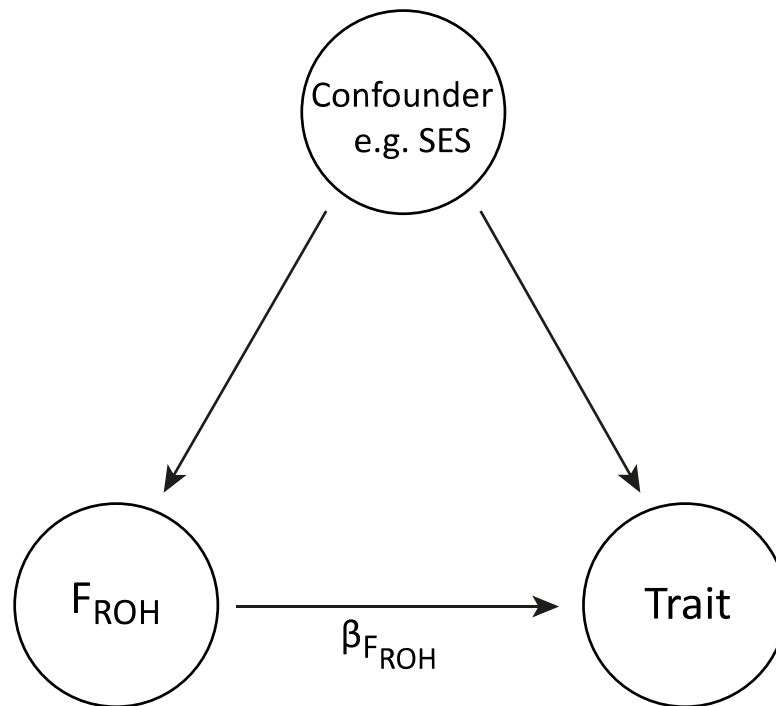


Figure 45: **Graphical representation of causal and confounder explanations of an association between F_{ROH} and a trait.**

However, the observed associations could equally be explained by any unmeasured variable which is itself correlated with both F_{ROH} and the trait of interest. For example, it is well known that many human traits are strongly correlated with socio-economic status (SES), and it is likely that average inbreeding levels also vary with SES in many populations.

4.1.1.1 Confounding by parental phenotypes

Since an individual's inbreeding coefficient is determined by the kinship of their parents (and random chance) the most natural confounders are parental phenotypes. An example is reported by Abdellaoui et al.¹¹² - in a cohort of ~2000 Dutch adults, the authors report that more highly educated adults migrate further from their birthplace and are therefore less likely to reproduce with close relatives.

Abdellaoui et al. find a strong association between a subject's F_{ROH} and the educational attainment of their parents (father: p -value $< 10^{-5}$ mother: p -value = 9×10^{-5}). Interestingly, they find a much weaker association between F_{ROH} and the subject's own educational attainment (p -value = 0.045), which illustrates that confounding by parental phenotypes requires the confounder to be transmissible to the offspring (either by shared genetic or environment). Confounding by parental phenotypes assumes that whichever factors make an individual more likely to reproduce with a relative also affect the phenotypes of their offspring. In the example from Abdellaoui et al., lower educational attainment in the parents results in slightly lower educational attainment in their offspring – broadly consistent with published estimates of educational attainment heritability around 40%¹³⁹. Abdellaoui et al. refer to this confounding as “*assortative mating*”, but the nomenclature is discussed further in section 4.1.2.

4.1.1.2 *Environmental confounding caused by parental relatedness*

Although parental phenotypes are likely confounders, they are not the only possibility. For example, having related parents might cause environmental changes which affect a subject's phenotypes. One might imagine, for example, that the offspring of related parents have reduced (or increased) extended family support, or act, or are treated, differently because their parents are related. Such ‘environmental confounding *caused* by parental relatedness’ seems less likely than ‘confounding by parental phenotypes’ (where concrete examples have been reported^{111–113}), but should be considered since some methods of accounting for confounders exclude only the latter.

4.1.1.3 *Accounting for confounders*

The problem of confounding has been well recognized by previous research, and various efforts have been made to mitigate potential bias. Much of the earliest research into inbreeding depression in humans involved experimental study design, where defined relative types, such as the offspring of first cousins, were compared to the offspring of unrelated parents¹⁰⁰. These studies typically undertook to match

cases with controls, and any conclusions are only as reliable as this matching. This method is subject to the same pitfalls as fitting covariates in statistical models – i.e., matching can only be done on measured variables and may overlook unknown confounders.

Other early studies of inbreeding estimated effects from pedigree inbreeding coefficients¹⁴⁰. Gathering large family histories is time-consuming and, importantly, rarely linked to rich phenotyping. Such studies have therefore focused on relatively small, highly inbred, communities and most often on phenotypes which are necessarily collected with a family history – e.g., reproductive success¹⁰². Since families are likely to have some degree of environmental consistency, small family studies may better control for confounders than larger population studies, but cannot account for ‘confounding *caused* by parental relatedness’ (see section 4.1.1.2), and are necessarily very limited in size.

More recently, the advent of genome-wide SNP genotyping has allowed inbreeding to be studied in population cohorts most often gathered for other research purposes. This has the great advantage of very large sample sizes which allow effects to be estimated in outbred human populations. However, such population cohorts have the disadvantage that they were often gathered for GWAS and, as such, recording phenotypes likely to be informative of confounders, especially religious, or cultural beliefs, was not always a priority. Nevertheless, a popular approach in these population cohorts has been to fit the available covariates – most often ancestry informative principal components, and occasionally measures of SES or religiosity. However, it is unrealistic to assume that these fully describe the complete effect of confounders for every individual. For example, in a phenotype model of height in UKB, separate measures of SES remain highly significant despite conditioning on each other (Supplementary Table 29). In this context, it is possible that another measure of SES, for example F_{ROH} , could also be significantly associated with height, and misinterpreted as having a causal effect.

Several previous studies have examined the sensitivity of $\hat{\beta}_{F_{ROH}}$ to the inclusion of covariates associated with potential confounders. In results from a global consortium, Joshi et al. reported that including educational attainment as a covariate in models of height, FEV1, and cognitive g reduced $\hat{\beta}_{F_{ROH}}$ by 17, 18 and 35% respectively¹³. The authors interpreted this as evidence that a large part of the observed effects can be attributed to genetics, although this could also be interpreted as evidence of confounding and that more complete SES covariates may have reduced these estimates further. Johnson et al. went further, using the UK Biobank data, by comparing $\hat{\beta}_{F_{ROH}}$ from models with and without 7 sociodemographic covariates⁷¹. They found that when these sociodemographic covariates were included, two out of five previously significant effect estimates (height and FEV1) reduced below significance. The results from these studies suggest two difficulties with this approach: firstly, although fitting appropriate covariates can prove that an effect is caused by confounding, it can never prove that one is not. And secondly, consideration needs to be given to whether the covariates themselves are affected by F_{ROH} – for example in Joshi et al., if the association between F_{ROH} and EA is entirely caused by genetics then fitting EA as a covariate underestimates the effect of F_{ROH} on cognition by 35%.

Another popular approach has been to compare the effects of short and long ROH^{13,71,72}. Since meiotic recombination reduces haplotype lengths every generation, longer ROH are likely to be caused by relatively recent parental relatedness (consanguinity) while shorter ROH are more likely to be caused by distant parental relatedness (which is more frequent in small effective population sizes). Since confounders associated with consanguinity might differ from those associated with a small effective population size, estimating $\beta_{F_{ROH}}$ separately for short and long ROH may give some insight into the relative contributions of confounding and genetics. Unfortunately, there are also theoretical predictions that longer ROH may have greater relative effects than shorter ROH³³, which can make drawing robust conclusions difficult. For example, Howrigan et al. find that ROH longer than 5Mb predict lower cognitive ability while shorter ROH do not - and

interpret this as evidence of selective pressure against cognition reducing alleles⁷⁰. In contrast, Johnson et al. interpret an effect of ROH length on alcohol drinking status as evidence of confounding⁷¹. These results notwithstanding, several studies have found a relative consistency of effect for most traits when comparing short and long ROH – although this should be interpreted in light of the reduced statistical power concomitant with dividing an analysis in two.

A third approach involves considering parental phenotypes either directly^{72,112} (in the rare cohorts where these are available), or indirectly, by calculating a polygenic score from SNP genotypes of the subject⁷². This latter approach relies on the theoretical expectation that an individual's expected polygenic score is equal to the average of their parents' polygenic score. Thus, if parents with lower genetic predisposition to educational attainment are more likely to reproduce with a relative, then this genetic predisposition should be inherited by the offspring. In a study of 54 extreme inbreeding cases (EI) in UKB⁷², Yengo et al. found no evidence that extreme F_{ROH} (>0.1) was associated with lower genetic scores for EA. However, the variance of EA explained by a polygenic predictor is likely to be low (and is not given by Yengo et al.) so it is not clear whether this failure to reject the null hypothesis (of no parental confounding) is informative about the alternative hypothesis. Yengo et al. also consider the number of diseases reported for the parents of EI cases and find no significant enrichment relative to EI controls.

4.1.2 Assortative mating

In the context of inbreeding, the difference between confounding and assortative mating is subtle and appears to mainly a question of nomenclature. Assortative mating is generally defined as: individuals with similar phenotypes mating together more often than would be expected under random mating. Since relatives have more similar traits than random members of the population (at least for heritable traits), assortative mating causes increased homozygosity compared to a randomly mating population. This increased homozygosity will occur at all levels of a phenotype, but will be more pronounced at *both* trait extremes. As such, assortative mating based on a trait should not lead to a directional association

between homozygosity and the trait. In contrast, the confounding that has been considered here is a directional relationship where one trait extreme is more likely to reproduce with a relative – and not necessarily because they share similar phenotypes. For this reason, this type of confounding is referred to in this thesis as “*confounding by parental phenotypes*” and not “*assortative mating*”.

4.1.3 Introduction summary

In summary, it is well recognised that studying inbreeding in humans is complicated by social factors influencing mate choice. Unmeasured social factors may confound F_{ROH} and therefore bias estimates of inbreeding depression, or even lead to false inferences. Several methods of accounting for confounders have been employed by previous authors, but, at best, these have provided an absence of evidence for confounding, rather than evidence of an absence. As ever larger human population cohorts are gathered, it seems inevitable that humans will become the species of choice in which to study inbreeding and, if conclusions drawn from these studies are to be reliable, it is critical to establish the relative roles of confounding and genetics for all studied traits.

4.2 Methods

4.2.1 Relationship between F_{ROH} and early-life SES

Consistent with a hypothesis of confounding, F_{ROH} is associated with several components of socio-economic status (SES) in UKB (e.g., Educational attainment, Income, Number of vehicles in household, and Townsend deprivation – See section 3.3.5). However, these traits are measured in adult participants, so it is unclear whether increased F_{ROH} causes reduced SES, or whether these traits are simply correlated with some wider measure of SES. It is clear, however, that an individual’s F_{ROH} cannot influence anything occurring before their birth, and is unlikely to have a substantial effect of the phenotypes of others. Confounding between F_{ROH} and SES can therefore be explored, without the complication of genetic effects, by considering early-life indicators of socioeconomic status (see Section 4.3.1.1).

4.2.1.1 *Index of Multiple Deprivation at birth address*

The UK Biobank contains birth coordinates for 444,730 participants born in the UK. Like the home coordinates available for almost all participants, these birth coordinates are rounded to the nearest kilometre for anonymity. These coordinates were used, along with postcode data from the Office of National Statistics (ONS), to calculate an Index of Multiple Deprivation at birth (*IMD_birth*) for each UK-born participant. For comparison, an Index of Multiple Deprivation at current home address (*IMD_home*) was also calculated for all UKB subjects. As expected, *IMD_home* is strongly correlated ($r = 0.54$) with the UKB-provided Townsend deprivation index, although the latter is derived from non-rounded home addresses and a different set of deprivation metrics.

4.2.2 *Married couples in UKB*

It has been previously noted that the UK Biobank contains a large number of married (or cohabiting) couples¹⁴¹. Although UKB does not explicitly define couples it does provide a list of participants sharing the same home address. From this cohabiting data, and participants' answers to life-history questions, it is possible to identify couples with a high degree of confidence. In this thesis, UKB participants were considered to be a reproductive couple if they met the following criteria:

- 1) Both shared an address with exactly one other participant who self-reported to be living with a "Husband, wife or partner".
- 2) Were of opposite sex.
- 3) Both reported the same number of children born or fathered.
- 4) Were born within 15 years of each other.

Of the 74,808 couples identified by this method, 108 have kinship between 0.044 and 0.10 and are therefore likely to be first-cousin marriages. (Kinship coefficients, calculated by KING, are provided by UKB for all subjects with an estimated coefficient of kinship > 0.044). This 1:690 rate of first-cousin marriage is about half the 1:360 rate at which F_{ROH} falls within the same bounds in UKB, but this could be explained by long-term movements away from cousin marriage (the UKB

participants are a generation older than the offspring of these couples) or by participation biases (there is a slight enrichment of genetically British couples).

If the offspring of any children produced by these couples are assumed to have F_{ROH} equal to the couple kinship (as expected, on average), and phenotypes equal to the couple average (as expected of a perfectly heritable trait or perfectly transmitted environmental deviations), then any confounding of F_{ROH} by parental phenotypes can be estimated by regressing mean couple phenotype on couple kinship. In practise, offspring F_{ROH} is a noisy measure of parental kinship and so $\hat{\beta}_{F_{ROH}}$ would be expected to underestimate parental confounding, due to attenuation bias.

For 155 significant UKB traits the mean couple phenotype residual was regressed on couple kinship and the t -statistics are compared to $t_{F_{ROH}}$ in section 4.3.1.2.

4.2.3 Combining the effects of multiple confounders

If some, or all, of the observed effects of F_{ROH} are caused by confounding it may be possible to identify the confounding factors. If these confounders could be perfectly measured, then fitting them as covariates in the phenotype model would explain the observed effects and reduce $\hat{\beta}_{F_{ROH}}$ to zero. In practice, no single variable, nor linear combination of several variables, is likely to fully describe the effects of a complex confounder and therefore different measures of the same confounder are likely to be simultaneously informative. In this way, F_{ROH} may be associated with a trait due to its association with a confounder, even when other measures of that confounder are fitted as covariates in the phenotype model. However, in this scenario the estimated effects of F_{ROH} would be expected to be consistent (in direction and relative magnitude) with the effects of the fitted covariates across all affected traits.

It is therefore interesting to compare $\hat{\beta}_{F_{ROH}}$ to the effect of various covariates in the phenotype models (Figures 54 to 56). The comparison is complicated by the fact that F_{ROH} and the confounder covariates are in different units and because a potential confounder may be described by several covariates. To overcome the difficulty of units, comparisons were made between t -statistics rather than between

effect estimates. To overcome the difficulty of several covariates describing the same confounder, a predicted trait value was created for each subject, as described below. Taking, for example, the phenotype model of height, where a comparison is desired between $\hat{\beta}_{F_{ROH}}$ and the effect of SES, as described by *edu* and *townsend_deprivation*. First, the full phenotype model was fitted as:

$$\begin{aligned} height \sim F_{ROH} + townsend_deprivation + edu \\ + all_other_covariates \end{aligned} \quad (28)$$

From this model $\hat{\beta}_{F_{ROH}}$ was recorded, and the effects of *townsend_deprivation* and *edu* used to calculate a predicted height for each sample (with F_{ROH} and *all_other_covariates* held constant). The effect of this predicted height was then estimated in model (29) below, in order to estimate the combined effect of *edu* and *townsend_deprivation* on height ($\hat{\beta}_{predicted_height}$).

$$height \sim F_{ROH} + predicted_height + all_other_covariates \quad (29)$$

$\hat{\beta}_{predicted_height}$ is by definition positive and the sign of the relationship between F_{ROH} and *predicted_height* is unknown, but can be determined from the model:

$$predicted_height \sim F_{ROH} \quad (30)$$

Thus, an effect of combined confounders, which can be compared in direction and magnitude to $\hat{\beta}_{F_{ROH}}$, can be calculated as

$$sign(\beta'_{F_{ROH}}) * \hat{\beta}_{predicted_height} \quad (31)$$

Since the units of equation (31) differ from the units of $\hat{\beta}_{F_{ROH}}$ it is convenient to make comparisons across several traits on *t*-statistics rather than effect estimates (as shown in Figures 54 to 56).

4.2.4 F_{ROH} -based methods to quantify the role of confounding

Previous studies highlight the need to find methods of estimating $\beta_{F_{ROH}}$ which adequately control for any bias caused by confounding. Several approaches applied in ROHgen and UKB are described below in sections 4.2.4.1 to 4.2.4.6.

4.2.4.1 *Estimating effects in different populations*

The genetic effects of autozygosity on complex traits are assumed to be relatively consistent across different populations¹³. In contrast, non-genetic effects may differ between populations if associations between confounders and F_{ROH} are culturally specific. Therefore, one simple method to assess confounding is to compare $\hat{\beta}_{F_{ROH}}$ from culturally distinct populations. The ROHgen consortium is well suited to this method since contributing cohorts represent a range of worldwide populations.

4.2.4.1.1 In ROHgen

ROHgen cohorts were allocated to one of seven ancestral groups (European, S&W Asian, African, East Asian, Hispanic, Japanese, and Mixed) based on the reported ethnicity of the cohort subjects (Supplementary Table 1). Within each of these groups $\beta_{F_{ROH}}$ was estimated for all traits by fixed-effects meta-analysis and results are reported in Supplementary Table 16.

Similarly, ROHgen cohorts were also allocated to one of four inferred demographic origins as determined by their location on a plot of $\bar{F}_{ROH}$ versus F_{IS} (Figure 59). Again, $\beta_{F_{ROH}}$ was estimated within each of these groups by fixed-effects meta-analysis and results are reported in Supplementary Table 17.

4.2.4.1.2 In UK Biobank

Although 83.9% of UK Biobank participants are genetically British (with mean $F_{ROH}=0.00352$) there are smaller numbers from ethnicities with higher average F_{ROH} , including: 1748 participants declaring Pakistani ethnicity (mean $F_{ROH}=0.0255$), 221 Bangladeshis (mean $F_{ROH}=0.010$) and 5716 Indians (mean $F_{ROH}=0.00928$). Because of their higher average F_{ROH} , the 78,659 non-British UKB participants have similar statistical power for estimating $\beta_{F_{ROH}}$ as the 409,629 British participants.

Estimates of $\beta_{F_{ROH}}$ in both the British and non-British groups were obtained from the full phenotype models by multiplying F_{ROH} by two dummy variables indicating membership (gb), or not (not_gb), of the genetically British group.

$$trait \sim F_{ROH}: gb + F_{ROH}: not_gb + covariates \quad (32)$$

The statistical significance of the difference between the British and non-British estimates was evaluated separately by testing the significance of the interaction $F_{ROH}: gb$ in the model:

$$trait \sim F_{ROH} + gb + F_{ROH}: gb \quad (33)$$

4.2.4.2 Plotting traits in bins of increasing F_{ROH}

If the effects of F_{ROH} are caused by increased homozygosity of causal variants, these effects might be assumed to increase proportionally with F_{ROH} , since increasing F_{ROH} will proportionally increase the expected number of homozygotes. In contrast, there is no reason to expect the effect of a confounder to be proportional to F_{ROH} across the full range of observed inbreeding coefficients. This is particularly apparent at the extremes: very high F_{ROH} caused by reproduction between close relatives is a very different situation to the common variation of short ROH, about which carriers might well be unaware. It would be coincidental for the effect of a confounder to be proportional to F_{ROH} in both situations.

It is therefore informative to plot the effect F_{ROH} across the range of observed values. To do this, subjects were allocated to bins of increasing F_{ROH} (specifically: {0, 0.002, 0.0041, 0.0067, 0.0108, 0.0186, 0.0333, 0.06, 0.10, 0.18 and 1.0}) and the mean trait residuals and standard errors were calculated in each bin.

When a trait was measured in several cohorts, a meta-analysis was calculated for each bin mean. Within each cohort the expectation of ϵ' is zero at the mean F_{ROH} , however as mean F_{ROH} varies between cohorts (Supplementary Table 3) it was necessary to express ϵ' relative to a common F_{ROH} before meta-analysis. Hence, for this analysis only, the trait residuals (ϵ') were expressed relative to the $F_{ROH}=0$ intercept, i.e., by subtracting μ from equation (19).

4.2.4.3 *Effect of F_{ROH} within full siblings*

Although full siblings share the same two parents, some variation in F_{ROH} is expected since, at many sites they inherit different haplotypes from one, or both parents. Mendelian segregation is believed to be entirely random and so the variation of F_{ROH} within full siblings should be completely uncorrelated with any confounders. Siblings might, of course, experience a very different environments, but there is no reason for these differences to be correlated with F_{ROH} . (Unless these environmental differences are caused by F_{ROH} , which would still be of interest). One very minor caveat to this rule concerns indirect genetic effects – the F_{ROH} of one sibling could potentially affect the phenotype of another, but this is expected to be of very minor importance for most traits.

The variance of F_{ROH} within siblings is very much reduced from the population variance (Supplementary Table 3), so the main difficulty with this method lies in gathering a sufficiently large sample of siblings. This method was developed after the initial analysis plan was distributed to the ROHgen consortium, but was subsequently applied in several of the most powered cohorts (including Biobank Japan, deCODE Iceland and Iran, Estonian biobank, Born in Bradford, and UK Biobank). Together, the cohorts which conducted the siblings analysis contributed 79.6% of the total statistical power in the ROHgen meta-analysis of height.

Close relatives were identified using KING (parameters: --related --degree 2) and full siblings were then selected as pairs with genomic kinship > 0.175 and IBS0 > 0.001 . These parameters were chosen from the histograms shown in Figures 46 and 47, and verified by the analyses described in section 4.2.4.3.1. Monozygotic twins (sample pairs with Kinship > 0.4), were also included in these full-sibling families. Although monozygotic twins are expected to have identical F_{ROH} , they may not have identical trait values. Including additional trait measurements decreases the sampling error of the within-family variance estimate, hence increasing statistical efficiency. Dizygotic twins were also included.

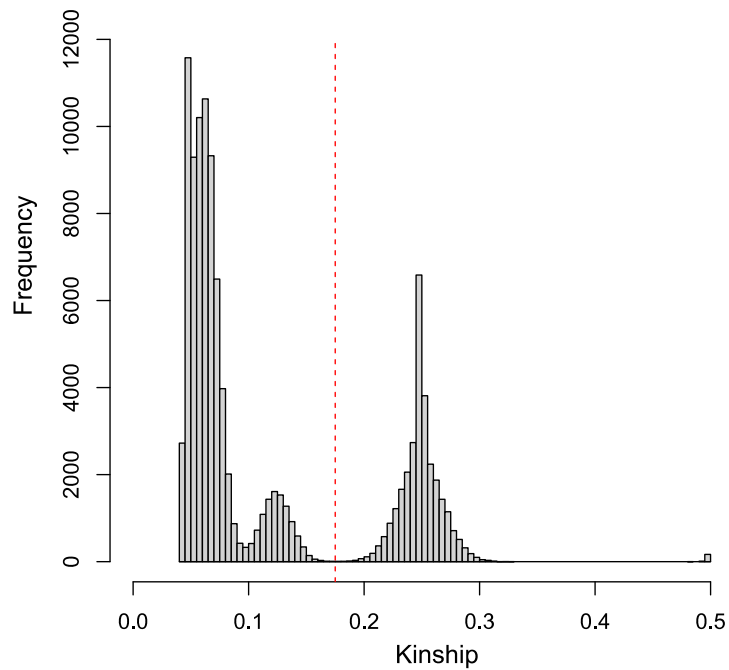


Figure 46: **Histogram of kinship coefficient for all close relatives ($\phi > 0.044$) in UK biobank.** Kinship coefficients for first-degree relatives have an approximately Gaussian distribution, centred on 0.25, which does not overlap significantly with other relative types. Based on this distribution, sample pairs with kinship coefficient > 0.175 were assumed to be first-degree relatives.

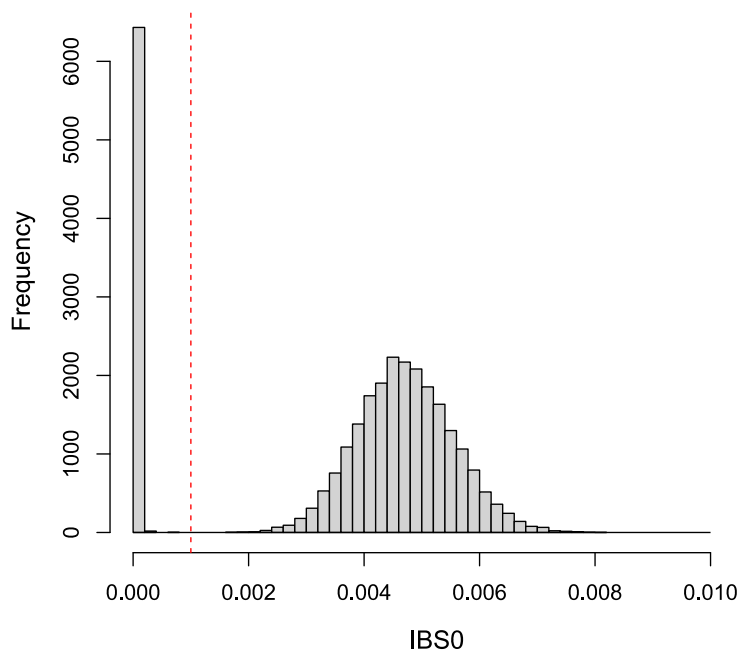


Figure 47: **Histogram of IBS0 for all relative pairs with Kinship coefficients > 0.175 .** Parent-offspring pairs (and MZ twins) have IBS0 close to the theoretical expectation of zero, with small deviations (< 0.001) presumably due to genotyping errors or de novo mutations. Sample pairs with IBS0 > 0.001 (and Kinship > 0.175) were assumed to be full siblings.

From the identified sibling pairs, full sibling families were identified by the function *max_cliques()* in the network analysis R package *iGraph*. This function identifies sub-networks where all vertices (samples, in this case) are connected to all others in the sub-network (families, in this case). I.e., to be included in a sibling family a sample must be identified as a sibling (Kinship > 0.175 and IBS0 > 0.001) of all other samples in that family. Finally, in UKB only, sibling families with mean $F_{ROH} < 0.19$ were excluded to minimise the chance of other close relative types being misidentified as full siblings (based on the results of section 4.2.4.3.1).

Each family was identified by a categorical variable, *sibling_family*, and the effect of F_{ROH} within full siblings ($\beta_{F_{ROH_wSibs}}$) was then estimated by fitting the model:

$$trait \sim F_{ROH} + age + age^2 + sex + sibling_family \quad (34)$$

Age and sex are not expected to confound F_{ROH} within sibling families, but were fitted because they may explain some of the trait variance between siblings and hence improve statistical power. Other covariates fitted in the full phenotype models (e.g., declared ethnicity or principal components) were not fitted because, although they do vary slightly between siblings, this variation is not thought to be informative. Also, because the sibling method is believed to be free from confounding, great care was taken not to introduce any methodological bias. Although it is not thought to be the case (see section 3.2.5), principal components are calculated from genotypes and could conceivably measure autozygosity, and fitting alternative measures of autozygosity in equation (34) would bias $\beta_{F_{ROH_wSibs}}$.

Fitting equation (34) by ordinary least squares is computationally time-consuming, since *sibling_family* is a categorical variable with many levels (20,294 full sibling families in UKB). This was resolved using the function *felm()* in the R package *lfe* which “uses the Method of Alternating projections to sweep out multiple group effects from the normal equations before estimating the remaining coefficients with OLS”^{142,143}.

4.2.4.3.1 Verifying that samples identified as siblings are indeed siblings

UK biobank contains a number of phenotypes which can help to confirm that samples identified as siblings are indeed siblings and not other relative types. These phenotypes include mother's age (or age at death), father's age (or age at death) and number of full siblings – all of which should be identical for full siblings. In practice, even MZ twins in UKB (where there is no uncertainty about their relationship based on the genetic data) occasionally report substantially different answers to these three questions. Therefore, for each relative pair, a degree of support for the hypothesis that the pair are full siblings was calculated from the absolute difference of their answers. For example, reporting identical mother's age is 15 times more common among the presumed siblings than it would be for two randomly chosen samples from the UKB cohort. Equally reporting a difference of 1 year is 5 times more likely among the siblings than it would be for two random samples (mother's age is recorded with resolution of one year, so although it can be corrected for date of attending the assessment centre a difference of up to 1 year is possible even when all answers are correct). Conversely, reporting a difference of 10 years is 20 times less likely among the presumed siblings than it would be for two random samples.

In addition to the three "family" traits, which should be identical for siblings, year of birth is also informative since it is rarely identical, nor greatly different, for siblings. By multiplying the likelihood ratios for age, mother's age (or age at death), father's age (or age at death), and number of full siblings an overall ratio of support was calculated for each relative pair. The level of support for being full siblings was found to be equally high for all sample pairs with kinship > 0.19 , but lower for pairs below this threshold (Figure 48).

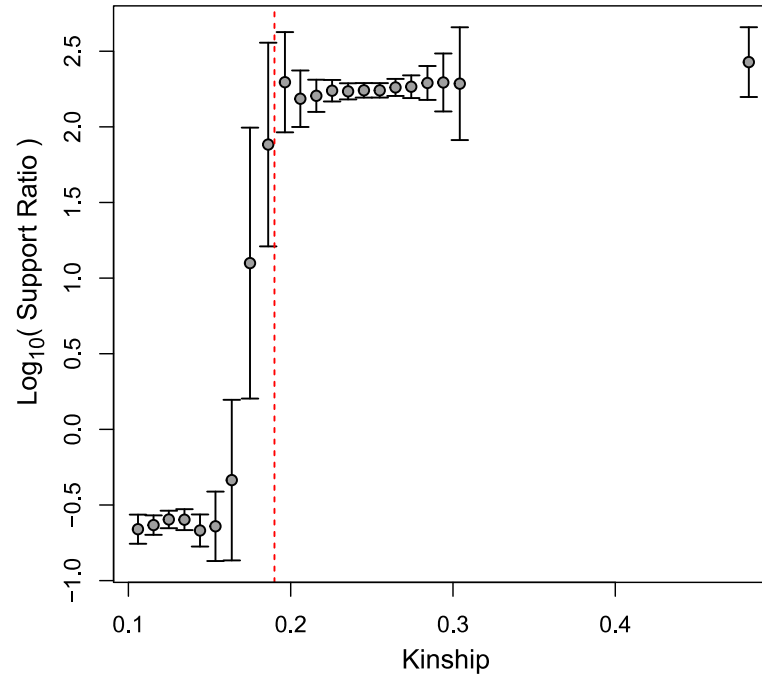


Figure 48: **Support for hypothesis of full siblings as a function of sample pair kinship.** The mean log-likelihood ratio for being full siblings is plotted in bins of increasing kinship for all pairs with kinship > 0.1. The average likelihood ratio is above 100 for pairs with kinship > 0.19, but starts to reduce below this threshold.

4.2.4.3.2 Estimating $\beta_{F_{ROH}}$ between sibling families

As discussed previously, $\hat{\beta}_{F_{ROH_wSibs}}$ is believed to give a confounder-free estimate of the genetic effect of F_{ROH} on a trait. $\hat{\beta}_{F_{ROH_wSibs}}$ can therefore be compared to a whole population estimate of $\beta_{F_{ROH}}$ to determine what fraction of observed effects can be attributed to genetics. However, if the sample used to estimate $\beta_{F_{ROH}}$ contains the siblings used to estimate $\beta_{F_{ROH_wSibs}}$ there is sample overlap and therefore covariance between the estimates. However, the variation of F_{ROH} within sibling families is typically a small fraction of the total population variance (in ROHgen: min 4% to max 21%, sample size weighted average = 12%, see Supplementary Table 3) and siblings are often a small part of the sample cohorts (in ROHgen: min 1% to max 80%, average = 10%, see Supplementary Table 3). The covariance between $\hat{\beta}_{F_{ROH_wSibs}}$ and $\hat{\beta}_{F_{ROH}}$ is therefore expected to be modest and the subsequent bias in the ratio $\hat{\beta}_{F_{ROH_wSibs}}:\hat{\beta}_{F_{ROH}}$ tolerable.

Nevertheless, since confounding is a central question of this thesis, the issue of sample overlap was addressed in UKB. A simple method of ensuring the independence of $\hat{\beta}_{F_{ROH_wSibs}}$ and $\hat{\beta}_{F_{ROH}}$ is to remove all siblings from the calculation of $\hat{\beta}_{F_{ROH}}$, but this has the disadvantage of reducing statistical power by discarding the variation of F_{ROH} between sibling families. Alternatively, the effect of F_{ROH} between families ($\hat{\beta}_{F_{ROH_bSibs}}$) can be calculated and used in comparisons with $\hat{\beta}_{F_{ROH_wSibs}}$.

$\hat{\beta}_{F_{ROH_bSibs}}$ can be calculated by the R function *plm()* with method="between" and index="sibling_family". Alternatively, identical estimates of $\hat{\beta}_{F_{ROH_bSibs}}$ can be obtained from simple linear regression by regressing the family mean phenotype on family mean F_{ROH} for each sibling family. In both methods, samples without identified siblings are included in the calculation by assigning them a unique *sibling_family* identifier. Although $\hat{\beta}_{F_{ROH_bSibs}}$ was calculated for all traits in UKB, the increase in power compared to removing all siblings from $\hat{\beta}_{F_{ROH}}$ was found to be negligible. Therefore, in this thesis $\hat{\beta}_{F_{ROH_wSibs}}$ is compared with $\hat{\beta}_{F_{ROH}}$ estimated in the cohort of UKB subjects without identified siblings. The between-sibling method is reported here solely because it may, in future, be useful in cohorts with a greater fraction of siblings.

4.2.4.3.3 Erratum

The method of estimating $\beta_{F_{ROH_wSibs}}$ that was published in Clark et al. (2019) is now understood to contain an error¹. Although the point estimates of $\beta_{F_{ROH_wSibs}}$ are unbiased, the standard errors are underestimated by a factor of approximately $\sqrt{2}$. The overall conclusions remain unchanged, although one trait (overall health) which was stated to reach Bonferroni corrected significance falls short of this threshold with corrected standard errors. In this thesis the corrected method described here was reapplied in three cohorts (UKB, BBJ and deCODE-Iceland) and all other standard errors of $\hat{\beta}_{F_{ROH_wSibs}}$ were multiplied by $\sqrt{2}$.

4.2.4.4 *Effect of F_{ROH} in uniparental disomy (UPD) cases*

Uniparental disomy (UPD) occurs when segregation errors result in both copies of a section of an individual's genome being inherited from the same parent. In the most extreme cases, one parent contributes both copies of one, or more, entire chromosomes. When both transmitted alleles are copies of the same parental haplotype (called isoUPD) the UPD section manifests as a long ROH in the offspring. A recent large study estimated the prevalence of UPD in a healthy population as 1 in 2000, with approximately half of these cases being isoUPD⁷⁹.

The phenotypic effects of isoUPD are not well documented, but are likely to combine the consequences of imprinting difficulties with the increased homozygosity of recessive variants in this most extreme form of autozygosity. Importantly in the context of confounding, the occurrence of isoUPD is thought to be a random event and is therefore likely to be independent of the social factors which may confound ROH caused by parental relatedness.

In theory, long isoUPD events are readily identifiable by ROH – so long as sufficiently stringent parameters are chosen to avoid confusing isoUPD with long autozygous segments caused by parental relatedness. Nakka et al. describe an algorithm for identifying isoUPD in European populations, which they used to identify 182 isoUPD cases in UKB subjects of Northern European ancestry⁷⁹. Here, the same algorithm was used to identify 181 isoUPD cases in UKB subjects of Northern European ancestry. (Note, UKB datasets may not be identical due to ongoing participant withdrawals).

Nakka et al. do not provide an algorithm to identify isoUPD in non-European samples, but to maximize isoUPD sample size in this study, conservative parameters were applied to non-European samples. Any sample with >100Mb of ROH on a single chromosome and no other autosomal ROH was considered to be a UPD case. This identified a further 19 likely isoUPD cases in the non-European UKB samples.

To estimate the effect of F_{ROH} within isoUPD cases, binary indicator variables (*upd* and *not_upd*) were defined for all UKB subjects. $\beta_{F_{ROH_wUPD}}$ and $\beta_{F_{ROH_wNotUPD}}$ were then estimated from the model:

$$trait \sim F_{ROH}:upd + F_{ROH}:not_upd + covariates \quad (35)$$

For $\hat{\beta}_{F_{ROH_wUPD}}$ to be free from confounding requires that the 200 isoUPD cases are indeed random, and not associated with any factors which also affect phenotypes. This was checked by the method described in section 4.2.7 as shown in Figure 87.

4.2.4.4.1 isoUPD and somatic mosaicism

Although somatic mosaicism is not mentioned in Nakka et al., it seems likely that a small number of the isoUPD segments identified by their predictor are actually somatic mutations associated with Haematological cancers – notably ten approximately 40Mb segments at the p end of chromosome 9 (See section 5.3.1.3.1). With the genotype data used here (and in Nakka et al.) it is not possible to distinguish somatic UPD from germline UPD and therefore all called UPD segments were used in this analysis. However, there is sufficient doubt about the direction of causality (i.e., whether UPD at 9p22 causes blood cancer, or whether blood cancer causes genomic instability which causes somatic UPD at 9p22) that four traits strongly associated with blood cancers (*icd10_d45_polycythaemia_vera*, *erythrocyte_distribution_width*, *immature_reticulocyte_fraction*, and *high_light_scatter_reticulocyte_percentage*) were excluded from the UPD analysis.

4.2.4.5 Effect of F_{ROH} conditioned on F_{ROH_cm}

This chapter considers the confounding of F_{ROH} by social factors correlated with parental relatedness. In theory, parental relatedness is estimated more accurately by a genetic map distance measure of F than by a physical distance measure such as F_{ROH} . Indeed, the variation of F_{ROH} conditioned on an accurate map distance measure should be independent of all confounding since only a subject's ancestry, the genetic recombination rate map, and chance, determine where haplotypes, and therefore ROH, start and end. ROH of a given centimorgan length may vary

considerably in physical length due to the varying relationship between cM and Mb throughout the genome.

Unlike confounding, it might be hypothesised that the genetic effects of autozygosity are distributed more evenly in physical distance than in genetic distance. This was tested by first calculating a centimorgan equivalent of F_{ROH} for each sample:

$$F_{ROH_cM} = \frac{\sum_{i=1}^n cM_i}{3550cM} \quad (36)$$

Where cM_i is the genetic length of each ROH and 3550cM is a nominal genetic length. The conditional effects of F_{ROH} and F_{ROH_cm} were then estimated for 155 significant traits by fitting bivariate models:

$$trait \sim F_{ROH} + F_{ROH_cM} + covariates \quad (37)$$

4.2.4.6 Effect of F_{ROH} in adopted subjects

7207 UKB participants report having been adopted as a child. In a typical outbred population this sample size would provide only very limited power to estimate $\beta_{F_{ROH}}$, however the variance of F_{ROH} in these adopted subjects is approximately five times greater than in the non-adopted. This is primarily due to a substantial enrichment in the number of presumed offspring of first-degree (N=16) and second-degree relatives (N=5).

It seems likely than some of the confounding associated with the parental environment would be eliminated by adoption, and therefore estimates of $\beta_{F_{ROH}}$ within these adopted samples should be relatively free of confounding. Effect estimates within adopted samples ($\hat{\beta}_{F_{ROH_wAdopted}}$) were estimated and compared to the corresponding effects in non-adopted samples ($\hat{\beta}_{F_{ROH_wNotAdopted}}$) by fitting the following model:

$$trait \sim F_{ROH:adopted} + F_{ROH:not_adopted} + covariates \quad (38)$$

4.2.5 Comparing $\beta_{F_{ROH}}$ to the effect of hemizygous deletions

Hemizygous deletions in UKB genotypes were previously identified for the purpose of removing them from F_{ROH} (Section 2.3.1.1). The effects of these deletions can be estimated in a very similar way to $\hat{\beta}_{F_{ROH}}$. For each sample the deleted genomic fraction (F_{DEL}) was calculated by summing the base-pair length of all deletions and dividing by a nominal genome length, as shown previously for F_{ROH} in equation (4). Then, for all 155 significant UKB traits, the effect of F_{DEL} ($\beta_{F_{DEL}}$) was estimated by fitting the model:

$$trait \sim F_{ROH} + F_{DEL} + covariates \quad (39)$$

4.2.6 Comparing effect estimates from different methods

This chapter frequently needs to compare estimates of $\beta_{F_{ROH}}$ from two different methods across many different traits. These comparisons are difficult to synthesise across a large number of traits (and to represent graphically) since the estimates are often in different units, and derive from both binary and quantitative traits. This difficulty was addressed in two different ways:

4.2.6.1 Effect estimate ratios

Traits can be reduced onto the same unitless scale by calculating the ratio of two corresponding estimates, for example $\hat{\beta}_{F_{ROH_wSibs}} / \hat{\beta}_{F_{ROH}}$. Often, as in this example, the numerator estimate will be free from confounding, and the denominator estimate will be derived from the whole population. In this scenario, a ratio of 1 would be expected for an entirely genetic effect and a ratio of 0 expected in the case of complete confounding. The ratio may be between 0 and 1 if an effect is caused by both confounding and genetics, or even outside this range, if the effects of genetics and confounding are in opposite directions.

As well as the point estimate of $\hat{\beta}_{F_{ROH_wSibs}} / \hat{\beta}_{F_{ROH}}$ it is useful to establish a confidence interval. Analytic formulae exist for the standard error of a ratio of estimates, but these break down if the denominator is insufficiently far from zero. Therefore, the median and 95% confidence interval of all effect estimate ratios

were calculated by resampling 10000 random values of both the numerator and denominator.

Meta-analysis of these effect estimate ratios was done by the usual inverse variance method, assuming the sampling distributions of the ratios are normally distributed with means equal to the resampled medians and standard errors equal to the 95% confidence intervals divided by 3.92 (i.e., as expected of a Gaussian distribution).

4.2.6.2 *TT plots*

The ratio method described above does not lend itself well to a graphical comparison of effect estimates; for a relatively small number of traits a forest plot remains readable (e.g., Figure 65), but less so for the 155 significant UKB traits. Ideally, the two effect estimates would be plotted on the x and y axes, but the traits must be normalised for this to be meaningful. Normalising by trait standard deviations was considered, but this does not allow meaningful comparison of quantitative and binary traits, and different effect estimates have different standard errors on this scale. Displaying standard errors for a large number of points makes a very busy plot.

The solution chosen was to plot *t*-statistics in place of the effect estimates (e.g., Figure 66) - i.e., to normalise the effect estimates by their standard errors. Compared to normalising by trait standard deviations this has a number of advantages: it allows comparison of binary and quantitative traits, the standard errors on these axes are all implicitly equal to one, and it even allows comparison of different predictor variables (e.g., Figure 54). However, there are also disadvantages: the expected gradient for equal effects depends on the relative power of the two methods and may even vary slightly from trait to trait. Where an expected gradient was required, this was calculated from the ratios of the standard errors of the two methods.

4.2.6.2.1 Deming regression

Simple linear regression assumes the predictor variables are known with certainty. This is not the case in a TT plot where both axes are normalised estimates. However, assuming both tests in the TT plot are unbiased, the variance of all points is known and equal to one in both dimensions. In this case, the maximum likelihood fit can be found by a special case of total least squares called Deming regression with equal error variances, or orthogonal regression. In orthogonal regression, errors are minimized perpendicular to the regression line rather than along the y-axis as in simple linear regression. Unless stated otherwise, best fit lines on TT plots were estimated by orthogonal regression using the R function *deming()*.

4.2.7 Testing confounding-free methods

In this chapter, four methods have been proposed to estimate $\beta_{F_{ROH}}$ by means which are expected to eliminate, or greatly reduce, the potential bias of confounders. These “confounding-free” methods will be referred to as: sibling method, UPD method, adopted method, and centimorgan method.

The effectiveness of these methods at eliminating confounding can be tested by applying them to traits that are known to be associated with F_{ROH} by non-genetic causes – i.e., known confounders. The 74 traits use as covariates in UKB phenotype models conveniently fit this description.

$\beta_{F_{ROH}}$ was therefore estimated for each of these “confounder traits” both by the whole population method of regressing $Trait \sim F_{ROH}$ and by the four “confounding-free” methods. No additional covariates were applied in any models since the objective was to evaluate how effectively the methods themselves reduce confounding. Four of the 74 UKB covariates are categorical variables and therefore cannot be directly used as dependent variables. These categorical covariates were converted to numeric by determining the mean F_{ROH} of all samples in each category and using that in place of the categorical value.

Effect ratios for each method (as described in section 4.2.6.1) are plotted for the significant confounder traits in Figures 86 to 89. Since these traits are believed to

be associated with F_{ROH} entirely by confounding, the ratio of

$\hat{\beta}_{F_{ROH_confounding_free_method}} : \hat{\beta}_{F_{ROH}}$ is expected to be zero. The sibling method and the UPD method both perform perfectly, with confidence intervals of the average effect ratio overlapping zero and no evidence of heterogeneity between traits. This may seem trivial for the sibling method, but does provide evidence that siblings are correctly identified, and the method is correctly applied. The UPD result (Figure 87) is perhaps more useful since it confirms that the identified UPD cases are not associated with any covariates – i.e., they are not, for example, enriched in ethnicities with high average F_{ROH} . The strength of this result is demonstrated by repeating the UPD analysis, but replacing each UPD case by the non-UPD case with the nearest F_{ROH} (Figure 88). For these equivalent non-UPD cases the effect ratio averages 1.19 [95% CI 1.15 to 1.23], although there is significant heterogeneity (p -value = 1×10^{-16}). This heterogeneity is believed to be caused by the non-linearity of effect versus F_{ROH} for these confounder traits; the F_{ROH} of UPD cases falls in a narrow band from 0.02 to 0.09, while the whole population estimates of $\beta_{F_{ROH}}$ are calculated from the entire range of F_{ROH} .

The performance of the centimorgan method is less satisfactory since, although the average effect ratio is close to zero 0.06 [95% CI 0.02-0.10], there is very significant heterogeneity between traits (p -value 4×10^{-48}). In particular several principal components (pc4, pc5, p7 and pc8) are significantly associated with F_{ROH} despite conditioning on F_{ROH_cm} . It therefore appears that the difference between F_{ROH} and F_{ROH_cm} is informative of some aspects of ancestry, perhaps by being informative of average ROH length. Nevertheless, this method was retained, but should not be interpreted as completely free of potential confounders.

Finally, this method of evaluation is believed to be less suitable in the adopted cohort. Many of the confounder traits, (for example ethnicity, PCs, etc.) are not altered by adoption and therefore a significant association between F_{ROH} and ethnicity (for example) only confirms that high F_{ROH} adoptees come from high average F_{ROH} ethnicities.

4.3 Results

4.3.1 Relationships between F_{ROH} and confounders in UKB

4.3.1.1 Relationship between F_{ROH} and early-life SES

Coordinates of birth address (in black) and current home address (in red) are plotted for UKB participants in Figure 49. Birth addresses are distributed throughout the country while home addresses are clustered around the 17 predominantly urban assessment centres. Birth addresses outside Great Britain were not recorded and are not shown.

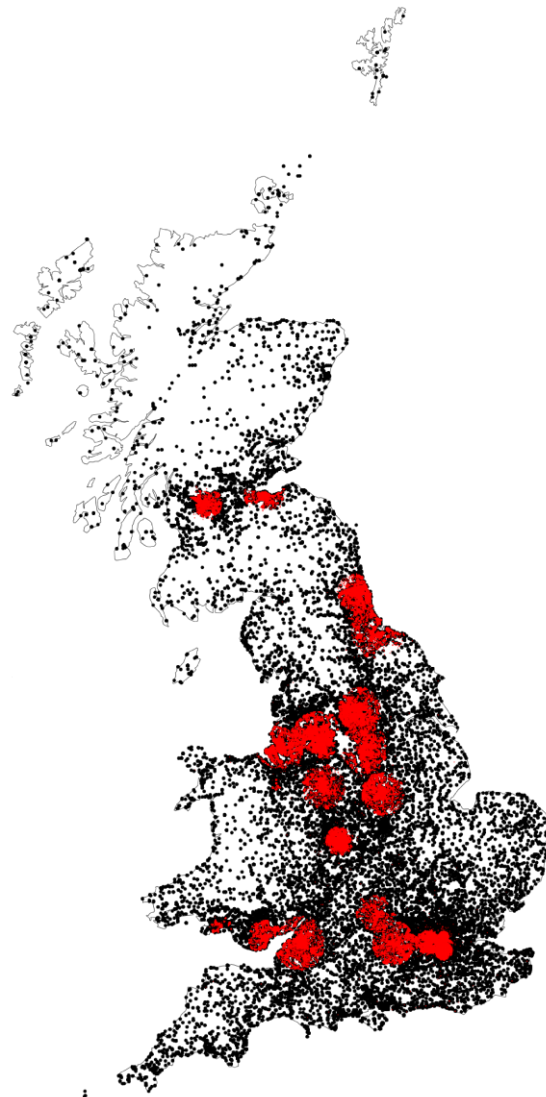


Figure 49: **UK Biobank participants' home and birth coordinates plotted on a map of Great Britain.** Home coordinates are shown in red and are clustered around UKB assessment centres. Birth coordinates are shown in black. Birth addresses outside Great Britain were not available at the same resolution and are not shown.

Indices of Multiple Deprivation (IMD) calculated from these birth and home addresses consequently have different distributions, with higher average deprivation and greater variability for *imd_birth* (Figure 50). On this scale lower values indicated less deprivation, so a majority of UK Biobank participants live in relatively favoured neighbourhoods, with smaller numbers in less privileged areas.

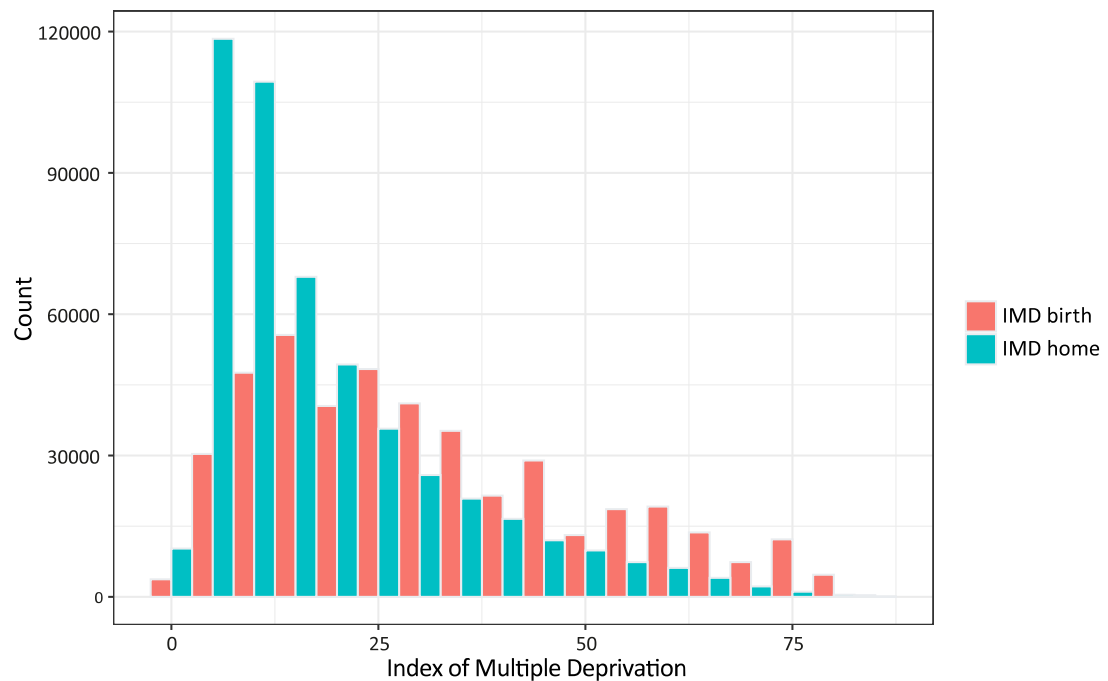


Figure 50: **Histograms of Indices of Multiple Deprivation (IMD) at home and birth addresses for UKB participants.** The mean and standard deviation of *imd_birth* are 30.9 and 19.4, respectively. The mean and standard deviation of *imd_home* are 18.3 and 14.9.

Figure 51 shows that *imd_birth* is strongly associated with four indicators of early-life socio-economic-status, however there is no significant association between *imd_birth* and F_{ROH} (Figure 52a). In contrast, there is a modest positive association between F_{ROH} and *imd_home* (Figure 52b) – i.e., higher F_{ROH} is associated with living in a deprived neighbourhood, but not with being born in one.

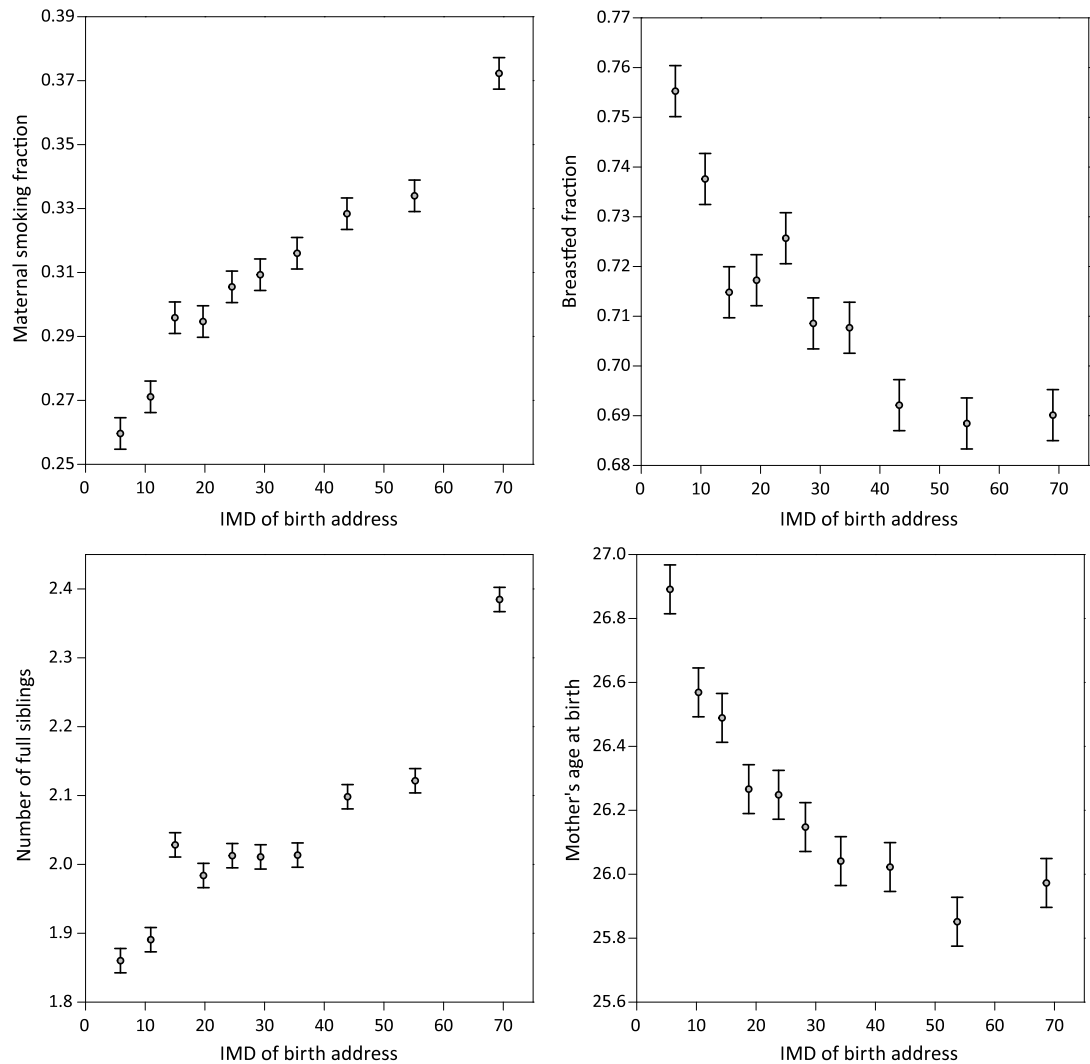


Figure 51: Strong associations between IMD at birth address and four markers of early-life socio-economic-status. a) Fraction of UKB participants reporting maternal smoking around time of birth. b) Fraction of UKB participants reporting to have been breastfed as a baby. c) Reported number of full siblings. d) Mother's age at birth calculated as difference between reported mother's age and participant's current age. All graphs are plotted in bins of increasing Index of Multiple Deprivation of birth address for 394,814 genetically British participants.

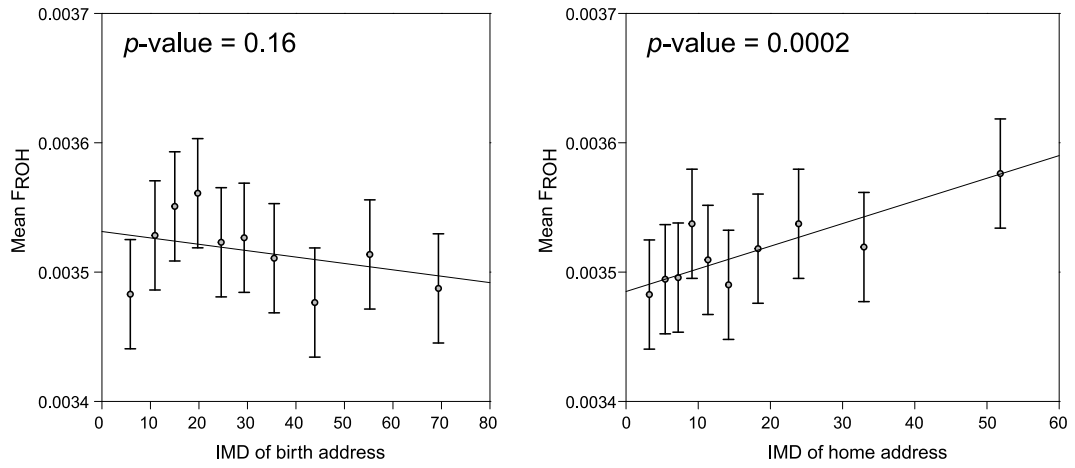


Figure 52: **Significant association between F_{ROH} and IMD at home address, but not at birth address.** a) Mean F_{ROH} of British-born, genetically British UKB participants in bins of increasing Index of Multiple Deprivation of birth address. b) Mean F_{ROH} of British-born, genetically British UKB participants in bins of increasing Index of Multiple Deprivation of home address.

4.3.1.2 Phenotypes associated with cousin marriage in UKB

Confounding by parental phenotypes can be directly explored in 74,808 UKB couples for whom estimates of kinship coefficient are available (See Methods section 4.2.2). For each of 155 significant UKB traits, the mean phenotype of each couple was regressed on the couple's kinship coefficient. The t -statistics from this couple phenotype regression are compared to the t -statistics of $\hat{\beta}_{F_{ROH}}$ in Figure 53

Only one trait (alcohol drinking status) has a significant association with couple kinship (higher kinship associated with less alcohol), and there is no significant trend across all 155 traits (p -value = 0.75).

In contrast, in a subset of 6 traits likely associated with confounding (alcohol drinker, pork and lamb consumption, religious group, education attainment and Townsend deprivation index), there is a significant association between the two t -statistics (p -value = 0.0006). The gradient of the best-fit line for this subset of traits is approximately twice that expected for equal effect sizes i.e., high kinship couples are approximately twice as likely to never drink alcohol as a subject with F_{ROH} equal to the couple's kinship. It therefore seems likely that the associations between these 6 confounder traits and F_{ROH} are caused by social confounding and that the

environmental heritability of this confounding is approximately 50%. For the remaining 149 traits, there is no evidence that the parental phenotypes of high kinship couples differ from those of unrelated couples.

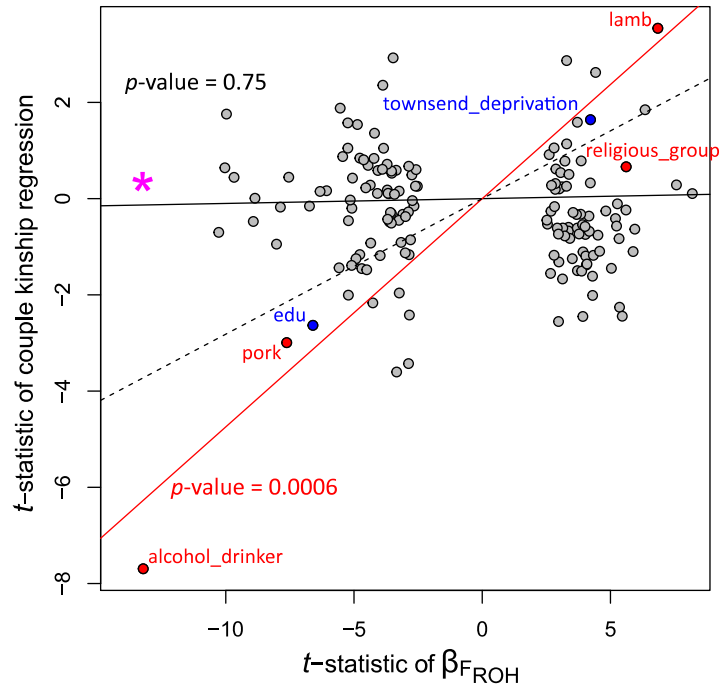


Figure 53: **t -statistic of the regression of mean couple phenotype on couple kinship plotted against the t -statistic of $\hat{\beta}_{F_{ROH}}$, for 155 significant traits in UK Biobank.** A least-squares best fit line through all traits is shown in black and a least-squares best fit line through the suspected confounders is shown in red. The expected gradient for equal effect estimates is shown in dashed black. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

4.3.1.3 Effects of F_{ROH} resemble the effects of known confounders

The phenotypic effects of possible confounders were explored by the method described in section 4.2.3.

In UKB the effect of Educational Attainment (EA) was found to be directionally consistent with $\hat{\beta}_{F_{ROH}}$ for 122 out 154 traits (Figure 54). Traits that are directionally inconsistent include: neb, ever had children, age at first sex, ever smoked and religious group participation. For example, increased F_{ROH} is associated with reduced reproductive success while reduced EA is associated with increased reproductive success.

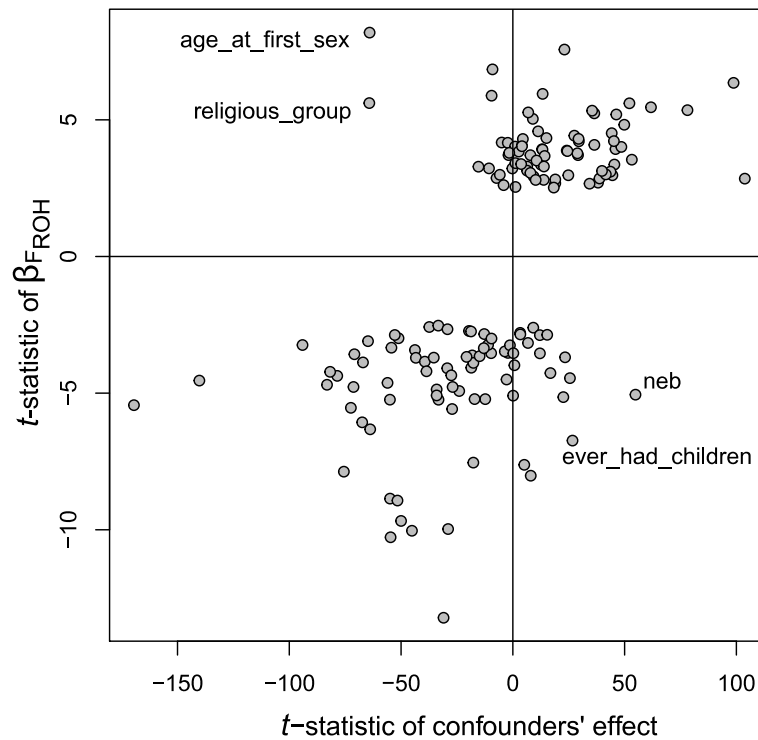


Figure 54: **t-statistic of $\hat{\beta}_{F_{ROH}}$ plotted against the t-statistic of the effect of Educational Attainment, for 154 significant traits in UKB.**

In addition to Educational Attainment, UKB contains other potential markers of SES. These include: Townsend deprivation index at home address, Index of Multiple Deprivation (IMD) at home and birth addresses, Maternal smoking at birth, Fathers age at birth, Father's current age, Father still alive, Mother's age at birth, Mother's current age, Mother still alive, Number of siblings, and Breastfed as a baby.

The combined effect of these 13 components of SES (including EA) was estimated on the remaining 153 significant UKB traits and plotted against the t -statistics of $\hat{\beta}_{F_{ROH}}$ in Figure 55.

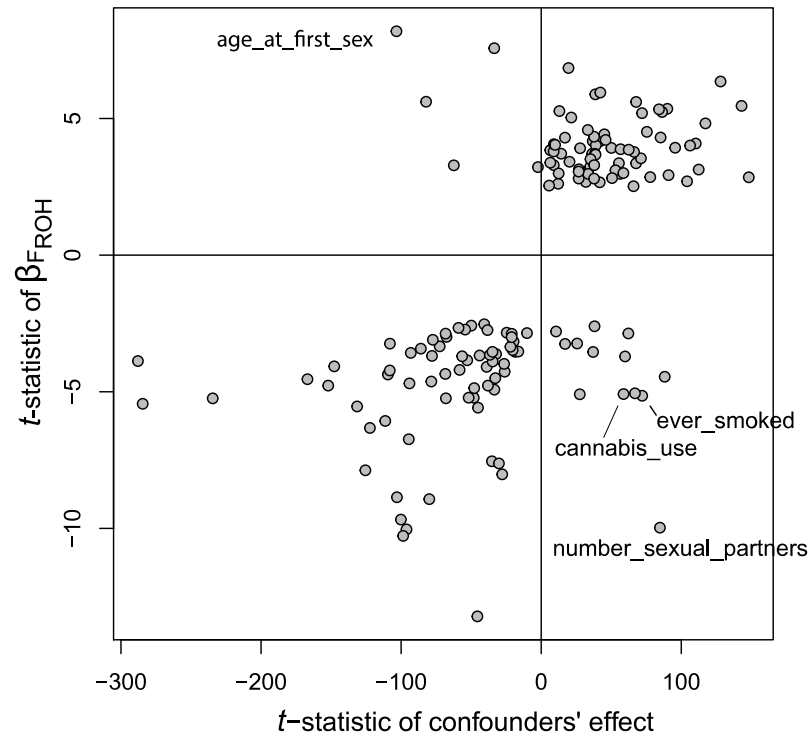


Figure 55: t -statistic of $\hat{\beta}_{F_{ROH}}$ plotted against the t -statistic of the combined effect of 13 components of SES, for 151 traits in UKB.

Taking account of these additional measures of SES reduces the number of discordant effects from 32 down to 18. Notably, the effect of SES is now consistent with the effect of F_{ROH} for the trait *ever_had_children* – i.e., reduced SES is associated with reduced reproductive success. This is partly explained by a strong association between increased Townsend Deprivation Index and childlessness – i.e., UKB participants living in economically deprived neighbourhoods are less likely to have children, although the direction of any causality is unknown.

Another suspected confounder of F_{ROH} , religious or cultural beliefs, is less well described in UKB, although several phenotypes can be considered as proxies: Participation in a religious group, alcohol drinker, and pork and lamb consumption. The result of adding these 4 traits to the 13 components of SES is shown in Figure 56.

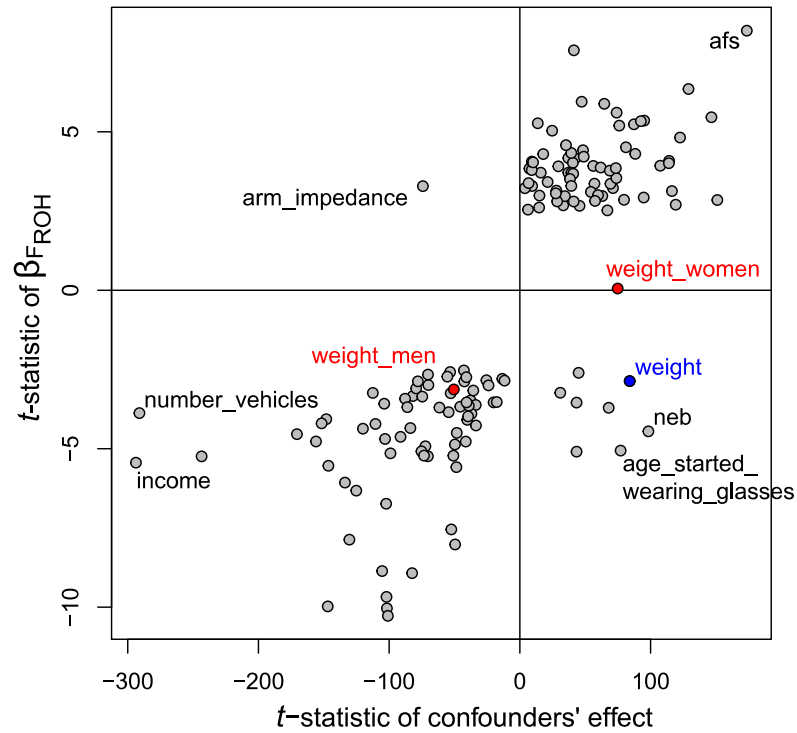


Figure 56: t -statistic of $\hat{\beta}_{F_{ROH}}$ plotted against the t -statistic of the combined effect of 17 potential confounders, covering components of SES and religious beliefs, for 149 significant traits in UKB. The effects on weight are also shown separately for males and females in red.

Accounting for these four proxies for religious beliefs reduces the number of directionally discordant traits from 18 to 9. Notably, the effect of these confounders on *age_at_first_sex*, *number_sexual_partners* and *ever_smoked* is now consistent with the effect of F_{ROH} – this is largely explained by strong associations between *religious_group* and these traits.

Participant weight is among the remaining 9 discordant traits – i.e., F_{ROH} is associated with reduced weight while reduced SES is associated with increased weight. However, if weight is split into two separate phenotypes, for men and women, the effect of SES is now concordant with the effects of F_{ROH} for both traits. I.e., in women, F_{ROH} and reduced SES both increase weight while in men the same changes reduce weight.

4.3.2 Accounting for confounders in $\beta_{F_{ROH}}$

4.3.2.1 Consistency of effect estimates between different populations

4.3.2.1.1 Stratifying ROHgen by ancestry

ROHgen cohorts were allocated to one of seven ancestral groups (European, S&W Asian, African, East Asian, Hispanic, Japanese, and Mixed) and $\beta_{F_{ROH}}$ was estimated by meta-analysis within these groups for all traits (Supplementary Table 16).

Although there is some heterogeneity, the effect estimates are mostly concordant, as shown for height in Figure 57, and for all ROHgen traits in Figure 58.

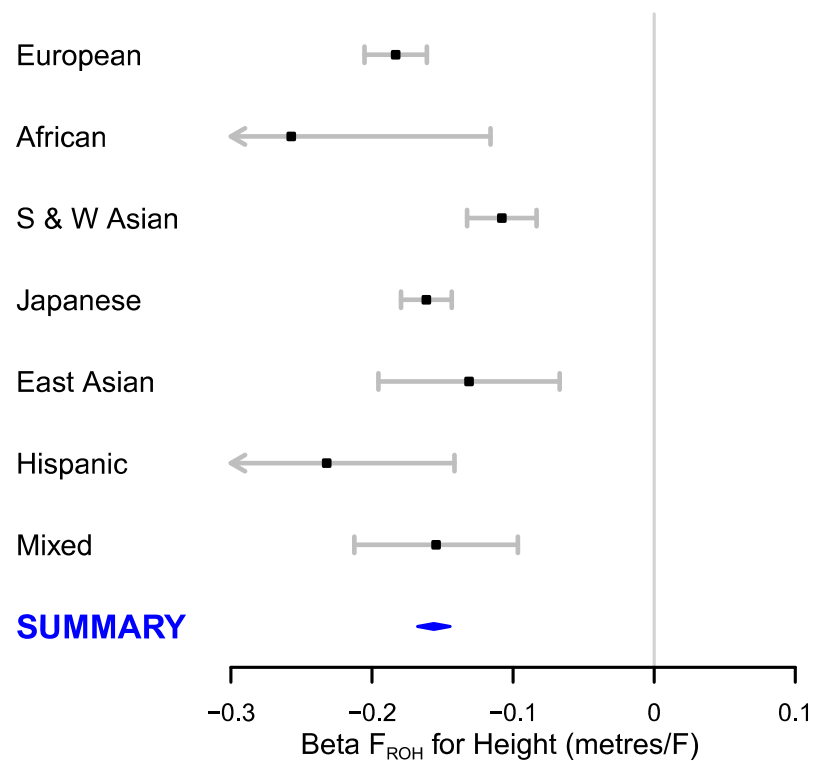


Figure 57: **Effect of F_{ROH} on height is robust to stratification by ancestral group.**

ROHgen cohorts were divided into seven broad ancestral groups and meta-analysed separately. Although some heterogeneity is observed (heterogeneity p -value = 3×10^{-4}), $\hat{\beta}_{F_{ROH}}$ is directionally consistent and differs significantly from 0 in all ancestral groups. All errors bars represent 95% confidence intervals.

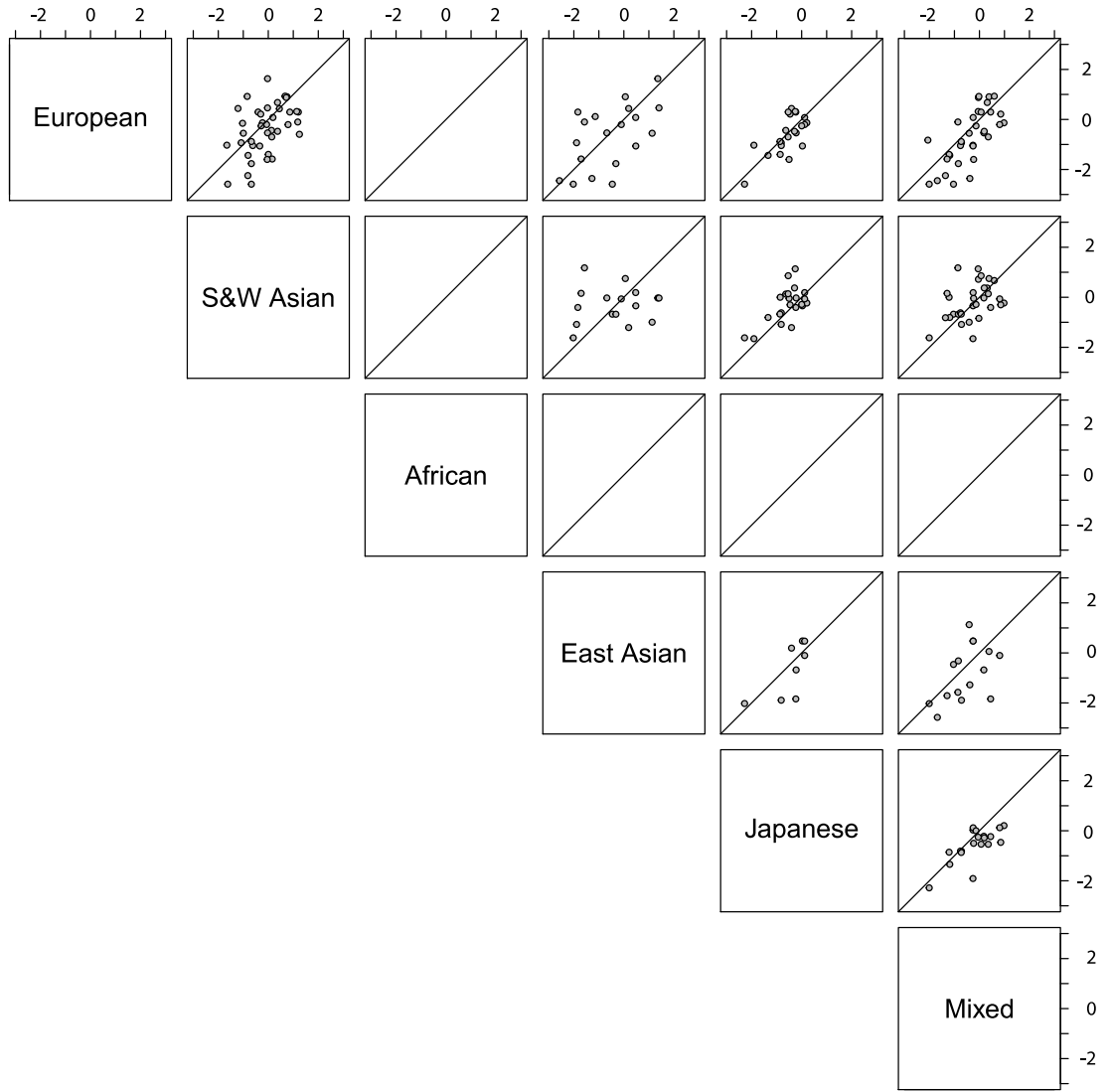


Figure 58: **Comparison of $\hat{\beta}_{F_{ROH}}$ for all ROHgen traits in cohort groups with different ancestral origins.** All $\hat{\beta}_{F_{ROH}}$ are normalised by within-sex trait standard deviations and only traits with estimation errors of less than 1 on this scale, in both cohorts, are shown.

4.3.2.1.2 Stratifying ROHgen by demographic origin of ROH

Similarly, ROHgen cohorts were allocated to one of four inferred demographic origins as determined by their location on a plot of $\bar{F}_{ROH}$ versus F_{IS} (Figure 59). Again, estimates of $\beta_{F_{ROH}}$ for show good consistency, both for height (Figure 60) and more generally for all ROHgen traits (Figure 61).

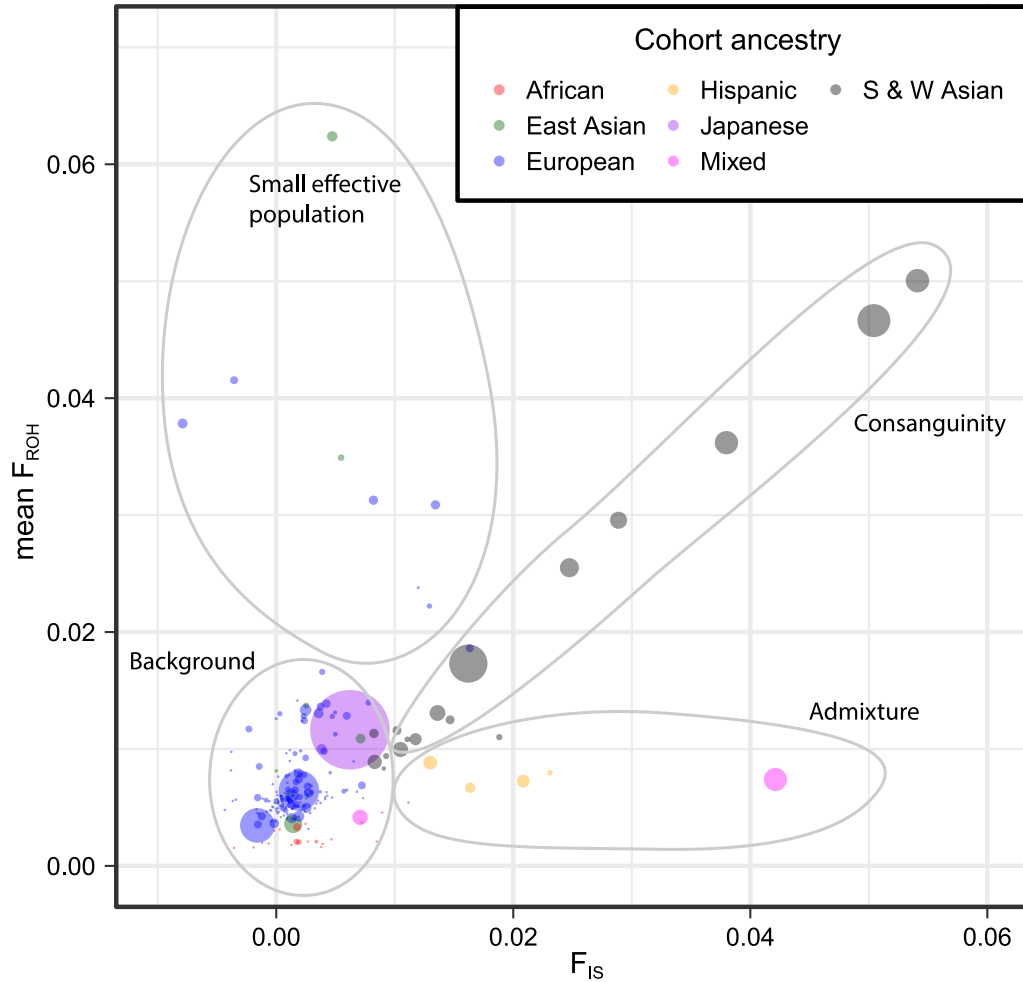


Figure 59: **Assignment of cohorts to one of four inferred demographic histories.** Figure 9 is replicated (see also Figure 9 legend) and used to empirically assign cohorts to one of four inferred demographic histories. In cohorts where $F_{IS} > 0.1$, but the Cartesian distance to the 1:1 line was < 0.005 , *consanguinity* was inferred to be the main origin of ROH. Cohorts which had not been defined as consanguineous but had mean $F_{ROH} > 0.02$ were considered to have a *small effective population*. Cohorts with $F_{IS} > 0.1$, but not consanguineous nor small effective population, were defined as *admixed* and the remaining cohorts were described as *background*.

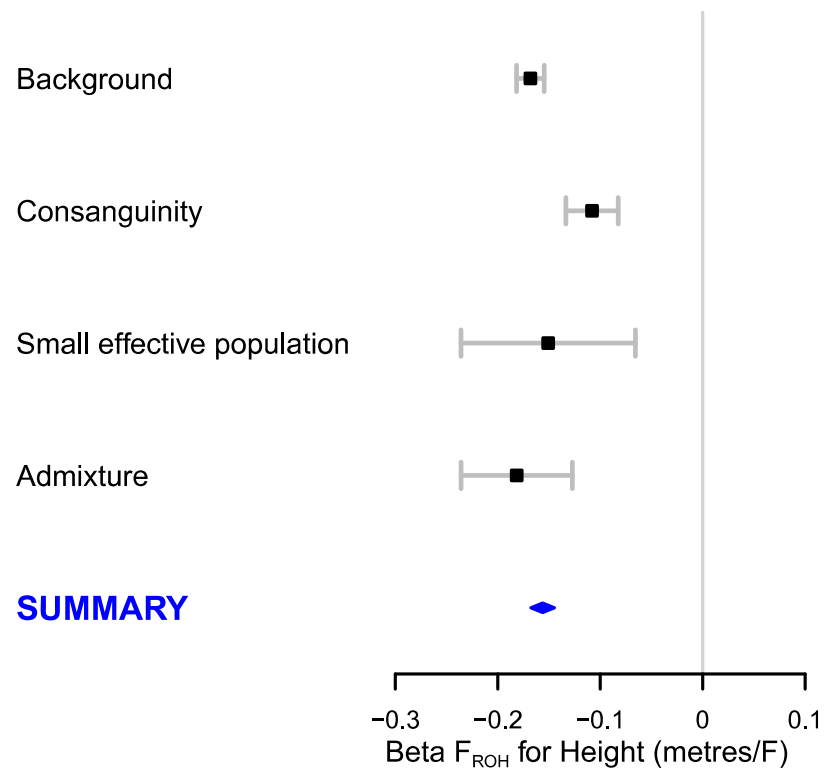


Figure 60: **Effect of F_{ROH} on height is robust to stratification by inferred demographic history.** Cohorts were divided by inferred demographic history (Figure 59, Supplementary Table 1) and meta-analysed separately. A small amount of heterogeneity is observed (heterogeneity p -value = 0.008), but $\hat{\beta}_{F_{ROH}}$ is directionally consistent and differs significantly from 0 in all groups. In particular, in the small effective population size cohorts, where the variation of F_{ROH} is believed to be caused by variations in cryptic relatedness between parents, $\hat{\beta}_{F_{ROH}}$ [-0.15, 95% CI -0.07 -0.23, p -value 3×10^{-4}] is consistent with the global estimate. All errors bars represent 95% confidence intervals.

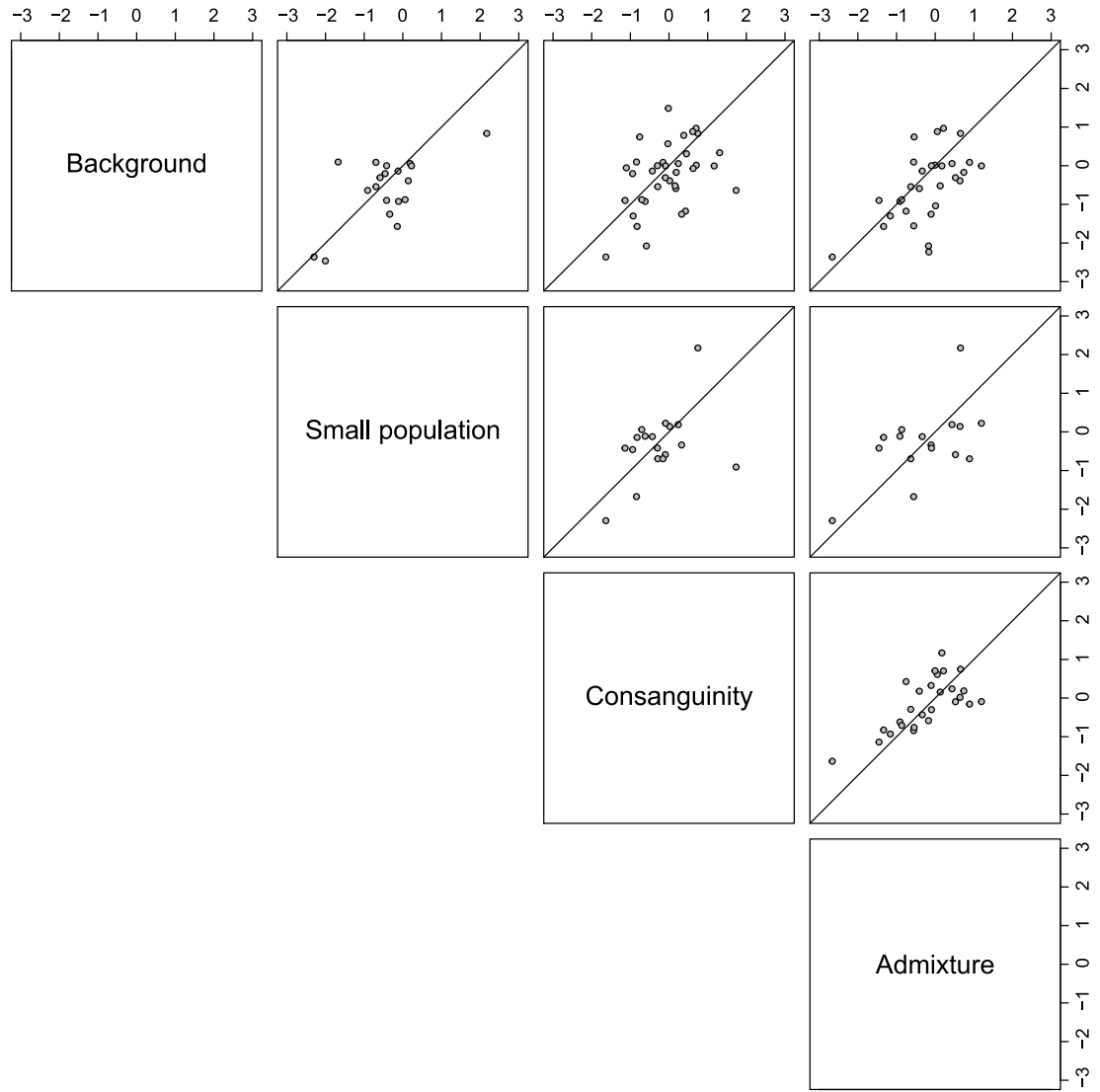


Figure 61: **Comparison of $\hat{\beta}_{F_{ROH}}$ for all ROHgen traits meta-analysed in groups of different ROH origin.** All $\hat{\beta}_{F_{ROH}}$ are normalised by within-sex trait standard deviations and only traits with estimation errors of less than 1 on this scale, in both cohorts, are show.

4.3.2.1.3 Stratifying UK Biobank by genetic ethnicity

In the UK Biobank $\beta_{F_{ROH}}$ was estimated for all 155 significant traits, in both the genetically British and genetically non-British cohorts. Comparing the t -statistics of these two estimates shows that there is generally good agreement, albeit with several prominent outliers (Figure 62). Of the 155 traits, only 7 have significantly differing estimates between the two groups, and the three of these outliers (lamb, pork and alcohol consumption) were previously suspected of being confounded by religious or cultural beliefs. Unexpectedly, the effect of F_{ROH} on “number of children even born” (neb) both differs significantly, and is directionally inconsistent, between the two groups. Increased F_{ROH} is associated with reduced neb in the British cohort, but not in the non-British.

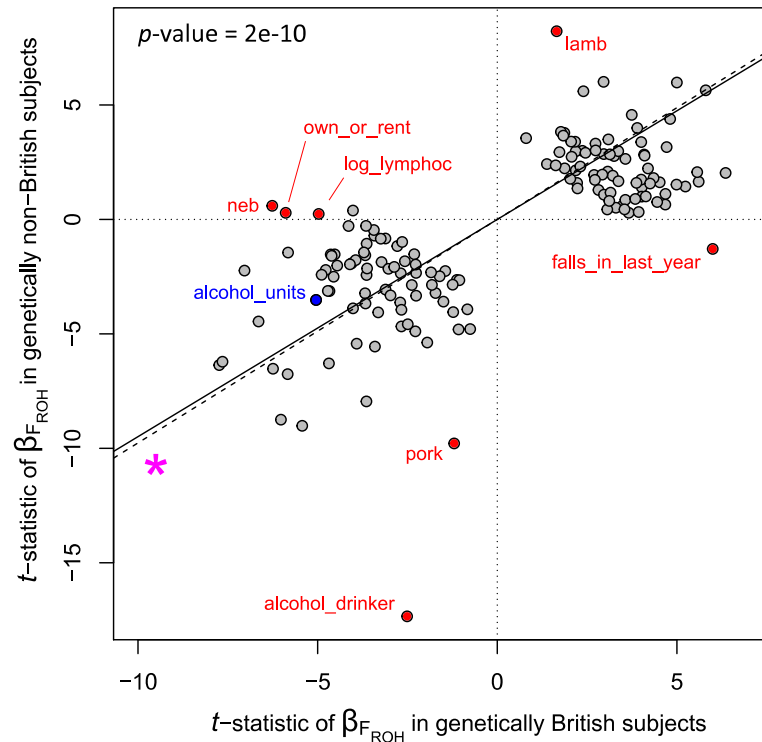


Figure 62: t -statistic of $\hat{\beta}_{F_{ROH}}$ in non-British subjects plotted against the t -statistic of $\hat{\beta}_{F_{ROH}}$ in British subjects, for 155 significant UK Biobank traits. Significant outliers are coloured red. Alcohol units per week in drinkers is shown in blue, as it is referred to in section 4.4.2 The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

4.3.2.2 Proportionality of effects across the range of observed F_{ROH}

In ROHgen, average trait values were calculated for all traits in bins of increasing F_{ROH} (See Methods section 4.2.4.2). Results for all traits are given in Supplementary Table 18 and are plotted for *height* and *ever_had_children* in Figure 63.

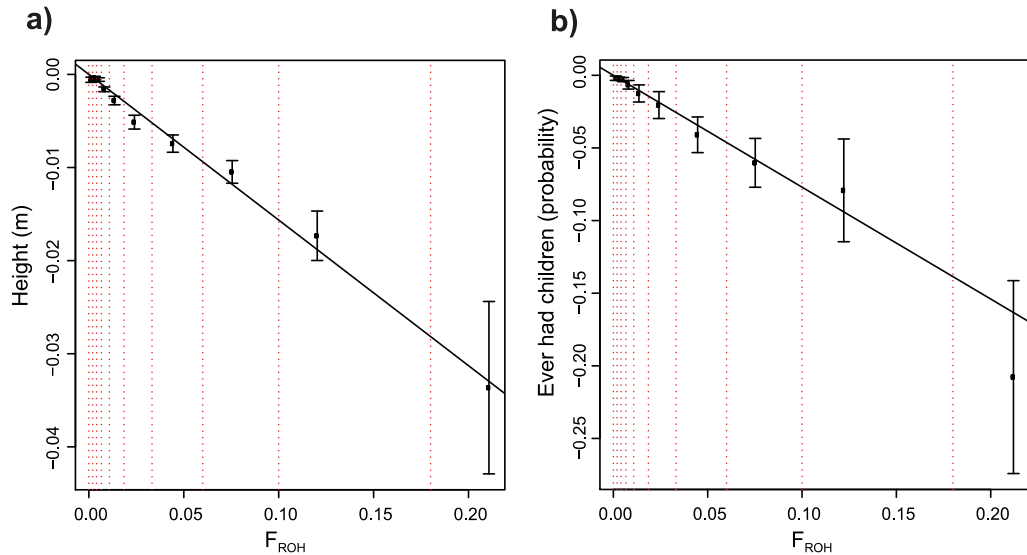


Figure 63: Height and reproductive success plotted in bins of increasing F_{ROH} . (a) Linear decrease in height with increasing F_{ROH} . Average height (in metres) is plotted in bins of increasing F_{ROH} . The limits of each bin are shown by red dotted lines, and correspond to the offspring of increasing degree unions left-to-right. The overall estimate of $\beta_{F_{ROH}}$ is shown as a solid black line. (b) Linear decrease in the fraction of subjects having children with increasing F_{ROH} .

Both average height and likelihood of having children reduced linearly over a wide range of F_{ROH} . Even within the highest F_{ROH} groups, subjects with kinship equal to the offspring of full-sibling or parent-child unions are significantly shorter and less likely to have children than those of avuncular or half-sibling unions, who in turn are significantly shorter than those of first-cousin unions.

The magnitudes of these effects were compared to the effects of Townsend deprivation (Figure 64). Townsend deprivation is associated with reduced height and likelihood of having children, but even in the most deprived neighbourhoods, the average height and the fraction of the population having children are substantially higher than in the offspring of first-degree relatives (shown by a dotted red line).

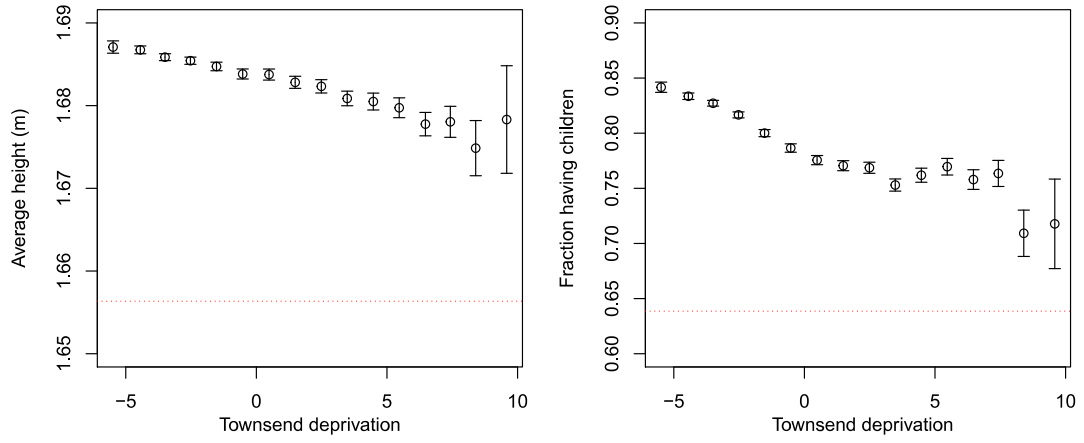


Figure 64: **Height and fraction of subjects having children plotted against Townsend Deprivation Index.** a) Average height of UKB subjects in bins of increasing Townsend deprivation. b) Average fraction of UKB subjects having children in bins of increasing Townsend deprivation. The corresponding averages in very high F_{ROH} samples (>0.175) are shown as dotted red lines.

4.3.2.3 Effect estimates within full sibling families ($\beta_{F_{ROH_wSibs}}$)

The effect of F_{ROH} within full siblings is believed to be free from almost all confounding and therefore gives an unbiased estimate of the genetic effects of autozygosity. In ROHgen a total of 150,568 subjects with at least one genotyped sibling were identified across 28 cohorts and $\beta_{F_{ROH_wSibs}}$ estimated for all available phenotypes (see Methods section 4.2.4.3). These estimates have larger standard errors than $\hat{\beta}_{F_{ROH}}$, (due to the smaller sample size and reduced variance of the independent variable), and only for one trait (ever had children) does $\hat{\beta}_{F_{ROH_wSibs}}$ differ significantly from zero (p -value = 0.003). However, meta-analysing the ratio $\hat{\beta}_{F_{ROH_wSibs}}:\hat{\beta}_{F_{ROH}}$ across all ROHgen traits (Figure 65) shows that, on average, the within-sibling estimates are a substantial fraction of the whole-population estimates - specifically 74% [95% CI 40% to 110%].

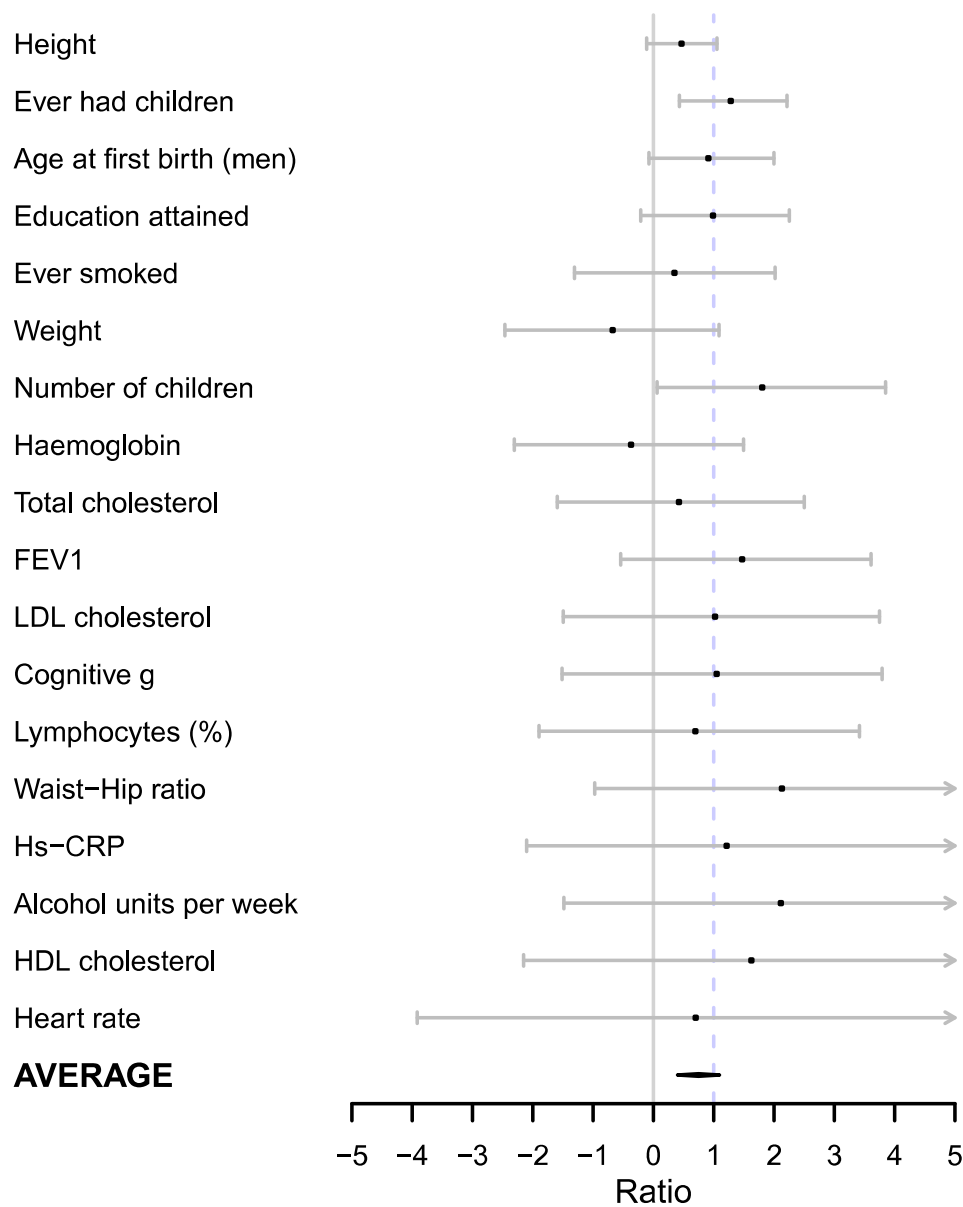


Figure 65: **The ratio of $\hat{\beta}_{F_{ROH_wSibs}} : \hat{\beta}_{F_{ROH}}$ for all 18 significant ROHgen traits.** 95% confidence intervals determined by permutation are shown in grey. A ratio of 1 would be expected of entirely genetic effects, while a ratio of 0 would suggest complete confounding. One trait (Ever had children) has a ratio which differs significantly from 0, and a meta-analysis across all traits has an average value of 0.74 [95% CI 0.40 to 1.10] which also differs significantly from 0 (p -value 1×10^{-5}), but not from 1.

In UKB, 41,798 individuals were identified with at least one genotyped sibling and the effect of F_{ROH} within siblings ($\beta_{F_{ROH_wSibs}}$) was estimated for all 155 significant UKB traits (Supplementary Table 31). Again, the standard errors of $\hat{\beta}_{F_{ROH_wSibs}}$ are large and no single trait reaches 5% Bonferroni corrected significance. However,

meta-analysing the ratio $\hat{\beta}_{F_{ROH_wSibs}}:\hat{\beta}_{F_{ROH}}$ across all traits gives an average value of 0.92 [95% CI 0.29 to 1.55, p -value = 0.002]. Note, to ensure statistical independence of the sibling and population estimates, the denominator ($\hat{\beta}_{F_{ROH}}$) is estimated in a separate cohort of 446,490 of samples without any genotyped siblings. A forest plot in the style of Figure 65 is difficult to read for 155 traits, so the within-sibling results are shown as a TT plot (Figure 66).

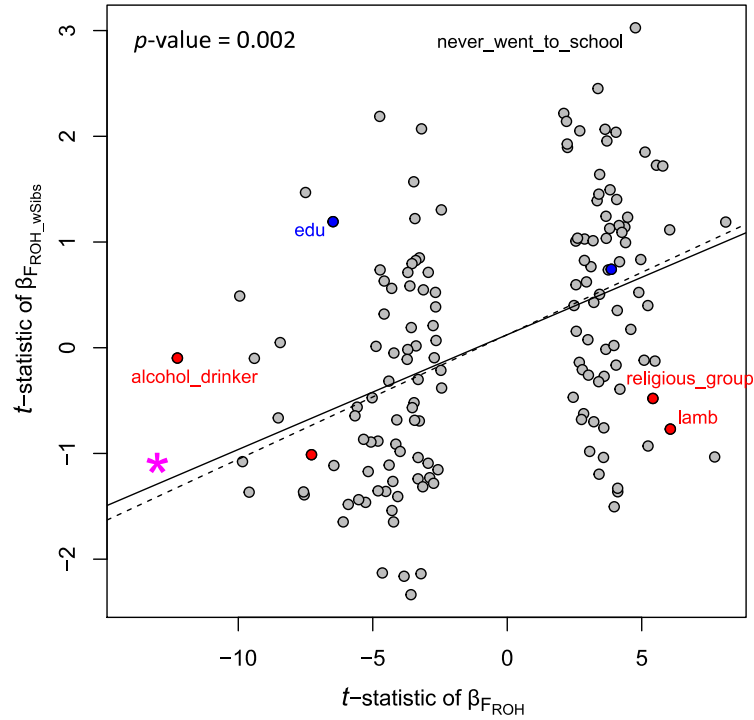


Figure 66: t -statistic of $\hat{\beta}_{F_{ROH_wSibs}}$ estimated within 41,798 full siblings plotted against the t -statistic of $\hat{\beta}_{F_{ROH}}$ estimated in a separate cohort of 446,490 samples, for 155 significant UKB traits. A Deming regression best-fit line is shown in solid black and has a gradient which differs significantly from zero (p -value 0.002). The expected regression gradient for equal effect sizes is shown in dashed black. Two traits possibly associated with socio-economic confounding (Educational attainment and Townsend deprivation index of home address) are shown in blue, and four traits likely associated with religious confounding (Alcohol drinker status, pork consumption, lamb consumption and participation in a religious group) are shown in red. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

4.3.2.3.1 Multivariate trait

To summarize the effects of F_{ROH} in a single, well-powered “trait” a multivariate prediction of F_{ROH} was calculated in two large cohorts (see Methods section 3.2.7). In UKB, $\hat{\beta}_{F_{ROH_wSibs}}$ for this *multivariate* trait is 102% [95% CI 10% – 193%, p -value = 0.029] of the whole population estimate. In deCODE-Iceland this figure is 56% [95% CI 25% – 87%, p -value = 4×10^{-4}]. Although these two *multivariate* traits are derived from different phenotypes (145 in UKB and 14 in deCODE-Iceland), meta-analysing gives an estimate of 60% [95% CI 32% to 88%, p -value = 2×10^{-5}].

4.3.2.4 Effect estimates in UPD cases ($\beta_{F_{ROH_wUPD}}$)

Uniparental disomy (UPD) segments were deduced in 200 subjects by identifying long ROH unlikely to have been caused by parental relatedness. As expected of a random event, the F_{ROH} of these subjects was not associated with any known confounders (Figure 87) - in contrast to non-UPD subjects with similar F_{ROH} (Figure 88).

The effect of F_{ROH} in UPD cases ($\beta_{F_{ROH_wUPD}}$) was estimated for 155 significant UKB traits (Supplementary Table 32) by fitting phenotype models with interactions between F_{ROH} and UPD status ($F_{ROH}: upd + F_{ROH}: not_upd$). Despite the small number of UPD cases, one trait (Ever had children) has a significant association with UPD F_{ROH} at Bonferroni corrected significance of 5%. Furthermore, the t -statistics of UPD F_{ROH} are significantly correlated with the t -statistics of non-UPD F_{ROH} (p -value = 5×10^{-8} , Figure 67). Meta-analysing across all traits, the average effect of UPD F_{ROH} is 1.43 times greater [95% CI 0.90 to 1.96] than the effect of non-UPD F_{ROH} .

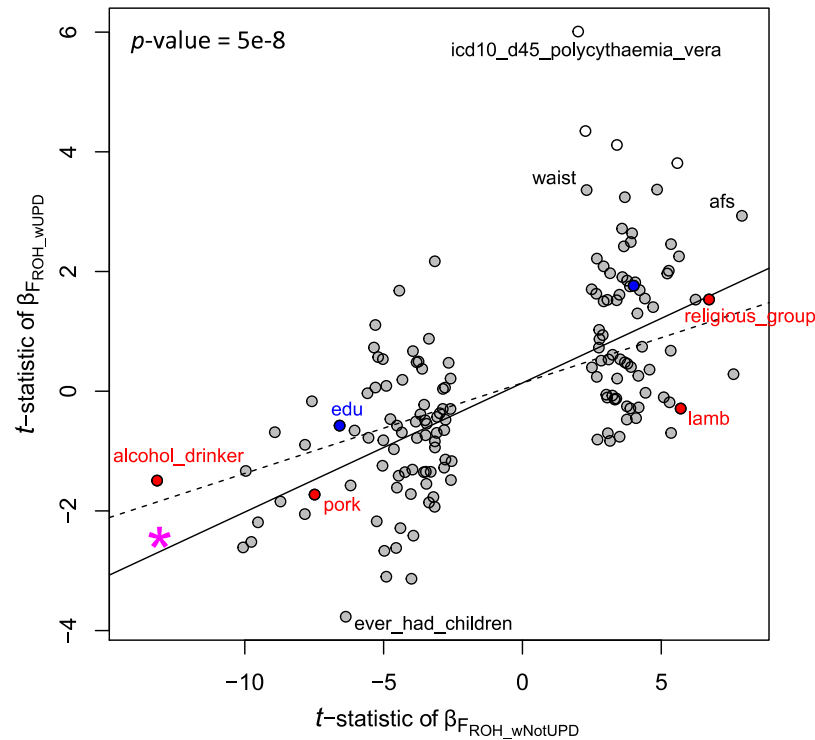


Figure 67: **t -statistic of the effect of UPD F_{ROH} plotted against the t -statistic of non-UPD F_{ROH} , for 155 significant UKB traits.** A Deming regression best-fit line is shown in solid black and has a gradient which differs significantly from zero (p -value = 5×10^{-8}). The expected regression gradient for equal effect sizes is shown in dashed black. Four haematology traits shown as empty circles were left out of the regression due to concerns about the direction of causality (see Methods section 4.2.4.4.1). Two traits possibly associated with socio-economic confounding (Educational attainment and Townsend deprivation index of home address) are shown in blue, and four traits likely associated with religious confounding (Alcohol drinker status, pork consumption, lamb consumption and participation in a religious group) are shown in red. The *multivariate* trait (not included in the regression) is shown as $\star$ but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

4.3.2.5 Effect of F_{ROH} conditioning on F_{ROH_cm}

The variation of F_{ROH} which is independent of F_{ROH_cm} is believed to be free from confounding associated with parental relatedness. The effect of F_{ROH} conditioning on F_{ROH_cm} ($\beta_{F_{ROH}|F_{ROH_cm}}$) was therefore estimated in bivariate models for 155 significant UKB traits (Supplementary Table 33 and Figure 68). For one trait (age at first sex) the conditional effect of F_{ROH} reaches 5% Bonferroni corrected significance and across all traits there is a significant association (p -value = 0.0016) between the

t -statistics of $\beta_{F_{ROH}|F_{ROH_CM}}$ and $\beta_{F_{ROH_CM}}$. Meta-analysing all traits the average ratio of $\hat{\beta}_{F_{ROH}|F_{ROH_CM}}:\hat{\beta}_{F_{ROH_CM}}$ is 0.79 [95% CI 0.12 – 1.46].

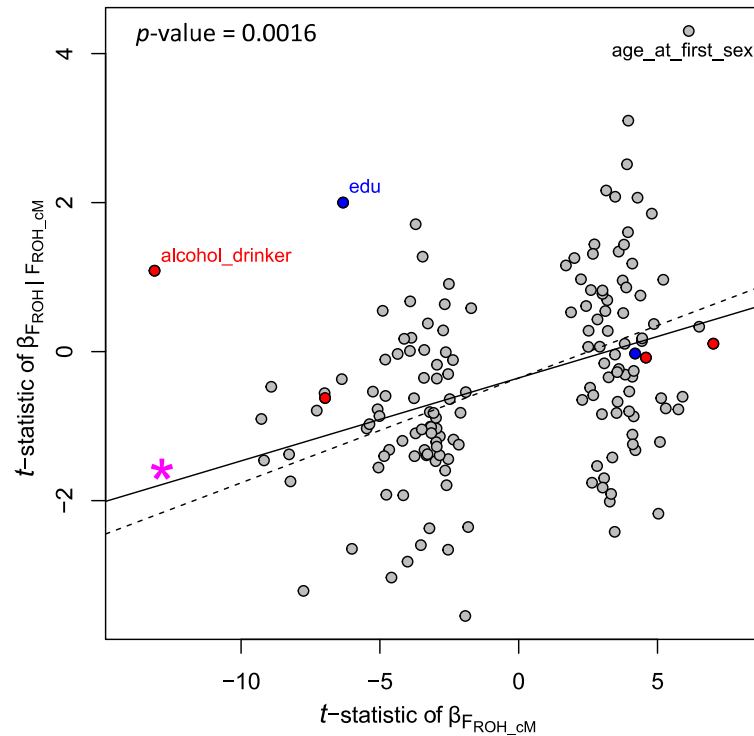


Figure 68: t -statistic of $\hat{\beta}_{F_{ROH}|F_{ROH_CM}}$ plotted against the t -statistic of $\hat{\beta}_{F_{ROH_CM}}$ for 155 significant UKB traits. A Deming regression best-fit line shown in solid black has a gradient which differs significantly from zero (p -value = 0.0016). The expected regression gradient for equal effect sizes is shown in dashed black. Two traits possibly associated with socio-economic confounding (Educational attainment and Townsend deprivation index of home address) are shown in blue, and four traits likely associated with religious confounding (Alcohol drinker status, pork consumption, lamb consumption and participation in a religious group) are shown in red. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

4.3.2.6 Effect estimates in adopted subjects ($\beta_{F_{ROH_wAdopted}}$)

$\beta_{F_{ROH}}$ was estimated for 155 significant UKB traits in a cohort of 7207 adopted subjects, and compared to the corresponding effects in 479,104 non-adopted subjects (Figure 69 and Supplementary Table 34). In the adopted cohort six traits (fluid intelligence, addition test, word interpolation, number of sexual partners, volume set in hearing test and paraplegia) reach 5% Bonferroni corrected significance – all in a direction consistent with the non-adopted cohort. In addition,

there is a strong association between the t -statistics of $\hat{\beta}_{F_{ROH_{wAdopted}}}$ and $\hat{\beta}_{F_{ROH_{wNotAdopted}}}$ (p -value = 2×10^{-12} , Figure 69), and meta-analysing across all traits the average ratio of $\beta_{F_{ROH_{wAdopted}}} : \beta_{F_{ROH_{wNotAdopted}}}$ is 0.95 [95% CI 0.68 – 1.21].

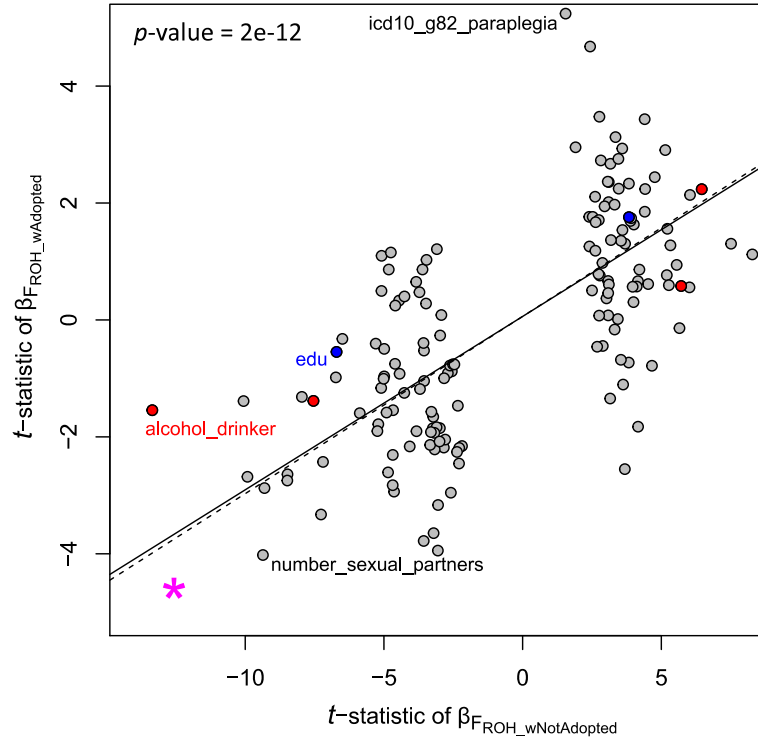


Figure 69: t -statistic of $\hat{\beta}_{F_{ROH}}$ in a cohort of 7207 adopted samples plotted against the t -statistic of $\hat{\beta}_{F_{ROH}}$ in 479,104 non-adopted samples, for 155 significant UKB traits. A Deming regression best-fit line shown in solid black has a gradient which differs significantly from zero (p -value = 2×10^{-12}). The expected regression gradient for equal effect sizes is shown in dashed black. Two traits possibly associated with socio-economic confounding (Educational attainment and Townsend deprivation index of home address) are shown in blue, and four traits likely associated with religious confounding (Alcohol drinker status, pork consumption, lamb consumption and participation in a religious group) are shown in red. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

4.3.3 Comparing $\beta_{F_{ROH}}$ to effect of hemizygous deletions

A full description of the medical consequences of the hemizygous deletions identified in UKB is beyond the scope of this thesis, and has already been addressed by published work^{88,89}. Never-the-less, it is interesting to note that, of the 1382 subjects who carry a deletion, only 9% receive disability benefits and 40% report

having a longstanding illness (compared to 6% and 34% in the full UKB cohort). Surprisingly, not one of 1382 deletion cases has a hospital ICD10 code (Q9) that would indicate a genetic abnormality (compared to 0.03% in the full cohort). This is surprising as almost two-thirds of deletions (919 out of 1448) occur at one of the 47 genomic loci deleted in 5 or more samples, and a large majority of these repeatedly deleted loci are associated with recognised microdeletion syndromes⁸⁹.

Characterizing the less commonly deleted loci is limited by small sample sizes, but carrying a relatively rare deletion (seen in less than 5 samples) is similarly detrimental to Educational attainment (beta = -0.66 years; se = 0.15) to carrying a relatively more common deletion (beta = -0.88 years; se=0.15). For the *multivariate* trait (which summarises the effect of F_{ROH} across 145 traits), rare deletions have a greater effect (beta = -0.051; se = 0.004) than common deletions (beta = -0.029; se = 0.005). In summary, a majority of the deletions identified in UKB are likely to have substantial phenotypic consequences similar to recognized microdeletion syndromes - although they seem to be surprisingly underdiagnosed in this cohort.

The phenotypic effects of hemizygous deletions are typically attributed to haploinsufficiency (where the level of a gene product is insufficient to permit normal cell function) and thus have a different causal mechanism to that hypothesized for ROH. Never-the-less, hemizygous deletions have an indisputably genetic effect and do cause a loss of heterozygosity analogous to autozygosity. Also, unlike ROH, deletions are thought to be a mostly random event and therefore their effects are not expected to be biased by confounders. It is therefore interesting to compare the effects of hemizygous deletion to the effects of F_{ROH} .

For all UKB participants the deleted genomic fraction (F_{DEL}) was calculated and fit in bivariate models with F_{ROH} for all 155 significant UKB traits (Supplementary Table 35 and Figure 70). The correspondence between $\hat{\beta}_{F_{DEL}}$ and $\hat{\beta}_{F_{ROH}}$ is striking, with only 8 estimates of $\beta_{F_{DEL}}$ discordant with the corresponding effect of $\beta_{F_{ROH}}$ – and none of these 8 estimates is significant. Unexpectedly, the most significant effects of F_{DEL} are on *fev1*, *grip_strength*, *pef*, *ever_had_children* and *height* – all traits among those most affected by F_{ROH} . In addition, there are significant effects of F_{DEL} on many

sociological traits also affected by F_{ROH} (age at first sex, risk tolerance, motorway speeding, Townsend deprivation index, etc.). Although this does not prove that the effects of F_{ROH} are genetic, it does demonstrate that traits which could very plausibly be confounded by social factors can also be substantially affected by non-additive genetics. The effect of F_{DEL} on the *multivariate* trait is 64.2 times greater [95% CI 56.5 to 72.0] than the effect of F_{ROH} , so although qualitatively similar, hemizygous deletions have a much greater average effect per Mb than ROH.

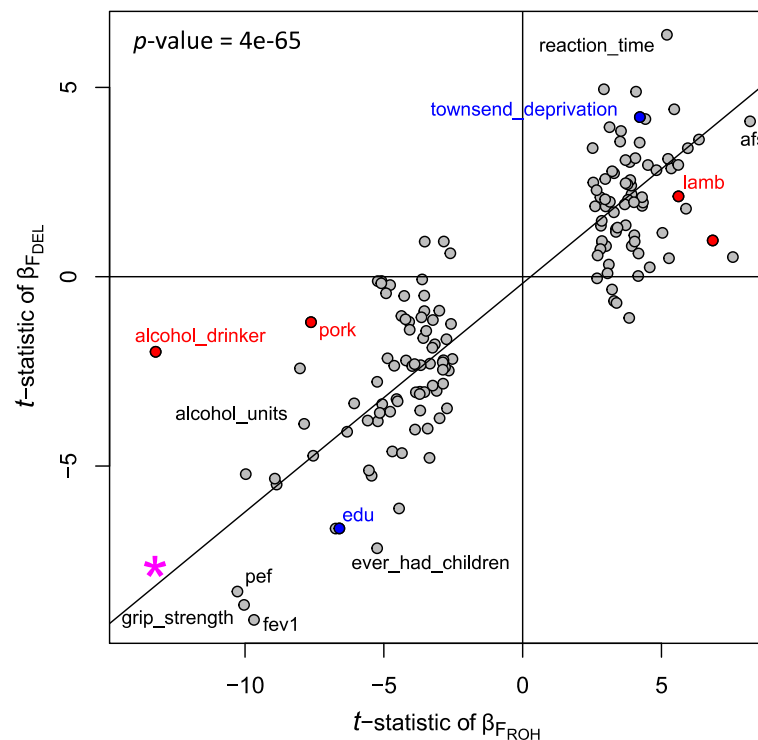


Figure 70: t -statistic of $\hat{\beta}_{F_{DEL}}$ plotted against the t -statistic of $\hat{\beta}_{F_{ROH}}$, for 155 significant UKB traits. A Deming regression best-fit line is shown in solid black and has a gradient which differs significantly from zero (p -value = 4×10^{-65}). Two traits possibly associated with socio-economic confounding (Educational attainment and Townsend deprivation index of home address) are shown in blue, and four traits likely associated with religious confounding (Alcohol drinker status, pork consumption, lamb consumption and participation in a religious group) are shown in red. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

4.3.3.1 Comparing confounding-free estimates of $\beta_{F_{ROH}}$

The previous sections reported estimates of $\beta_{F_{ROH}}$ from four methods thought to reduce, or eliminate, any bias caused by social confounding. In addition, $\hat{\beta}_{F_{ROH}}$ were compared to the effects of hemizygous deletion, which may share some common causal mechanisms. These five methods differ in approach, but were applied to the same UKB cohort. Therefore, to compare results from the different methods, it was necessary to eliminate any correlation caused by sample overlap.

To compare $\hat{\beta}_{F_{ROH_wUPD}}$ to the other methods, all UPD cases were removed from the sibling analysis ($N_{UPD}=19$), the centimorgan analysis ($N_{UPD}=200$) and the adopted analysis ($N_{UPD}=0$). To compare $\hat{\beta}_{F_{ROH_wAdopted}}$ to the remaining other methods all adopted samples were removed from the sibling analysis ($N_{Adopted}=72$), and the centimorgan analysis ($N_{Adopted}=7207$). Finally, to compare $\hat{\beta}_{F_{ROH_wSibs}}$ and $\hat{\beta}_{F_{ROH}|F_{ROH_cM}}$ all siblings were removed from the centimorgan analysis ($N_{siblings}=41,798$). It was not thought necessary to modify the estimates of $\beta_{F_{DEL}}$ since deletions had already been removed from F_{ROH} , and there is no correlation between F_{ROH} and F_{DEL} (p -value = 0.43). With these exclusions, estimates from the five methods are thought to be entirely separate and independent.

Comparisons of the five methods are shown for 155 significant UKB traits (Figure 71). For 10 out of 12 combinations there is a nominally significant association between methods plotted on the x and y axes (p -values shown in the lower triangle). Differences in the strength of association largely reflect the statistical power of the five methods. The statistical power of each method is sample dependent but, in UKB, the deletions method has the highest average power (albeit to estimate $\beta_{F_{DEL}}$ not $\beta_{F_{ROH}}$), followed by the UPD method, Adopted method, the Centimorgan method and finally the Sibling method.

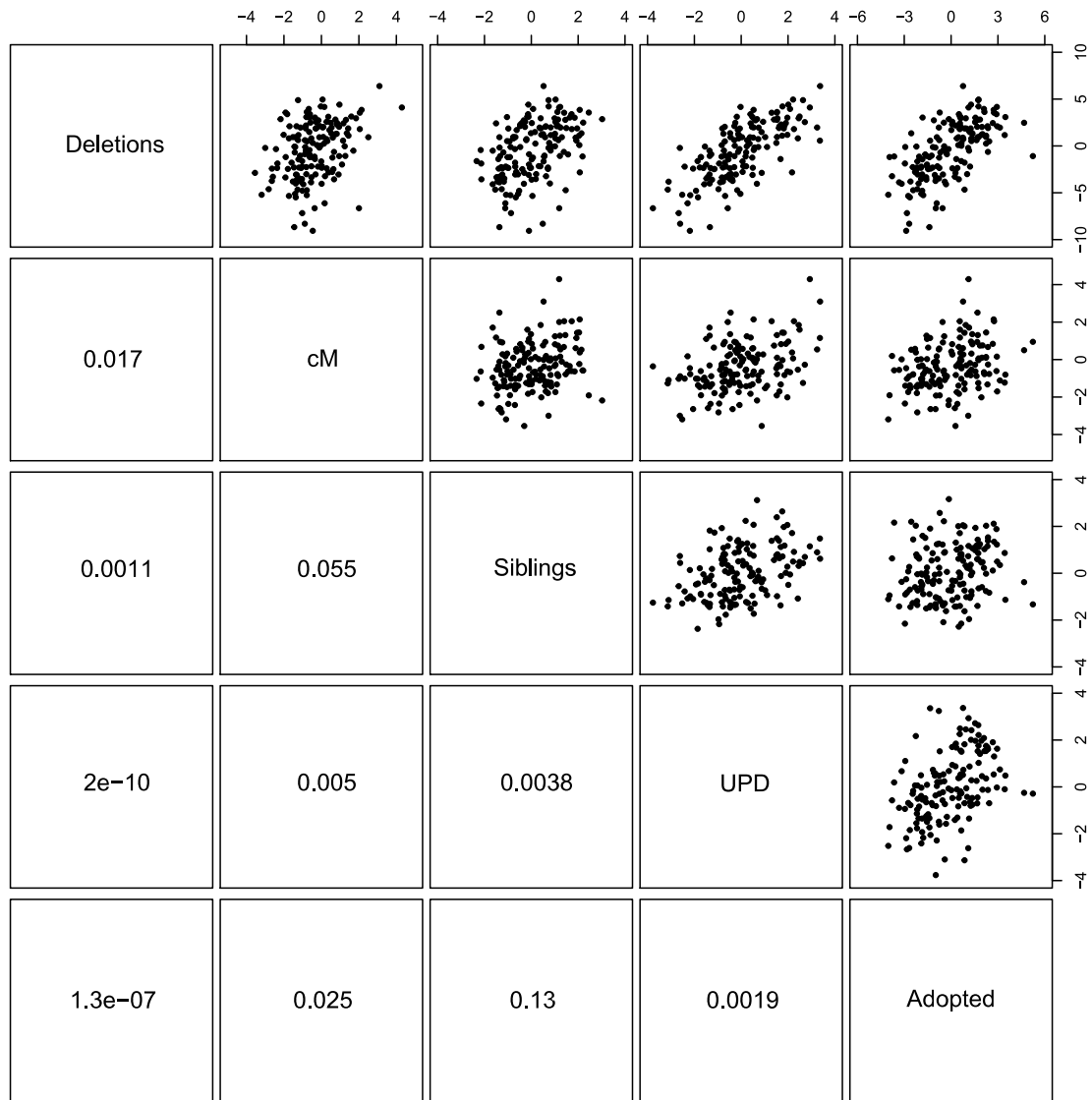


Figure 71: **Scatter plots of t -statistics from 4 independent and confounding-free methods of estimating $\beta_{F_{ROH}}$ and one method of estimating $\beta_{F_{DEL}}$.** The diagonal panes describe the methods plotted on the x and y axes of the corresponding scatter plot shown in the upper triangle. Each point represents estimates for one of 155 significant UKB traits. P -values of non-zero association between the x and y axis values are shown in the lower triangle. As in Figure 67, four haematology traits were excluded from the UPD analysis because of concerns about the direction of causality.

4.4 Discussion

4.4.1 Introduction

Confounding of F_{ROH} by unmeasured environmental factors has long frustrated the interpretation of putative genetic associations between F_{ROH} and human traits. This chapter attempts to resolve this issue: firstly, by investigating the relationship between F_{ROH} and possible confounders, and more generally by introducing methods which aim to estimate $\beta_{F_{ROH}}$ with reduced confounder bias.

4.4.2 Relationships between F_{ROH} and possible confounders

Several previous studies have reported associations between F_{ROH} and markers of SES, however these traits were measured in adult participants which makes it difficult to disentangle causal effects from confounding^{71,112,113}. Here, a measure of SES in adult participants (*IMD_home*) was found to be significantly associated with F_{ROH} , while an identically derived measure at birth (*IMD_birth*) was not (Figure 52). This suggests that socioeconomic deprivation develops throughout the lifetime of high F_{ROH} subjects – strongly implying causality.

In contradiction to this result, there is some weak evidence that cousin marriage is more common in less educated UKB subjects (see Figure 53). This is consistent with Abdellaoui et al. who report that F_{ROH} is confounded by education in a cohort of 2000 Dutch subjects¹¹². In UKB, as in Abdellaoui et al., the association between F_{ROH} and parental education is relatively stronger than the association between F_{ROH} and education of the subjects themselves - as expected when education is only partially heritable.

The reason for the apparent contradiction between *IMD_birth* and parental education in cousin marriages is not clear, although the association between cousin marriage and education (p -value = 0.007) does not reach experiment-wise significant, and thus may not be repeatable. Also, first-cousin marriage is only one source of the variation of F_{ROH} – the offspring of other relative types may have a different relationship with SES. Finally, the *IMD_birth* analysis considers only UK-

born, genetically British subjects, while the relationship between SES and F_{ROH} may be different in the non-British.

Educational attainment notwithstanding, the large majority of UKB traits show no evidence of confounding by parental phenotypes (see Figure 53). For example, there is no association between height and cousin marriage (p -value = 0.97), despite a strong relationship between height and F_{ROH} (p -value = 1×10^{-16}). One exception to this is alcohol drinking status – high kinship couples are less likely to drink alcohol – as are high F_{ROH} subjects. Given the prohibition of alcohol by some religions which encourage cousin marriage, this seems likely to reflect confounding by religious beliefs. Interestingly, the effect of F_{ROH} on whether a subject drinks alcohol is approximately half as strong as the effect of couple kinship on whether the couple drink alcohol – which implies a plausible 50% transmissibility of alcohol drinking. Consistent with explanation, the effect of F_{ROH} on alcohol drinking is much smaller in the genetically British (who presumably include fewer Muslims) than in the non-British (Figure 62). Interestingly however, the effect of F_{ROH} on weekly alcohol units consumed by alcohol drinkers is very significant in both groups (Figure 62). Other UKB traits which are probably associated with similar religious confounding (pork and lamb consumption) are also among those nominally associated with cousin marriage.

In summary, a small number of traits in UKB, directly relating to religious beliefs, do appear to be confounded with cousin marriage, and therefore with F_{ROH} . This is consistent with Abdellaoui et al. (*Behav Genet* 2013) where an association was observed between religious affiliation and F_{ROH} (in the Netherlands)¹¹³. There is also some weak evidence that cousin marriage is associated with reduced education in UKB, but this confounding does not appear to explain the associations between F_{ROH} and other traits. Also, in the UK-born, genetically British samples of UKB there is no evidence that F_{ROH} is associated with lower SES at birth.

4.4.3 Effects of F_{ROH} resemble effects of suspected confounders

Johnson et al. explained clearly that sociodemographic confounders may bias estimates of $\beta_{F_{ROH}}$, and that available covariates are unlikely to capture the full effects of complex confounders⁷¹. However, this thesis appears to be the first to explicitly compare the effect of potential confounders to the effect of F_{ROH} across a wide range of traits (Figures 54 to 56).

The concordance between the combined effects of suspected confounders (SES and religiosity) and F_{ROH} is alarming (Figure 56); for only 9 out of 149 traits is the effect of F_{ROH} in the opposite direction to the combined effect of the confounder covariates. For these 9 discordant traits, this is strong evidence that these the observed effects of F_{ROH} cannot be explained by SES and religiosity. For example, reduced SES and increased religiosity (the presumed directions of F_{ROH} confounding) are both associated with increased fecundity (*neb*), while F_{ROH} causes reduced *neb*. Equally, the discordance of *age_at_first_wearing_glasses* seems to validate the effects of F_{ROH} on eyesight – reduced SES is strongly associated with starting to wear glasses later in life, while increased F_{ROH} causes earlier glasses wearing.

However, for the remaining 140 traits the concordance can be interpreted in two ways: either F_{ROH} is just a further measure of SES and religiosity, or the effects of F_{ROH} are remarkably similar to the effects those suspected confounders. Overall, the evidence presented in this chapter does not support the conclusion that F_{ROH} is a strong correlate of early-life SES (in the populations investigated), nor that the observed effects of F_{ROH} are primarily caused by confounding. Instead, it must be concluded that the causal effects of F_{ROH} are remarkably similar to the effects of reduced SES and increased religiosity.

Perhaps the simplest explanation for this similarity is that increased F_{ROH} causally reduces cognition which consequently affects a large number of other traits. However, this explanation does not fit well, as the effect of F_{ROH} on cognition appears to be relatively modest and not sufficient to explain the magnitude of effects observed on other traits. Indeed, educational attainment was fitted as a

covariate in all UKB phenotype models and therefore this direct causal pathway can be excluded. Alternatively, it may be that both increased F_{ROH} and reduced SES are deleterious to evolutionary fitness which manifest as similar phenotypic changes. And finally, although apparently far-fetched, it cannot be excluded that the causal effects of autozygosity have played a role in shaping society. One might imagine, for example, that religious beliefs which promote cousin marriage (and therefore autozygosity) could themselves become associated with the effects of autozygosity. I.e., if the most devout practice cousin marriage, which causally increases “conservatism” (age at first sex, number of sexual partners, alcohol units, ever smoked, answered sex questions, etc.), then these traits could themselves become associated with devotion.

4.4.4 Constancy of F_{ROH} effects between different populations

Joshi et al. introduced the idea of stratifying cohorts by geographic origin to qualitatively assess the stability of $\beta_{F_{ROH}}$ in different populations¹³. This idea was repeated here, with a similar conclusion – i.e., that, although there is some heterogeneity, for most traits the effects in different cohorts are consistent with a common cause.

Extending upon stratification by geography, ROHgen cohorts were also stratified by their inferred demographic history. In particular, there are a small number of ROHgen cohorts (East and West Polynesia, Hutterites, Amish), which do not practice consanguinity, but do have elevated burdens of ROH caused by a small effective population size. In these populations, it seems likely that much of the variation in F_{ROH} is caused by random variation in the inheritance of haplotypes, and is therefore independent of confounders. Again, effects stratified by demographic origin are broadly consistent, in keeping with a genetic explanation.

In UKB, $\beta_{F_{ROH}}$ was estimated independently in the genetically British and genetically non-British samples (Figure 62). Overall, there is good agreement between the two groups and only three traits previously suspected markers of religiosity (alcohol_drinker, pork, and lamb) have significantly greater effects in the non-

British cohort. Four traits (*neb*, *own_or_rent*, *log_lymphoc* and *falls_in_last_year*) do have a significantly greater effect in the British-cohort, and the reason for this is unclear. However, some heterogeneity does not necessarily mean that the observed effects are entirely caused by confounding. For example, a small part of the effect might be confounding, or there may cultural reasons why the same underlying genetic changes result in different phenotypic changes.

4.4.5 Linearity of effects with F_{ROH}

Several previous studies have compared the effects of short and long ROH, as different confounding might be associated with consanguinity (long ROH) and small effective population size (short ROH). Here, this idea was extended by plotting trait values in 10 bins of increasing F_{ROH} . Although, this method cannot prove the absence of confounding, the observed proportionality is consistent with a proportional increase in the number of homozygous genotypes (Figure 63). This proportionality continues to the two highest F_{ROH} groups (offspring of 2nd and 1st degree relatives, respectively) where there is little evidence for proportionally substantial confounding. Also, the magnitude of changes seen in the highest F_{ROH} group is substantial – significantly greater than the difference between the most, and least deprived neighbourhoods, for example.

4.4.6 Methods to reduce confounding in $\beta_{F_{ROH}}$

The importance of determining *with certainty* that the observed effects of F_{ROH} are caused by autozygosity has already been stressed, both in this thesis and by previous work⁷¹. The methods described so far in this discussion (exploring the relationship of F_{ROH} with possible confounders, or stratifying effects by ancestry, or by ROH length) can only increase (or decrease!) confidence that the observed effects are caused by autozygosity. The possibility of alternative explanations, however unlikely, will always leave a grain of doubt. Ultimately, certainty will only be achieved by a method of estimating $\beta_{F_{ROH}}$ which is indisputably free of confounding.

4.4.6.1 Within siblings

One such a method, estimating $\beta_{F_{ROH}}$ within full-sibling families, was proposed in Clark et al. and has been further refined in this thesis. Unfortunately, the variation of F_{ROH} between full siblings is particularly small, and large sample sizes are required to disprove the null hypothesis (of no F_{ROH} effect). The 131,523, mostly outbred siblings (mean $F_{ROH}=0.006$) available in this study were sufficient to confirm one trait (*ever had children*) as genetically associated with F_{ROH} . Meta-analysing across all traits, both in ROHgen and UKB, is strongly suggestive that a majority of traits will be associated with F_{ROH} within siblings when larger samples sizes become available, but currently it is not possible to say which. The ROHgen sibling meta-analysis (Figure 65) is currently the best powered estimate, and suggests that 74% [95% CI 40% to 110%] of the population effects of F_{ROH} are confirmed in siblings. Unfortunately, it is currently not possible to determine whether this means that 26% of traits are fully confounded, or whether a larger fraction of traits are partially confounded.

Although *ever_had_children* is the only individual trait to reach experiment-wise significance in the ROHgen siblings analysis, a compound trait summarising all traits (*multivariate* – see section 3.2.7) is significantly associated with F_{ROH} within siblings in both UKB (p -value = 0.01) and deCODE-Iceland (p -value = 0.0002). The *multivariate* trait is thus a convenient single trait for which the association with F_{ROH} is proven to have a substantial genetic component.

4.4.6.2 Other methods

Although, the within-sibling method of estimating $\beta_{F_{ROH}}$ is thought to be almost entirely free of potential confounding (and, as such, is the recommend method for future studies), its statistical power is not fully satisfactory in the datasets available here. Recommended sample sizes for future studies are explored in Appendix E.4, but to strengthen the evidence against confounding, with the data available here, other methods of reducing the influence of confounders were explored in this chapter.

4.4.6.2.1 isoUPD

ROH occur when two copies of the same haplotype are inherited from a single ancestor. The fact that this ancestor may be more or less distant in time has been exploited by previous studies comparing the effects of short and long ROH. The usual assumption being that longer ROH, most likely associated with consanguinity, are more likely to be confounded with social factors. However, this overlooks the closest theoretical source of identical haplotypes – a parent. It has recently become clear that the inheritance of both maternal and paternal alleles from a single parent (uniparental disomy, or UPD) is not as rare as once believed⁷⁹, and when both alleles are copied from one parental homologue (isoUPD) this results in very long ROH. Importantly, UPD is caused by chromosomal non-disjunction, which is believed to be a random event, and therefore ROH of this origin should be disassociated from confounders. Indeed, it seems likely that most carriers of UPD ROH are completely unaware of their condition.

isoUPD may be seen as a unique form of autozygosity where phenotypic effects are caused both by the homozygosity of recessive variants and by disorders of genomic imprinting occurring when both copies of the chromosome are inherited from one parent. Notably, homozygous variants in UPD segments have been exposed to a minimum of purifying selection and might therefore be more deleterious than variants which have survived multiple ancestral generations. Consistent with this explanation, the observed effects of UPD F_{ROH} on 155 UKB traits are, on average, 1.41 times greater than the effects of F_{ROH} in the cohort of non-UPD subjects. Despite the relatively small number of UPD cases, this analysis has sufficient statistical power such that one trait (*ever_had_children*) is significantly associated with UPD F_{ROH} – partly because of the increased effect sizes and partly because the average F_{ROH} of UPD cases is relatively high (mean F_{ROH} = 0.03).

If this UPD method is to be applied in future work, either very large population cohorts, or carefully matched UPD case/control studies will be required. The algorithm published by Nakka et al. was used to identify isoUPD, but this algorithm was not validated in all ethnicities and does risk misidentifying some common

ancestry as UPD. Here it was possible to confirm that the identified UPD cases were not enriched in any groups (Figure 87), but ideally parent-offspring trios would be gathered for an unequivocal diagnosis. Finally, the importance of imprinting disorders in UPD is unknown, so even with these conditions, estimates of $\beta_{F_{ROH}}$ within UPD cases will never be exactly comparable with population estimates.

4.4.6.2.2 Effect of F_{ROH} conditioning on F_{ROH_CM}

Realised F_{ROH} depends only on a subject's genealogy (summarised by F_{PED}) and on the randomness of segregation and recombination. Confounders are unlikely to be associated with segregation or recombination, and hence conditioning $\hat{\beta}_{F_{ROH}}$ on parental relatedness (either realised or expected) appears an attractive method of obtaining confounding-free estimates. Unfortunately, insufficient parent-offspring trios were available to condition $\beta_{F_{ROH}}$ on realised parental relatedness. Equally, conditioning on F_{PED} was not attempted since complete pedigrees were only available in relatively small cohorts - and any efforts are likely to be sensitive to pedigree misspecification. Instead, efforts were concentrated on estimating parental relatedness from subjects' genotypes.

It is interesting to note that in the offspring of any type of relatives, the number of ROH (NROH) is relatively less variable than F_{ROH} (which depends on both the number of ROH and the random length of each). NROH should therefore be a better predictor of parental relatedness than F_{ROH} and, as a first-approximation, should not affect the number of homozygous variants - nor therefore the magnitude of inbreeding depression (ignoring selection arguments - i.e., assuming ROH of all lengths have similar relative effects). Conditioning $\hat{\beta}_{F_{ROH}}$ on NROH therefore seems an attractive method for reducing confounding. This was attempted (and shows some promise - results not reported), but unfortunately F_{ROH} does remain informative about pedigree when conditioned on NROH. I.e., the equivalent plot to Figure 89 has a large number of significant associations. Instead, it was necessary to estimate parental relatedness more accurately, from a genetic map-based measure of autozygosity, F_{ROH_CM} , - although this substantially reduces statistical power

because very little of the variation in F_{ROH} is independent of F_{ROH_CM} . Theoretically, the variation of F_{ROH} which is independent of an accurate genetic map-based measure of F should have no association with parental relatedness, nor therefore any confounders. Figure 89 shows that this is not perfectly true when $\beta_{F_{ROH}}$ is conditioned on F_{ROH_CM} , possibly because shorter ROH are enriched in low recombination regions (Figure 17), and the burden of shorter ROH is a marker of ancestry. Nevertheless, the evaluation of $\hat{\beta}_{F_{ROH}|F_{ROH_CM}}$ is broadly satisfactory (Figure 89), and this method was applied in UKB.

Estimates of $\hat{\beta}_{F_{ROH}|F_{ROH_CM}}$ for 155 traits in UKB correlate significantly with $\hat{\beta}_{F_{ROH_CM}}$ (Figure 68) which gives further evidence that these effects have a genetic basis. Interestingly, this also suggests that inbreeding effects are more evenly distributed in physical distance than in map distance, which had not been demonstrated previously.

4.4.7 Effects of F_{ROH} resemble effects of hemizygous deletions

McQuillan et al. described how hemizygous deletions can be mistakenly called as ROH, but also showed that these deletions contribute only a small fraction of the variation of F_{ROH} , and do not substantially bias effect estimates⁸². Many other studies have assessed the medical consequences of specific deletions, but this thesis appears to be the first to quantify the effect of hemizygous deletion burden (F_{DEL}) on a wide range of complex traits, and then compare these effects to the effects of F_{ROH} .

This may seem a surprising comparison to make, since the causal mechanisms are thought to be different, but the concordance in effects is striking (Figure 70).

Although this is, in a sense, a comparison of apples and oranges, the fact that the indisputably genetic effect of hemizygous deletions, can affect the same traits as F_{ROH} , seems to be strong circumstantial evidence that the F_{ROH} effects themselves have a genetic origin.

Considering the *multivariate* trait, which can be considered a summary of all traits, the effects of hemizygous deletions are 64.2 times greater than the effects of ROH.

To put this in context, the average effect of ROH expected in the offspring of 1st cousins is equivalent to 3 Mb of hemizygous deletion.

As part of future work, it would be interesting to explore further the effect of deletion burden on complex traits. In particular, it is unclear whether the burden of shorter deletions (which occur in almost all UKB samples⁸⁸) has the same effect as the burden of the relatively rare, longer (>1.5Mb) deletions considered here.

4.4.8 Discussion summary

Confounding of F_{ROH} by social variables has frustrated researchers for many decades. Previous work has identified examples where confounding does bias $\hat{\beta}_{F_{ROH}}$, but there is little agreement about the extent of this problem. As the price and availability of human genotyping tumble, it seems inevitable that *Homo sapiens* will become the organism of choice for studying many aspects of evolutionary genetics, including inbreeding depression. In this context it is vital to determine which effects of F_{ROH} are caused by genetics and which are caused by social confounding.

This chapter has introduced several methods which demonstrate, almost beyond doubt, that autozygosity causes the majority of complex trait changes associated with F_{ROH} - in both ROHgen and UKB. However, in UKB, increased F_{ROH} is associated with markers of religiosity and (to a lesser extent) SES, so confounding cannot be ignored. Most importantly, a method of estimating $\beta_{F_{ROH}}$ within full sibling families has been introduced, which is believed to be entirely free from confounding, and will resolve the question of confounding for specific traits when larger sample sizes become available. For now, the best estimate from this method suggests that 74% [95% CI 40% to 110%] of the effects observed in ROHgen have a genetic basis. In addition, one trait - a reduced likelihood of having children - is proven to have a genetic component.

5 Effects of ROH in specific genomic locations

5.1 Introduction

Homozygosity mapping is a common method used to identify genomic regions which may harbour recessive disease variants in consanguineous families¹⁴⁴. Typically, autozygous regions shared by all (or a significantly raised fraction) of affected family members are identified, and then prioritised for follow-up studies such as resequencing genes in the candidate region. Given sufficiently large consanguineous families this can be powerful method for identifying monogenic disease genes, but the power falls for variants with incomplete penetrance, or when multiple loci contribute to disease susceptibility¹⁴⁴.

Complex traits are, by definition, associated with many genomic loci, and very large consanguineous families would be required to detect modest effects. Instead, more recent homozygosity mapping efforts have taken place in population cohorts, where much larger samples are available^{80,99}. However, a complication of ROH mapping in population cohorts is that recombination breakpoints occur at different positions in different samples, creating a complex pattern of distinct, but sometimes overlapping ROH. This complexity has been addressed in different ways by different studies. One simple method is to assess the relationship between traits and ROH at regularly spaced loci – for example, every 500 kb⁶⁹. Since average ROH length are much greater than 500 kb, this captures all large effects, but may be conservative due to excessive multiple test correction.

Alternatively, partial ROH overlaps may summarised in “consensus ROH”, identified either by PLINK or by algorithms originally developed for amalgamating CNVs. The consensus ROH method has been used by two recent studies reporting significant associations, but both studies identify, and test, tens of thousands of consensus ROH^{80,99}. Given the strong correlation between ROH present at loci even 1 Mb apart, this method also seems to exceed the possible number of independent tests. In any case, the reported associations reported (homozygosity at *GIL4C* and

cognition⁸⁰, and homozygosity at *HS3ST1* and Alzheimer's disease⁹⁹) would benefit from replication.

It might also be noted that overlapping ROH do not necessarily represent autozygosity of the same haplotypes. Different haplotypes are likely to contain different recessive variants, and amalgamating their effects may cause a dilution of any association signal. To this end, a recent preprint proposed a method to identify long ROH diplotypes (distinct homozygous haplotypes) shared by multiple subjects¹⁴⁵. Applied to UK Biobank this method identifies numerous associations, most overlapping with known GWAS hits. The authors of this method state that *"While the association signals we found mostly overlap with existing GWAS hits, we are testing different genetic effects. The existing GWAS are mainly testing the additive effects of single SNPs, while we are testing the recessive effects of relatively long haplotypes"*. However, this statement, and indeed other studies reporting local ROH associations, seem to overlook the fact that, if associations occur within ROH islands, they are likely to capture the additive effect of the homozygous haplotypes, as well as possible recessive effects (see section 5.3.1.1).

In any case, very few, if any, recessive effects on human complex traits have been firmly established at specific genomic loci. As previous authors have suggested, this implies that the known effects of inbreeding are distributed among a relatively large number of causal loci, and that localised effects are generally small. In this context, an interesting alternative may be to attempt to discover genomic features or annotations that are associated with the genome-wide variation of ROH effects. This approach was taken by a recent study which fitted bivariate models of F_{ROH} , and F_{ROH} weighted by 44 different annotations, to 11 quantitative traits measured in UK Biobank⁷³. Averaging across 11 traits, the authors report that inbreeding depression is enriched in regions of high recombination rate and low nucleotide diversity. A similar analysis is performed in this thesis (section 5.3.2), and the results are compared in section 5.4.1.

In summary, a small number of studies have extended homozygosity mapping to population cohorts, but the few reported associations are not proven to be truly

recessive effects. Investigating the regional variation of ROH effects may elucidate the genetic architecture of inbreeding depression, and may help target future homozygosity mapping efforts⁷³.

5.2 Methods

5.2.1 Estimating local effects of ROH in 3 Mb wide windows

To estimate the effect of ROH at specific genomic locations, the autosomes and X chromosome were divided into approximately one thousand, non-overlapping, 3 Mb wide windows. The start and end coordinates of each window (on human genome build GRCh37/hg19) are given in Supplementary Table 20. Since all ROHgen cohorts were required to provide genotypes on build GRCh37/hg19, these windows should represent the same physical positions in all cohorts.

In each ROHgen cohort, and for every sample, the fraction of each window covered by a ROH was calculated. These fractions, ranging between 0 and 1, were denoted as $window_i_j$ for the i^{th} window in the j^{th} sample. Then, for all i , the effect of $window_i$ was estimated by the regression:

$$\epsilon' \sim window_i + (F_{ROH} - \frac{window_i}{1000}) \quad (40)$$

Where ϵ' are the trait residuals from equation (17). It is necessary to fit F_{ROH} as a covariate in equation (40), since $window_i$ is correlated with F_{ROH} both by its own homozygosity and its correlation with the homozygosity of nearby windows.

Subtracting $\frac{window_i}{1000}$ from F_{ROH} , removes $window_i$'s contribution to F_{ROH} , so that the effect of homozygosity in this window is not counted twice.

The effects of windows on the X chromosome can only be estimated in women, but were estimated in a similar manner, noting that F_{ROH_X} was calculated by dividing the summed Mb length of X chromosome ROH by a nominal X chromosome length (156 Mb), and not by 3000 Mb as for F_{ROH} .

$$\epsilon' \sim window_i + F_{ROH} + (F_{ROH_X} - \frac{3 * window_i}{nominal_X_length}) \quad (41)$$

In this way, the effect of having ROH entirely overlapping a window was estimated for all windows in all ROHgen cohorts. These individual estimates were then meta-analysed, by a fixed-effect inverse variance method, to obtain overall estimates for each window.

5.2.1.1 Removing additive effects caused by ROH islands

Unfortunately, the method described in section 5.2.1 was found to be confounded by the additive effects of the common haplotypes which cause ROH islands (see results section 5.3.1.1) and, since the ROHgen analyses were performed by scripts distributed to many analysts, it was not feasible to fix this in the full consortium. However, the problem was addressed, in two ways, in UK Biobank where individual level data was available.

5.2.1.1.1 Conditioning the effect of significant windows on SNPs

In results section 5.3.1.1, five windows were found to have significant associations with 12 traits. Unfortunately, all of these windows overlap both ROH islands and GWAS hits for the relevant traits, so additive genetic effects are a plausible explanation of the observed associations (see explanation in section 5.3.1.1).

This was investigated by fitting all SNPs in each significant window, and the window autozygosity, as covariates in a multivariate model (equation (42)). For example, height was found to be significantly associated with the homozygosity of window_948 (33 to 36 Mb on chromosome 20). For each of the 236 genotyped SNPs in window_948 a dosage (0, 1, or 2) was extracted for all samples, and denoted snp_{ij} for the i^{th} SNP in the j^{th} sample. Then the effect of ROH in window_948, conditional on all SNPs in the window, was estimated from the model:

$$height_residuals \sim snp_1 + snp_2 + \dots + snp_236 + window_948 \quad (42)$$

$$+ (F_{ROH} - \frac{window_948}{1000})$$

Since the SNPs have not been pruned for Linkage Disequilibrium (LD) their dosages are occasionally highly correlated but, although this makes the individual SNP effect estimates unreliable, it does not affect the estimated effect of window_948.

This method was used to account for additive effects in four of the five significantly associated windows, but one window (window_373) covers the HLA region on chromosome 6. The HLA locus is by far the most polymorphic region of the human genome, and window_373 contains 3524 SNPs, which proved to be too many to fit in equation (42). However, UK Biobank provides a separate imputation of HLA alleles at 11 classical HLA genes⁸¹, so instead of SNPs, the imputed dosage of each of these 362 HLA alleles were fit as covariates in equation (42).

5.2.1.1.2 Filtering ROH by genetic length to remove ROH islands

Section 2.5.5.1 showed that ROH islands could be effectively removed by removing all ROH of genetic length less than 3 cM. The resulting genome wide distribution of ROH is approximately constant (Figure 19), as would be expected of autozygosity defined as identity-by-descent from a fixed base generation. This constant distribution of ROH implies that any base generation allele has an equal probability of being made autozygous by inbreeding, which eliminates all concerns about additive effects.

The calculations of section 5.2.1 were therefore repeated in UK Biobank with all ROH less than 3 cM removed. Unfortunately, in UK Biobank, 84% of total ROH length occurs in ROH of less than 3 cM, so the statistical power is greatly reduced - and this is particularly true at ROH islands. For example, before filtering window_131 is homozygous 10.0% of UKB subjects, while after filtering this falls to 0.085%.

5.2.2 Summarising genomic annotations in 3 Mb-wide windows

Summaries of several genetic annotations were calculated in the same 3 Mb-wide windows described in section 5.2.1 (Supplementary Table 20). These summaries were subsequently used to investigate trends in the window effects estimated in ROHgen.

5.2.2.1 Gene density

Gene locations were downloaded from the UCSC SQL table browser (<https://genome.ucsc.edu/cgi-bin/hgTables>) in the knownGene table from the UCSC

genes track for assembly “Feb 2009 GRCh37/hg19”¹⁴⁶. This table contains start and end coordinates for 19,672 protein-coding genes, of which 18,768 are located on the autosomes and 833 on the X chromosome (the remainder are on the Y chromosome, or unplaced contigs, or in the Mitochondrial genome). The number of protein-coding genes in each of the 995 previously defined autosomal, and 53 X chromosome windows was then counted. Since only a very small fraction of genes overlap window boundaries, genes were considered to be located at their mid-point to simplify this analysis.

5.2.2.2 Additive variance explained

For the best-powered ROHgen trait, height, the approximate additive variance explained by each autosomal window was calculated, by estimating independent SNP effects in UK Biobank.

SNP genotypes for 784,256 autosomal markers in 409,629 UKB subjects determined (by UKB) to be genetically British were quality controlled by removing all SNPs with $MAF < 1\%$, SNPs with $> 3\%$ missingness and finally 1,189 samples with $> 3\%$ missing genotypes. To remove linkage disequilibrium (LD) between correlated SNPs, stringent LD pruning was performed in PLINK^{67,68} using the parameters `-indep-pairwise 1000kb 10 0.1`. This greedily removes all pairs of SNPs (within 1000 kb of each other) with r^2 greater than 0.1 until no such pairs remain. This LD-pruned dataset contains 152,467 autosomal SNPs for 408,440 samples and with no two markers correlated above $r^2 = 0.1$.

A genome wide association test was then performed by PLINK, on the residuals of height calculated in equation (17). This gives effect estimates ($\hat{\beta}_{SNP_i}$) for all SNPs from $i = 1$ to 152,467. The additive variance explained by a locus is given by $2pqa^2$, where p and q are the major and minor allele frequencies and a is the additive effect of one allele¹⁶. Since $\hat{\beta}_{SNP_i}$ is an estimate of a , it is necessary to remove the sampling variance, so the additive variance explained by each SNP (V_{A_i}) can be estimated as:

$$V_{A_i} = 2 * p * q * (\hat{\beta}_{SNP_i}^2 - se(\hat{\beta}_{SNP_i})^2) \quad (43)$$

The total additive variance explained by each autosomal window was then calculated by summing V_{A_i} for all SNPs contained by the relevant window.

5.2.3 Effect of ROH on each chromosome

In UK Biobank, the effect of ROH on the *multivariate* trait was estimated for each chromosome by multiple regression. The summed length of ROH (in Mb) on each autosome, and the X chromosome, was calculated, and denoted chr_i_j for the i^{th} chromosome in the sample. The effect of ROH on each autosome was then estimated from the linear model fitted in all samples:

$$multivariate \sim chr_1 + chr_2 + chr_3 + \dots + chr_22 \quad (44)$$

The effect of ROH on the X chromosome was estimated separately, and only in women, by the model:

$$multivariate \sim F_{ROH} + chr_X \quad (45)$$

5.2.4 Calculating F_{ROH} at specific genomic annotations

5.2.4.1 Protein-coding genes

In ROHgen, the effect of ROH in each of 995 autosomal windows was first estimated, and then compared to gene density (results section 5.3.2.1). Since each window effect is estimated separated, this is unlikely to be the most statistically efficient method of determining whether there is a relationship between gene density and ROH effects.

Alternatively, if the genetic variants causing inbreeding depression are mostly located in, or near to, protein-coding genes, then the fraction of protein-coding genes in ROH should be more strongly associated with the autozygosity of causal variants (and therefore inbreeding depression) than F_{ROH} . This was tested in UK Biobank by calculating, F_{GENES} , the estimated fraction of exonic bases in ROH, for each sample. Exons were downloaded as described in section 5.2.2.1 and, since exons are mostly small compared to ROH, assumed to be entirely located at their

mid-point. Hence, for each sample the summed length of exons with a mid-point falling in a ROH was calculated. These summed lengths are proportional to F_{ROH} , but on a basepair scale rather than a genomic fraction, so were normalised by the regression coefficient from:

$$F_{GENES} \sim F_{ROH} \quad (46)$$

The conditional effect of F_{GENES} ($\beta_{F_{GENES}}$) was then estimated for each trait by fitting the full phenotype model:

$$trait \sim F_{ROH} + F_{GENES} + covariates \quad (47)$$

5.2.4.2 Other genomic annotations

In section 5.2.4.1, F_{GENES} was calculated from exon locations defined by their location (chromosome and base-pair mid-point) and length. This same calculation can be applied to any genomic annotation that can be summarised by a point location and a length (or weight). An equivalent of F_{GENES} , called $F_{ANNOTATION}$, was thus calculated for 7 different genomic annotations or features. The following sections briefly describes how the locations and weights of these features were obtained.

5.2.4.2.1 Copy number variations (CNVs)

The qualitative similarity between the effects of ROH and the effect of long hemizygous deletions (Figure 70) suggests the possibility that inbreeding depression is caused by the homozygosity of the short deletions that are ubiquitous in healthy populations⁸⁸. Hemizygous deletions, and more generally CNVs, are reported to occur most frequently at specific genomic locations, “*with breakpoints most common near the ends of chromosomes, and at known regions of segmental duplication*”⁸⁸.

A complete list of CNVs called in UK Biobank (by PennCNV⁹⁰) was downloaded from Aguirre et al.⁸⁸ Start and end coordinates, as well as frequency of occurrence, are provided for 275,180 CNVs ranging from 2 kb to 60.8 Mb (median length = 99.8 kb). This was converted to a table of 202,490 unique breakpoint positions with

corresponding occurrence frequencies, which was then used to weight $F_{ANNOtATION}$, as described in section 5.2.4.2.

5.2.4.2.2 Loss-of-function (LoF) intolerant genes

Results recently published by the Genome Aggregation Database (gnomAD) estimate the degree of constraint imposed by selection on 19,704 protein-coding genes¹⁴⁷. In summary, 3063 genes with a significant depletion (compared to the expectation) of predicted loss-of-function (pLoF) variants in 141,456 exome sequences are classified as pLoF intolerant (pLI). Although this primarily indicates selection against heterozygous pLoF variants, these are likely to be genes with critical functions which may also be affected by homozygosity. The enrichment of ROH effects at pLoF genes was therefore investigated by calculating F_{GENES} at a set of 2773 pLI genes which could be uniquely identified by gene symbol in the previously downloaded exon table (section 5.2.2.1).

5.2.4.2.3 Recessive disease genes

A list of 1,178 well-characterised, autosomal recessive disease genes was downloaded from Lek et al.¹⁴⁸, who had combined sets previous published in Blekhman 2008¹⁴⁹ and Berg 2013¹⁵⁰. Again, the enrichment of ROH effects at recessive disease genes was investigated by calculating F_{GENES} at a set of 1,163 recessive genes which could be uniquely identified by gene symbol in the previously downloaded exon table (section 5.2.2.1).

5.2.4.2.4 Nuclear DNA (nDNA) genes involved in mitochondrial processes

A list of 1,529 autosomal genes with a gene ontology term GO:0005739 (mitochondrion) were download from ensembl BioMart (<https://grch37.ensembl.org/biomart/>), and F_{GENES} at these locations was calculated as previously described.

5.2.4.2.5 Imprinted autosomal genes

A list of 273 autosomal genes with at least some evidence for imprinting was downloaded from geneimprint (geneimprint.com)¹⁵¹ and used to calculate F_{GENES} as described previously.

5.2.4.2.6 De novo mutation rate

Since inbreeding depression is thought to be caused by the homozygosity of rare variants, regional variation in the de novo mutation rate might plausibly influence the distribution of inbreeding depression effects throughout the genome.

An approximation for the local mutation rate was obtained from Besenbacher et al.¹⁵², who identified 17,812 de novo mutations by whole genome sequencing of 283 parent-offspring trios. The location of these de novo mutations was downloaded, and converted from build GRCh38 to GRCh37 using the R package *liftover*. $F_{ANNOtATION}$ was then calculated at 17,560 SNP locations which could be mapped onto build Chr37, with each location given equal weight.

5.2.4.2.7 1000 genomes SNP diversity

Inbreeding depression at a biallelic locus can be described by $2pqFd$, where p and q , are the major and minor allele frequencies, F is the inbreeding coefficient and d is the heterozygote trait value, relative to the mid-point of the two homozygote values¹⁶. Since F is expected to be the same at all loci, and d is currently unknown at all loci, it may be informative to explore the variation of inbreeding depression with local nucleotide diversity (π) which can be estimated from allele frequencies as $2 \sum pq$ ¹⁵³.

Whole genome sequence genotype calls for 2504 worldwide samples were downloaded in vcf format from the 1000 genomes project (<ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>)¹⁵⁴, and converted to bed format by PLINK v1.9. Allele frequencies for 81,271,745 autosomal variants were then calculated using the PLINK command `–freq`. Expected heterozygosity ($2pq$) was calculated at each locus and then summed in 268,326 bins of 10 kB width. The mid-point of each bin, and corresponding summed heterozygosity were then used as the position and weight for calculating $F_{ANNOtATION}$.

5.2.5 Simulation of inbreeding depression at known causal loci

To better understand the results expected from the model $multivariate \sim F_{ROH} + F_{GENES}$, a trait with statistical properties similar to the *multivariate* trait was simulated for a range of different assumptions about the nature of causal variants.

The main inputs to the model are:

- 1) The number of causal variants located in, or near, protein-coding exons.
- 2) The maximum distance between an “exonic” causal variant and its corresponding exon (the “jitter”).
- 3) The number of causal variants unrelated to genes, and therefore randomly distributed throughout the genome.
- 4) The distribution of effect sizes for simulated causal variants.

Firstly, the location of all causal variants was simulated. Causal variants were specified in two groups: the first group being associated with genes, the second randomly distributed throughout the genome. For each causal variant associated with a gene, an exon was randomly chosen (with replacement) from the list of 193,881 exons previously downloaded. The base-pair position of the causal variant was then randomly selected between the start and end position that exon. If a “jitter” had been specified, a random value between -jitter and +jitter was added to the base-pair position. The positions of variants unrelated to genes were selected by first randomly selecting a 3 Mb window (see section 5.2.1) and then adding a random value between 0 and 3,000,000 to the selected window start position. The effect of each variant being in ROH was assigned according to either a uniform or exponential distribution: for the uniform distribution each variant was given an effect equal to $-1/number_of_causal_variants$, and for the exponential distribution each variant was first assigned a value (β_i) randomly chosen from an exponential distribution, and then divided by $-1 * \sum_{i=1}^N \beta_i$. In this way, for both distributions, the effect of having all causal variants in ROH is -1, which is equal to $\beta_{F_{ROH}}$ for the *multivariate* trait.

Subsequently, for each UK Biobank sample, the simulated causal variant locations overlapped by a ROH were identified, and the sample given a simulated trait value equal to the sum of the corresponding effect sizes. This gives a simulated trait with $\hat{\beta}_{FROH} = -1$, like the *multivariate* trait, but with almost no error variance. To match the variance of the *multivariate* trait, a value randomly drawn from a normal distribution with mean=0 and $sd=sd(multivariate)$ was added to each sample. In this way, each simulated trait, regardless of the inputs selected, has the same $\hat{\beta}_{FROH}(-1)$ and variance (0.1513) as the *multivariate* trait.

5.3 Results

5.3.1 ROH effects in specific 3 Mb wide windows

5.3.1.1 Local effects of ROH in 3 Mb-wide windows

The local effects of ROH in 941 autosomal windows, and 53 X-chromosome windows, were estimated for all ROHgen traits, as described in section 5.2.1 (54 autosomal windows have no ROH and therefore no estimated effects). The large number of separate tests (49 traits * 994 windows) necessitates correction for multiple testing, and an experiment-wise p -value of 1×10^{-6} ($\approx 0.05 / 49 * 994$) was used to assess whether any of these localised effects differ significantly from zero. For example, ROH in 3 separate windows are associated with height at a p -value of less than 1×10^{-6} (Figure 72).

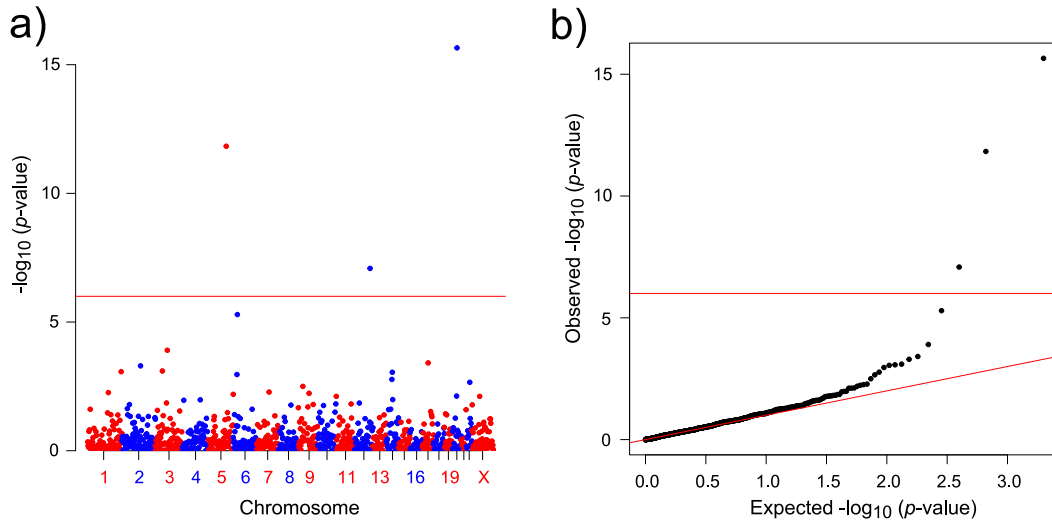


Figure 72: **Significant associations between height and ROH in specific 3 Mb-wide windows.** (a) Manhattan plot of $-\log_{10} p$ -value against window location for 941 autosomal and 53 X-chromosome windows. (b) Quantile-quantile (QQ) plot of observed versus expected p -values for 994 window associations with height. Bonferroni-corrected significance (p -value = 1×10^{-6}) is shown as a horizontal red line in both plots.

In total, 24 window-trait combinations reach experiment-wise significance (Table 2). Of these, 10 affect traits which are not significantly associated with (genome-wide) F_{ROH} . Among the remaining 14, 11 effects are directionally consistent with $\hat{\beta}_{F_{ROH}}$, while 3 are not – a split slightly unlikely to occur by chance (one-sided binomial test p -value = 0.029).

Table 2: **Genomic windows significantly associated with ROHgen traits.** To aid comparison, effect estimates ($\hat{\beta}$) are normalised by the intra-sex trait standard deviation.

Window	Chr	Locus	Trait	$\hat{\beta}$ (trait SD)	<i>p</i> -value
window_131	2	Lactase persistence (<i>LCT</i>)	Total cholesterol	-0.028	4.9 x 10 ⁻¹⁰
window_344	5	Unknown	Height	0.026	1.5 x 10 ⁻¹²
window_344	5	Unknown	Mean platelet volume	0.035	1.4 x 10 ⁻⁷
window_373	6	HLA	White blood cell count	-0.111	1.8 x 10 ⁻³⁸
window_373	6	HLA	Triglycerides	-0.070	6.1 x 10 ⁻¹⁸
window_373	6	HLA	Lymphocytes (%)	-0.056	6.3 x 10 ⁻¹¹
window_373	6	HLA	Haemoglobin	-0.045	1.4 x 10 ⁻⁸
window_373	6	HLA	Total cholesterol	-0.040	1.5 x 10 ⁻⁸
window_373	6	HLA	Platelet count	-0.046	1.8 x 10 ⁻⁸
window_373	6	HLA	Mean platelet volume	0.056	7.0 x 10 ⁻⁸
window_373	6	HLA	Gamma-Glutamyl Transferase	-0.043	1.7 x 10 ⁻⁷
window_373	6	HLA	Waist-Hip ratio	-0.036	4.3 x 10 ⁻⁷
window_373	6	HLA	Diastolic blood pressure	-0.035	9.4 x 10 ⁻⁷
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	Platelet count	0.110	1.1 x 10 ⁻⁵⁶
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	Haemoglobin	0.069	2.8 x 10 ⁻²⁴
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	White blood cell count	0.068	9.9 x 10 ⁻²³
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	Gamma-Glutamyl Transferase	-0.057	5.6 x 10 ⁻¹⁵

window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	HDL cholesterol	-0.046	2.8×10^{-12}
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	Weight	-0.031	2.0×10^{-9}
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	Total cholesterol	-0.032	5.0×10^{-8}
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	Height	-0.028	8.2×10^{-8}
window_705	12	Aldehyde dehydrogenase (<i>ALDH2</i>)	Alcohol units per week	-0.029	5.0×10^{-7}
window_948	20	Unknown	Height	-0.055	2.2×10^{-16}
window_948	20	Unknown	Weight	-0.031	6.8×10^{-7}

Only 5 distinct windows are associated with localised effects, with most affecting two or more traits. At least three of the significant windows overlap well known genomic loci (window_131 = Lactase persistence locus, window_373 = human leukocyte antigen, window_705 = aldehyde dehydrogenase (*ALDH2*) gene), and all have average ROH levels well above the genome-wide average – indicating that they overlap ROH islands (Table 3).

Table 3: **Elevated autozygosity of significantly associated windows.** ROH islands may be less than 3 Mb wide, so as well as the average ROH fraction of each window in ROHgen (returned as a summary statistic by each cohort), the height of the ROH island in UK Biobank British is reported.

Window	Chr	Start (Mb)	End (Mb)	Mean ROH fraction of window in ROHgen (%)	Peak ROH fraction in UKB British (%)
Window_131	2	135	138	8.7	16.9
Window_344	5	129	132	8.1	8.2
Window_373	6	30	33	2.7	3.5
Window_705	12	111	114	5.1	8.1
Window_948	20	33	36	3.1	3.2

The elevated autozygosity of these significantly associated windows is perhaps unsurprising, since increased ROH rates increase statistical power, but it also reveals a major, unforeseen, shortcoming of this method:

ROH islands are thought to be caused by long, common haplotypes. For example, if, say, 9% of samples are homozygous for the same 1.5 Mb haplotype, then assuming Hardy-Weinberg equilibrium, approximately 30% ($q = \sqrt{0.09}$) of alleles at the locus will be that haplotype, and approximately 42% ($2pq$) of samples will be carriers. This study has not investigated how many distinct haplotypes contribute to each ROH island, but other studies have found there is often one predominant haplotype¹⁴⁵, and even when there are several, the concerns discussed in the following paragraph remain.

When ROH are caused by one or more common haplotypes (as opposed to any allele having an equal probability of being made autozygous by inbreeding) then the effect of autozygosity is confounded by the additive effects of the common haplotypes. In the most extreme case, the presence of a ROH island indicates that an individual carries two copies of a common haplotype, and testing the effect of the ROH island compares samples with two copies to those with one, or none. If the common haplotype has an additive effect (i.e., two copies have twice the effect of one copy - as typically found in GWAS), then testing the ROH island effectively becomes testing (almost twice) the additive effect. While discovering additive effects of genomic regions is a legitimate aim, it is not the purpose of this analysis (which aims to find non-additive dominance effects), and it is likely that traditional GWAS will be better powered to find most additive effects. Indeed, all 24 window-trait associations overlap known GWAS hits for the relevant traits.

Despite this complication at ROH islands, the method remains valid in the remainder of the genome where ROH islands do not occur. ROH rates vary continuously throughout the genome, so delimiting ROH islands is necessarily arbitrary, but, in ROHgen, 54% of autosomal windows have average ROH fractions between 0.18 and 0.4% and can be considered mostly unaffected by ROH islands (Figure 73). There are no significant associations between the homozygosity of

these unconfounded windows and any traits (even when the significance threshold is lowered to account for the smaller number of windows), and this puts an upper limit on the magnitude of any localised ROH effects. Taking height as an example, the median standard error of these window effect estimates is 1.5 mm, and a t -statistic of 4.6 corresponds to the significance threshold, so ROH covering a typical window would need to change average height by 6.9 mm to be statistically significant. This represents an effect corresponding to 11% of the intra-sex standard deviation of height, or 5% of the total effect of F_{ROH} .

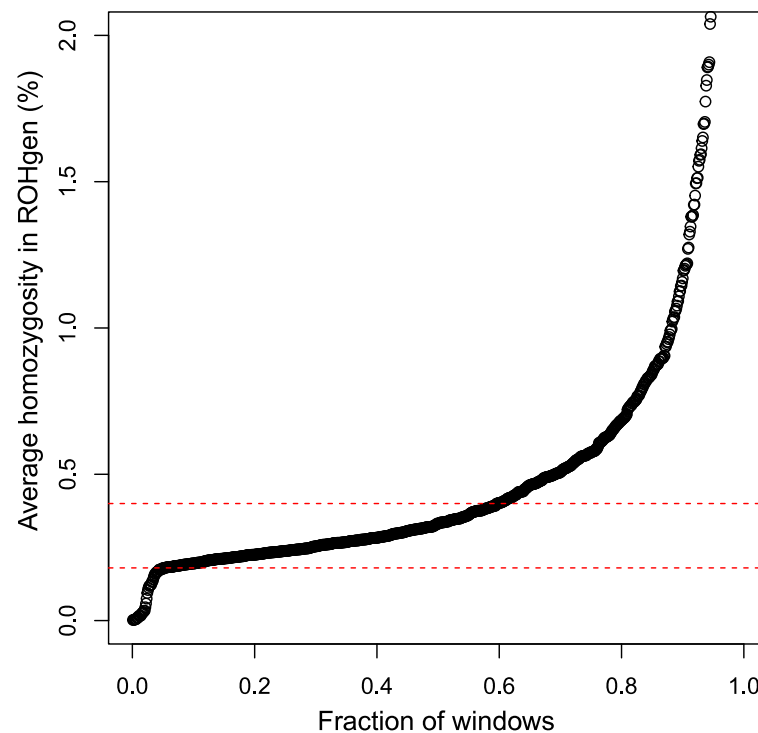


Figure 73: **Autosomal windows ordered by average homozygosity in ROHgen.** The y-axis is truncated at 2% to emphasise the region of interest. Horizontal lines shown in dashed red (at 0.18% and 0.4%) correspond to the approximate range of windows not confounded by ROH islands, or affected by genotyping difficulties. The approximately 5% of windows which have average autozygosity less than 0.18%, are located in regions of the genome which are poorly represented on current SNP microarrays – such as centromeres or telomeres.

Unfortunately, because the ROHgen analyses were achieved by scripts distributed to numerous analysts, changing the method to account for additive effects was not feasible in the full consortium. However, many of the associations remain significant (see Figure 74) in UK Biobank British alone (for which individual level data is available), and this cohort was used to investigate further in the following section.

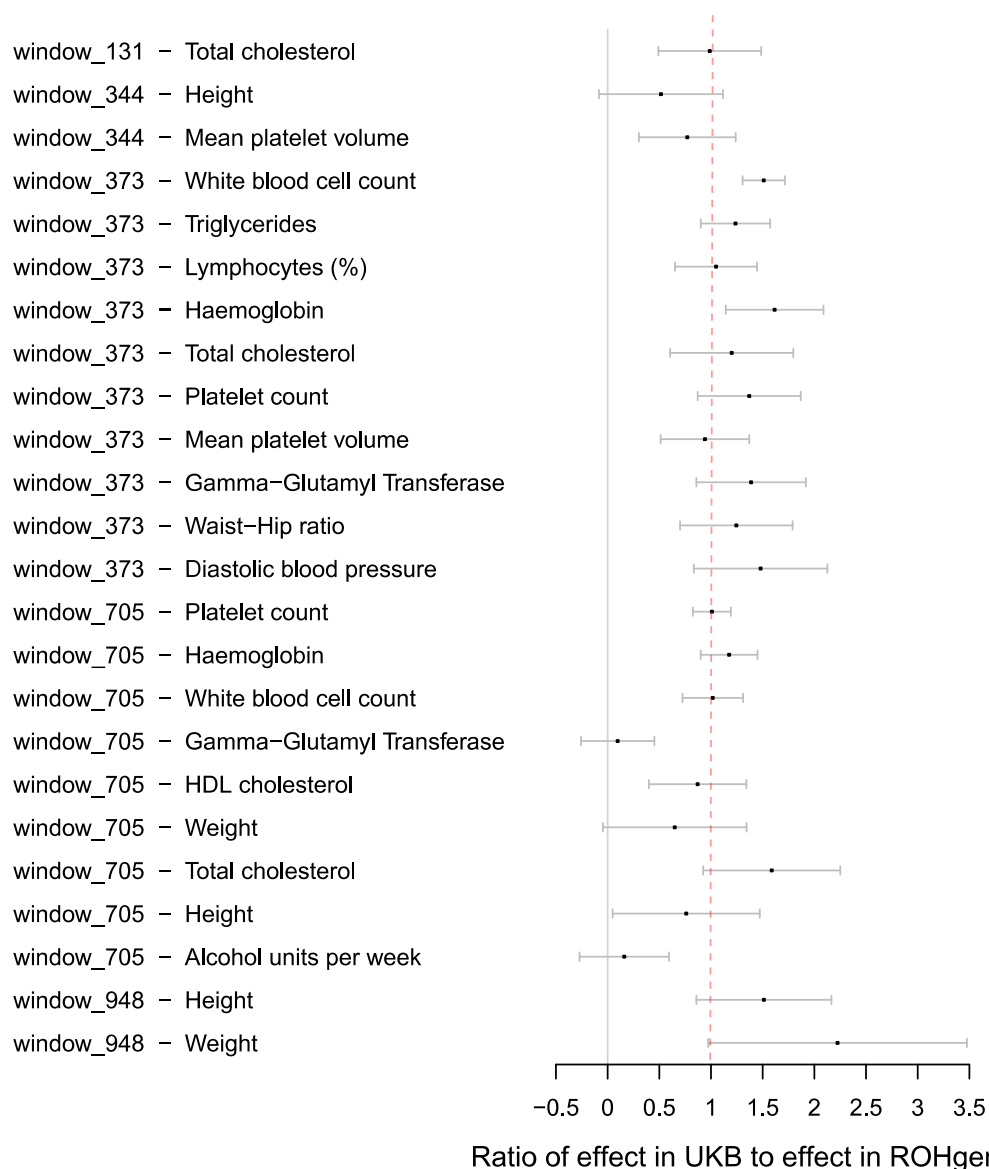


Figure 74: **Ratio of window effect in UK Biobank British to effect in ROHgen for 24 significant window-trait pairs.** The expected ratio of 1 is shown as a dashed red line.

Although most trait-window associations remain significant (p -value $< 0.05/24$) in UK Biobank British alone, five do not (window_344 with height, and window_705

with height, weight, alcohol units and Gamma-Glutamyl Transferase (GGT)). For these exceptions, the effect estimates were examined in individual cohorts:

The association between window_344 and height is significant in deCODE-Iceland (p -value = 3×10^{-6}), but not in UK Biobank British (p -value = 0.09), nor in any other cohort. Genetic drift and/or selection may cause haplotypes to segregate at different frequencies in different populations, so this association may be caused by a variant enriched in the Icelandic population - which will therefore be difficult to investigate in UK Biobank.

The associations between window_705 and height, weight, alcohol units and GGT are significant in Biobank Japan (p -values 1×10^{-9} , 3×10^{-7} , $< 1 \times 10^{-300}$, and 7×10^{-76} respectively), but not in UK Biobank British (p -values, 0.03, 0.07, 0.46, and 0.59). The effect on alcohol units is particularly remarkable: In Biobank Japan (BBJ), 1 Mb of ROH in window_705 is associated with drinking 5.2 fewer units of alcohol per week [95% CI = 5.0 to 5.4]. The *ALDH2* gene, a variant of which is associated with an adverse flushing reaction to alcohol consumption in east Asian populations, lies within this window and is presumed to be responsible for the observed effects. Unfortunately, without the individual data it is not possible to determine whether this ROH method is simply detecting the well-known additive effect, or whether there is also a recessive component.

5.3.1.2 Effect of ROH in 3 Mb-wide windows, accounting for additive effects

For each of the 19 window-trait pairs that are significant in UK Biobank British, covariates were fit to describe the additive effect of common haplotypes (see methods section 5.2.1.1.1). The window effect estimates conditioning on these additive covariates are shown in Figure 75, below.

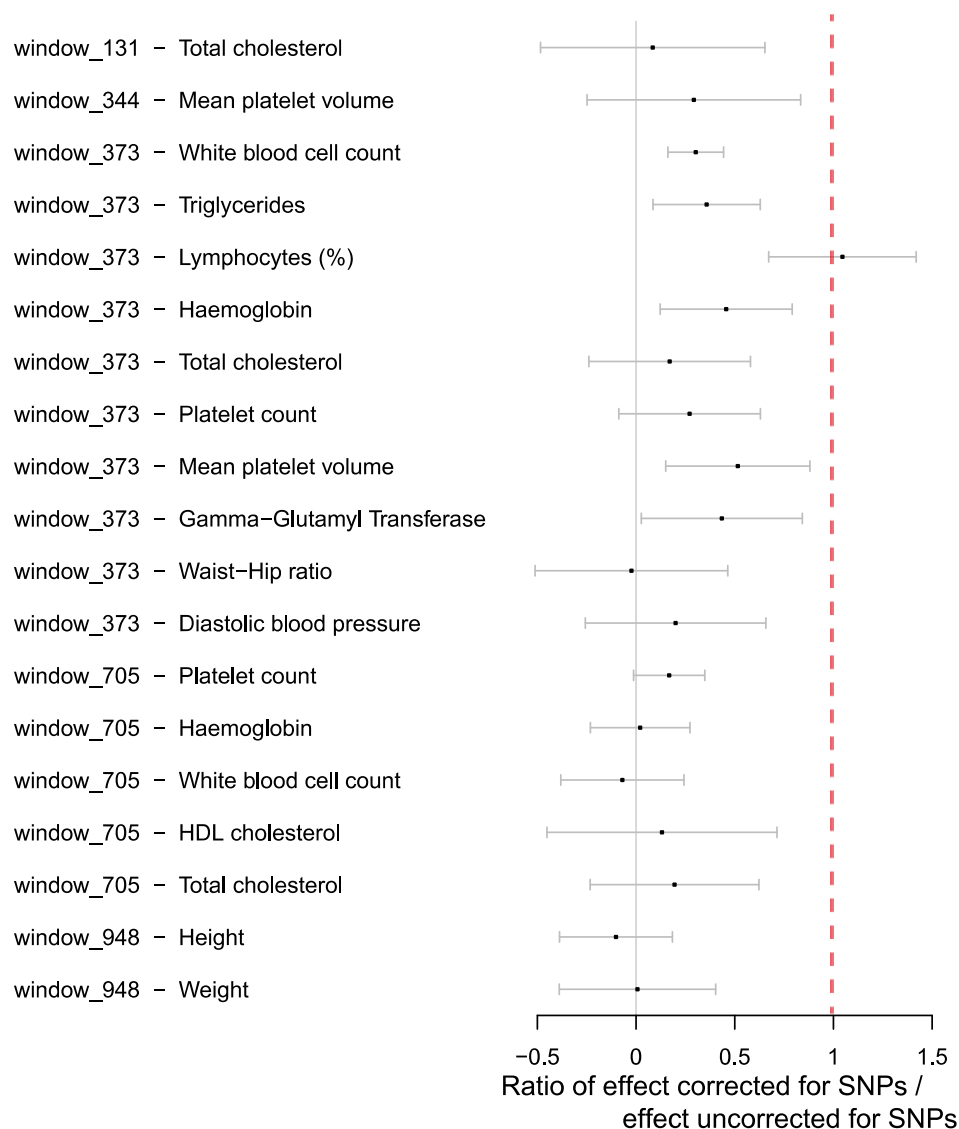


Figure 75: **Effect estimates in UK Biobank British for 19 window-trait combinations, conditioning on additive effects.** All estimates are normalised by the equivalent estimate without conditioning on the additive effects – i.e., a ratio of 0 is expected if the association is explained entirely by the additive effects, and a ratio of 1 if the association is caused only by window homozygosity. 95% confidence intervals are shown in grey.

Conditioning on the SNPs contained within each window reduces the estimated effect of ROH for all but one window-trait combination, and only two traits (Lymphocyte percentage and White blood cell count) remain significantly associated (p -value $< 0.1 / 19$) with the homozygosity of window_373 (the HLA region), once additive effects are accounted for.

In conclusion, the additive effects of common haplotypes do confound the sought-after recessive effects of ROH and must be accounted for if ROH islands are present. Never-the-less, homozygosity in the HLA region appears to reduce lymphocyte percentage and total white blood cell count, independently of any additive effects.

5.3.1.3 Regional associations in UK Biobank with ROH islands filtered out

In the full UK Biobank cohort, ROH islands were removed by excluding all ROH less than 3 cM in length. This results in a relatively even distribution of ROH (Figure 19), as expected of inbreeding defined as identity by decent from a fixed base generation. By this definition, all base generation alleles have an equal probability of appearing in ROH, and therefore, for additive alleles, inbreeding does not cause changes to the trait mean. Therefore, window-trait associations in this filtered dataset are expected to reflect genuinely recessive effects. Unfortunately, the average homozygosity in UK Biobank is greatly reduced by removing short ROH, and only 12 window-trait combinations reach experiment wise significance (p -value $< 0.05/994*155$) in this filtered dataset (Table 4).

Table 4: **Significant window-trait associations in UKB with ROH islands removed.**

Window	Chr	Start (Mb)	End (Mb)	Trait	Effect of 3 Mb window in ROH (Trait SD)	p -value
window_526	9	0	3	Polycythaemia vera	3.57	1.1×10^{-8}
window_526	9	0	3	Lymphocytes (%)	-0.63	2.7×10^{-7}
window_527	9	3	6	Polycythaemia vera	3.07	1.1×10^{-8}
window_527	9	3	6	Lymphocytes (%)	-0.59	2.5×10^{-8}

window_528	9	6	9	Polycythaemia vera	2.78	3.3×10^{-8}
window_528	9	6	9	Lymphocytes (%)	-0.56	9.1×10^{-9}
window_529	9	9	12	Polycythaemia vera	2.46	4.7×10^{-8}
window_529	9	9	12	Lymphocytes (%)	-0.47	5.9×10^{-8}
window_530	9	12	15	Polycythaemia vera	2.75	3.0×10^{-8}
window_530	9	12	15	Lymphocytes (%)	-0.54	6.2×10^{-8}
window_531	9	15	18	Polycythaemia vera	2.15	2.9×10^{-7}
window_532	9	18	21	Polycythaemia vera	2.38	1.7×10^{-7}

Two traits, Polycythaemia vera and Lymphocyte %, are significantly associated with the homozygosity of several windows near the end of the short (p) arm of chromosome 9. Finding several adjacent windows almost equally associated with the same phenotypes seems to defy the hypothesis of autozygosity revealing the recessive effects of a specific locus, and therefore these associations are investigated further in the following section.

The absence of significant associations with other traits may be partly explained by the severity of multiple testing ($155 \times 994 = 154,070$ tests), and also the low average autozygosity in UK Biobank (mean $F_{ROH} = 0.0006$ when ROH < 3cM removed). Nevertheless, the absence of locally significant ROH effects (for almost all traits) implies that the effects of autozygosity are at least moderately evenly distributed throughout the genome.

5.3.1.3.1 Associations with ROH at short (p) end of Chromosome 9

228 samples in UK Biobank have ROH starting at the first genotyped SNP on chromosome 9, ranging between 1.52 Mb and 40.8 Mb in length (mean = 9.3 Mb). Among these 228 samples there is a significant enrichment of a rare blood cancer, Polycythaemia vera (PV): 27 have an ICD10 code for PV (12%) compared to 0.1% in the remainder of the cohort. The previously estimated effect of F_{ROH} on

Polycythaemia vera ($\hat{\beta}_{F_{ROH}}$) is largely due to this locus, since, if these 228 samples are excluded, $\hat{\beta}_{F_{ROH}}$ is reduced to 46.1% of its original estimate, and is no longer statistically significant. In addition, having ROH at this locus is significantly associated (p -value < 0.05/155) with changes in several other blood, and disease, traits (Table 5).

Table 5: Traits associated with ROH at the p-end of chromosome 9. Effect estimates and standard errors are normalised by trait intra-sex standard deviations. The percentage of the total effect explained was calculated by re-estimating $\beta_{F_{ROH}}$ while excluding the 228 samples with ROH at this locus.

Trait	Effect of ROH at 9p24.3 (Trait SD)	Standard Error (Trait SD)	p -value	Fraction of $\beta_{F_{ROH}}$ explained by this locus (%)
Polycythaemia vera	3.43	0.63	4.2×10^{-8}	53.9
Lymphocyte %	-0.60	0.12	6.1×10^{-7}	7.5
Erythrocyte distribution width	0.88	0.18	1.4×10^{-6}	4.6
High light scatter reticulocyte percentage	0.45	0.10	2.7×10^{-6}	4.9
Immature reticulocyte fraction	0.37	0.08	8.1×10^{-6}	5.9
Number of ICD10 codes	0.37	0.09	3.2×10^{-5}	2.6
Cystatin C	0.31	0.08	2.0×10^{-4}	1.8

Polycythaemia vera (PV) is a rare blood cancer affecting the bone marrow, and causing it to make too many red blood cells. At least 95% of PV cases are associated with mutations in the Janus Kinase gene (*JAK2*),¹⁵⁵ which is located approximately 5 Mb from the end of the short (p) arm of chromosome 9. Initially, the association between PV and ROH at this locus may suggest exposed recessive effects in *JAK2*, however, copy-neutral somatic loss of heterozygosity (cn-LoH) is known to be associated with haematological cancers.¹⁵⁶ Specifically, “*The JAK2 V617F mutation is frequently observed in classic MPN [myeloproliferative neoplasia], and disease*

*progression is associated with biallelic acquisition of the mutation through mitotic recombination and aUPD [acquired uniparental disomy]”.*¹⁵⁶ It therefore seems likely that the associations observed here are a consequence of PV causing cn-LoH (in blood), rather than the more interesting case of germline autozygosity causing disease.

5.3.2 Associations between window effects and genomic features

5.3.2.1 Gene density

Despite the paucity of locally significant window-trait associations, it may be possible to identify factors underlying any variation of ROH effects throughout the genome. Specifically, if inbreeding depression is principally caused by the homozygosity of variants in, or near, protein-coding genes, then the estimated effects of autozygosity may be greater in gene-dense regions.

This was initially explored in the meta-analysed ROHgen window-trait results described in section 5.3.1.1. The number of protein-coding genes contained by each window was determined, as described in section 5.2.2.1, and plotted on the x-axis against estimated window effect on height, on the y-axis (Figure 76).

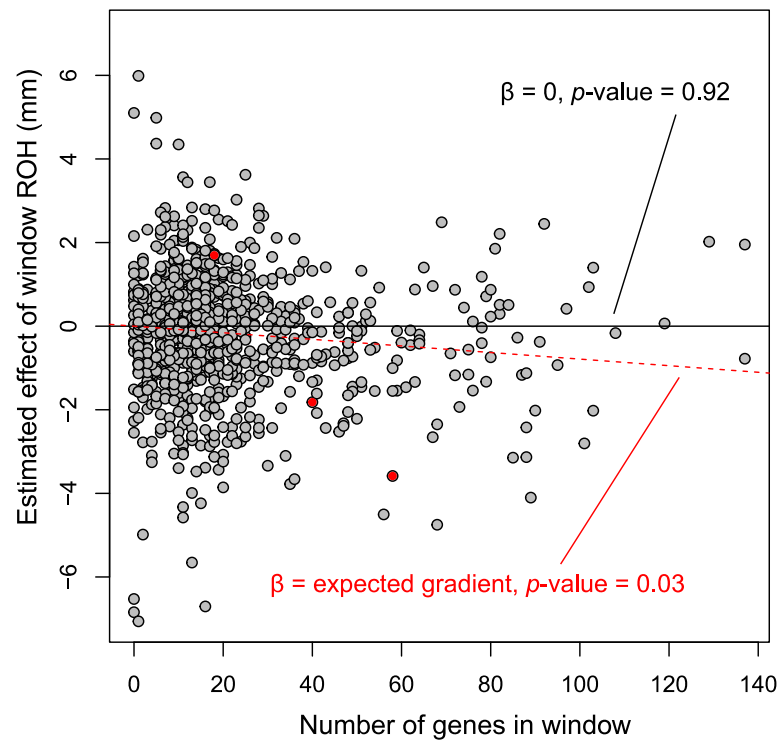


Figure 76: **Estimated effect on height of 941 autosomal windows plotted against gene count in the window.** Standard errors are not shown, but were used to weight the best-fit regression line shown in black. The expected gradient assuming window effects are, on average, proportional to the number of genes contained is shown in dashed red. A p -value for the hypothesis that the best fit gradient is equal to 0 is shown in black, and p -value for the hypothesis that the best fit gradient is equal to the expected gradient is shown in red. Three windows with statistically significant effects (Figure 72) are shown in red.

There is no significant association between the number of genes in a window and the effect of that window on height (p -value = 0.92). In contrast, the best fit gradient does differ significantly from the gradient expected if effects are proportional to the number of genes in a window. This slightly surprising result implies that the effect of inbreeding depression on height is not concentrated in gene dense regions of the genome.

5.3.2.2 Additive variance explained

Although there appears to be no association between inbreeding depression (on height) and gene density, this analysis can be refined since many of genes underlying the variation in height are known. The approximate additive variance

explained by each window was calculated (5.2.2.2) and plotted against estimated the estimated effect on height for 941 autosomal windows (Figure 77).

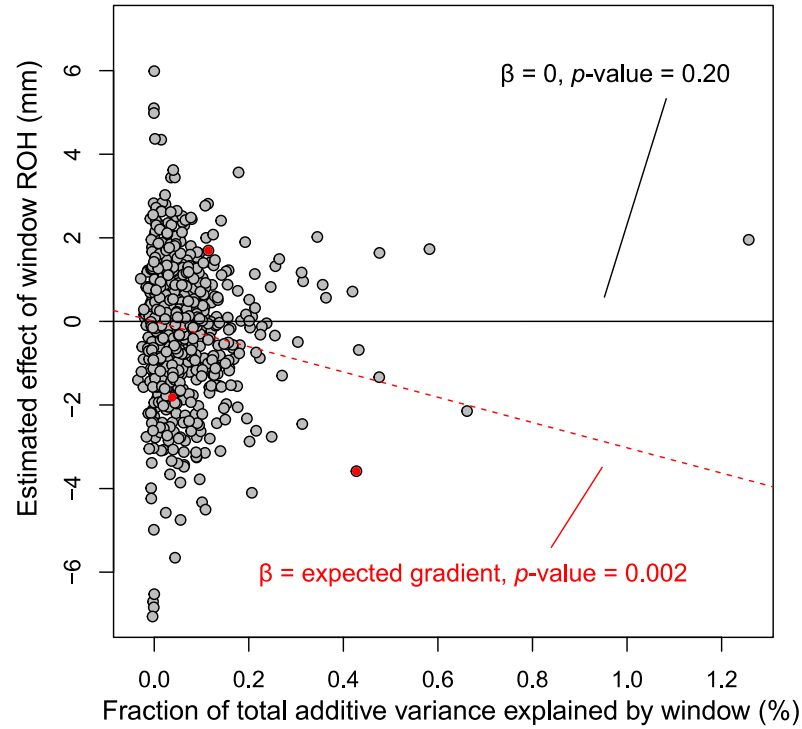


Figure 77: **Estimated effect on height of 941 autosomal windows plotted against the estimated additive variance explained by that window.** Standard errors are not shown, but were used to weight the best-fit regression line shown in black. The expected gradient assuming window effects are, on average, proportional to the additive variance explained is shown in dashed red. A p -value for the hypothesis that the best fit gradient is equal to 0 is shown in black, and p -value for the hypothesis that the best fit gradient is equal to the expected gradient is shown in red. Three windows with statistically significant effects (Figure 72) are shown in red.

Again, there is no significant association between the estimated window effect on height and the additive variance explained by SNPs in that window. If all variants affecting height have the same dominance coefficient (h), the magnitude of inbreeding depression (β_F) explained by a window should be approximately proportional to the additive variance (V_A), since $V_A = 2pq\alpha^2$ and $\beta_F = -2dpqF$, but the best fit gradient differs significantly from this hypothesis. In summary, it appears that the genetic variants causing inbreeding depression on height are not the same as those responsible for most of additive variation of height.

In contrast to Figure 76, it is interesting to note that the additive variance of height explained by a window is strongly associated ($p\text{-value} = 3 \times 10^{-30}$) with the number of genes contained by that window (Figure 78).

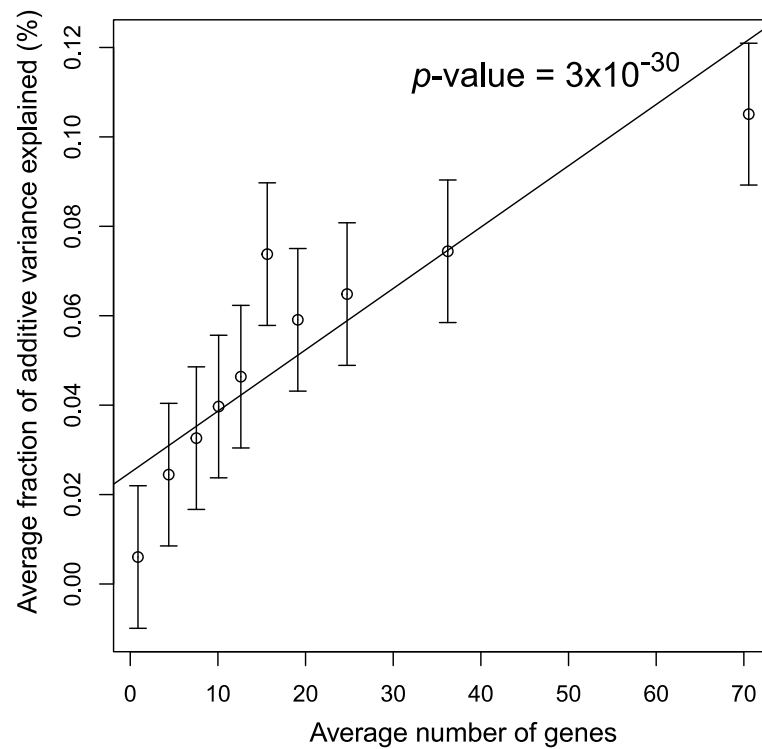


Figure 78: **Average additive variance of height explained by a window plotted against number of genes contained.** Each window was assigned to one of 10 equally sized bins and the mean additive variance calculated in each bin. 95% confidence intervals are shown for each bin.

5.3.3 Distribution of ROH effects in UK Biobank

5.3.3.1 By Chromosome

In UK Biobank, the effect of ROH on the *multivariate* trait was estimated for each chromosome, as described in section 5.2.3. The estimated effect sizes, per Mb of ROH, are shown in Figure 79, below.

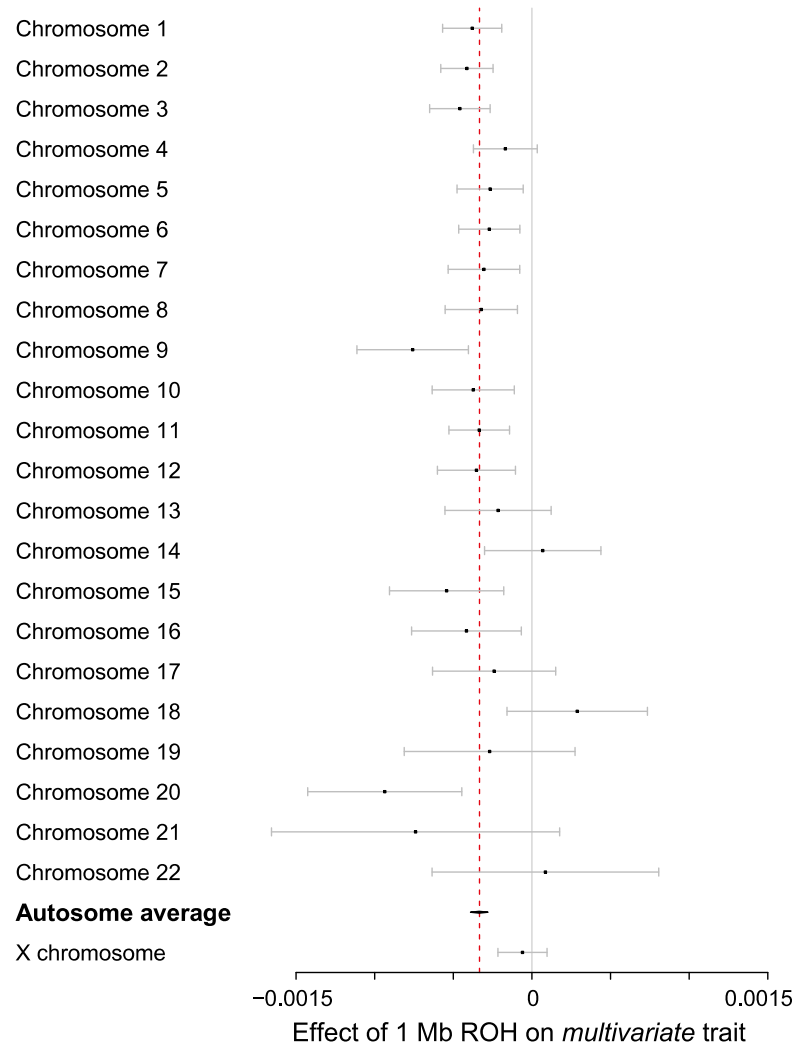


Figure 79: **Effect of 1 Mb of ROH on the UKB *multivariate* trait estimated by chromosome.** The genome-wide average effect of 1 Mb of ROH is shown as a dashed red line. 95% confidence intervals are shown in grey.

ROH on 7 out of 23 chromosomes are independently associated with a significant (p -value $< 0.05/23$) reduction in the *multivariate* trait and, as expected, the meta-analysed average of the autosome effects (-0.000335 , p -value $= 3 \times 10^{-29}$) is close to $\hat{\beta}_{F_{ROH}}/3000$ (-0.000333 , p -value $= 2 \times 10^{-148}$), since F_{ROH} is the total length of ROH in

Mb divided by 3000, equation (4). Although there is a small amount of heterogeneity between the autosome estimates (p -value = 0.04), none of these differ significantly (p -value < 0.05/22) from the genome wide average. The effect of ROH on the X chromosome, however, is significantly less than the average autosome effect (p -value = 0.0006) and does not differ significantly from 0 - perhaps because X-linked recessive variants are exposed to selection in males.

This relative consistency in ROH effect across the autosomes comes despite a ten-fold variation in gene density. For example, chromosome 19 has an average of 23.4 genes per Mb compared to 2.7 genes per Mb on chromosome 13. Again, if inbreeding depression were caused by the homozygosity of genes, an association between estimated effect and gene density might have been expected (Figure 80).

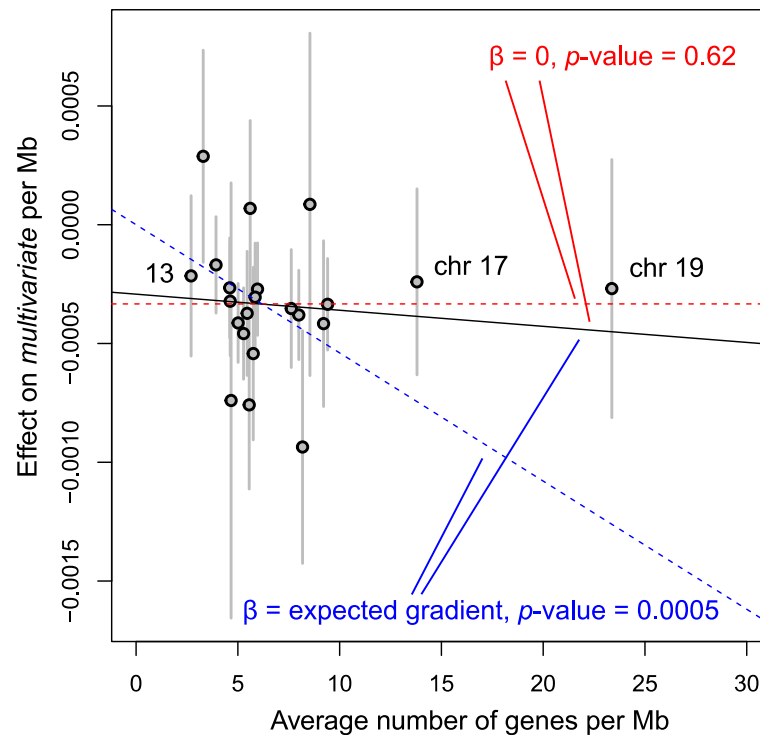


Figure 80: **Estimated effect of 1 Mb ROH plotted against gene density for all 22 autosomes.** 95% confidence intervals are shown in grey. A best fit regression line, weighted by standard errors, is shown in black. The genome wide average effect of 1 Mb ROH is shown in dashed red, and the gradient expected if effects are proportional to gene density is shown in dashed blue. A p -value for the hypothesis that the best fit gradient is equal to 0 is shown in red, and p -value for the hypothesis that the best fit gradient is equal to the expected gradient is shown in blue.

Chromosomes 19 and 17 have gene densities substantially above average, and yet ROH on these chromosomes have effects similar to the genome wide average. A best fit regression line is consistent with the hypothesis that gene density has no effect on estimated effect (p -value = 0.62), but is not consistent with the hypothesis that effects are proportional to gene density (p -value = 0.0005).

5.3.3.2 X chromosome effects

The effect of X Chromosome ROH ($\beta_{F_{ROH_X}}$) was estimated (in women only) for all 153 significant UKB traits measured in women, as described in section 5.2.3 (Figure 81).

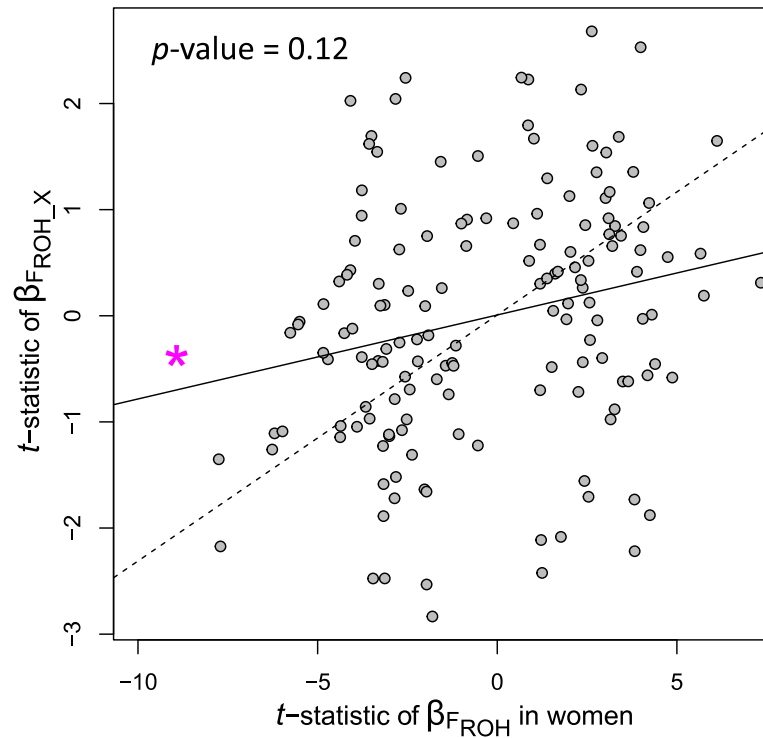


Figure 81: **t-statistic of $\hat{\beta}_{F_{ROH_X}}$ plotted against the t-statistic of $\hat{\beta}_{F_{ROH}}$ for 153 significant UKB traits.** A Deming regression best-fit line is shown in solid black and has a gradient which does not differ significantly from zero (p -value = 0.12). The expected regression gradient for equal effect sizes per Mb is shown in dashed black. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t-statistics have been divided by two in order to make best use of the display area.

$\hat{\beta}_{F_{ROH_X}}$ does not differ significantly from zero for any of the 153 individual traits, and the t-statistics of $\hat{\beta}_{F_{ROH_X}}$ are not significantly correlated with the t-statistics of

$\hat{\beta}_{F_{ROH}}$ (p -value = 0.12). For the *multivariate* trait, the effect of 1 Mb ROH on the X chromosome is 18% [95% CI = -29% to 65%, p -value = 0.45] of the effect of 1 Mb ROH on the autosomes.

In summary, the effect of X chromosome ROH is consistent with zero for the traits tested and, on average, is significantly smaller than the effect of autosomal ROH (p -value = 0.0006).

5.3.3.2.1 Effect of F_{ROH_X} without deletions removed

The effect of X chromosome ROH was also estimated for all ROHgen traits, and the results initially appear to conflict with the UKB results reported in section 5.3.3.2.

Indeed, $\hat{\beta}_{F_{ROH_X}}$ differs significantly from zero for seven ROHgen traits, all with effect sizes greater than $\hat{\beta}_{F_{ROH}}$ on the autosomes (Figure 82).

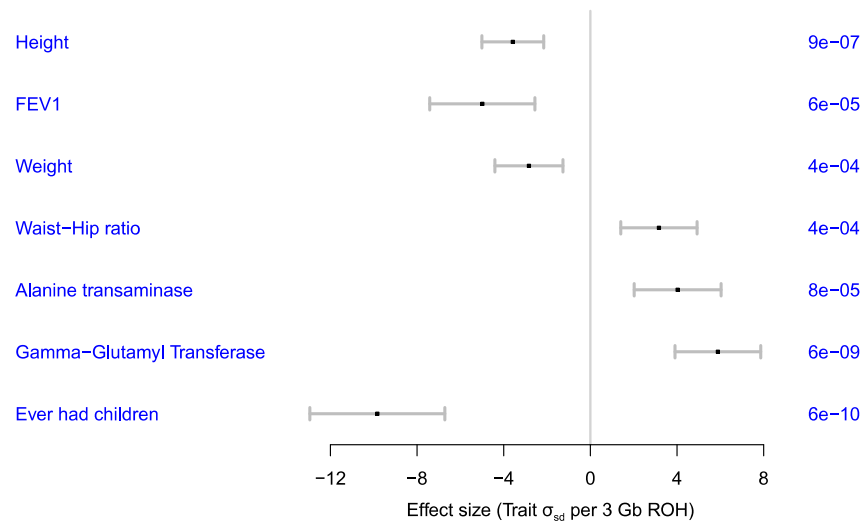


Figure 82: **Seven ROHgen traits significantly associated with F_{ROH_X} .** To facilitate comparison with Figure 30, effect sizes are shown in units of trait standard deviation (in women) per 3 Gb of ROH. 95% confidence intervals are shown in grey.

However, hemizygous deletions were not explicitly removed in ROHgen, and although section 3.3.5.3 shows that this makes very little difference for the autosomes, the traits affected in Figure 82 give a hint that deletions do make an important contribution to X chromosome ROH:

Turner syndrome is a genetic disorder, affecting approximately 1 in 2000 women, which results when one copy of the X chromosome is either partially or entirely

missing. Women with Turner syndrome may have a wide range of symptoms including: short stature, infertility, hearing loss¹⁵⁷, and elevated liver enzymes¹⁵⁸. In UKB, 424 women (1 in 440) were identified as having any deletion on the X chromosome, and 81 (1 in 2280) as having a deletion greater than 20 Mb in length. In the hospital admissions phenotype, 37 women have an ICD10 code indicating Turner syndrome, of which 24 were identified here as carrying X chromosome deletions (ranging from 43 to 149 Mb - for reference the entire X chromosome length is 156 Mb).

All deletions were excluded from the calculation of F_{ROH_X} in section 5.3.3.2, although not from UKB's contribution to ROHgen (as per the ROHgen analysis plan). If Figure 81 is reproduced without removing X chromosome deletions, very different conclusions would be drawn (Figure 83).

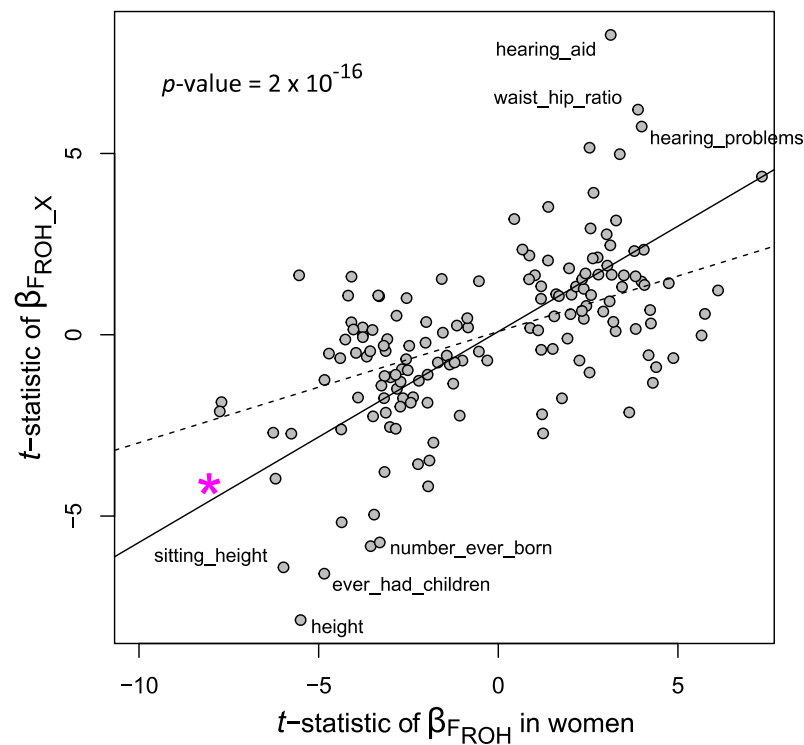


Figure 83: **t -statistic of $\hat{\beta}_{FROH_X}$ plotted against the t -statistic of $\hat{\beta}_{FROH}$ for 153 significant UKB traits, without removing hemizygous deletions.** A Deming regression best-fit line is shown in solid black and has a gradient which differs significantly from zero (p -value = 2×10^{-16}). The expected regression gradient for equal effect sizes per Mb is shown in dashed black. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

When X chromosome deletions are not removed, F_{ROH_X} is significantly associated with 17 UKB traits in the domains of: body size and strength, reproductive success, hearing, and wellbeing (Figure 83). Considering the *multivariate* trait, ROH on the X chromosome now have an average effect which is 1.71 times [95% CI = 1.30 to 2.11] greater than the effect of ROH on the autosomes (per Mb). In view of the traits affected, it is likely that these associations with F_{ROH_X} are caused by Turner syndrome cases, but, in any case, it is clearly important to remove hemizygous deletions to accurately estimate the effect of X chromosome autozygosity.

Although it is not possible to be certain without returning to the raw data, it seems likely that the trait associations with F_{ROH_X} in ROHgen (Figure 82) are confounded by the inclusion of Turner syndrome cases.

5.3.4 F_{ROH} calculated at specific genomic annotations

5.3.4.1 F_{ROH} calculated at protein-coding genes

Since additive genetic effects are concentrated in gene-dense regions (Figure 78), the enrichment of inbreeding depression at protein-coding genes was investigated in UK Biobank. Appendix E.3 shows that inbreeding depression is proportional to the effect-size weighted average of the homozygosity of causal variants. If these causal variants are enriched at, or near genes, then the homozygosity of genes, F_{GENES} (section 5.2.4.1), should be a better predictor inbreeding depression than F_{ROH} . This was assessed in bivariate models for all 155 significant UKB traits (Figure 84).

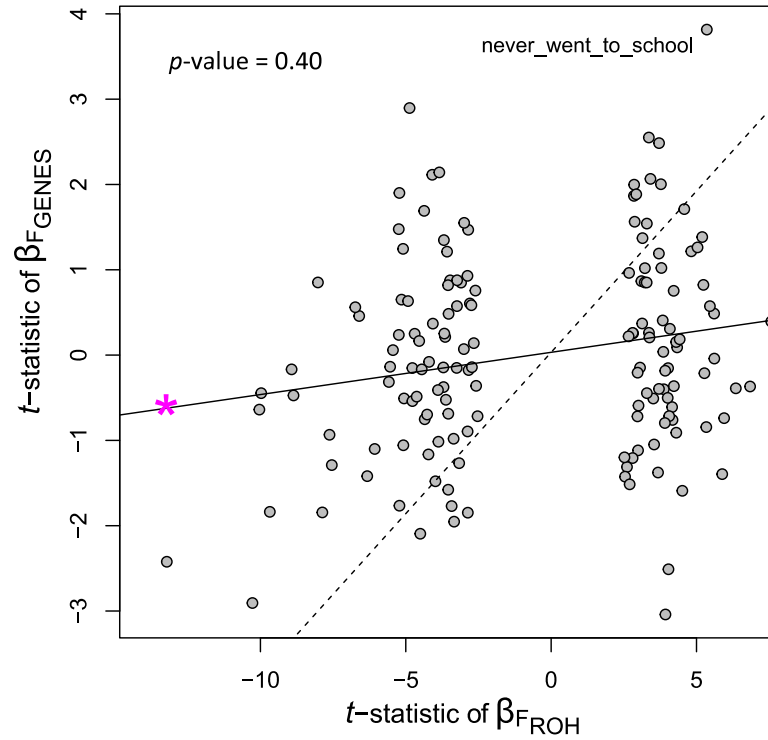


Figure 84: **t -statistic of $\hat{\beta}_{F_{GENES}}$ plotted against the t -statistic of $\hat{\beta}_{F_{ROH}}$ for 155 significant UKB traits.** A Deming regression best-fit line is shown in solid black and has a gradient which does not differ significantly from zero (p -value = 0.40). The expected regression gradient for equal effect sizes is shown in dashed black. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

In bivariate models with F_{ROH} , $\hat{\beta}_{F_{GENES}}$ is not significantly different from zero for 154 out of 155 traits. The regression of t -statistics of $\hat{\beta}_{F_{GENES}}$ (Figure 84) implies that average effect sizes are just 9% [95% CI = -15 to 33%] of $\hat{\beta}_{F_{ROH}}$. In addition, the effect of F_{GENES} on the well-powered *multivariate* trait (which is itself an average of all traits) is non-significant (p -value = 0.45), while effect of F_{ROH} remains significant (p -value = 3.4×10^{-13}) in the bivariate model. One trait (*never_went_to_school*), does have a significant association (p -value = 1.5×10^{-4}) with F_{GENES} and, although the reason for this is unknown, chance cannot be completely excluded (significance p -value = $3.2 \times 10^{-4} = 0.05/155$).

In summary, there is no evidence of an enrichment of inbreeding depression effects at protein coding genes. Although the null hypothesis of no enrichment can never be proven, the results disprove the hypothesis of an enrichment proportional to gene density: if causal variants are only in randomly selected genes, $\hat{\beta}_{F_{GENES}}$ should equal the univariate $\hat{\beta}_{F_{ROH}}$ (see simulation in section 5.3.4.4), which is significantly not the case (p -value = 3.4×10^{-14}) for the *multivariate* trait.

5.3.4.2 F_{DEL} calculated at protein-coding genes

It was previously shown, in section 4.3.3, that the effects of hemizygous deletions (measured by F_{DEL}) are qualitatively similar to the effects of ROH (measured by F_{ROH}). To determine whether the effects of hemizygous deletions can be attributed to genes, the F_{GENES} analysis described above was repeated for hemizygous deletions. In exactly the same way as F_{GENES} corresponds to F_{ROH} , a deleted gene fraction F_{DEL_GENES} was calculated to correspond to F_{DEL} . Bivariate models were then fitted for all 155 significant UKB traits, and results are plotted in Figure 85.

$$trait \sim F_{DEL} + F_{DEL_GENES} + covariates \quad (48)$$

Unlike for ROH (Figure 84), there is a significant association (p -value = 3×10^{-14}) between the t -statistics of $\hat{\beta}_{F_{DEL_GENES}}$ and the univariate t -statistics of $\hat{\beta}_{F_{DEL}}$. The gradient of the best-fit regression line is almost identical to the expectation if $\beta_{F_{DEL_GENES}}$ equals the univariate $\beta_{F_{DEL}}$, implying that the effects of deletions are entirely caused by genes. In addition, 13 individual traits, and the *multivariate* trait, are significantly associated with F_{DEL_GENES} .

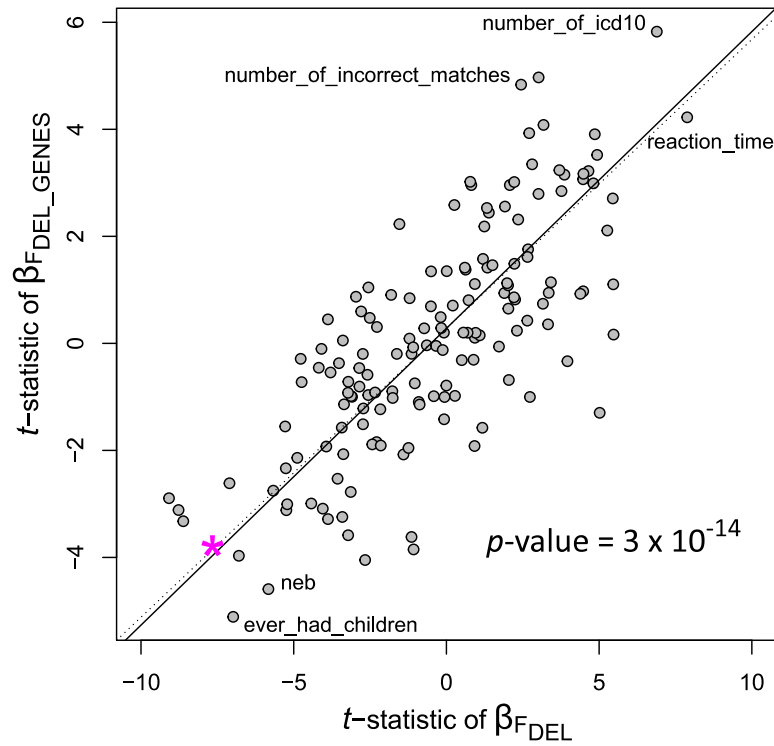


Figure 85: **t -statistic of $\hat{\beta}_{FDEL_GENES}$ plotted against the t -statistic of $\hat{\beta}_{FDEL}$, for 155 significant UKB traits.** A Deming regression best-fit line is shown in solid black and has a gradient which differs significantly from zero (p -value = 3×10^{-14}). The expected regression gradient for equal effect sizes is shown in dashed black. The *multivariate* trait (not included in the regression) is shown as * but, for this point only, both x and y axis t -statistics have been divided by two in order to make best use of the display area.

These results not only show that haploinsufficiency of protein-coding genes is likely responsible for the observed effects of hemizygous deletions, but also confirm that the bivariate method ($trait \sim F_{ROH} + F_{GENES}$) works as expected. Since the same method was used for ROH, and F_{ROH} has slightly greater power than F_{DEL} in UKB (Figure 70), the conclusion that ROH effects are not associated with gene density is strengthened, while never-the-less contrasting with the conclusion for deletions.

5.3.4.3 F_{ROH} calculated at other genomic annotations

Although no enrichment of ROH effects was found with gene density, the effect of several other genomic features was estimated (see methods section 5.2.4.2).

Results of bivariate models of *multivariate* $\sim F_{ROH} + F_{ANNOTATION}$ are shown in Table 6.

Table 6: Estimated effects on the *multivariate* trait of F calculated at 8 genomic annotations ($F_{ANNOTATION}$), in a bivariate model with F_{ROH} .

Genomic annotation	F_{ROH}			$F_{ANNOTATION}$		
	$\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value
Protein coding genes	-0.91	0.13	3.4×10^{-13}	-0.092	0.12	0.45
pLoF intolerant genes	-0.90	0.10	2.7×10^{-16}	-0.098	0.10	0.31
Imprinted genes	-1.00	0.06	1.1×10^{-60}	-0.004	0.04	0.91
nDNA Mitochondrial genes	-0.98	0.08	4.7×10^{-29}	-0.022	0.08	0.78
Autosomal recessive disease genes	-0.97	0.08	3.4×10^{-30}	-0.026	0.07	0.71
Copy Number Variants	-0.98	0.08	2.5×10^{-33}	-0.022	0.07	0.76
1000 genomes SNP diversity	-1.04	0.32	0.001	0.038	0.31	0.90
De novo mutation rate	-0.75	0.46	0.10	-0.254	0.46	0.58

For all eight annotations, $\hat{\beta}_{F_{ANNOTATION}}$ does not differ significantly from zero, and conditioning $\hat{\beta}_{F_{ROH}}$ on $F_{ANNOTATION}$ does not significantly change its estimated value

(univariate $\hat{\beta}_{F_{ROH}} = -1$). In the bivariate model, $\hat{\beta}_{F_{ROH}}$ remains significant for all annotations, apart from the de novo mutation rate, where standard errors are too large to draw any conclusions (because F_{ROH} and $F_{ANNOATION}$ are very strongly correlated for this feature, $r=0.996$).

In summary, there is no evidence of any enrichment of ROH effects at any of the genomic annotations tested. In addition, for seven out of eight annotations the effects of F_{ROH} cannot solely be explained by effects proportional to that annotation.

5.3.4.4 *Simulated inbreeding depression at gene locations*

The slightly unexpected finding that inbreeding depression effects are not enriched at known gene locations raises the question whether the results from the bivariate model, $trait \sim F_{ROH} + F_{GENES}$, are being correctly interpreted. It seems intuitive that $\hat{\beta}_{F_{ROH}}$, conditional on F_{GENES} , should be zero when causal variants are only located inside genes - if the homozygosity of genes is known and accounted for, the fraction of the remainder of the genome in ROH should be of no importance. What is less clear, however, is what should be expected if causal variants are some distance from the coding exons used in the calculation of F_{GENES} , as is the case for promoters and enhancers, for example. To help resolve this uncertainty, a trait with statistical properties similar to the *multivariate* trait was simulated for a range of different assumptions about the location, and effect size distribution, of causal variants. The simulation method is described in section 5.2.5, and the results are summarised in Table 7, below.

Table 7: **Estimated effects of F_{ROH} and F_{GENES} on a simulated *multivariate* trait, under a range of different assumption.** Each set of assumptions was simulated several times (column 7), since the results depend on the location (relative to known gene positions) of the randomly chosen causal loci.

Number of randomly distributed loci	Number of exonic loci	Effect size distribution	Jitter on exonic loci (Mb)	$\bar{\beta}_{F_{ROH}}$	$\bar{\beta}_{F_{GENES}}$	Number of sets of simulated causal loci	S.D. of $\hat{\beta}_{F_{GENES}}$
20	0	Uniform	0	-0.96	-0.04	199	0.50
100	0	Uniform	0	-1.06	0.05	54	0.26
500	0	Uniform	0	-1.02	0.01	10	0.12
2000	0	Uniform	0	-0.97	-0.02	2	0.05
0	20	Uniform	0	0.13	-1.24	192	1.02
0	100	Uniform	0	-0.06	-1.04	54	0.44
0	500	Uniform	0	0.11	-1.21	11	0.27
0	2000	Uniform	0	-0.21	-0.98	2	0.13
20	20	Uniform	0	-0.54	-0.52	104	0.56
100	100	Uniform	0	-0.51	-0.54	26	0.26
500	500	Uniform	0	-0.54	-0.53	4	0.18
2000	2000	Uniform	0	-0.45	-0.63	1	0.02
0	500	Uniform	0	0.11	-1.21	11	0.27
0	500	Uniform	1	-0.04	-1.03	10	0.22
0	500	Uniform	2	-0.09	-0.99	10	0.19
0	500	Uniform	5	-0.42	-0.65	9	0.17
0	500	Uniform	10	-0.40	-0.61	10	0.13
0	20	Exponential	0	-0.02	-1.07	150	1.29
0	100	Exponential	0	0.09	-1.18	49	0.66
0	500	Exponential	0	-0.08	-0.99	9	0.17
0	2000	Exponential	0	0.12	-1.23	2	0.10

As expected, when causal loci are randomly distributed throughout the genome F_{ROH} is a better predictor of inbreeding depression than F_{GENES} , no matter how many (or few) causal loci are simulated. However, it should be noted that $\bar{\beta}_{F_{ROH}}$ and $\bar{\beta}_{F_{GENES}}$ are averages from many sets of simulated causal loci (column 7 in Table 7), and when the number of simulated loci is small (e.g., twenty, as in row 1 of Table 7),

the standard deviation of $\hat{\beta}_{F_{GENES}}$ becomes large, because a small number of randomly distributed causal loci may, by chance, be located in unexpectedly gene dense regions.

Similarly, when causal loci are simulated only within protein-coding exons, F_{GENES} is a better predictor of inbreeding depression than F_{ROH} , again, no matter how many (or few) causal loci are simulated. The average $\hat{\beta}_{F_{GENES}}$ is slightly greater than the summed effect of all simulated causal loci (-1) since the variance of F_{GENES} is approximately 12% greater than the variance of F_{ROH} . Also, when causal loci are simulated in both exons and randomly distributed, $\bar{\beta}_{F_{ROH}}$ and $\bar{\beta}_{F_{GENES}}$ are consistent with the simulated split.

Gene expression promoters and enhancers may be located some distance from the genes themselves and, when this is simulated by adding a random jitter to the location of all causal loci, a small reduction in $\bar{\beta}_{F_{GENES}}$ is seen. However, when causal loci are up to 1 Mb away from exons, F_{GENES} is still a much better predictor of inbreeding depression than F_{ROH} . Even when causal loci are up to 10 Mb away from genes, far greater than most likely separations, F_{GENES} and F_{ROH} are of similar utility for predicting inbreeding depression.

Finally, the impact of non-uniform effect size distributions was investigated by simulating causal loci with exponentially distributed effect sizes. Again, F_{GENES} remains a better predictor of inbreeding depression than F_{ROH} when effects are located at, or near genes. However, for the same number of causal loci, the standard deviation of $\hat{\beta}_{F_{GENES}}$ is increased by the non-uniform distribution of effects. This is similar to the consequence of reducing the number of causal loci, since the location of the largest effects is of inflated importance.

In summary, F_{GENES} is a better predictor of simulated inbreeding depression than F_{ROH} , even when causal loci are relatively distant from gene locations. If causal genes are few in number, they may, by chance, be located outside gene dense regions, but this is unlikely to happen for a larger number of causal genes (above approximately a few hundred).

5.4 Discussion

5.4.1 ROH effects at specific genomic locations

Identifying the individual genetic variants causing inbreeding depression would improve our understanding of genetic architecture and, in order to progress towards that goal, this chapter has investigated the effect of ROH at specific genomic locations.

However, in light of several recent papers reporting significant local associations with ROH^{80,99}, an important conclusion reached in this chapter is that the method commonly used for testing localised recessive effects is potentially confounded by the additive effects of common haplotypes forming ROH islands. In UK Biobank, 19 out of 21 local ROH associations which were identified in the naïve analysis, were rendered non-significant once additive effects were accounted for. Given that many of the previously reported ROH associations occur near GWAS hits for the same traits, it seems unproven that these are caused by the claimed recessive effects.

In this chapter, two methods are suggested to eliminate confounding by additive effects. Either additive effects may be accounted for by fitting genetic markers as linear covariates or, alternatively, the bias caused by additive effects can be mitigated by removing ROH islands. The first method retains the high frequency of ROH in islands and therefore keeps good statistical power at these loci. However, there may remain some concern that additive effects have not been fully accounted for by the SNPs used as covariates, and any recessive effects remaining at ROH islands are most likely caused by non-additive effects of a common haplotype. As discussed previously, since the homozygosity of common variants is caused largely by chance, effects associated with this chance variation might be better described as directional dominance, rather than inbreeding depression. The second method of controlling additive effects (i.e., filtering out ROH islands), has the advantage that significant effects can safely be described as inbreeding depression, since the remaining ROH are attributable to consanguinity. However, removing ROH islands greatly reduces statistical power in outbred cohorts (such as UK Biobank), although

the impact is likely to be much smaller in consanguineous cohorts. Ideally, the localised effects of ROH should be investigated in well-powered cohorts (i.e., consanguineous, or extremely large) with ROH islands removed.

Despite the complications described, two traits (lymphocyte percentage, and white blood cell count) are significantly reduced by ROH at the human leukocyte antigen (HLA) locus, even once additive effects are accounted for. The effects of ROH > 3 cM (which are believed to be attributable to recent consanguinity) are directionally consistent with the significant effects (which are mostly attributable to the homozygosity of relatively common HLA haplotypes), but are themselves non-significant. These effects appear to be consistent with the observation of inbreeding depression on disease resistance in other species, and with the hypothesis that HLA heterozygosity plays a role in immune defence against infectious disease¹⁵⁹.

As well as considering additive effects, any future attempts at homozygosity mapping in large population cohorts should be aware of the possibility of reverse causality. In UKB, the only significant associations found outside of ROH islands, appear to be caused by a rare blood cancer, polycythaemia vera. In this scenario, the disease causes a loss-of-heterozygosity, which is detected by ROH, rather than the alternative hypothesis of germline ROH causing disease. It is also recommended that future studies control for genome wide ROH (F_{ROH}) while estimating localised effects. This has typically not been done in previous studies but, if not accounted for, correlations between local ROH and F_{ROH} may inflate local effect estimates.

5.4.2 Genome wide distribution of inbreeding depression

5.4.2.1 X Chromosome effects

Although previous studies of ROH in humans have measured X chromosome autozygosity⁷⁹, this thesis appears to be the first to report the effects of X chromosome autozygosity on human complex traits. This may be partly a consequence of sample size, since the effects of X chromosome ROH can only be estimated in half the population, i.e., women.

Theory predicts that the effect of autozygosity may be lower on the X chromosome than on the autosomes because X-linked recessive variants are exposed to selection in males, and are therefore selected against more strongly than autosomal recessive variants³³. There is empirical support for this prediction in *Drosophila melanogaster*, where egg-to-adult viability is substantially affected by autosomal homozygosity, but not by X chromosome homozygosity^{26,160}.

In section 5.3.3.1, X-chromosome autozygosity is shown to have a significantly smaller average effect than an equal length of autozygosity on the autosomes. In UK Biobank, the average effect of X chromosomal autozygosity is 18% [95% CI = -29% to 65%, p -value = 0.45] of the effect of autosomal autozygosity. While this estimate does not differ from zero, larger studies will be required to reduce the confidence interval, and determine if any individual traits are affected by X chromosomal autozygosity. It might be noted that apparently deleterious X-linked recessive conditions (e.g., colour-blindness, haemophilia) do occur at substantial frequencies in modern populations, so the existence of X chromosome inbreeding depression cannot be precluded.

From a practical perspective, future studies should take care to remove samples with X-linked deletions. In particular, the large hemizygous deletions associated with Turner syndrome may substantially bias estimates of X chromosome autozygosity.

5.4.2.2 *Distribution of effects within autosomes*

Although inbreeding depression has been recorded in many species, including humans, the causal genetic variants remain largely unidentified. In the absence of specific variants, discovering trends in the genome-wide distribution of inbreeding effects would help elucidate the genetic architecture. Also, as other authors have noted, any correlation between inbreeding depression and a genomic feature would reduce the possibility of confounding⁷³.

However, the analyses in this chapter reveal a remarkably even distribution of inbreeding effects throughout the autosomes. There is no evidence any non-

uniformity of inbreeding effect per unit of physical genome length (within the statistical power of the available sample). Most notably, there appears to be no association between gene density and inbreeding depression, despite sufficient statistical power to detect the expected association. Initially, at least, this seems very surprising, since protein-coding genes are typically considered the functional units of genetic variation, and additive genetic variation is strongly associated with genes (Figure 78). However, the theory of mutation-selection balance offers a possible explanation.

Section 1.2.2.1 shows that, in mutation-selection balance, the magnitude of inbreeding depression does not actually depend on the strength of selection (s). Deleterious variants in highly constrained regions, e.g., genes, reach lower equilibrium frequencies than those in less constrained regions, resulting in equal inbreeding depression for all levels of constraint. This provides a theoretical explanation for the observed uniformity of effects, although for this theory to hold there does need to be some degree of selection at all loci. This initially appears inconsistent with reports that a large fraction of the human genome is non-functional, although selection at linked sites may extend constraint some distance from causal variants. Similarly, the results in this chapter do not necessarily exclude the possibility that all ID causal variants occur in protein coding genes. If causal variants in gene dense regions are kept at lower frequency than those in gene sparse regions, there may be no association between ID and gene density, even if the causal variants themselves occur within genes.

In apparent contrast to the findings here, a very recent study (also in UK Biobank) reported that inbreeding depression is enriched in regions of high recombination rate and, to a lesser extent, low nucleotide diversity (Yengo et al. 2021)⁷³. In this thesis, recombination rate was not included amongst the genomic annotations used to weight F_{ROH} , in light of the confounding concerns discussed in Chapter 4. Yengo et al. state that “*differential contribution of genomic regions to ID would not be expected*” if ID were caused by confounding, but this is only true if regions are defined on a genetic map scale. Indeed, when regions are defined on a physical

scale (Mb) it is precisely the regions of higher recombination rate which are more informative of parental relatedness, and therefore confounding. It seems, therefore, that confounding should at least be considered as a possible explanation for an enrichment of ID in regions of higher recombination rate.

Confounding aside, the results of Yengo et al, might also be explained in terms of ROH islands. Figure 44 shows a relationship between recombination rate and ID effect size, mainly caused by the reduced effect of $ROH < 1$ cM. The associations reported by Yengo et al. might therefore be explained in terms of the reduced effect of short ROH, which have lower average recombination rate per Mb, and occur almost exclusively in ROH islands. Short ROH occurring in ROH islands could have reduced effects for any number of reasons, including an increased opportunity for selection to remove deleterious haplotypes. Suggesting that inbreeding depression is enriched in regions of high recombination rate implies that, all else being equal, a ROH falling in in a region of high recombination rate will have a greater effect than one falling in in a region of low recombination rate. However, it seems possible that the association reported by Yengo et al. may be generated because the ROH in different regions are not equal, and not because the effect of autozygosity in different regions is unequal.

6 Conclusions

As described in the introduction, the main aims of this study were to:

- 1) Estimate the effect of inbreeding on a broad catalogue of human complex traits.
- 2) Determine whether any significant effects are manifestations of inbreeding depression, or artifacts of social confounding.
- 3) Investigate the genetic architecture of inbreeding depression in humans.
- 4) Make recommendations relevant to future studies.

The first three points have been discussed in Chapters 3, 4, and 5 respectively, and will be only briefly summarised here. Instead, this chapter will attempt to put the main findings in a wider context and, in particular, discuss future directions.

6.1 Main findings

6.1.1 Cause or Confounding?

In other species, it is well established that most inbreeding load is caused by small deleterious effects, and only a lesser amount by rare mutations with large effects^{26,34,35}. In humans, consanguinity increases the risk of many recessive Mendelian disorders, but the smaller effects of inbreeding on complex traits have been harder to establish with certainty. In particular, uniquely human associations between inbreeding and social variables raise the very real possibility of mistaking social correlations for genetic causality. It is therefore of foremost importance to establish that the observed associations between F_{ROH} and human complex traits are caused by genetic homozygosity, and a large part of this thesis has been devoted to that objective.

The results of Chapter 4 seem to establish that most of the effects described in Chapter 3 are likely to be caused by homozygosity, rather than by environmental confounding. Perhaps the best single estimate, a meta-analysis of the sibling method applied to 17 traits in ROHgen, suggests that, on average, 74% [95% CI 40% to 110%] of the observed effects are caused by autozygosity. Unfortunately, the

sample sizes currently available are insufficient to determine whether some effects are entirely caused by confounding, or whether confounding has a small influence on many estimates. However, the within siblings method introduced here will doubtless resolve this uncertainty when large sample sizes become available (see section 6.2.1 for a discussion of required samples).

6.1.2 Complex traits affected by inbreeding

With the genetic origin of the effects mostly established, it is interesting to consider which traits are affected by inbreeding. Many associations have already been reported by other authors, but this thesis has examined a wider range of traits than previous studies, and has attempted to understand the effects in the context of selection. In this study over 150 complex traits are significantly affected by inbreeding, with evidence that almost all traits considered (thousands) are affected to a lesser degree. The traits most affected by inbreeding can be broadly grouped into categories of:

- 1) Reduced well-being.
- 2) Biochemical changes associated with reduced overall health.
- 3) Increased disease prevalence.
- 4) Reduced socio-economic status.
- 5) Lighter skin colour.
- 6) Smaller body size and strength.
- 7) Reduced reproductive success.
- 8) Reduced cognitive ability.
- 9) Poorer eyesight.
- 10) Poorer hearing.
- 11) Reduced risk-taking and increased conservatism.

It is interesting to note that, despite the uniqueness of human society, many of the observed effects have parallels in other species. For example, inbreeding also reduces body size, disease resistance and reproductive success in many other species. In this context, it is perhaps unsurprising that autozygosity reduces

reproductive success in humans, but that it might do so through behavioural traits (e.g., age at first sex, number of sexual partners, alcohol consumption, and risk-taking behaviour) is less well known. However, this possibility is consistent with experimental studies in other species which report lower sexual motivation in inbred animals¹²⁶. In humans, Gardener et al.¹²⁴, report that the *heterozygous* burden of PTVs is associated with lower reproductive success, and similarly conclude that behavioural traits, rather than fertility, are likely to be the intermediate phenotypes.

However, why some traits are significantly affected (e.g., cognition, height) and others are not (e.g., blood pressure), appears largely unexplained. Affected traits are often described as “*fitness-related*”, but it is sometimes unclear what is meant by “*related*”. Related could imply correlation but, in contrast to many of the traits reported here, fitness-correlated traits are expected to have low genetic variance. A correlation with fitness implies a response to selection, which seems unlikely to be sustained without depleting genetic variance. Indeed, for cognition, there is evidence that the current selection response is in the wrong direction¹¹⁵. A heuristic explanation is discussed in Appendix F, suggesting that the traits affected by inbreeding depression are an inherent property of the human genome, which has itself been shaped by selection.

6.1.2.1 *Advantages of mating with kin*

Although the phenotypic effects of inbreeding are almost always harmful, reproduction with a relative may bring other important benefits, in both humans and other species. Indeed, there is evidence that most animal species show no preference for unrelated mates, despite a widespread assumption to the contrary. Moreover, in humans, consanguineous unions are favoured in many parts of the world, due to their undoubted social advantages¹⁶¹. However, to focus on the social and individual advantages, obscures the fact that evolution is best understood at the gene level^{162,163}, and here inbreeding confers a substantial advantage.

Sex and meiotic recombination are believed to have evolved because they promote genetic diversity, which is essential to a species' ability to adapt to, and survive through, changing environments. However, to an individual allele, inbreeding confers an evolutionary advantage because it increases the probability of being transmitted to the subsequent generation along with an identical copy¹⁶⁴.

Therefore, any mutation which promotes inbreeding is well positioned to spread through a species unless there is some opposing fitness disadvantage. Inbreeding also reduces genetic diversity, undoing some of the benefits of sex, which reduces the adaptability of the species, and is likely to have consequences for its long-term survival. In this context, inbreeding depression may be a necessary property to avoid the erosion of genetic diversity which inbreeding entails. Furthermore, inbreeding can be undone in a single generation, so inbreeding depression must act in a single generation - to counter the strategy of inbreeding and outbreeding in alternate generations. In this context, it is interesting to note that, in ROHgen, the effect of F_{ROH} on number of children ever born (neb) is -1.9 [95% CI = -2.40 to -1.40] – consistent with the minimum value theoretically required (-2) to ensure that inbreeding does not confer a selective advantage to an individual's genes.

6.1.3 Genetic architecture of inbreeding depression

In the past 20 years, despite the great advances made towards understanding the genetic architecture of human complex traits, relatively little has been learned about the genetic variants causing inbreeding depression. Indeed, most of what is known can be attributed to experiments on the fruit fly, *Drosophila melanogaster*. This thesis aimed to exploit the large human datasets now available, to make some initial progress towards understanding the nature of the genetic variants causing inbreeding depression in humans.

In Chapter 3, F_{ROH} was shown to be a good predictor of inbreeding depression, better than estimates based on common SNP homozygosity, implying that inbreeding effects are mainly caused by rare variants. However, even very short ROH (< 1cM), attributable to common ancestors perhaps 50 generations ago, cause significant inbreeding depression, implying that most causal variants are unlikely to

be ultra-rare mutations unique to single families. Surprisingly, no association was found with gene density, showing that the effects of inbreeding are evenly distributed in physical length throughout the genome. This lack of association with gene density contrasts with what is known about the distribution of additive genetic effects, and understanding this discrepancy should be a priority of future work. Together with the large number of traits affected, these results imply an omnigenic architecture for inbreeding depression, where rare mutations in many genes affect many traits through highly interconnected networks.

6.2 Future directions

6.2.1 Sibling studies

A high priority for future studies should be to establish definitively which inbreeding effects are caused by homozygosity, and which can be attributed to social confounding. This could potentially be achieved by at least two methods. Firstly, any correlation (on a genetic map scale) between inbreeding effects and any genomic feature should be independent of all confounders. Secondly, the variation of homozygosity between siblings is also believed to be independent of confounders. Since the first method requires some hitherto elusive insight into a pertinent genomic feature, the within sibling method is most likely to resolve the remaining uncertainty.

It may therefore be useful to estimate what size of future sibling study will be required. Based on the calculations in Appendix E.4, Supplementary Table 42 estimates the number siblings requires to have 80% power to confirm a significant effect, for a range of plausible effect sizes. For a trait, such as height, with a moderately large effect ($\beta'_{FROH} = 2$) approximately 465,000 siblings would be required in a cohort similar to UK Biobank. However, in a cohort comprised entirely of the offspring of first cousins (as simulated in section 2.6.4), only 4,000 siblings would be required to achieve the same power. Although this clearly demonstrates the potential of consanguineous cohorts, it might also be noted that the former cohort is probably already gathered. In 2019, Nakka et al. described a dataset

gathered by 23andme with 4,400,363 genotyped subjects including 916,712 parent-offspring pairs⁷⁹. Although the number of genotyped siblings is not mentioned, it seems likely to be of a similar magnitude.

It should also be noted that in cohorts with relatively low rates of consanguinity (such as UKB British), a material fraction of the variation in F_{ROH} may occur in ROH islands. The concerns about additive effects discussed in Chapter 5 therefore potentially affect the within sibling analysis. Although there is no reason to expect the average additive effect of common haplotypes to mirror the effects of F_{ROH} , it would be preferable if additive effects could be completely excluded. This could either be achieved by conducting the analysis in cohorts with a high rate of consanguinity, or by adjusting ROH calling methods to identify only autozygous segments attributable to relatively recent consanguinity.

6.2.2 ROH calling methods

Even in relatively outbred cohorts, much of the statistical power to estimate inbreeding effects is derived from long, readily identified ROH (section 2.5.4). For this reason, evaluating, or even improving, ROH calling methods has not been a particular focus of this study. However, the results reported here may provide some guidance to future work in this area.

In particular, the realisation that, at ROH islands, additive effects may confound recessive effects has implications for ROH calling. Ideally, ROH should represent segments autozygous with respect to the same base generation, such that the distribution of ROH is uniform throughout the genome, and the presence of a ROH is therefore not informative about the haplotype. For shorter ROH, this ideal will be unachievable because mutation rates and selection vary throughout the genome, but a step in the right direction could be made by calling ROH with respect to a minimum genetic map length (cM) rather than physical distance (Mb), as currently implemented in the most popular algorithms. Calling ROH in centimorgans also has the advantage that confounding can be excluded if the distribution of a genomic feature is found to be correlated with the distribution of inbreeding effects.

Although the simple method implemented in PLINK is suitable for identifying relatively long ROH, the results of section 3.3.8 show that even $ROH < 1\text{ cM}$ may have significant effects. In many parts of the genome, $ROH < 1\text{ cM}$ will be shorter than the 1.5 Mb minimum length threshold used here, and will thus be unidentified. Therefore, in future studies, Hidden-Markov-Model methods will be necessary to reliably identify such short autozygous segments. Such methods, which account for allele frequency, will also have significant advantages when applied to whole genome sequence, as discussed in section 6.2.3.

Ultimately, the best estimate of inbreeding coefficients is the one which best predicts the effects of inbreeding – i.e., the estimate most closely correlated with the excess homozygosity of causal variants. Previous comparisons of ROH calling methods simulated reasonable assumptions about the nature of these causal variants, but comparisons could now also be made on the observed effects of inbreeding.

6.2.3 Whole genome sequence

Although this study uses only SNP microarray data, some thought should be given to the future potential of large whole genome sequence (WGS) datasets. Although most WGS datasets are currently orders of magnitude smaller than the samples size used here, the genotyping of all variants (especially rare and structural variants) offers novel opportunities.

In terms of ROH calling, WGS may not offer much improvement for long ROH, which are already unambiguously called in dense SNP microarray data. However, large numbers of rare variants in WGS should allow shorter ROH to be more reliably called, especially using HMM methods which take account of allele frequencies. For example, the homozygosity of rare variants in ROH islands should differentiate between autozygous haplotypes from a recent common ancestor, and more distant (or recombinant) haplotypes. In general, leveraging information from rare variants (and perhaps calling ROH in centimorgans as discussed in section 6.2.2) should

enable ROH to correspond more closely to the definition of autozygosity as identity-by-descent from a fixed base generation.

Beyond any improvement in ROH calling, WGS offers the hope of identifying individual causal variants, or annotations in which causal variants are enriched. Chapter 5 concludes that the effects of autozygosity are relatively evenly distributed throughout the genome, and mostly caused by rare variants. However, it is not known whether these rare variants are comparatively few in number, with large effects, or more numerous with smaller effects. The former situation may offer some hope of identifying specific recessive variants in non-additive GWAS, although it is perhaps more realistic to envisage burden tests, such as recently reported for heterozygous variants¹²⁴.

6.2.4 Trait variance

This study has investigated how inbreeding affects mean trait values, however changes to the trait variance may also be informative about genetic architecture. Specifically, for additive effects, the trait variance is expected to increase in direct proportion to the inbreeding coefficient. In contrast, at overdominant loci, inbreeding may cause a reduction in genetic variance, while at highly recessive loci inbreeding may dramatically increase variability. Of course, as for changes to the mean, other factors may explain observed changes to the variance, and some care will be needed to correctly infer causality. Notably, a reduction in heterozygosity caused by inbreeding may increase an organism's sensitivity to environmental change, as initially proposed by Lerner's theory of genetic homeostasis^{165,166}.

6.2.5 Unexplained results from this study

At least two findings in this thesis could be described as unexpected, and might merit further investigation. Most notably, no association was found between inbreeding depression and gene density, in contrast to the known distribution of additive variance. Although mutation-selection theory may offer a partial explanation (see section 5.4.2.2), it would be interesting to understand whether ID causal variants occur in genes (but with effect sizes inversely proportional to local

gene density), or whether causal variants are genuinely distributed without regard to gene locations.

Secondly, the concordance between the effects of ROH and the effects of large hemizygous deletions is remarkable (Figure 70) and largely unexplained.

Furthermore, the effect of a deletion depends on the number of genes deleted (Figure 85), while the effect of a ROH does not (Figure 84). Most deletions are likely to be deleterious and partially recessive, so it is likely that their homozygosity contributes to inbreeding depression. It would be interesting to explore in WGS whether the homozygosity of smaller deletions (which are ubiquitous in healthy populations) contributes disproportionately to inbreeding depression.

6.2.6 Recommendations for future studies

- Large sibling studies should be used to estimate the true genetic effect of ROH on individual traits. Ideally, ROH islands would be removed to avoid any potential confounding by additive effects.
- ROH islands should be removed before attempting homozygosity mapping, to avoid confounding by additive effects.
- ROH calling algorithms should be modified to call ROH above a minimum genetic length threshold. This should help to reduce the height of ROH islands, and to call ROH which correspond more closely to the definition of autozygosity as identity-by-descent from a fixed base generation.
- Hemizygous deletions should be removed from X chromosome ROH, however doing so is less important on the autosomes.
- In low autozygosity cohorts, uniparental disomy (UPD) can make a significant contribution to ROH. Although the effects of UPD are similar to the effects of 'common-ancestry' ROH it may be desirable to remove UPD cases.
- Some blood cancers may cause a somatic loss-of-heterozygosity which may easily be mistaken for germline autozygosity. Beware the direction of causality.

6.3 Social context

In most societies, reproduction between close relatives (1st or 2nd degree) is taboo, and cousin-marriage is a sensitive topic in many cultures. In this context it is perhaps important to stress that, although many traits are negatively affected by autozygosity, the magnitude of effect on any one trait is small. For example, the average height expected in the offspring of 1st degree relatives ($E(F_{ROH})=0.25$), is only 3.6 cm shorter [95% CI = 3.4 to 3.9] than the population average.

Approximately 30% of the population would be expected to be shorter than this due to other genetic and environmental factors. Indeed, the observation that many very high F_{ROH} samples are healthy, and indistinguishable from outbred samples, demonstrates that even the most extreme inbreeding is often of relatively minor consequence to the individual. Cousin marriage brings many important social benefits to a large number of people throughout the world, and the results reported here are in no way intended to stigmatise inbreeding. On the contrary, this thesis should be understood as an investigation of human genetic architecture, and how the inherent evolutionary benefits of inbreeding are compensated at a species level.

6.4 Final words

The unprecedented quantities of human genetic and phenotypic data becoming available offer exciting possibilities to answer research questions which were unimaginable just a few years ago. Although using ROH to estimate inbreeding coefficients only became feasible about 15 years ago, much has been learned in that time, and it is becoming increasingly clear that human society is a valid context in which to study the evolutionary and medical consequences of homozygosity. However, for the pace of progress to be sustained, it will be important to ensure that future datasets represent the full range of human diversity.

References

1. Clark, D. W. & et al. Associations of autozygosity with a broad range of human phenotypes. *Nature communications* (2019).
2. Darwin C. R. *The Effects of Cross and Self Fertilization in the Vegetable Kingdom*. (John Murray, London, 1876).
3. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nature reviews. Genetics* **10**, 783–96 (2009).
4. Bittles, A. H. *Consanguinity in Context. Cambridge Studies in Biological and Evolutionary Anthropology* **63**, (2012).
5. Hamamy, H. Consanguineous marriages preconception consultation in primary health care settings. *Journal of Community Genetics* (2012). doi:10.1007/s12687-011-0072-y
6. Broman, K. W. & Weber, J. L. Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain. *The American Journal of Human Genetics* **65**, 1493–1500 (1999).
7. Ostrer, H. & Skorecki, K. The population genetics of the Jewish people. *Human Genetics* (2013). doi:10.1007/s00439-012-1235-6
8. Bellwood, P., Chambers, G., Ross, M. & Hung, H. Are 'Cultures' Inherited? Multidisciplinary Perspectives on the Origins and Migrations of Austronesian-Speaking Peoples Prior to 1000 bc. in *Investigating Archaeological Cultures* (2011). doi:10.1007/978-1-4419-6970-5_16
9. Rousset, F. Inbreeding and relatedness coefficients: What do they measure? *Heredity* (2002). doi:10.1038/sj.hdy.6800065
10. Wright, S. Genetical structure of populations. *Nature* (1950). doi:10.1038/166247a0
11. Keller, L. F. & Waller, D. M. Inbreeding effects in wild populations. *Trends in Ecology and Evolution* **17**, 230–241 (2002).
12. Charlesworth, B. & Charlesworth, D. *Elements of Evolutionary Genetics*. (Roberts and Company, 2010).
13. Joshi, P. K. *et al.* Directional dominance on stature and cognition in diverse human populations. *Nature* **523**, 459–462 (2015).
14. Crow, J. F. Dominance and overdominance. in *The Genetics and Exploitation of Heterosis in Crops* (2015). doi:10.2134/1999.geneticsandexploitation.c5
15. Ohta, T. Associative overdominance caused by linked detrimental mutations. *Genetical Research* (1971). doi:10.1017/S0016672300012684
16. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics (4th Edition)*. *Trends in Genetics* **12**, (1996).

17. Bourguet, D. The evolution of dominance. *Heredity* **83**, 1–4 (1999).
18. Fisher, R. A. The Possible Modification of the Response of the Wild Type to Recurrent Mutations. *The American Naturalist* (1928). doi:10.1086/280193
19. Wright, S. Fisher's Theory of Dominance. *The American Naturalist* **63**, 274–279 (1929).
20. Wright, S. Physiological and Evolutionary Theories of Dominance. *The American Naturalist* **68**, 24–53 (1934).
21. Kacser, H. & Burns, J. A. The molecular basis of dominance. *Genetics* (1981).
22. Charlesworth, B. Evidence against Fisher's theory of dominance [7]. *Nature* (1979). doi:10.1038/278848a0
23. Orr, H. A. A test of Fisher's theory of dominance. *Proceedings of the National Academy of Sciences of the United States of America* (1991). doi:10.1073/pnas.88.24.11413
24. Tolwinski, N. S. Introduction: Drosophila-A model system for developmental biology. *Journal of Developmental Biology* (2017). doi:10.3390/jdb5030009
25. Castle, W. E., Carpenter, F. W., Clark, A. H., Mast, S. O. & Barrows, W. M. The Effects of Inbreeding, Cross-Breeding, and Selection upon the Fertility and Variability of Drosophila. *Proceedings of the American Academy of Arts and Sciences* (1906). doi:10.2307/20022155
26. Charlesworth, D. & Charlesworth, B. Inbreeding depression and its evolutionary consequences. *Annual review of ecology and systematics*. Vol. 18 (1987). doi:10.1146/annurev.es.18.110187.001321
27. McBride, G. & J'Ames, J. W. Effects of inbreeding and bristle number on reproductive success in selected lines of Drosophila melanogaster. *Heredity* (1965). doi:10.1038/hdy.1965.59
28. Biémont, C. Inbreeding effects: Evidence for a genetic system which regulates viability in Drosophila melanogaster populations. *Mechanisms of Ageing and Development* (1978). doi:10.1016/0047-6374(78)90004-0
29. Vermeulen, C. J. & Bijlsma, R. Changes in mortality patterns and temperature dependence of lifespan in Drosophila melanogaster caused by inbreeding. *Heredity* (2004). doi:10.1038/sj.hdy.6800412
30. Van Oosterhout, C. *et al.* Inbreeding depression and genetic load of sexually selected traits: How the guppy lost its spots. *Journal of Evolutionary Biology* (2003). doi:10.1046/j.1420-9101.2003.00511.x
31. Opatová, P. *et al.* Inbreeding depression of sperm traits in the zebra finch Taeniopygia guttata. *Ecology and Evolution* (2016). doi:10.1002/ece3.1868
32. Miller, D. E., Cook, K. R. & Scott Hawley, R. The joy of balancers. *PLoS Genetics* (2019). doi:10.1371/journal.pgen.1008421

33. Haldane, J. B. S. A mathematical theory of Natural and Artificial Selection. *Mathematical Proceedings of the Cambridge Philosophical Society* (1927). doi:10.1017/S0305004100011750
34. Crow, J. F. & Simmons, M. J. The mutation load in *Drosophila*. in *The genetics and biology of Drosophila vol. 3c* (1983).
35. Keightley, P. D. Nature of deleterious mutation load in *Drosophila*. *Genetics* (1996). doi:10.1093/genetics/144.4.1993
36. Nakajima, M. & Taniguchi, N. Genetics of the guppy as a model for experiment in aquaculture. *Genetica* (2001). doi:10.1023/A:1013717516234
37. Johnson, A. M. *et al.* Inbreeding Depression and Inbreeding Avoidance in a Natural Population of Guppies (*Poecilia reticulata*). *Ethology* (2010). doi:10.1111/j.1439-0310.2010.01763.x
38. Mariette, M., Kelley, J. L., Brooks, R. & Evans, J. P. The effects of inbreeding on male courtship behaviour and coloration in guppies. *Ethology* (2006). doi:10.1111/j.1439-0310.2006.01236.x
39. Nakadate, M., Shikano, T. & Taniguchi, N. Inbreeding depression and heterosis in various quantitative traits of the guppy, *Poecilia reticulata*. *Aquaculture* (2003). doi:10.1016/S0044-8486(02)00432-5
40. Pitcher, T. E., Rodd, F. H. & Rowe, L. Female choice and the relatedness of mates in the guppy (*Poecilia reticulata*). *Genetica* (2008). doi:10.1007/s10709-008-9246-x
41. Sheridan, L. & Pomiankowski, A. Fluctuating asymmetry, spot asymmetry and inbreeding depression in the sexual coloration of male guppy fish. *Heredity* (1997). doi:10.1038/hdy.1997.191
42. Zajitschek, S. R. K. & Brooks, R. C. Inbreeding depression in male traits and preference for outbred males in *Poecilia reticulata*. *Behavioral Ecology* (2010). doi:10.1093/beheco/arq077
43. Van Oosterhout, C. *et al.* The guppy as a conservation model: Implications of parasitism and inbreeding for reintroduction success. *Conservation Biology* (2007). doi:10.1111/j.1523-1739.2007.00809.x
44. Smallbone, W., van Oosterhout, C. & Cable, J. The effects of inbreeding on disease susceptibility: *Gyrodactylus turnbulli* infection of guppies, *Poecilia reticulata*. *Experimental Parasitology* (2016). doi:10.1016/j.exppara.2016.04.018
45. Pusey, A. & Wolf, M. Inbreeding avoidance in animals. *Trends in Ecology and Evolution* (1996). doi:10.1016/0169-5347(96)10028-8
46. Frankham, R. Conservation genetics. *Annual Review of Genetics* (1995). doi:10.1146/annurev.ge.29.120195.001513
47. Caro, T. M. & Laurenson, M. K. Ecological and genetic factors in conservation: A cautionary tale. *Science* (1994). doi:10.1126/science.8290956

48. Crnokrak, P. & Roff, D. A. Inbreeding depression in the wild. *Heredity* (1999). doi:10.1038/sj.hdy.6885530
49. Byers, D. L. & Waller, D. M. Do plant populations purge their genetic load? Effects of population size and mating history on inbreeding depression. *Annual Review of Ecology and Systematics* (1999). doi:10.1146/annurev.ecolsys.30.1.479
50. Neaves, L. E. *et al.* The fitness consequences of inbreeding in natural populations and their implications for species conservation - A systematic map Children's Environmental Health. *Environmental Evidence* (2015). doi:10.1186/s13750-015-0031-x
51. Huisman, J., Kruuk, L. E. B., Ellis, P. A., Clutton-Brock, T. & Pemberton, J. M. Inbreeding depression across the lifespan in a wild mammal population. *Proceedings of the National Academy of Sciences* **113**, 3585–3590 (2016).
52. Rijks, J. M., Hoffman, J. I., Kuiken, T., Osterhaus, A. D. M. E. & Amos, W. Heterozygosity and lungworm burden in harbour seals (*Phoca vitulina*). *Heredity* (2008). doi:10.1038/hdy.2008.18
53. Coltman, D. W., Pilkington, J. G., Smith, J. A. & Pemberton, J. M. Parasite-mediated selection against inbred Soay sheep in a free-living, island population. *Evolution* (1999). doi:10.1111/j.1558-5646.1999.tb04538.x
54. Leroy, G. Inbreeding depression in livestock species: Review and meta-analysis. *Animal Genetics* (2014). doi:10.1111/age.12178
55. Nani, J. P. & Peñagaricano, F. Whole-genome homozygosity mapping reveals candidate regions affecting bull fertility in US Holstein cattle. *BMC Genomics* (2020). doi:10.1186/s12864-020-6758-y
56. Peripolli, E. *et al.* Runs of homozygosity: current knowledge and applications in livestock. *Animal Genetics* (2017). doi:10.1111/age.12526
57. Szpiech, Z. A. *et al.* Long runs of homozygosity are enriched for deleterious variation. *American Journal of Human Genetics* (2013). doi:10.1016/j.ajhg.2013.05.003
58. Wright, S. Coefficients of Inbreeding and Relationship. *The American Naturalist* (1922). doi:10.1086/279872
59. Li, C. C. & HORVITZ, D. G. Some methods of estimating the inbreeding coefficient. *American journal of human genetics* (1953).
60. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76–82 (2011).
61. Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* **15**, 789–795 (2006).
62. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *American journal of human genetics* **91**, 275–292 (2012).

63. Li, C. C. & Horvitz, D. G. Some Methods of Estimating the Inbreeding Coefficient INTRODUCTION: POPULATION MODELS. *The American Journal of Human Genetics* **5**, (1953).
64. Kardos, M., Luikart, G. & Allendorf, F. W. Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity* (2015). doi:10.1038/hdy.2015.17
65. Carothers, A. D. *et al.* Estimating human inbreeding coefficients: Comparison of genealogical and marker heterozygosity approaches. *Annals of Human Genetics* (2006). doi:10.1111/j.1469-1809.2006.00263.x
66. Nietlisbach, P. *et al.* Pedigree-based inbreeding coefficient explains more variation in fitness than heterozygosity at 160 microsatellites in a wild bird population. *Proceedings of the Royal Society B: Biological Sciences* (2017). doi:10.1098/rspb.2016.2763
67. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559–575 (2007).
68. Purcell, S. & Chang, C. PLINK 1.9. Available at: www.cog-genomics.org/plink/1.9/.
69. Keller, M. C. *et al.* Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genet* **8**, (2012).
70. Howrigan, D. P. *et al.* Genome-wide autozygosity is associated with lower general cognitive ability. *Molecular Psychiatry* (2016). doi:10.1038/mp.2015.120
71. Johnson, E. C., Evans, L. M. & Keller, M. C. Relationships between estimated autozygosity and complex traits in the UK Biobank. *PLoS Genetics* (2018). doi:10.1371/journal.pgen.1007556
72. Yengo, L., Wray, N. R. & Visscher, P. M. Extreme inbreeding in a European ancestry sample from the contemporary UK population. *Nature Communications* (2019). doi:10.1038/s41467-019-11724-6
73. Yengo, L. *et al.* Genomic partitioning of inbreeding depression in humans. *The American Journal of Human Genetics* (2021). doi:10.1016/j.ajhg.2021.06.005
74. Howrigan, D. P., Simonson, M. A. & Keller, M. C. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* (2011). doi:10.1186/1471-2164-12-460
75. Browning, S. R. & Browning, B. L. High-Resolution Detection of Identity by Descent in Unrelated Individuals. *American Journal of Human Genetics* (2010). doi:10.1016/j.ajhg.2010.02.021
76. Narasimhan, V. *et al.* BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw044

77. Wang, S., Haynes, C., Barany, F. & Ott, J. Genome-wide autozygosity mapping in human populations. *Genetic Epidemiology* (2009). doi:10.1002/gepi.20344
78. Szpiech, Z. A., Blant, A. & Pemberton, T. J. GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx102
79. Nakka, P. *et al.* Characterization of Prevalence and Health Consequences of Uniparental Disomy in Four Million Individuals from the General Population. *American Journal of Human Genetics* (2019). doi:10.1016/j.ajhg.2019.09.016
80. Kornilov, S. A., Tan, M., Aljughaiman, A., Naumova, O. Y. & Grigorenko, E. L. Genome-Wide Homozygosity Mapping Reveals Genes Associated With Cognitive Ability in Children From Saudi Arabia. *Frontiers in Genetics* (2019). doi:10.3389/fgene.2019.00888
81. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* (2018). doi:10.1038/s41586-018-0579-z
82. McQuillan, R. *et al.* Runs of Homozygosity in European Populations. *American Journal of Human Genetics* **83**, 359–372 (2008).
83. Kirin, M. *et al.* Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5**, (2010).
84. McQuillan, R. *et al.* Evidence of inbreeding depression on human height. *PLoS Genet* **8**, (2012).
85. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
86. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
87. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *American Journal of Epidemiology* (2017). doi:10.1093/aje/kwx246
88. Aguirre, M., Rivas, M. A. & Priest, J. Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *American Journal of Human Genetics* (2019). doi:10.1016/j.ajhg.2019.07.001
89. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: Analysis of the UK Biobank. *Journal of Medical Genetics* (2019). doi:10.1136/jmedgenet-2018-105477
90. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* (2007). doi:10.1101/gr.6861907
91. Wright, S. Genetics of Populations. *encyclopedia britannica* vol. 10: 111-A-D-112 (1948).

92. WRIGHT, S. The genetical structure of populations. *Annals of eugenics* (1951). doi:10.2307/2407273
93. Benton, M. *et al.* Complete mitochondrial genome sequencing reveals novel haplotypes in a polynesian population. *PLoS ONE* (2012). doi:10.1371/journal.pone.0035026
94. Nothnagel, M., Lu, T. T., Kayser, M. & Krawczak, M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Human Molecular Genetics* **19**, 2927–2935 (2010).
95. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* (2007). doi:10.1038/nature06258
96. Campbell, H. *et al.* Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Human Molecular Genetics* **16**, 233–241 (2007).
97. Verweij, K. J. H. *et al.* The association of genotype-based inbreeding coefficient with a range of physical and psychological human traits. *PLoS ONE* (2014). doi:10.1371/journal.pone.0103102
98. Yengo, L. *et al.* Detection and quantification of inbreeding depression for complex traits from SNP data. *Proceedings of the National Academy of Sciences* **114**, 8602–8607 (2017).
99. Moreno-Grau, S. *et al.* Long runs of homozygosity are associated with Alzheimer's disease. *Translational Psychiatry* (2021). doi:10.1038/s41398-020-01145-1
100. Bashi, J. Effects of inbreeding on cognitive performance. *Nature* **266**, 440–442 (1977).
101. Morton, N. E. Effect of inbreeding on IQ and mental retardation. *Proceedings of the National Academy of Sciences of the United States of America* (1978). doi:10.1073/pnas.75.8.3906
102. Alvarez, G., Ceballos, F. C. & Berra, T. M. Darwin was right: Inbreeding depression on male fertility in the Darwin family. *Biological Journal of the Linnean Society* **114**, 474–483 (2015).
103. Robert, A., Toupance, B., Tremblay, M. & Heyer, E. Impact of inbreeding on fertility in a pre-industrial population. *European Journal of Human Genetics* **17**, 673–681 (2009).
104. Rudan, I. *et al.* Inbreeding and risk of late onset complex disease. *Journal of Medical Genetics* (2003). doi:10.1136/jmg.40.12.925
105. Shami, S. A., Qaisar, R. & Bittles, A. H. Consanguinity and adult morbidity in Pakistan. *The Lancet* (1991). doi:10.1016/0140-6736(91)91828-I
106. Rudan, I. *et al.* Inbreeding and the genetic complexity of human hypertension. *Genetics* (2003). doi:10.1093/genetics/163.3.1011

107. Rudan, I. *et al.* Inbreeding and susceptibility to osteoporosis in Croatian island isolates. *Collegium Antropologicum* (2004).
108. Lebel, R. R. & Gallagher, W. B. Wisconsin consanguinity studies. II: Familial adenocarcinomatosis. *American Journal of Medical Genetics* (1989). doi:10.1002/ajmg.1320330102
109. Abaskuliev, A. A. & Skoblo, G. V. [Inbreeding, endogamy and exogamy among relatives of schizophrenia patients]. *Genetika* (1975).
110. Bulayeva, K. B. *et al.* Mapping genes of complex psychiatric diseases in Daghestan genetic isolates. *American Journal of Medical Genetics - Neuropsychiatric Genetics* (2005). doi:10.1002/ajmg.b.30073
111. Johnson, E. C. *et al.* No Reliable Association between Runs of Homozygosity and Schizophrenia in a Well-Powered Replication Study. *PLoS Genetics* (2016). doi:10.1371/journal.pgen.1006343
112. Abdellaoui, A. *et al.* Educational attainment influences levels of homozygosity through migration and assortative mating. *PLoS ONE* **10**, (2015).
113. Abdellaoui, A. *et al.* Association between autozygosity and major depression: Stratification due to religious assortment. *Behavior Genetics* **43**, 455–467 (2013).
114. Keller, M. C., Visscher, P. M. & Goddard, M. E. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* **189**, 237–249 (2011).
115. Kong, A. *et al.* Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences of the United States of America* (2017). doi:10.1073/pnas.1612113114
116. Pulanić, D. *et al.* Effects of isolation and inbreeding on human quantitative traits: An example of biochemical markers of hemostasis and inflammation. *Human Biology* (2008). doi:10.3378/1534-6617-80.5.513
117. Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G. & Allendorf, F. W. Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications* (2016). doi:10.1111/eva.12414
118. Gazal, S. *et al.* Inbreeding coefficient estimation with dense SNP data: Comparison of strategies and application to HapMap III. *Human Heredity* (2014). doi:10.1159/000358224
119. Belonogova, N. M., Svishcheva, G. R., van Duijn, C. M., Aulchenko, Y. S. & Axenovich, T. I. Region-Based Association Analysis of Human Quantitative Traits in Related Individuals. *PLoS ONE* **8**, (2013).
120. Cook, J. P., Mahajan, A. & Morris, A. P. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *European journal of human genetics : EJHG* **25**, 240–245 (2017).

121. Pain, O., Dudbridge, F. & Ronald, A. Are your covariates under control? How normalization can re-introduce covariate effects. *European Journal of Human Genetics* (2018). doi:10.1038/s41431-018-0159-6
122. Jakobsen, L. H., Sorensen, J. M., Rask, I. K., Jensen, B. S. & Kondrup, J. Validation of reaction time as a measure of cognitive function and quality of life in healthy subjects and patients. *Nutrition* (2011). doi:10.1016/j.nut.2010.08.003
123. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* (2005). doi:10.1038/sj.hdy.6800717
124. Gardner, E. J. *et al.* Sex-biased reduction in reproductive success drives selective constraint on human genes. *bioRxiv* (2020). doi:10.1101/2020.05.26.116111
125. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics* (2019). doi:10.1038/s41588-018-0309-3
126. Valenstein, E. S., Riss, W. & Young, W. C. Sex drive in genetically heterogeneous and highly inbred strains of male guinea pigs. *Journal of Comparative and Physiological Psychology* (1954). doi:10.1037/h0061296
127. Margulis, S. W. & Altmann, J. Behavioural risk factors in the reproduction of inbred and outbred oldfield mice. *Animal Behaviour* (1997). doi:10.1006/anbe.1996.0422
128. Janicke, T., Vellnow, N., Lamy, T., Chapuis, E. & David, P. Inbreeding depression of mating behavior and its reproductive consequences in a freshwater snail. *Behavioral Ecology* (2014). doi:10.1093/beheco/art122
129. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* (2017). doi:10.1016/j.cell.2017.05.038
130. Ebel, E. R. & Phillips, P. C. Intrinsic differences between males and females determine sex-specific consequences of inbreeding. *BMC Evolutionary Biology* (2016). doi:10.1186/s12862-016-0604-5
131. Sakaluk, S. K. *et al.* Effects of inbreeding on life-history traits and sexual competency in decorated crickets. *Animal Behaviour* (2019). doi:10.1016/j.anbehav.2019.05.027
132. Chang, S. M. Gender-specific inbreeding depression in a gynodioecious plant, *Geranium maculatum* (Geraniaceae). *American Journal of Botany* (2007). doi:10.3732/ajb.94.7.1193
133. Wilson, M. L., Miller, C. M. & Crouse, K. N. Humans as a model species for sexual selection research. *Proceedings of the Royal Society B: Biological Sciences* (2017). doi:10.1098/rspb.2017.1320

134. Wilder, J. A., Mobasher, Z. & Hammer, M. F. Genetic evidence for unequal effective population sizes of human females and males. *Molecular Biology and Evolution* (2004). doi:10.1093/molbev/msh214
135. Ceballos, F. C., Hazelhurst, S. & Ramsay, M. Runs of homozygosity in sub-Saharan African populations provide insights into complex demographic histories. *Human Genetics* (2019). doi:10.1007/s00439-019-02045-1
136. Darwin, C. The descent of man and selection in relation to sex, in Charles Darwin, *The origin of species and The descent of man* (combined volume). *Journal of anatomy and physiology* (1871).
137. Andersson, M. & Iwasa, Y. Sexual selection. *Trends in Ecology and Evolution* (1996). doi:10.1016/0169-5347(96)81042-1
138. Fisher, R. A. Statistical methods in genetics. *Heredity* **6**, 1–12 (1952).
139. Branigan, A. R., Mccallum, K. J. & Freese, J. Variation in the heritability of educational attainment: An international meta-analysis. *Social Forces* (2013). doi:10.1093/sf/sot076
140. Postma, E., Martini, L. & Martini, P. Inbred women in a small and isolated Swiss village have fewer children. *Journal of Evolutionary Biology* (2010). doi:10.1111/j.1420-9101.2010.02013.x
141. Tenesa, A., Rawlik, K., Navarro, P. & Canela-Xandri, O. Genetic determination of height-mediated mate choice. *Genome Biology* (2016). doi:10.1186/s13059-015-0833-8
142. Gaure, S. Lfe: Linear group fixed effects. *R Journal* (2013). doi:10.32614/rj-2013-031
143. Gaure, S. OLS with multiple high dimensional category variables. *Computational Statistics and Data Analysis* (2013). doi:10.1016/j.csda.2013.03.024
144. Lander, E. S. & Botstein, D. Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. *Science* (1987). doi:10.1126/science.2884728
145. Naseri, A., Zhi, D. & Zhang, S. Discovery of runs-of-homozygosity diplotypes clusters and their associations with diseases in UK Biobank. *medRxiv* (2020).
146. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* (2002). doi:10.1101/gr.229102
147. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2020). doi:10.1038/s41586-020-2308-7
148. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* (2016). doi:10.1038/nature19057
149. Blekhman, R. *et al.* Natural Selection on Genes that Underlie Human Disease Susceptibility. *Current Biology* (2008). doi:10.1016/j.cub.2008.04.074

150. Berg, J. S. *et al.* An informatics approach to analyzing the incidentalome. *Genetics in Medicine* (2013). doi:10.1038/gim.2012.112
151. Falls, J. G., Pulford, D. J., Wylie, A. A. & Jirtle, R. L. Genomic imprinting: Implications for human disease. *American Journal of Pathology* (1999). doi:10.1016/S0002-9440(10)65309-6
152. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLoS Genetics* (2016). doi:10.1371/journal.pgen.1006315
153. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* (1979). doi:10.1073/pnas.76.10.5269
154. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
155. Furtado, L. V. & Betz, B. L. Molecular Testing in Myeloproliferative Neoplasms. in *Diagnostic Molecular Pathology* (2017). doi:10.1016/b978-0-12-800886-7.00034-0
156. O’Keefe, C., McDevitt, M. A. & Maciejewski, J. P. Copy neutral loss of heterozygosity: A novel chromosomal lesion in myeloid malignancies. *Blood* (2010). doi:10.1182/blood-2009-10-201848
157. Parkin, M. & Walker, P. Hearing loss in Turner syndrome. *International Journal of Pediatric Otorhinolaryngology* (2009). doi:10.1016/j.ijporl.2008.10.012
158. Larizza, D. *et al.* Serum liver enzymes in Turner syndrome. *European Journal of Pediatrics* (2000). doi:10.1007/s004310050038
159. Zinkernagel, R. F. & Doherty, P. C. Immunological surveillance against altered self components by sensitised T lymphocytes in lymphocytes choriomeningitis. *Nature* (1974). doi:10.1038/251547a0
160. Eanes, W. F., Hey, J. & Houle, D. Homozygous and hemizygous viability variation on the X chromosome of *Drosophila melanogaster*. *Genetics* (1985). doi:10.1093/genetics/111.4.831
161. Bhinder, M. A. *et al.* Consanguinity: A blessing or menace at population level? *Annals of Human Genetics* (2019). doi:10.1111/ahg.12308
162. Hamilton, W. D. The Evolution of Altruistic Behavior. *The American Naturalist* (1963). doi:10.1086/497114
163. Goodhart, C. B. & Williams, G. C. Adaptation and Natural Selection. *The Journal of Animal Ecology* (1968). doi:10.2307/3103
164. Kokko, H. & Ots, I. WHEN NOT TO AVOID INBREEDING. *Evolution* (2006). doi:10.1554/05-613.1
165. Lerner, I. M. Genetic Homeostasis. *Population (US Edition)* **10**, 759 (1956).

166. Lerner, I. M. Phenodeviants and Genetic Homeostasis. *American journal of human genetics* (1961).
167. Kardos, M., Nietlisbach, P. & Hedrick, P. W. How should we compare different genomic estimates of the strength of inbreeding depression? *Proceedings of the National Academy of Sciences* **115**, E2492–E2493 (2018).
168. Helgason, A., Pálsson, S., Guobjartsson, D. F., Kristjánsson, P. & Stefánsson, K. An association between the kinship and fertility of human couples. *Science* **319**, 813–816 (2008).

Appendix A Supplementary Figures

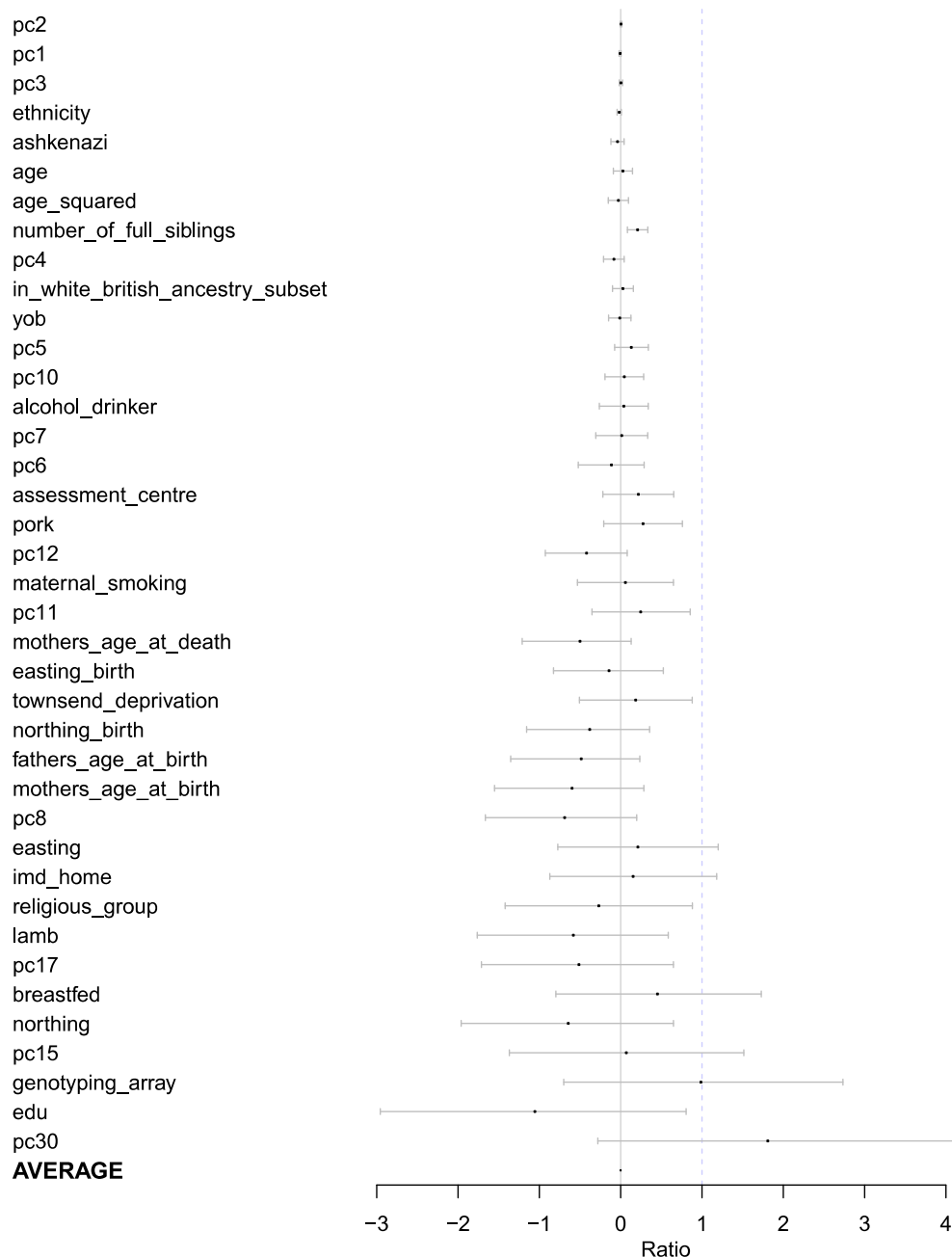


Figure 86: The ratio of $\hat{\beta}_{FROH_wSibs} : \hat{\beta}_{FROH}$ for all 40 significant UKB covariates. 95% confidence intervals determined by permutation are shown in grey. A ratio of 1 would be expected of entirely genetic effects, while a ratio of 0 would indicate complete confounding. The average effect ratio of -0.0003 [95% CI -0.0007 to 0.0006] and absence of heterogeneity (p -value = 0.28) show that the within siblings method is effective at eliminating confounding.

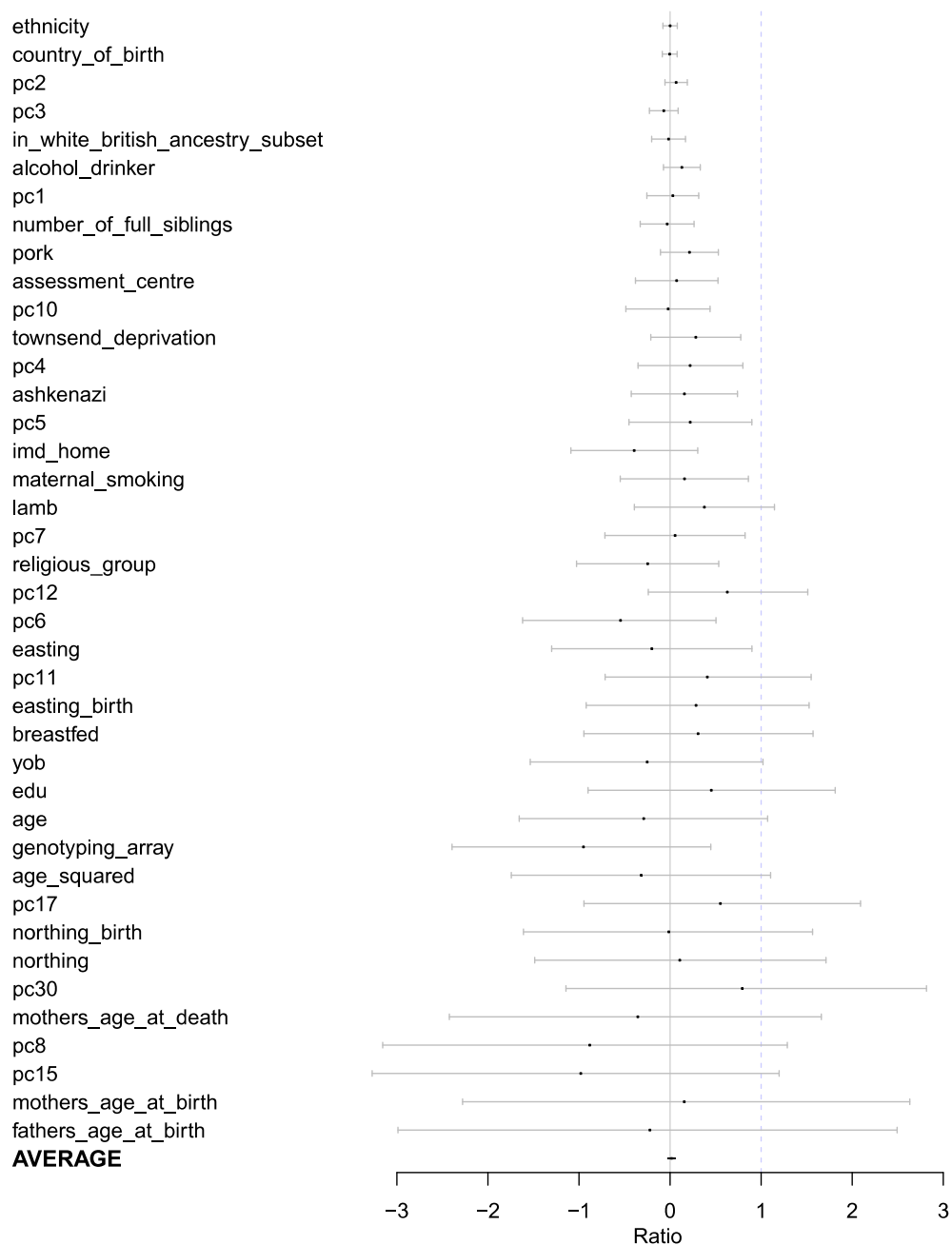


Figure 87: **The ratio of $\hat{\beta}_{FROH_wUPD} : \hat{\beta}_{FROH}$ for all 40 significant UKB covariates.** 95% confidence intervals determined by permutation are shown in grey. A ratio of 1 would be expected of entirely genetic effects, while a ratio of 0 would indicate complete confounding. The average effect ratio of 0.02 [95% CI -0.03 to 0.06] and absence of heterogeneity (p -value = 0.99) show that the UPD method is effective at eliminating confounding.

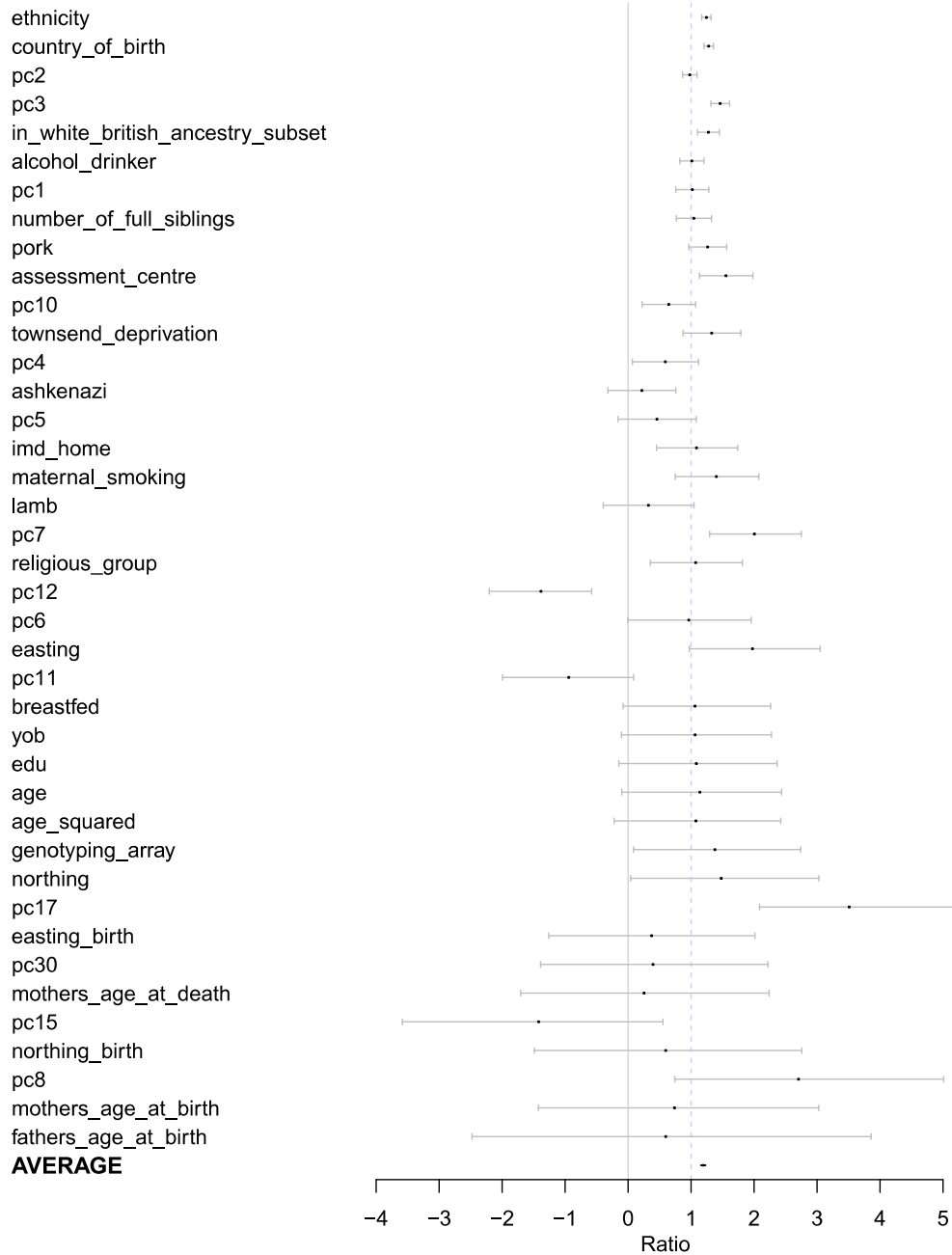


Figure 88: The ratio of $\hat{\beta}_{F_{ROH_WUPD}}:\hat{\beta}_{F_{ROH}}$ for all 40 significant UKB covariates in equivalent non-UPD cases. 95% confidence intervals determined by permutation are shown in grey. Here $\hat{\beta}_{F_{ROH_WUPD}}$ was estimated by the UPD method but replacing each UPD case with the non-UPD subject which had the nearest F_{ROH} . The average effect ratio of 1.19 [95% CI 1.15 to 1.23] shows that non-UPD subjects with F_{ROH} equivalent to UPD cases would be expected to be associated with many covariates.

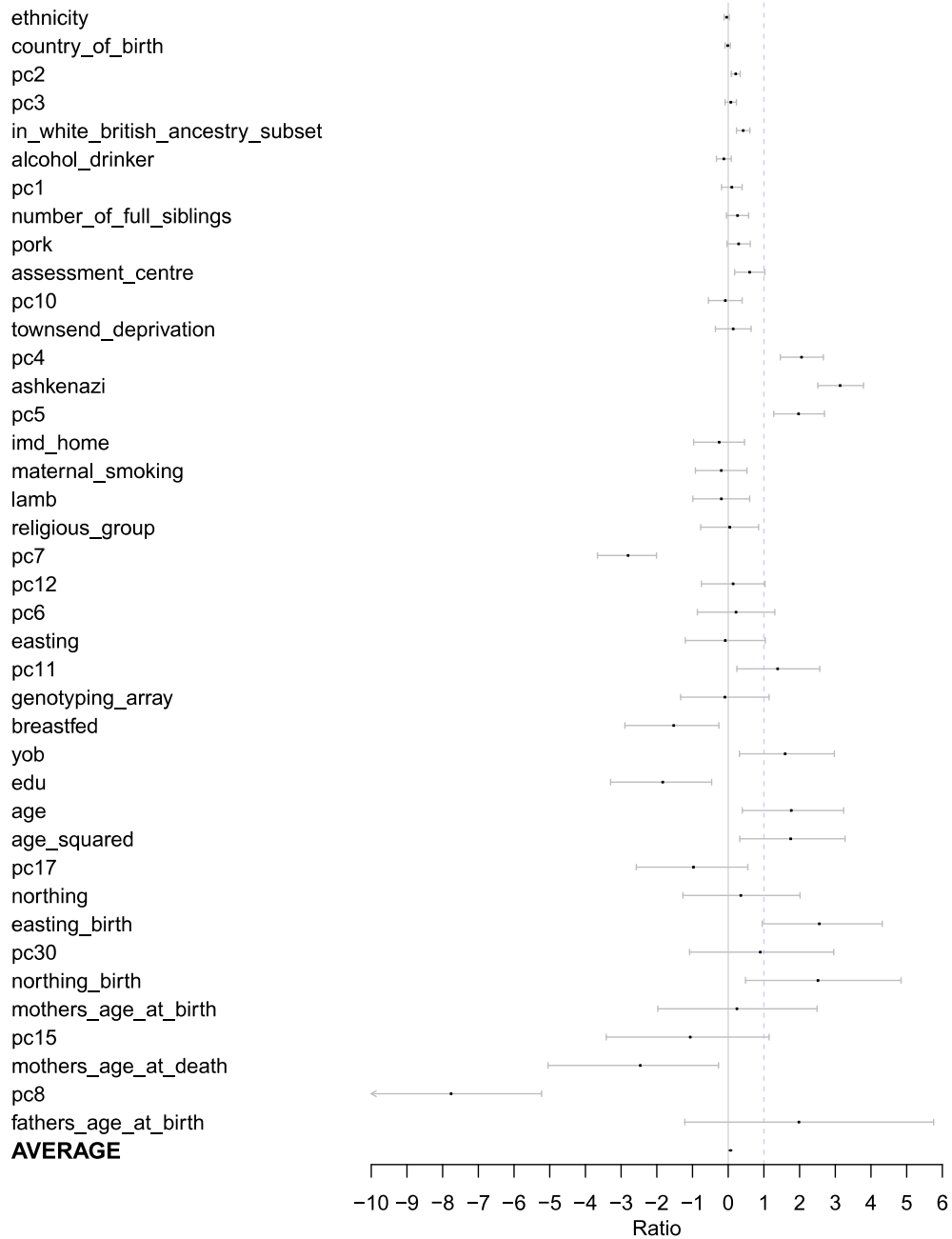


Figure 89: **The ratio of $\hat{\beta}_{FROH:FROH_cm} : \hat{\beta}_{FROH}$ for all 40 significant UKB covariates.** 95% confidence intervals determined by permutation are shown in grey. A ratio of 1 would be expected of entirely genetic effects, while a ratio of 0 would indicate complete confounding. The average effect ratio of 0.06 [95% CI 0.02 to 0.10] shows that the $FROH_cm$ method is partially effective at eliminating confounding, but the significant heterogeneity between traits (p -value 4×10^{-48}) indicates that confounding remains on some ancestry informative traits.

Appendix B Supplementary Table legends

Several tables referenced in the text are too large to print in this document. These “Supplementary Tables” are provided in an Excel spreadsheet which can be downloaded from <https://doi.org/10.7488/ds/3156>. Supplementary Table legends are provided below.

Supplementary Table 1: Summary of 234 sub-cohorts contributing to the ROHgen meta-analysis. *Ancestral group* is given for each sub-cohort, from one of seven broad groups (European, African, South & West Asian, East Asian, Hispanic, Japanese & Mixed). *Number genotyped* is the number of samples in each sub-cohort which pass genotypic Quality Control (QC). *Number of SNPs* is the number of SNPs which pass QC. *Number of samples phenotyped* is the maximum number of samples passing QC for any trait. *Number of phenotypes* is the number of traits (out of a possible 49) contributed by each sub-cohort.

Supplementary Table 2: Details of 234 sub-cohorts contributing to the ROHgen meta-analysis. Information on cohort recruitment, genotyping, quality control, acknowledgements and funding are presented. PMID, PubMed ID; MAF, minor allele frequency; HWE, Hardy-Weinberg equilibrium *p*-value threshold; SNP, single nucleotide polymorphism.

Supplementary Table 3: Homozygosity summary statistics by ROHgen sub-cohort. Measures of homozygosity are presented for each sub-cohort. *Relative per sample power*; is the approximate power of a sample relative to a single sample from the UK Biobank British cohort (UKBB). This is estimated by $\frac{\sigma_{F_{ROH}}^2}{\sigma_{F_{ROH_UKBB}}^2}$. *Percent total power*; is an estimate of the percentage of the total study power attributable to each cohort. This was calculated from the relative per sample power multiplied by the number of samples phenotyped.

Supplementary Table 4: ROHgen phenotype counts. Number of samples passing QC contributed by each cohort for each trait.

Supplementary Table 5: Covariates applied in each ROHgen sub-cohort. For each sub-cohort information is presented about the covariates applied to all traits. *Mixed model*; Although F_{ROH} is not narrow-sense heritable, a mixed model was fitted where significant family structure was known to exist. *Number of PCs*; the number of principal components of the Genomic Relationship Matrix (GRM) fitted as covariates. *Other covariates*; any other covariates specific to each cohort, such as assessment centre or genotyping batch.

Supplementary Table 6: **Summary of all ROHgen traits.** Full name and units are presented for all 49 traits analysed, along with the total number of samples measured and the number of sub-cohorts contributing each trait.

Supplementary Table 7: **ROHgen trait modelling.** Details of the linear models used are presented for all traits. A trait was only analysed where all *required covariates* were available, however if additional *optional covariates* were also available, they were included. A small number of traits were *rank normalised* before analysis, but then multiplied by the original trait standard deviation to return to trait units.

Supplementary Table 8: **ROHgen trait summary statistics.** Intra-sex means and standard deviations are presented for both sexes, for all traits.

Supplementary Table 9: $\hat{\beta}_{F_{ROH}}$ **for all ROHgen traits.** Effect size estimates, standard errors and associated p -values are presented for all traits meta-analysed. Traits are ordered by ascending p -value. Effect estimates are presented in trait units.

Supplementary Table 10: $\hat{\beta}_{F_{ROH}}$ **for all ROHgen traits normalised by intra-sex standard deviation.** To facilitate comparison effect size estimate have been normalised by intra-sex standard deviation. Normalised effect size estimates, normalised standard errors and associated p -values are presented for all traits meta-analysed. Traits are ordered by descending absolute magnitude of normalised effect size.

Supplementary Table 11: $\beta_{F_{ROH}}$ **estimated for each sex separately, for all ROHgen traits.** Estimates, standard errors and p -values of $\beta_{F_{ROH}}$ are presented for each sex separately. Effect size estimates and standard errors have been normalised by the corresponding intra-sex standard deviation. Also shown is the p -value for heterogeneity of effect between sexes, and traits are presented in ascending order of this p -value. Traits only measured in one sex are not shown.

Supplementary Table 12: **Estimates of $\beta_{F_{ROH}}$, $\beta_{F_{SNP}}$ and $\beta_{F_{GRM}}$ for all ROHgen traits.** Meta-analysed estimates, standard errors and p -values are shown for three different estimators of the inbreeding coefficient: F_{ROH} , F_{SNP} and F_{GRM} . For all significant traits, estimates of $\beta_{F_{ROH}}$ are larger than the corresponding estimates of $\beta_{F_{SNP}}$ and $\beta_{F_{GRM}}$. Figure 93 shows that this attenuation is a function of cohort homozygosity.

Supplementary Table 13: **ROHgen effect estimates from bivariate models of $Trait \sim F_{ROH} + F_{SNP}$.** Meta-analysed estimates, standard errors and p -values are shown for two inbreeding coefficient estimates ($F_{ROH} + F_{SNP}$) in a bivariate model.

Supplementary Table 14: **ROHgen effect estimates from bivariate models of $Trait \sim F_{ROH} + F_{GRM}$** . Meta-analysed estimates, standard errors and p -values are shown for two inbreeding coefficient estimates ($F_{ROH} + F_{GRM}$) in a bivariate model.

Supplementary Table 15: **Estimates of $\beta_{F_{ROH}<3Mb}$, $\beta_{F_{3Mb}<ROH<5Mb}$ and $\beta_{F_{ROH}>5Mb}$ for all ROHgen traits**. Meta-analysed estimates, standard errors and p -values are shown for three measures of homozygosity: $F_{ROH}<3Mb$, $F_{3Mb}<ROH<5Mb}$ and $F_{ROH}>5Mb}$. estimated in a multivariate model.

Supplementary Table 16: **ROHgen effect estimates stratified by cohort ancestry**. Effect size estimates, standard errors, p -values and sample sizes are shown for seven groupings of cohort ancestry. The cohort groupings are defined in Supplementary Table 1.

Supplementary Table 17: **ROHgen effect estimates stratified by inferred demographic history**. Figure 59 was used to infer the demographic history of each cohort, and effect estimates were meta-analysed within each group. Effect size estimates, standard errors, p -values and sample sizes are shown for each of the four inferred demographic history groups.

Supplementary Table 18: **Mean ROHgen trait residuals in bins of increasing F_{ROH}** . Samples were grouped in bins of similar F_{ROH} , and mean trait values were calculated within each group. The mean, standard error, p -value and sample size of each group are shown for all traits.

Supplementary Table 19: **Estimates of the effect of F_{ROH} within full-sibling families ($\beta_{F_{ROH-wSibs}}$), for all ROHgen traits**. Effect size estimates, standard errors, p -values and sample size are shown for all traits. Sample size is the number of individuals for whom at least one full sibling with a corresponding phenotype was identified.

Supplementary Table 20: **Location of 996 autosomal and 53 X chromosome windows used in local ROH analyses**. Windows are each 3 Mb long and non-overlapping.

Supplementary Table 21: **Average F_{ROH} by country of birth in UK Biobank**. Self-reported country of birth is available for 502,536 samples, born in 202 countries. The number of samples born in each country (N), and the average F_{ROH} of genotyped samples are given.

Supplementary Table 22: **Associations between 77 potential covariates and F_{ROH} in UK Biobank**.

Supplementary Table 23: **Associations between 1786 unprocessed phenotypes and F_{ROH} in UK Biobank.** Field_ID is the reference ID used by the UKB data showcase (<https://biobank.ndph.ox.ac.uk/ukb/>). Manually curated traits, retained for use in subsequent analyses, are shown in blue and given a short name in column “Thesis trait name”. Unlike in Supplementary Table 26, p -values, standard errors and t -statistics, are direct outputs from a linear regression model and have not been adjusted by resampling – they are therefore liable to be bias in traits with skewed distributions.

Supplementary Table 24: **Associations between 13,646 hospital ICD10 codes and F_{ROH} in UK Biobank.** Manually curated traits, retained for use in subsequent analyses, are shown in blue and given a short name in column “Thesis trait name”. P -values, standard errors and t -statistics were estimated by resampling, as described in section 3.2.6.1.

Supplementary Table 25: **Associations between 446 self-reported illnesses and F_{ROH} in UK Biobank.** Manually curated traits, retained for use in subsequent analyses, are shown in blue and given a short name in column “Thesis trait name”. P -values, standard errors and t -statistics were estimated by resampling, as described in section 3.2.6.1.

Supplementary Table 26: **Significant associations between 155 manually curated traits and F_{ROH} in UK Biobank.** P -values, standard errors and t -statistics were estimated by resampling, as described in section 3.2.6.1. The UKB.Field.ID column describes the UKB fields used in each trait.

Supplementary Table 27: **Normalised $\hat{\beta}_{F_{ROH}}$ for 155 significant traits in UK Biobank.** Effect sizes from Supplementary Table 26 are normalised by the trait intra-sex standard deviation. Traits are ordered by absolute magnitude of normalised effect size.

Supplementary Table 28: **ANOVA table for F_{ROH} modelled by 74 covariates.** Covariates are ordered by p -value of association with F_{ROH} .

Supplementary Table 29: **Linear model of height described by F_{ROH} and 74 covariates in UK Biobank.** Significant covariates which might be considered markers of socio-economic status are labelled with an asterisk. Height is in units of metres.

Supplementary Table 30: **Simulated ROH metrics in the offspring of various types of relative.** All columns, unless stated otherwise, are calculated only from ROH longer than 1.5 Mb, as this is the minimum length called in real data.

Supplementary Table 31: **Effect of F_{ROH} estimated within full-sibling families ($\hat{\beta}_{F_{ROH_wSibs}}$) in UK Biobank.** Estimates, standard errors, and p -values are given for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 32: **Effect of F_{ROH} estimated in 200 uniparental disomy (UPD) cases ($\hat{\beta}_{F_{ROH_wUPD}}$) in UK Biobank.** Estimates, standard errors, and p -values are given for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 33: **Effect of F_{ROH} conditioned on F_{ROH_cm} in UK Biobank.** Estimates, standard errors, and p -values are given for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 34: **Estimated effect of F_{ROH} in a cohort of 7207 adopted samples in UK Biobank.** Estimates, standard errors, and p -values are given for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 35: **Estimated effect of hemizygous deletions (F_{DEL}) in UK Biobank.** Estimates, standard errors, and p -values are given for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 36: **Estimates of $\beta_{F_{ROH}<3Mb}$, $\beta_{F_{3Mb}<ROH<5Mb}$ and $\beta_{F_{ROH}>5Mb}$ in UK Biobank.** Estimates, standard errors, and p -values are given for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 37: **$\beta_{F_{ROH}}$ estimated for each sex separately in UK Biobank.** Estimates, standard errors and p -values of $\beta_{F_{ROH}}$ are presented for each sex separately. Also shown is the p -value for heterogeneity of effect between sexes, and traits are presented in ascending order of this p -value. Traits only measured in one sex are not shown.

Supplementary Table 38: **Effect estimates in UKB from bivariate models of $Trait \sim F_{ROH} + F_{SNP}$.** Effect estimates, standard errors and p -values are shown for F_{ROH} and F_{SNP} in a bivariate model for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 39: **Effect estimates in UKB from bivariate models of $Trait \sim F_{ROH} + F_{GRM}$.** Effect estimates, standard errors and p -values are shown for F_{ROH} and F_{GRM} in a bivariate model for 155 significant individual traits and for the *multivariate* trait.

Supplementary Table 40: **All 2383 available UK Biobank phenotype fields.** In addition to the UKB field identifier and description, each field has been manually assigned a “phenotype type” to identify whether it is a potential phenotype or covariate.

Supplementary Table 41: **Estimated sample sizes required to have 80% power to find a significant effect.** Required sample sizes are given for a range of normalised effect sizes ($\beta'_{F_{ROH}}$) and standard deviations of cohort homozygosity ($sd(F_{ROH})$). Bonferroni correction for 20 traits is assumed. The standard deviation of F_{ROH} is given for all ROHgen cohorts in Supplementary Table 3, but a few notable cohorts are referenced in the table.

Supplementary Table 42: **Estimated sample sizes required to have 80% power to find a significant effect in a within sibling analysis.** Required sample sizes are given for a range of normalised effect sizes ($\beta'_{F_{ROH}}$), trait heritabilities, and within sibling standard deviations of F_{ROH} . Bonferroni correction for 5 traits is assumed. The within sibling standard deviation of F_{ROH} is given for all ROHgen cohorts in Supplementary Table 3, but a few notable cohorts are referenced in the table.

Appendix C Comparison of inbreeding coefficient estimates

This Appendix is reproduced verbatim from “Associations of autozygosity with a broad range of human phenotypes” by Clark et al.¹ but is entirely my own work and writing. It was written in response to reviewers’ comments at *Nature Communications* which raised concerns that F_{ROH} was an unsuitable measure of inbreeding, citing Yengo et al. (*PNAS* 2017)⁹⁸.

C.1 Introduction

Accurate estimation of Inbreeding Depression (β_F) requires precise and unbiased estimates of individual inbreeding coefficients (F). Numerous methods have been proposed for estimating inbreeding coefficients from dense SNP marker data. Broadly, these can be grouped into two categories: methods that consider the excess homozygosity at a large number of, preferably independent, markers (e.g., F_{HOM} , F_{UNI}) and methods that identify autozygous genomic segments from unbroken tracts of homozygous genotypes, which are unlikely to occur by chance (e.g., F_{ROH}).

To maximize statistical power and minimize bias in $\hat{\beta}_F$, an estimator of F should be both unbiased, ($E[\hat{F}] = F$) and precise, $MSE(\hat{F} - F) \ll var(F)$. Extensive comparisons of different estimators have been made by previous studies. Many of these conclude that F_{ROH} calculated from appropriately parameterized ROH calling gives estimates of F with minimal bias and lower variance than independent SNP methods^{82,114,118,167}. In contrast, Yengo et al. (*PNAS* 2017) claim that $\hat{\beta}_{F_{ROH}}$ may be upwardly biased by as much as 162%, and provide apparent evidence of this in both simulated and real data⁹⁸. Furthermore, Yengo et al. show, in both theory and simulation, that two independent SNP methods (F_{UNI} and F_{HOM}) give unbiased estimates of β_F when causal SNPs are a random subset of all genotyped SNPs. This would cast doubt on the validity of using F_{ROH} to estimate β_F , and suggests F_{UNI} might be a more appropriate measure. To understand the apparent contradictions between different studies, we have repeated and extended the investigations described in Yengo et al.

C.2 Method

Yengo et al. base their study on genotype data derived from the first phase of the UK Biobank (UKB) imputation. 9,493,148 SNPs with a minor allele frequency (MAF) of >1%, INFO score of > 0.3 and HWE p -value > 1×10^{-6} were selected from the imputation of 140,720 British individuals. SNP dosages were rounded to the nearest whole genotype and LD pruning $r^2 > 0.9$ was performed to reduce the number of SNPs to 3,857,369. F_{UNI} was calculated by the formula presented by Yang et al. (2011)²⁰ as $\hat{F}_{HI}$ (implemented in PLINK by parameters `--ibc Fhat3`) and ROH were called by two PLINK parameterizations, including the values proposed in Joshi et al. (2015)¹³. Joshi et al. validated the PLINK parameters for moderately dense SNP chips, but not for the different characteristics (SNP density, error rate) of genotypes called from imputed dosages. In particular, the PLINK method allows only one heterozygote or 5 missing genotypes within each ROH.

C.3 F_{ROH} calculated from SNP chip genotypes agrees well with F_{UNI} indicating minimal bias

We followed the method described in Yengo et al. to calculate both F_{UNI} and F_{ROH} from the UKB imputation, but initially compared Yengo's F_{UNI} to our F_{ROH} calculated from SNP chip data available for the same phase one individuals. From now on, we use lowercase “roh” to refer to measures based on imputed data and uppercase “ROH” to refer to those from SNP-chip genotypes. We find good correspondence between F_{UNI} and F_{ROH} (Figure 90).

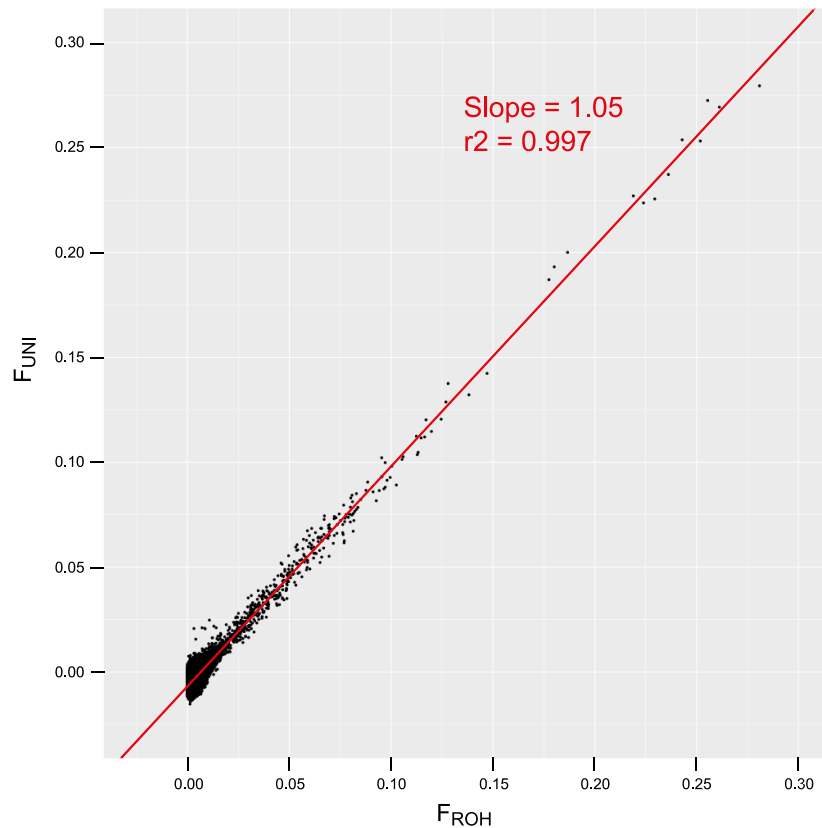


Figure 90: **Good correspondence between F_{ROH} and F_{UNI} .** For 141,774 British samples in UK Biobank F_{UNI} , calculated from excess homozygosity of imputed genotypes, is plotted against F_{ROH} , calculated from SNP array genotypes. A weighted linear regression line is shown in red. Because average inbreeding coefficients are low ($F_{ROH}=0.003$ in this population), it is high F individuals who contribute most of the statistical power to estimates of β_F . Weighting the regression by an estimate of power contribution $(F_i^{ROH} - \bar{F}_{ROH})^2$, we estimate $\beta_{F_{UNI}, F_{ROH}} = 1.05$ and $r^2 = 0.997$. The good correspondence between F_{UNI} and F_{ROH} suggests both have minimal bias in estimating F .

Since F_{UNI} is believed to be an unbiased estimator of F , the correspondence between F_{UNI} and F_{ROH} places limits on the possible bias of F_{ROH} . Specifically, if we model F_{ROH} as $F_i^{ROH} = \theta F_i + \epsilon_i$, then:

$$\frac{r^2}{\beta_{F_{UNI}, F_{ROH}}} \leq \theta \leq \frac{1}{\beta_{F_{UNI}, F_{ROH}}} \quad (49)$$

(see Appendix E.1). Substituting the regression values from Figure 90 gives $0.94 \leq \theta \leq 0.95$. I.e., the good correspondence between F_{UNI} and F_{ROH} limits the potential error in F_{ROH} to a small downward bias. This very small underestimate of F may be caused by autozygosity not captured in F_{ROH} , namely ROH of less than 1.5 Mb length or ROH in sparsely genotyped regions.

C.4 Calculating F_{roh} from imputed genotypes can introduce large downward bias

To call ROH from the UKB imputation we followed the method described by Yengo et al., but identified only 9,067,605 SNPs (out of a total of 72,355,667) matching their inclusion criteria (MAF > 1%, INFO > 0.3, HWE exact p -value > 1×10^{-6}). We then used PLINK 2.0 to convert the imputed genotype probabilities to the hard-called genotypes required for LD pruning and ROH calling. PLINK first converts genotype probabilities to an estimated dosage which is then rounded to the nearest genotype $\in [0 \ 1 \ 2]$. By default, a genotype is recorded if the estimated dosage is within 0.15 of any of $[0 \ 1 \ 2]$, otherwise the genotype is recorded as missing. Using these default parameters introduces 3.9% missingness into our dataset. We nevertheless proceeded to LD-prune this dataset, both with these missing data (method 2), and having removed all SNPs with a missing fraction > 0.03 (method 1). However, Yengo et al. state that “*Imputed SNPs were called to the genotypes having the largest posterior probability*” which is achieved by changing the PLINK hardcall parameter from 0.15 to 0.499999 (method Yengo). After LD pruning 3,061,484 SNPs remain for 141,774 British individuals.

For all three methods of data preparation, we called ROH and calculated F_{roh} using the Joshi et al. PLINK parameters also used by Yengo et al. We find that calling ROH

from hard-called imputed dosages (method Yengo) gives F_{roh} with an expected value of just 0.39 of the F_{ROH} obtained from SNP chip genotypes (Figure 91). More stringent treatments of uncertain genotype probabilities (methods 2 and 1) give progressively less downwards bias.

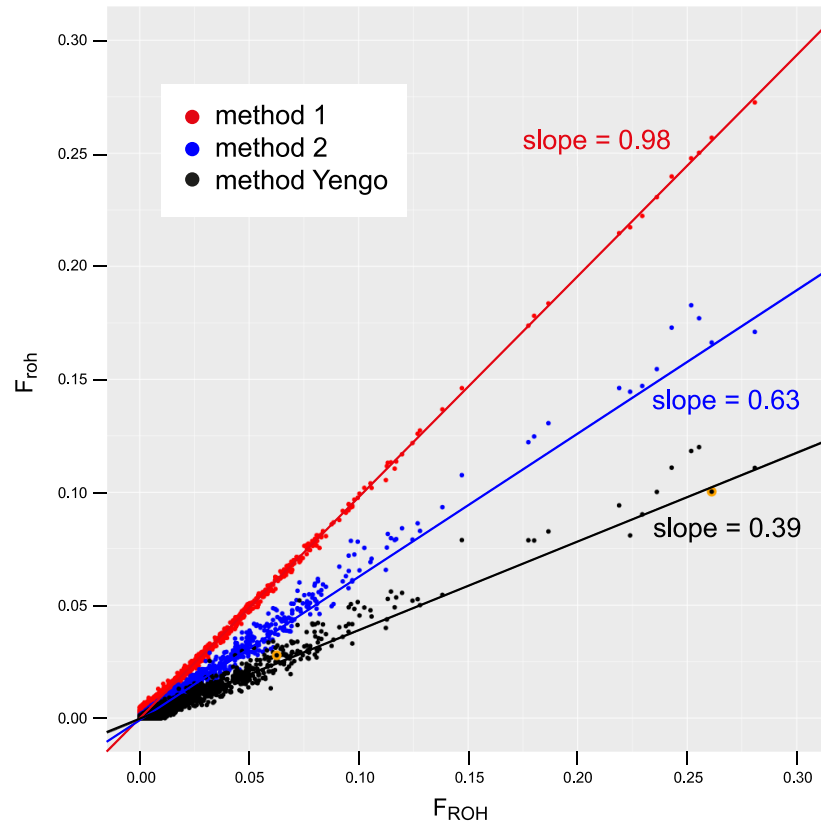


Figure 91: **Comparison of F_{roh} from imputed data and F_{ROH} from SNP array genotypes.** For 141,774 British samples in UK Biobank, F_{ROH} calculated from imputed genotypes is plotted against F_{ROH} calculated from SNP array genotypes for three methods of imputed genotype preparation. In method 1, in red, uncertain genotypes are removed. In method 2, in blue, uncertain genotypes are set to missing and in method Yengo, in black, no genotype filtering is performed. Increasingly permissive treatments of uncertain genotypes introduce increasing downward bias in F_{ROH} . Two high F_{ROH} samples are highlighted in orange and further explored in Figure 92.

To understand the cause of this downward bias in F_{roh} we plotted the genome-wide distribution of ROH for two high F individuals, highlighted in orange in Figure 91. For these two individuals all ROH called from SNP chip genotypes (in blue) and the Yengo et al. imputed data method (in red) are shown in Figure 92. Individual 1 has an $F_{ROH} = 0.261$, amongst the highest observed in UKB, and is

most likely the progeny of 1st degree relatives⁷² where $E[F] = 0.25$. Individual 2 has $F_{ROH} = 0.0626$ and is most likely the offspring of first cousins (3rd degree relatives) where $E[F] = 0.0625$. Calling ROH from hard-called imputed dosages fragments, and consequently fails to identify, many of the ROH found in SNP chip data. The resultant downward bias in F_{roh} is sufficient to explain an upward bias of up to 156% ($1/0.39$) in $\hat{\beta}_{F_{roh}}$.

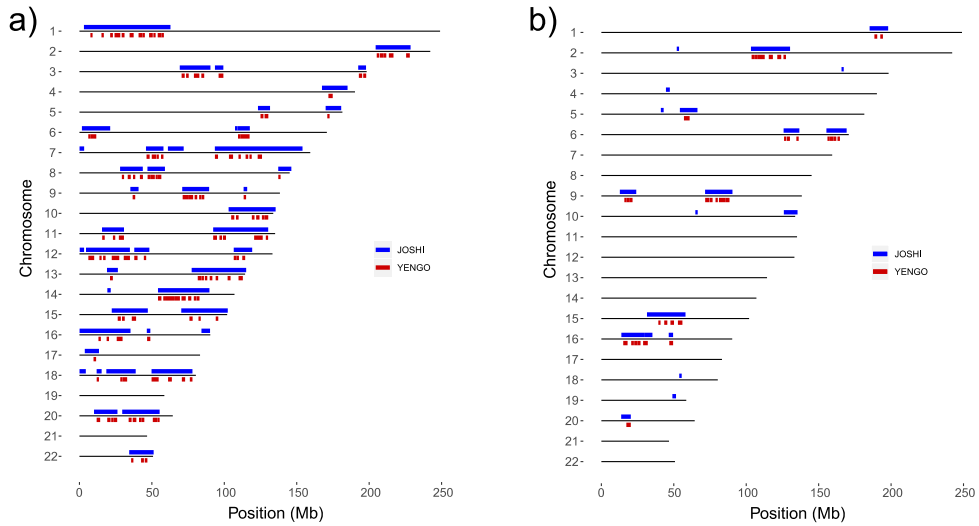


Figure 92: **Comparing ROH calling from SNP array genotypes and imputed dosages.** For two high F_{ROH} individuals, the locations of called ROH are compared for two methods. The method shown in blue calls ROH from SNP array genotypes using the parameters used in Joshi et al. (2015). The method shown in red calls ROH from hard called imputed dosages following the method described in Yengo et al. (2017). In both individuals the long ROH detected in SNP array genotypes, and which are thought to be autozygous segments, are broken up in the imputed data method by the presence of miscalled heterozygotes. **(a)** Individual with $F_{ROH} = 0.261$ thought to be offspring of first-degree relatives. **(b)** Individual with $F_{ROH} = 0.0626$ thought to be offspring of third-degree relatives.

In summary, calculating F_{ROH} from dense SNP chip genotypes, with the parameters used in this study, gives valid estimates of inbreeding coefficients. In contrast, calculating F_{roh} from unfiltered imputed genotypes, as done in Yengo et al., introduces a large bias which appears to be responsible for the poor performance of F_{roh} in that study.

C.5 $\hat{\beta}_{F_{GRM}}$ is downwardly biased in real data

Yengo et al. also show, in both theory and simulation, that $\hat{\beta}_{F_{UNI}}$ is an unbiased estimate of β_F in certain conditions, for example, when causal SNPs are a random subset of all observed SNPs. In real UKB data they find that $\hat{\beta}_{F_{roh}}$ is systematically of greater magnitude than $\hat{\beta}_{F_{UNI}}$, which they therefore interpret as empirical evidence that $\hat{\beta}_{F_{ROH}}$ is upwardly biased. We have already shown that calling ROH from imputed data may cause an upward bias of $\hat{\beta}_{F_{roh}}$, however, interestingly, we also observe that estimates obtained from unbiased SNP-chip genotypes ($\hat{\beta}_{F_{ROH}}$) are systematically larger than estimates obtained from frequency-based measures ($\hat{\beta}_{F_{SNP}}$ and $\hat{\beta}_{F_{GRM}}$) (Supplementary Table 12). Note, we use the nomenclature F_{GRM} to refer the $\hat{F}_{III}$ calculation first proposed by Yang et al⁶⁰. Although F_{GRM} and F_{UNI} are identical calculations (PLINK –ibc Fhat3), F_{GRM} is calculated from SNP-chip genotypes with a minimum MAF of 5%, while Yengo et al. calculated F_{UNI} from hard called imputed dosages with a minimum MAF of 1%. We explain the differences between $\hat{\beta}_{F_{ROH}}$ and $\hat{\beta}_{F_{GRM}}$ below.

C.5.1 *Causal variants for Inbreeding Depression are not in strong LD with common SNPs*

For all traits, we fit bivariate models with F_{ROH} and F_{GRM} as explanatory variables. For all 18 traits that were significant in the univariate analysis, we find that $\hat{\beta}_{F_{ROH}|F_{GRM}}$ is of greater magnitude than $\hat{\beta}_{F_{GRM}|F_{ROH}}$ in the conditional analysis (Supplementary Table 14). Furthermore, for 16 of these traits $\hat{\beta}_{F_{GRM}|F_{ROH}}$ does not differ significantly from zero. I.e., for many traits, the variation of F_{GRM} which is independent of F_{ROH} is not associated with any change in trait values. In Appendix D we show that these results are consistent with inbreeding depression caused by rare, but not common, variants. Furthermore, we observe that the downward bias of $\hat{\beta}_{F_{GRM}}$ is proportional to the ratio $\frac{var(F_{ROH})}{var(F_{GRM})}$ (Figure 93), as expected when the difference between F_{GRM} and F_{ROH} can be considered as estimation error (See Appendix E.2 below).

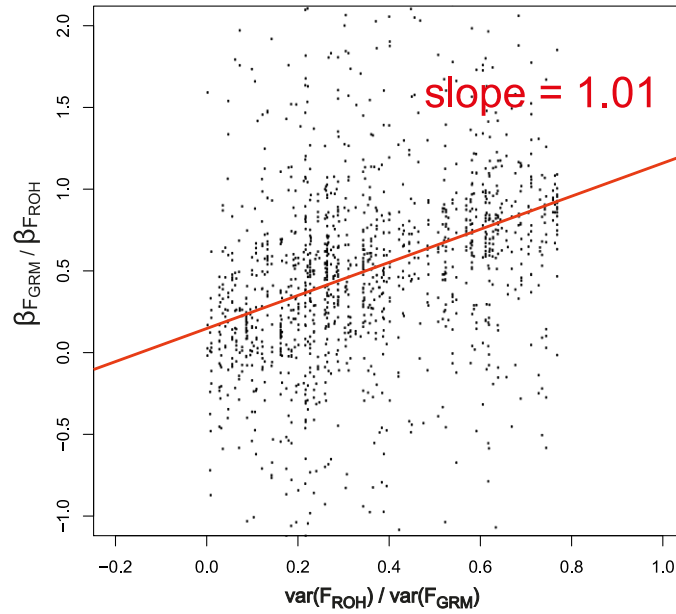


Figure 93: F_{ROH} is a lower variance estimator of the inbreeding coefficient than F_{GRM} . The ratio of $\beta_{F_{GRM}} : \beta_{F_{ROH}}$ is plotted against $var(F_{ROH})/var(F_{GRM})$ for all traits in all cohorts. When the variation of F_{GRM} which is independent of F_{ROH} has no effect on traits, $\hat{\beta}_{F_{GRM}}$ is downwardly biased by a factor of $var(F_{ROH})/var(F_{GRM})$ (Appendix E.2). A linear maximum likelihood fit, shown in red, has a gradient consistent with unity [1.01; 95% CI 0.84 – 1.18], as expected when the difference between F_{GRM} and F_{ROH} is not informative about the excess homozygosity at causal variants. (Appendix D).

In summary, Yengo et al. showed that $\hat{\beta}_{F_{UNI}}$ is unbiased when causal variants are a random subset of the observed SNPs. Although we agree with this statement, we find the evidence does not support the assumption of a random sample, but reveals the importance of rare variants, whose excess homozygosity is better predicted by F_{ROH} (Figure 95).

C.6 Comparison of genomic measures of inbreeding with genealogy

As a further assessment of the relative abilities of F_{ROH} , F_{SNP} and F_{GRM} to capture inbreeding, we analysed Pearson's product-moment correlations between the genomic inbreeding measures and pedigree inbreeding (F_{PED}) for 47,927 Icelanders with mostly complete (info score > 0.6)¹⁶⁸ 10 generation pedigrees. To decrease the confounding effects of pedigree misspecification, a small number of individuals (n=20) with extreme discrepancies between genetics and genealogy ($F_{ROH} > 0.05$ & $F_{PED} < 0.001$) were removed. The correlation was highest for F_{ROH} ($r = 0.779$),

lowest for F_{SNP} (0.632) and intermediate for F_{GRM} (0.682), further validating the utility of F_{ROH} as the most accurate of these genomic measures of inbreeding.

Appendix D Interpretation of $Trait \sim F_{ROH} + F_{GRM}$ models

This Appendix is reproduced verbatim from “Associations of autozygosity with a broad range of human phenotypes” by Clark et al.¹ but is entirely my own work and writing.

D.1 Are inbreeding effects caused by rare or common variants?

F_{ROH} is an estimate of autozygosity, which increases the homozygosity of all variants, both common and rare. In contrast, F_{GRM} is calculated from common SNPs (>5% MAF) and correlates well with the homozygosity of common SNPs, but less well with rare SNPs which may be in weak Linkage Disequilibrium (LD). We therefore performed bivariate models of all traits in real data ($Trait \sim F_{ROH} + F_{GRM}$) to establish whether the observed inbreeding effects associate more strongly with F_{ROH} or F_{GRM} . For all significant traits, we find the observed associations more attributable to F_{ROH} (Supplementary Table 14) suggesting inbreeding effects are caused by rare genetic variants. A recent study found evidence for a similar conclusion⁷¹, but to further support this interpretation we investigate below how both F_{ROH} and F_{GRM} predict the excess homozygosity of SNPs at a range of allele frequencies.

D.2 Relationships between F_{ROH} , F_{GRM} and excess homozygosity at different allele frequencies

For any trait exhibiting inbreeding depression, the degree of depression will be related to the excess homozygosity (above Hardy-Weinberg expectation) of the causal variants. In section E.3, below, we show that inbreeding depression, which is equal to the sum of the dominance deviations at the causal loci, is proportional to the inbreeding coefficient (F_{QTL}) defined in equation (51) overleaf.

$$ID_i = \sum_{i=1}^m \delta_i = \beta_F * F_{QTL} \quad (50)$$

Where

$$F_{QTL} = \frac{1}{m} \sum_{i=1}^m \frac{w_i(x_i^2 - (1 + 2p_i)x_i + 2p_i^2)}{2p_iq_i} \quad (51)$$

And

$$w_i = \frac{2p_iq_id_i}{\frac{1}{m} \sum_{i=1}^m 2p_iq_id_i} \quad (52)$$

We note that the unweighted form of equation (51) is identical to $\hat{F}_{III}$ introduced by Yang et al. (2011)⁶⁰, and implemented in PLINK by the parameters `-ibc Fhat3`. This is the same formula used to calculate F_{GRM} and F_{UNI} from different sets of marker SNPs. In summary, if the causal loci and effect sizes are known, a weighted calculation of $\hat{F}_{III}$ at the causal loci is directly proportional to the degree of inbreeding depression. We have used this, below, to simplify the simulation of inbreeding depression caused by variants at specific allele frequencies.

If we imagine inbreeding depression caused exclusively by variants at one allele frequency then, in the absence of strong selection or assortative mating on the causal loci in the current generation, the expectation of F_{QTL} will be equal to $\hat{F}_{III}$ calculated at marker variants of the same allele frequency (henceforth called F_{MAF}).

To calculate F_{MAF} across a range of allele frequencies we extracted SNPs at seven frequencies (MAF=0.01, 0.025, 0.05, 0.1, 0.2, 0.4 & 0.5) from 402,559 genetically British samples in the phase 2 UKB imputation. Selected SNPs were required to have a minor allele frequency (AF) within 10% of the specified MAF ($0.9 * \text{MAF} < \text{AF} < 1.1 * \text{MAF}$) and HWE p -value $> 1 \times 10^{-6}$. The numbers of SNPs retained at each MAF are reported in Table 8. F_{ROH} and F_{GRM} had previously been calculated, from SNP-chip genotypes, as part of the ROHgen meta-analysis.

Table 8: **Number of SNPs extracted from UKB imputation by allele frequency.** The excess homozygosity of SNPs at seven allele frequencies (F_{MAF}) was calculated for 402,559 genetically British samples in the phase 2 UKB imputation. The number of SNPs used at each frequency is shown.

MAF	Number of SNPs used in calculation of F_{MAF}
0.01	84835
0.025	122498
0.05	198310
0.1	261504
0.2	369777
0.4	253826
0.5	301191

To investigate the relationships between F_{ROH} , F_{GRM} and F_{MAF} we fit univariate ($F_{MAF} \sim F_{ROH}$ and $F_{MAF} \sim F_{GRM}$) and bivariate models ($F_{MAF} \sim F_{ROH} + F_{GRM}$) at each allele frequency. In the univariate models we find F_{ROH} to be an unbiased predictor of F_{MAF} across the entire frequency spectrum, while F_{GRM} is downwardly biased, particularly at low MAF (Figure 94a). Despite this downward bias, F_{GRM} is more strongly correlated than F_{ROH} at all MAF > 5% (Figure 94b).

In the bivariate model F_{GRM} is a stronger predictor of the homozygosity of common SNPs (>10%), but F_{ROH} is a stronger predictor for rare SNPs (Figure 95).

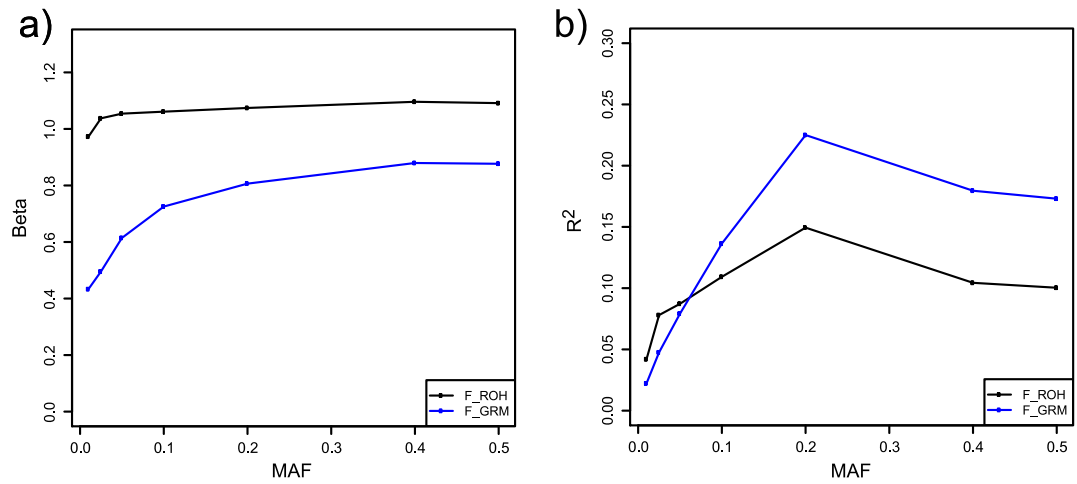


Figure 94: **Univariate relationships between estimates of inbreeding coefficient (F_{ROH} , F_{GRM}) and the excess homozygosity at specific allele frequencies (F_{MAF}).** The excess homozygosity of SNPs at seven allele frequencies (F_{MAF}) was calculated for 402,559 genetically British samples in the phase 2 UKB imputation. (a) Effect estimates of F_{ROH} and F_{GRM} on F_{MAF} . Univariate models of $F_{MAF} \sim F_{ROH}$ and $F_{MAF} \sim F_{GRM}$ were fitted at each allele frequency. A one unit increases in F_{ROH} is associated with a one unit increase in F_{MAF} across all allele frequencies. In contrast, the $\hat{\beta}_{F_{GRM}}$ is downwardly biased at all allele frequencies. (b) Correlations between of F_{ROH} , F_{GRM} and F_{MAF} . Univariate models of $F_{MAF} \sim F_{ROH}$ and $F_{MAF} \sim F_{GRM}$ were fitted at each allele frequency. Despite the downward bias of its effect estimate, F_{GRM} is more strongly correlated with F_{MAF} at most allele frequencies.

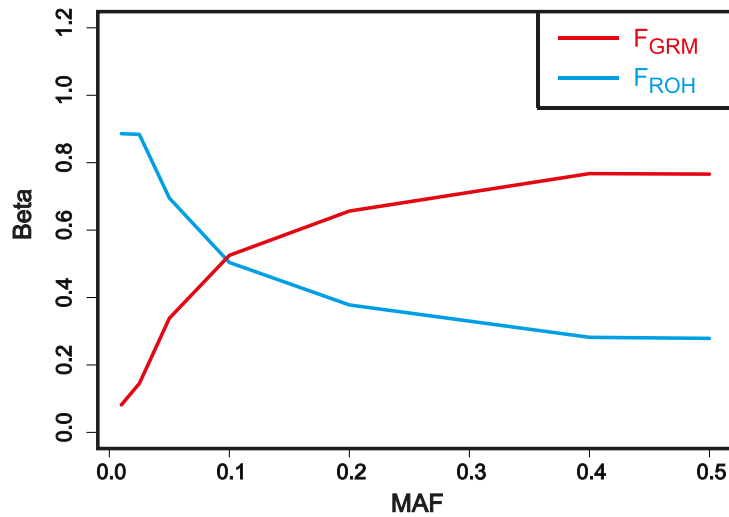


Figure 95: **F_{ROH} is a better predictor of rare variant homozygosity than F_{GRM} .** The excess homozygosities of SNPs, extracted from UK Biobank imputed genotypes, were calculated at seven discrete minor allele frequencies (F_{MAF}), and regressed on two estimators of inbreeding in a bivariate statistical model. The homozygosity of common SNPs is better predicted by F_{GRM} , but rare variant homozygosity is better predicted by F_{ROH} . The results from real data (Figure 40) are consistent with those simulated here, if the causal variants are predominantly rare.

D.3 Observed associations consistent with the homozygosity of rare, not common, SNPs

In real data models of $Trait \sim F_{ROH} + F_{GRM}$, we consistently find the observed associations are preferentially attributed to F_{ROH} rather than F_{GRM} (Figure 40). In light of Figure 95, these results are compatible with the action of rare, not common, causal variants.

Appendix E Proofs

E.1 Limits of bias in F_{ROH}

If F_{UNI} is an unbiased estimate of F then it can be expressed as

$$F_{UNI} = F + \varepsilon \quad (53)$$

If F_{ROH} is a potentially biased estimate of F then it can be expressed as

$$F_{ROH} = \theta F + \theta \varepsilon' \quad (54)$$

The regression slope (β) of F_{UNI} on F_{ROH} is known, and

$$\beta = \frac{cov(F_{UNI}, F_{ROH})}{var(F_{ROH})} \quad (55)$$

Substituting (53) and (54) in (55) and assuming independent errors gives:

$$\beta = \frac{\theta var(F)}{\theta^2 var(F) + \theta^2 var(\varepsilon')} \quad (56)$$

Rearranging (56) gives:

$$\theta = \left(\frac{1}{\beta}\right) \left(1 + \frac{var(\varepsilon')}{var(F)}\right)^{-1} \quad (57)$$

The range of $\frac{var(\varepsilon')}{var(F)}$ is limited by the correlation between F_{UNI} and F_{ROH} , and we

can put $\left(1 + \frac{var(\varepsilon')}{var(F)}\right)^{-1}$ in terms $\frac{var(\varepsilon)}{var(\varepsilon')}$ of by considering that

$$\frac{var(F_{ROH})}{var(F_{GRM})} = \frac{r^2}{\beta^2} \quad (58)$$

Again substituting (53) and (54) in equation (58) gives

$$\frac{\theta^2 var(F) + \theta^2 var(\varepsilon')}{var(F) + var(\varepsilon)} = \frac{r^2}{\beta^2} \quad (59)$$

Rearranging (59) gives:

$$\frac{var(\varepsilon')}{var(F)} = \frac{\theta^2 \beta^2 - r^2}{r^2 \frac{var(\varepsilon)}{var(\varepsilon')} - \theta^2 \beta^2} \quad (60)$$

Substituting equation (60) into equation (57)

$$\theta^2 = \left(\frac{1}{\beta}\right) \left(\frac{r^2 \frac{var(\varepsilon)}{var(\varepsilon')} - \theta^2 \beta^2}{r^2 \left(\frac{var(\varepsilon)}{var(\varepsilon')} - 1 \right)} \right) \quad (61)$$

As $\frac{var(\varepsilon)}{var(\varepsilon')} \rightarrow 0$, i.e., if F_{UNI} is precise and estimation errors entirely on F_{ROH} then equation (61) $\rightarrow$

$$\theta = \left(\frac{1}{\beta}\right) \left(\frac{\theta^2 \beta^2}{r^2} \right) \quad (62)$$

$$\theta = \frac{r^2}{\beta} \quad (63)$$

As $\frac{var(\varepsilon)}{var(\varepsilon')} \rightarrow \infty$, i.e., if F_{ROH} is precise and estimation errors entirely on F_{UNI} then equation (61) $\rightarrow$

$$\theta = \frac{1}{\beta} \quad (64)$$

Therefore, from the bounds of $\frac{var(\varepsilon)}{var(\varepsilon')}$ and equations (63) and (64)

$$\frac{r^2}{\beta_{F_{UNI}, F_{ROH}}} \leq \theta \leq \frac{1}{\beta_{F_{UNI}, F_{ROH}}} \quad (65)$$

E.2 Expected attenuation bias in $\hat{\beta}_{F_{GRM}}$

If F_{GRM} varies around F_{ROH} and the difference (ε) has no effect on the trait (y) then

$$F_{GRM} = F_{ROH} + \varepsilon \quad (66)$$

And

$$\beta_{F_{GRM}} = \frac{cov(F_{GRM}, y)}{var(F_{GRM})} \quad (67)$$

Substituting (66) in equation (67) gives:

$$\beta_{F_{GRM}} = \frac{cov(F_{ROH} + \varepsilon, y)}{var(F_{GRM})} \quad (68)$$

Because ε has no effect on the trait (y)

$$\beta_{F_{GRM}} = \frac{cov(F_{ROH}, y)}{var(F_{GRM})} \quad (69)$$

$$\beta_{F_{GRM}} = \beta_{F_{ROH}} * \frac{var(F_{ROH})}{var(F_{GRM})} \quad (70)$$

$$\frac{\beta_{F_{GRM}}}{\beta_{F_{ROH}}} = \frac{var(F_{ROH})}{var(F_{GRM})} \quad (71)$$

E.3 Calculation of F_{QTL} at known causal loci

If $x_i \in [0, 1, 2]$ is the number of copies of the reference allele at locus i of m causal loci, then the number of reference homozygotes at the locus is $\frac{x_i(x_i-1)}{2}$, the number of heterozygotes is $-x_i(x_i - 2)$, and the number of alternate homozygotes is $\frac{(x_i-2)(x_i-1)}{2}$.

If p_i is the frequency of the reference allele, and q_i is the frequency of the alternate allele, then the inbreeding depression with complete inbreeding (β) is

$$\beta = - \sum_{i=1}^m 2p_i q_i d_i \quad (72)$$

Where d_i is the difference between the heterozygote and mean homozygote value. We wish to define an inbreeding coefficient (F_{QTL}) which is directly proportional to realised inbreeding depression (the sum of the dominance deviations).

$$\beta F_{QTL} = \sum_{i=1}^m \delta_i \quad (73)$$

The dominance deviations (δ_i) for the three genotypes at a locus can be written in terms of d_i : $\delta_i \in [-2q_i^2 d_i, 2p_i q_i d_i, -2p_i^2 d_i]$. Substituting these dominance deviations and the genotype counts into equation (73) gives

$$\begin{aligned} \beta F_{QTL} = \sum_{i=1}^m & -\frac{x_i(x_i - 1)}{2} 2q_i^2 d_i - x_i(x_i - 2) 2p_i q_i d_i \\ & - \frac{(x_i - 2)(x_i - 1)}{2} 2p_i^2 d_i \end{aligned} \quad (74)$$

Rearranging equation (74) gives

$$\beta F_{QTL} = \sum_{i=1}^m -d_i (x_i^2 - (1 + 2)p_i x_i + 2p_i^2) \quad (75)$$

Substituting for β from equation (72) gives

$$F_{QTL} = \frac{1}{m} \sum_{i=1}^m w_i \frac{x_i^2 - (1 + 2p_i)x_i + 2p_i^2}{2p_i q_i} \quad (76)$$

Where,

$$w_i = \frac{2p_i q_i d_i}{\frac{1}{m} \sum_{i=1}^m 2p_i q_i d_i} \quad (77)$$

E.4 Statistical power calculations

The expected test statistic in the regression of a trait on F_{ROH} is given by

$$E[\hat{T}_{F_{ROH}}] = \frac{\beta_{F_{ROH}}}{se(\hat{\beta}_{F_{ROH}})} \quad (78)$$

Where $\beta_{F_{ROH}}$ is the true effect of F_{ROH} on the trait. The standard error of the regression estimate is given by

$$se(\hat{\beta}_{F_{ROH}}) = \sqrt{\frac{\sum_i \varepsilon_i^2}{(n-2) \sum_i (F_{ROH}^i - \bar{F}_{ROH})^2}} \quad (79)$$

Where n is the number of samples, and ε_i is the trait residual of the i^{th} sample. If n is large, and F_{ROH} explains a small fraction of trait variance, equation (79) is approximately equal to

$$se(\hat{\beta}_{F_{ROH}}) = \frac{sd(trait)}{\sqrt{n} * sd(F_{ROH})} \quad (80)$$

Substituting equation (80) into equation (78) then gives:

$$E[\hat{T}_{F_{ROH}}] = \frac{\beta'_{F_{ROH}}}{sd(trait)} * \sqrt{n} * sd(F_{ROH}) \quad (81)$$

So, the expected test statistic depends on the normalised effect size ($\beta'_{F_{ROH}}$), the standard deviation of F_{ROH} , and the square root of the sample size. The statistical power of a test is defined as the probability that test statistic reaches a defined significance threshold. The defined significance threshold will depend on the number of traits tested and the required confidence. If, for example, twenty traits are tested, a test statistic of 2.81 (calculated in R by `-qnorm(0.05/20)`) is required to reach a p -value of 0.05. To have an 80% probability of reaching this threshold requires the expected test statistic to be 0.84 (calculated in R by `qnorm(0.8)`) higher. Thus, to have an 80% statistical power to detect a significant effect, when testing twenty traits, requires the expected test statistic to equal 3.65 (2.81 + 0.84). Substituting these values into equation (81) and rearranging gives a required sample size equal to

$$n = \left(\frac{qnorm(power) - qnorm(\frac{p_value}{n_traits})}{\beta'_{F_{ROH}} * sd(F_{ROH})} \right)^2 \quad (82)$$

Supplementary Table 41 applies this formula to calculate the required sample size, assuming 80% power and twenty traits tested, for a range of plausible $\beta'_{F_{ROH}}$ and $sd(F_{ROH})$.

Power calculations for the within sibling method are more complicated, because power depends both on the number of siblings in each family and, crucially, on the heritability of the trait. For highly heritable traits, such as height, the power of the within sibling method is greater than would be naively expected, because the variance of height within siblings is substantially lower than variance within the population. The within sibling power calculation can be simplified by assuming exactly two siblings per family, in which case the method reduces to a paired t -test where $\beta'_{F_{ROH}}$ and $sd(F_{ROH})$ refer to the difference between sibling pairs, and n is the required number of pairs. The effect of trait heritability can be accounted for by increasing $\beta'_{F_{ROH}}$, since heritability multiplies the residuals (ε^2) in equation (79) by a factor of $(1 - h^2/2)$. The estimated required sibling sample sizes are therefore given in Supplementary Table 42 for high ($h^2 = 1$), medium ($h^2 = 0.5$) and low ($h^2 = 0$) heritability traits. If these heritability values seem slightly high, note that they include the correlation caused by common environment. In UKB the within sibling variance of height is almost exactly half the population variance, implying $h^2 = 1$.

Appendix F Traits affected by F_{ROH}

The theory of mutation-selection balance provides a useful theoretical framework for explaining the existence of inbreeding depression. However, mutation-selection balance is described in terms of evolutionary fitness (w) - a theoretical construct which cannot be directly observed. This makes it difficult to understand the relationship between evolutionary fitness and the measurable traits on which inbreeding depression is observed. Traits with an unavoidable effect on reproductive success (e.g., fertility, or viability) are commonly called components of fitness, and inbreeding depression on these is often explained by their inevitable correlation with evolutionary fitness. However, many traits that exhibit inbreeding depression (say height, or cognition, in humans) have no certain effect on reproductive success, making it slightly unclear why they are affected. Some authors suggest that these traits are correlated with evolutionary fitness (i.e.,

subject to directional selection), but this makes predictions which do not fit the data. For a trait such as human height, with abundant genetic variation, directional selection should cause a rapid response to selection and eventual fixation of height increasing alleles. It seems more plausible that many of traits affected by inbreeding depression (e.g., height) are currently variable around some equilibrium value. Other authors choose to describe traits affected by inbreeding depression as *"fitness-related"*, and while perhaps true, it is unclear what is meant by *"related"* if these traits are not correlated with fitness.

In this context, it is suggested here, that the traits affected by inbreeding depression might be interpreted as an inherent property of the ancestral genome. Mutation-selection theory describes inbreeding depression in terms of variants segregating in a population, and while it is undeniable that the currently segregating variants reveal inbreeding depression (no variants = no ID), it might be hypothesised that a completely different set of variants would have a qualitatively similar effect. For example, imagine it were possible to make a population of individuals who are all homozygous for the ancestral (or reference) genome. Now imagine a second population, composed of individuals who are all homozygous for different, randomly mutated copies of the reference genome. It seems plausible that changes seen in the second population would be the same as those seen in current inbred individuals (i.e., reduced stature, lower cognition, lighter skin tone, etc.). If true, the traits affected by inbreeding depression are not explained by currently segregating variants, nor by recent selection pressures, but are an inherent property of the genome which has itself been shaped by selection over billions of years.

This speculative prediction can be explained by an analogy with the evolution of the motor car. To many people a desirable car is one that is both fast and fuel efficient – two antagonistic requirements. In the 1960s a range of vehicles co-existed in the marketplace, from the fast, but thirsty Jaguar e-type to the slow, but frugal, Morris Minor. Both were justifiable choices at the time - as testified by their popularity. Over time, technology improved, and modern cars include the speedy Porsche 911 and the more fuel-efficient Toyota Prius. At least judging on these two traits alone,

no driver today would reasonably choose a Jaguar e-type or Morris Minor, since faster *and* more fuel-efficient models are available. However, these diverse cars all have at least one thing in common - if random, uninformed, changes are made to any of them (say, adjusting the ignition timing, or putting a few dents in the bodywork) the consequences are likely to be the same – i.e., slower *and* less fuel-efficient.

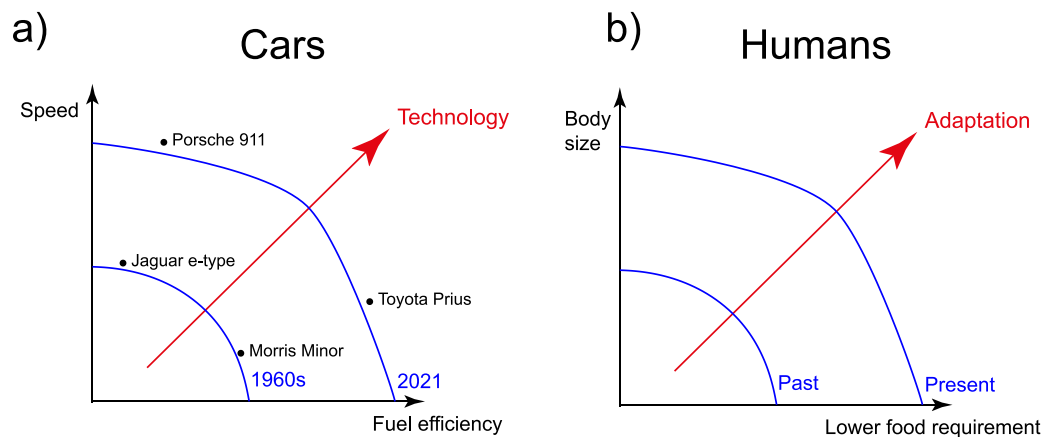


Figure 96: Technology improving the motor car as an analogy for adaptation improving the human phenotype. a) The desirability of a car is modelled as a compromise between the antagonistic properties of speed and fuel efficiency. Lines of equal commercial viability are shown in blue. Over time, technology has improved the compromise by introducing faster and more fuel-efficient models. b) Antagonistic traits in humans are represented by body size (bigger assumed to be better) and food requirement (less assumed to be better). Most segregating variants that increase body size are likely to increase food requirement by a corresponding amount and are therefore selectively neutral (blue lines). However, occasionally a variant may arise that improves the compromise between body size and food requirement (say by increasing metabolic efficiency). With a selective advantage this variant is more likely to reach fixation, resulting in adaptation.

In an analogous way, adaptation might be thought to have gradually “improved” the human genome by preferentially fixing advantageous mutations. Again, any random changes (i.e., homozygous mutations, in this case) to any version of the human genome are likely to change traits in the opposite direction to the one favoured by adaptation. By this model, while most random mutations are thought to be (on average) deleterious, those that are segregating at high frequency are likely to be (mostly) selectively neutral due to pleiotropic effects on many antagonistic traits.

In summary, I suggest that the traits affected by inbreeding depression are more easily understood as an inherent property of the human genome, than as a property of segregating variants, or current selective pressures. I speculate that the traits reduced by inbreeding have tended to be increased by adaptation, suggesting that, over evolutionary timescales, they have had a positive partial correlation with fitness.

Appendix G *Simulation R script*

```
# First read the recombination rate map into a data table called
df_map
library(data.table)
chromosome_genetic_lengths <- numeric(22)
chromosome_physical_lengths <- numeric(22)
df_map <- data.frame(Map=numeric(), Pos=numeric())
# Loop through each of the 22 autosomes
for(i in 1:22){
  strFilename <- paste("./genetic_map_GRCh37_chr", i, ".txt", sep="")
  df <- fread(strFilename)
  colnames(df)[c(2,4)] <- c("Pos", "Map")
  df$Pos <- (df$Pos - min(df$Pos))/1000000
  df$Map <- (df$Map - min(df$Map))/100
  chromosome_genetic_lengths[i] <- max(df$Map)
  chromosome_physical_lengths[i] <- max(df$Pos)
  if(i!=1){
    df$Pos <- df$Pos + max(df_map$Pos)
    df$Map <- df$Map + max(df_map$Map)
  }
  df_map <- rbind(df_map, df[, c("Map", "Pos")])
  print(i)
}
# Sort the recombination rate map table
df_map <- as.data.table(df_map)
setattr(df_map, "sorted", "Map")
setkey(df_map, Map)
# Define a few more map summary variables
max_genetic_length <- max(df_map$Map)
max_physical_length <- max(df_map$Pos)
chromosome_genetic_breaks <- c(0, cumsum(chromosome_genetic_lengths))
# We now have a sorted data table with 2 columns:
# map distance in Morgans (Map) and physical position in Mb (Pos)

# Define the number and type of samples to simulate
n_samples_to_simulate = 1000
n_meiosis_maternal = 3
n_meiosis_paternal = 3
n_common_relatives = 2

simulated_samples <- list()
for(k in 1:n_samples_to_simulate){
  # simulate where haplotype transitions occur in genetic distance
  n_segments <- rpois(1, (n_meiosis_maternal + n_meiosis_paternal)
    * max_genetic_length)
```

```

segment_breaks <- runif(n_segments,0,max_genetic_length)
segment_breaks <- c(chromosome_genetic_breaks, segment_breaks)
segment_breaks <- segment_breaks[order(segment_breaks)]
# convert genetic distance breakpoints to physical distance
i <- df_map[J(segment_breaks),roll=T,mult="last",.I,by=.EACHI]$I
segment_physical_breaks <-
  (((segment_breaks - df_map$Map[i])/
    (df_map$Map[i+1] - df_map$Map[i]))
    *(df_map$Pos[i+1] - df_map$Pos[i])) + df_map$Pos[i]
# get length and number of segments in this sample
segment_genetic_lengths <- diff(segment_breaks)
segment_physical_lengths <- diff(segment_physical_breaks)
nseg <- length(segment_genetic_lengths)
# Calculate expected inbreeding coefficient
Inbreeding_coeff <- (2*n_common_relatives) /
  (2^(n_meiosis_maternal + n_meiosis_paternal))
# Each segment has an probability F of being autozygous
j <- which(rbinom(nseg,1,Inbreeding_coeff)==1)
# get length and number of autozygous segments in this sample
roh_genetic_lengths <- 100 * segment_genetic_lengths[j]
roh_physical_lengths <- segment_physical_lengths[j]
nroh <- length(j)
# Calculate realised F in genetic and physical distance
f_cm <- sum(roh_genetic_lengths) / (100*max_genetic_length)
f_roh <- sum(roh_physical_lengths) / max_physical_length
# Finally, add this simulated sample to the list
simulated_samples[[k]] <- list(nroh=nroh,
  f_roh=f_roh,
  f_cm=f_cm,
  roh_physical_lengths=roh_physical_lengths,
  roh_genetic_lengths=roh_genetic_lengths)
}

# Extract number of ROH and F
nroh <- unlist(lapply(simulated_samples,"[[",1))
f_roh <- unlist(lapply(simulated_samples,"[[",2))
f_roh_cm <- unlist(lapply(simulated_samples,"[[",3))
# plot a histogram of F
hist(f_roh)

```

Appendix H Other publications

During the course of this PhD, I have contributed as an author to the following published papers.

1. Crawford, A. A. et al. Variation in the SERPINA6/SERPINA1 locus alters morning plasma cortisol, hepatic corticosteroid binding globulin expression, gene expression in peripheral tissues, and risk of cardiovascular disease. *Journal of Human Genetics* (2021). doi:10.1038/s10038-020-00895-6
2. Fairfield, C. et al. Genome-Wide Association Study of Non-Alcoholic Fatty Liver Disease Identifies Association with Apolipoprotein E. (2021). doi:10.1101/2021.05.05.21256592
3. Bretherick, A. D. et al. Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS Genetics* (2020). doi:10.1371/journal.pgen.1008785
4. Ceballos, F. C. et al. Autozygosity influences cardiometabolic disease-associated traits in the AWI-Gen sub-Saharan African study. *Nature Communications* (2020). doi:10.1038/s41467-020-19595-y
5. Clark, D. W. & et al. Associations of autozygosity with a broad range of human phenotypes. *Nature communications* (2019).
6. Gilbert, E. et al. The genetic landscape of Scotland and the Isles. *Proceedings of the National Academy of Sciences of the United States of America* (2019). doi:10.1073/pnas.1904761116
7. Timmers, P. R. et al. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife* (2019). doi:10.7554/elife.39856
8. Zeng, Y. et al. Parent of origin genetic effects on methylation in humans are common and influence complex trait variation. *Nature Communications* (2019). doi:10.1038/s41467-019-09301-y
9. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics* (2018). doi:10.1038/nrg.2017.109
10. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* (2018). doi:10.1038/s41588-018-0147-3
11. Karlsson Linnér, R. et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics* (2019). doi:10.1038/s41588-018-0309-3
12. Joshi, P. et al. Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nature Communications* 8, (2017).