UNIVERSITY OF CALIFORNIA,
IRVINE

Gene-based Dominance and Stabilizing Selection on Human Complex Traits

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Jaleal Salah Sanjak

Dissertation Committee:
Associate Professor Kevin R. Thornton, Chair
Professor Anthony D. Long
Assistant Professor Kirk G. Lohmueller

2018

DEDICATION

To my Granpap, Allan J. Melmed, for telling me not to believe everything I read in text books.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like first to thank my funding sources. My work was supported by NIH grant R01-GM115564 to Kevin R. Thornton as well as by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1321846 to myself. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscripts. Research presented herein was conducted using the UK Biobank Resource under project 12505.

I would also like to thank G3:Genes, Genomes, Genetics, PLoS Genetics and Proceedings of the National Academy of Sciences of the United States of America for publishing the work contained in this document. Their copyright policies allow me to reprint my articles here, provided full citations are presented.

Most importantly I would express my gratitude to my adivsor and mentor, Kevin R. Thornton. He leads by example and the lessons I have learned from him will be with me for the rest of my life.

ABSTRACT OF THE DISSERTATION

Gene-based Dominance and Stabilizing Selection on Human Complex Traits

By

Jaleal Salah Sanjak

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2018

Associate Professor Kevin R. Thornton, Chair

Most physical traits of agricultural and medical importance are complex, meaning that they are determined by multiple genetic and environmental factors. For decades it has a been a major goal in biology to be able to understand and predict complex phenotypes based on an individuals genomic sequence. Modern genotyping technologies have enabled the collection of massive samples of paired genotype-phenotype data. Despite this deluge of data, the genetic basis of complex traits remains unclear. Here I attempt to address this problem through a detailed simulation study based on explicit population genetic models for the maintenance of heritable phenotypic variation for a complex disease trait. The main conclusion of this study is that gene-based recessivity, under which compound heterozygotes can have excess disease risk, should be a leading candidate to explain some otherwise perplexing statistical properties of complex diseases. I complement this simulation study with an implementation of a statistical method found to be more powerful under a gene-based recessive genetic architecture. I then further support the assumed fitness model utilized in the simulation study through an empirical analysis of selection in a contemporary human population. Through studying the relationships between phenotypes and lifetime reproductive success, I showed that weak stabilizing selection is common on human traits and that many traits of clinical significance are under directional selection.

Chapter 1

Introduction

1.1 Chapter description

Introduction

1.2 The genetic architecture of human complex traits and disease

Natural phenotypic variation amongst individuals of a species is both striking and ubiquitous. Darwin showed us that these patterns of variation could be explained by the process of natural selection acting on stochastically generated biological variations. Mendel elucidated the beautiful laws of genetic inheritance through the study of simple traits in peas. Yet not all variation is genetic and very few traits of practical importance in medicine or agriculture follow the simple patterns found in Mendel's peas. Rather, as Fisher and Wright initially stated, most traits of practical importance follow complex patterns of inheritance and are

likely determined by several genetic and environmental factors[63, 257, 261, 260, 259]. So to what extent is this natural variation attributable to genetic causes? And what can we say about the specific nature of those genetic causes? These are the fundamental questions that motivate the study of quantitative genetics[60, 146].

After the sequencing of the human genome[128, 243] it became a major goal in biology find the genetic variants underlying common complex diseases. In an influential study, Risch and Merikangas suggested that genome wide association studies (GWAS) amongst unrelated individuals would be the best way to uncover the genetic basis of complex disease[201]. Specifically, they proposed the collection of a few thousand disease cases and healthy controls, and genotyping them at known common genetic markers. With enough markers throughout the genome it might be possible to tag disease risk alleles through linkage. This approach differed considerable from the previous paradigm in statistical genetics which involved the collection of closely related family members[129, 127]. The GWAS approach would enable easy collection of large study samples. And Risch and Merikangas further argued that variants effecting common disease were likely to be at intermediate frequency in the population and of smaller effect; GWAS is well suited to deal with genetic architectures of this sort.

With the goal of enabling GWAS[202], major consortium including the International HapMap project[76, 99] and the 1000 Genomes project[159] began to catalog common natural genetic variation in humans. The success of these population sequencing projects meant that genotyping population samples at known common variants became feasible. For the first time it would be possible to perform a GWAS along the lines of that proposed by Risch and Merikangas.

The Wellcome Trust Case Control Consortium (WTCCC) was the first to publish a ground breaking GWAS, which used thousands of cases and controls for each of seven common disorders and genotyping micro-arrays with around 500,000 common genetic variants[252]. The WTCCC results were generally striking for two reasons. Firstly, genome-wide inflations

of the association test statistics were observed for all seven diseases. Specifically the median of test statistics were far higher than expected under the null hypothesis of no association between genotype and phenotype. This was partially attributed to population structure, but that alone could not explain the extent of genome-wide inflation[197, 152, 268]. Secondly, in contrast to the inflation of the median test statistic, there were very few genome-wide significant associations. And those genome-wide significant associated variants explained very little genetic variance for disease risk. The field continued to publish GWAS for hundreds of diseases and established thousands of replicated associations[253]. Yet the fraction of genetic variance explained by disease associated variants remained extremely small[151]; this gap between explained variance and known heritability was aptly named the missing heritability problem.

Many hypotheses were proposed to explain why large GWAS failed to fully elucidate the genetic basis of disease risk[151]. An important class of hypotheses focused on the distribution of risk allele frequencies and their statistical effect sizes, i.e. the population genetic architecture of complex disease. A particularly spirited debate ensued surrounding whether common or rare variants were most relevant to complex disease risk[211, 77]. There was a popular belief that common large effect variants were likely to explain a large portion of the genetic variance underlying common diseases[198], the common disease common variant hypothesis(CDCV). The CDCV hypothesis specifically predicts that there are a few major important common variants. This stands in contrast to the infinitesimal model[64, 246], under which disease risk is determined by a very large number of causal variants each with extremely small effect sizes[246]. The CDVC hypothesis and the infinitesimal model differ quite substantially in their predictions on the power of GWAS as a function of sample size. In particular, the infinitesimal model requires that GWAS be performed with extremely large samples sizes to obtain statistically significant associations at single genetic variants[221]. While both hypothesis posit the importance of common genetic variation, the CDCV hypothesis has been largely discredited strictly on the basis of first-generation

GWAS results[151, 77]. However, the virtues of the infinitesimal model have been expounded from theoretical[10] and biological[19] perspectives and it remains firmly in contention as an explanation for the missing heritability problem.

A third, population genetics based hypothesis supposes that rare alleles of large effect (RALE) could drive the heritability of complex disease and produce associations in a common variant GWAS[39, 48]. The arguments in favor of the RALE hypothesis were predicated on arguments first put forward by Pritchard[187] prior to the publication of any large GWAS. Pritchard argued that if disease risk alleles were under negative selection, which might be a reasonable assumption for many diseases, then they would likely be rare in the population. A second line of argument in favor of the RALE hypothesis is based on the concept of allelic heterogeneity[158]. The allelic heterogeneity argument supposes that the relevant genomic loci are likely to harbor a variety of possible causal sites. This implies that any particular individual with elevated disease risk could have any one of a large number of possible multi-variant risk alleles. This model of allelic heterogeneity is supported by observations from studies of human Mendelian disease[158] and model organisms[115, 32]. However, most models of RALE predict the appearance of many low-frequency disease associated variants in a GWAS[255, 77], which has not typically been observed.

In addition to negative selection and allelic heterogeneity, recent human demographic history may have increased the relevance of rare variants. Early theoretical work used coalescent theory to characterize the relationship between a deterministically changing population size and expected genetic variation[81, 81, 206]. This theoretical work was quickly followed by the development of statistical methods for differentiating population expansion and neutral evolution[72, 14, 5], as well as for the inference of demographic history[83, 136]. Empirical study of human genomic data from European populations has revealed evidence consistent with a bottleneck corresponding to the Out-of-Africa(OOA) event and a recent exponential population expansion[196, 80, 229]. Evidence of an excess of rare variants in humans is

strongly supported by population genomic data[42, 109, 73, 75].

Yet the question of whether this excess of rare variants corresponds to a increased contribution of RALE to heritability of complex traits and disease remains an open question. Several authors have tried to address the relationship between human demography and the genetic basis of complex traits[74, 217, 141, 278, 242, 209, 8]. These studies showed that it is possible for the recent population expansion to have increased the role of rare variants in complex disease risk. However, no strong empirical conclusion has been reached to date because the predictions from theoretical models are very sensitive to modeling assumptions. Importantly, there is a balance between the effects of the OOA bottleneck[217, 141], the dominance of risk alleles[8, 209] and intensity of selection[278, 242].

Did the failure of the first generation of GWAS to elucidate the genetic basis of complex disease[151] have a silver lining? The development of theoretical models of complex disease might be viewed in that light. Prior to the WTCCC studies, attempts at modeling complex disease were fairly rudimentary[201, 198, 187]. After the missing heritability problem was posed[151] there were a wide range of attempts to model the genetic architecture of complex traits with both analytical[59] and computational approaches.

There is considerable diversity amongst the computational approaches. Some early work modeled risk allele frequencies as independent random variables and related those allele frequencies to power of GWAS[135, 147, 12]. Another class of simulation based approaches chose to generate individual level genetic data from a coalescent model and apply trait effects in a post-hoc manor[48, 145, 262]. Others chose to perform forward-in-time simulations so that genetic variation can be sampled from populations evolving under mutation-selection balance. A major benefit of forward simulation is the ability to simulate selection on variants arising in a large recombining region; in contrast, coalescent methods are incapable of simulating multiple linked selected variants. One popular way to perform forward simulations is to draw the fitness effects of risk alleles from a empirically motivated distribu-

tion and then leverage explicit models of the relationship between fitness effects and trait effects[114, 183, 21, 1, 217, 141, 278, 240, 173, 31, 242]. Another way to do forward simulation relies on drawing trait effects from empirically motivated distributions and making explicit assumptions about how selection acts directly of traits[232, 209].

Despite considerable effort to model complex diseases for the purposes of understanding their genetic architecture, no clear picture has emerged regarding best practices and what models are most consistent with empirical data. The only real agreement is that any such model should be based on evolutionary principles. And it is this fundamental supposition that motivates the studies presented later in this document. Given the centrality of the evolutionary process in understanding the genetic architecture of human complex disease, it is important to take a step back and consider the literature on how genetic variation for complex traits is generally maintained in populations.

1.3 The maintenance of heritability in populations

Those familiar with population genetics are aware of one of the most longstanding questions referred to as Lewontin's paradox of variation. It asks why should there be so much genetic variation when simple models of selection predict there to be much less[134]? An analogous, but not identical, question exists in quantitative genetics which is why are quantitative traits heritable despite being apparently subjected to strong stabilizing selection[64, 203, 237]? The paradox of quantitative genetic variation can be more generally stated as a question regarding what evolutionary forces maintain heritable variation in populations. The most simple, and thus popular, class of models for the maintenance of heritability focus on the potential balance between mutation, recombination and stabilizing selection[28].

It is helpful to lay out the central mathematical model of a phenotype in quantitative genetics

before delving into how phenotypes evolve. The simplest model says that individuals in a population have phenotypes which are composed of independent genetic and environmental effects, which combine additively.

$$P = G + E$$

The environmental component is typically treated as a normal random variable.

$$E \sim N(0, \sigma_e^2)$$

The genetic component, G, is the genotype to phenotype map and it maps a discrete genome-wide genotype to a real valued number. Under this model the total phenotypic variance, σ_p^2, can be decomposed into the sum of genetic variance, σ_g^2, and environmental variance, σ_e^2.

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2$$

The genetic variance can be further decomposed into additive and non-additive components following the same principles that underlie the statistical analysis of variance (ANOVA) [60, 146]. For example, if a trait is determined by loci with some degree of dominance or recessivity then Fisher[63] showed that the genetic variance is separable into additive, σ_a^2, and dominance, σ_d^2, components .

$$\sigma_p^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2$$

Another important concept in the study of genetic variance is heritability, which can be defined in a broad and narrow sense. Broad sense heritability, H^2, is the percent of total variance which is due to genetic variation, while heritability in the narrow sense, h^2, only considers the additive genetic variance. Heritability is a linear function of genetic variance and can also be decomposed into additive, dominance and interaction components. Although

the latter is rarely useful.

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2}$$

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

$$\delta^2 = \frac{\sigma_d^2}{\sigma_p^2}$$

Fisher[63] and Wright[256, 261, 260, 259, 258] also pointed out that the contribution of non-additive genetic effects such as dominance and epistasis to total genetic variance would depend variant frequencies. Under a neutral site-frequency spectrum, non-additive functional effects would typically contribute to additive genetic variance[44, 93]. This implies that one can not determine the genotype to phenotype map based on estimates of genetic variance components. But, variance components can still be very informative when combined with other forms of statistical genetic information.

With these basic concepts at hand, we can turn to the question of how heritability is maintained. Fisher was the first to point out that stabilizing selection for an intermediate optimum should reduce genetic variance and thus proposed that fitness related traits should have low heritabilities[64]. Empirically, traits with strong correlation to fitness do have slightly lower heritability then traits uncorrelated to fitness[95]. Yet, when fitness traits are scaled appropriately it has been shown that the reduced heritability is primarily due to relatively large environmental variance rather then an absence of genetic variance[95]. Further, empirical evidence for the prevalence of stabilizing selection has been observed in many species[103, 117]. These two observations motivate the study of a possible balance between mutation and stabilizing selection.

Haldane first proposed using a Gaussian function as a mathematical model for a stabilizing selection fitness function[85]. Soon after Haldane's work, several authors pursued mathematical analysis of polygenic traits evolving under Haldane's fitness function subject to

mutation at many biallelic loci[130, 25, 26] or at a few loci each with an infinite number of alleles[43, 112, 125, 237]. The two most influential analyses of mutation-stabilizing selection balance come from Lande[125] and Turelli[237], who explored equilibrium approximations to the continuum of alleles model[43, 112], introduced first for a single locus by Crow and Kimura [43, 112], in the small and large mutation effect size limits respectively.

These standard mutation-stabilizing selection models typically assume an additive genetic model, under which an individuals genetic value is a simple sum of the effects of its constituent alleles.

$$G = \sum_{i=1}^{n} (x_i + x_i')$$

The allelic effect values are determined by a random-walk mutation model. The random-walk mutation model assumes that when a mutation occurs on an allele with value x, the allele will take on a new effect value of $x + \xi$. On a more fine-grained level, this type of mutational model corresponds to assuming that when mutations occur at a locus they do not erase the presence of prior mutations, but instead could be considered as new variable sites. The effects of the mutant alleles at new variable sites is determined by the distribution of ξ. The distribution of mutational effect sizes, $p_\xi(\xi)$, can be treated as a Gaussian distribution, exponential or gamma distribution. The variance of the distribution of mutations effect sizes is often written as γ^2 and tends to play a pivotal role determining the properties of mutation-stabilizing selection balance models.

As previously mentioned, in the standard stabilizing selection model a Gaussian function is used[85]. Under this model fitness will decrease quadratically upon phenotypic deviation from the optimum P_o. The inverse selection intensity,V_s, is the critical parameter of this Gaussian fitness function.

$$W_P(P) = e^{-\frac{(P-P_O)^2}{2*V_s}}$$

Whenever V_s is a positive value then this Gaussian fitness function results in what is called

Gaussian stabilizing selection. If V_s takes on negative value then this model results in disruptive selection. Empirically, stabilizing selection has been shown to be more common in many species, but disruptive selection is not entirely rare[117]. The models of phenotype, mutation process, recombination and selection can be combined to form a cohesive theory of how polygenic quantitative traits might evolve within a population[28]. Therefore this body of mathematical theory can provide a powerful platform for understanding empirical data relating to the genetic architecture of complex traits.

1.4 The following documents

The goal of my graduate research was to make a contribution to how we understand the genetic architecture of complex traits in humans. I tried to combine theoretical, computational and empirical approaches to simultaneously interpret the empirical data in light of theory and inform the theory in light of the data. The following documents discuss my contributions to the study of genetic architecture of complex traits in significant detail. Three of my research chapters have been published in peer-reviewed journals. All of this work was completed during my time as a graduate student at the University of California, Irvine.

In chapter 2, I explore the effects of recent human population expansion and the genotype-phenotype map on the statistical properties of quantitative genetic variation. As mentioned earlier, forward in time evolutionary simulation has become the standard approach to generating samples of genetic variation for testing hypothesis regarding the genetic architecture of complex traits and disease. Using the fwdpp template library, developed by Kevin Thornton[231], I simulated a 100 kilobase region of the human genome evolving in a large population subject to mutations that affect a quantitative trait subject to Gaussian stabilizing selection. I utilize three different demographic models: a constant population size, recent exponential growth and a model inferred directly from human genomic data[229]. With this

I showed, in agreement with previous literature, that RALE can be critical to the genetic architecture of a complex trait when the population has recently expanded. Further, I explored the effect of dominance in the genotype-phenotype map. I explored two different conceptions of how dominance presents itself with respect to the function of a genetic element. One concept of dominance treats it as a property of the function of a gene[232], implying that the haplotype upon which a mutation arises is relevant for determining its effect, and the other concept treats dominance as a property of the mutation itself. Through several comparisons of these two models of dominance and a purely additive model with empirical results from human statistical genetics, including variance component analyses and GWAS, I showed that the gene-based conception of dominance is much more realistic. Importantly, under gene-based dominance, population expansion and stabilizing selection I showed that rare variants can be important without creating a statistical excess of low-frequency disease associated markers.

In addition to the contribution regarding interpretation of recent data from statistical genetics, chapter 2 also explores the statistical power of several genetic association methods. As was found in Thornton, et al 2013[232], the Excess of Significant Markers (ESM) test showed more power than standard methods[169, 131]. Chapter 3 showcases an efficient implementation and empirical validation of the ESM test. I implemented the ESM test as a C++ command-line tool and packaged it into a robust computational pipeline. The test was empirically validated on the WTCCC dataset within which associations were discovered that were only previously discovered in much larger datasets.

Throughout the simulation work in chapter 2, I used values for intensity of selection, V_s, based on assumptions from observations in model organisms and natural populations of non-human species[103, 117]. Specifically, I had assumed that selection on common complex human disease traits was not very strong compared to that found on traits of obvious ecological relevance in natural populations, as is typically studied in the literature. This was

a valid approach, but given the scale of human genetic datasets being collected it became a apparent that we could do better[222]. Thus chapter 4 provides an empirical analysis of contemporary selection in humans. Using data from the UK Biobank[228] I estimate the the strength of stabilizing selection, V_s, in a wide range of human traits. Through a series of phenotypic regressions, based on the work of Lande and Arnold[126], I observed widespread stabilizing selection that was much weaker than that found in natural populations of non-human species[103, 117]. These observations largely validate the assumptions used in chapter 2 and further bolster the conclusions found therein. Further, through analyses of genetic data, I observed directional selection on several interesting biometric and life-history traits.

Finally, chapter 5 presents a simulation study of methods to estimate the number of alleles at a causal locus. At the core of the RALE hypothesis is the suggestion that causal loci will harbor many different causal variants. Each haplotype with a unique causal variant has the potential to be a different functional allele. Recently, empirical evidence has been presented in Drosophila melanogaster that supports the presence of multiple functional alleles at expression QTL[115]. Therefore, in chapter 5 I explore statistical approaches[18, 110] to estimating the number of functional alleles at a QTL under several mapping panel designs

All together, my contributions have deepened the specificity with which the genetic architectures of human complex traits are studied from evolutionary perspectives. I have explored the effect genotype-phenotype map theoretically and empirically estimated key fitness function parameters. I carried these insights through to characterize how evolutionary parameters affect important statistical properties of GWAS. The future of this field lies in continuing along the path set out here. We need to understand how all the aspects of our model relate the observed data and eventually build inferential frameworks that tie together theory and experiment.

Chapter 2

A model of compound heterozygous, loss-of-function alleles is broadly consistent with observations from complex-disease GWAS datasets

2.1 Article

A model of compound heterozygous, loss-of-function alleles is broadly consistent with observations from complex-disease GWAS datasets

Jaleal S. Sanjak, Anthony D. Long, Kevin R. Thornton

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, 92697
Center for Complex Biological Systems, University of California Irvine, Irvine, CA 92697, USA

Corresponding Author: Jaleal S. Sanjak

Email: jsanjak@uci.edu

2.2 Preface

This chapter was originally published in *PLoS Genetics* under the title "A model of compound heterozygous, loss-of-function alleles is broadly consistent with observations from complex-disease GWAS datasets" [209]. It is reprinted here in its original form. The simulation machinery used to generate the dataset was written by Kevin Thornton with input from myself. I performed the bulk of the simulations, the statistical analysis of the dataset, drew the primary conclusions, and wrote the text of the paper.

2.3 Abstract

The genetic component of complex disease risk in humans remains largely unexplained. A corollary is that the allelic spectrum of genetic variants contributing to complex disease risk is unknown. Theoretical models that relate population genetic processes to the maintenance of genetic variation for quantitative traits may suggest profitable avenues for future experimental design. Here we use forward simulation to model a genomic region evolving under a balance between recurrent deleterious mutation and Gaussian stabilizing selection. We consider multiple genetic and demographic models, and several different methods for identifying genomic regions harboring variants associated with complex disease risk. We demonstrate that the model of gene action, relating genotype to phenotype, has a qualitative effect on several relevant aspects of the population genetic architecture of a complex trait. In particular, the genetic model impacts genetic variance component partitioning across the allele

frequency spectrum and the power of statistical tests. Models with partial recessivity closely match the minor allele frequency distribution of significant hits from empirical genome-wide association studies without requiring homozygous effect-sizes to be small. We highlight a particular gene-based model of incomplete recessivity that is appealing from first principles. Under that model, deleterious mutations in a genomic region partially fail to complement one another. This model of gene-based recessivity predicts the empirically observed inconsistency between twin and SNP based estimated of dominance heritability. Furthermore, this model predicts considerable levels of unexplained variance associated with intralocus epistasis. Our results suggest a need for improved statistical tools for region based genetic association and heritability estimation.

2.4 Author Summary

Gene action determines how mutations affect phenotype. When placed in an evolutionary context, the details of the genotype-to-phenotype model can impact the maintenance of genetic variation for complex traits. Likewise, non-equilibrium demographic history may affect patterns of genetic variation. Here, we explore the impact of genetic model and population growth on distribution of genetic variance across the allele frequency spectrum underlying risk for a complex disease. Using forward-in-time population genetic simulations, we show that the genetic model has important impacts on the composition of variation for complex disease risk in a population. We explicitly simulate genome-wide association studies (GWAS) and perform heritability estimation on population samples. A particular model of gene-based partial recessivity, based on allelic non-complementation, aligns well with empirical results. This model is congruent with the dominance variance estimates from both SNPs and twins, and the minor allele frequency distribution of GWAS hits.

2.5 Introduction

Risk for complex diseases in humans, such as diabetes and hypertension, is highly heritable yet the causal DNA sequence variants responsible for that risk remain largely unknown. Genome-wide association studies (GWAS) have found many genetic markers associated with disease risk[253]. However, follow-up studies have shown that these markers explain only a small portion of the total heritability for most traits [151, 244].

There are many hypotheses which attempt to explain the 'missing heritability' problem [151, 244, 77, 205]. Genetic variance due to epistatic or gene-by-environment interactions is difficult to identify statistically because of, among other reasons, increased multiple hypothesis testing burden [56, 251], and could artificially inflate estimates of broad-sense heritability [277]. Well-tagged intermediate frequency variants may not reach genome-wide significance in an association study if they have smaller effect sizes [64, 246]. One appealing verbal hypothesis for this 'missing heritability' is that there are rare causal alleles of large effect that are difficult to detect [158, 39, 77]. These hypotheses are not mutually exclusive, and it is probable that a combination of models will be needed to explain all heritable disease risk [245].

The standard GWAS attempts to identify genetic polymorphisms that differ in frequency between cases and controls. A complementary approach is to estimate the heritability explained by genotyped (and imputed) markers (SNPs) under different population sampling schemes [264, 78]. Stratifying markers by minor allele frequency (MAF) prior to performing SNP-based heritability estimation allows the partitioning of genetic variation across the allele frequency spectrum to be estimated [263], which is an important summary of the genetic architecture of a complex trait [187, 59, 177, 1, 217, 141, 263, 242]. This approach has inferred a contribution of rare alleles to genetic variance in both human height and body mass index (BMI) [263], consistent with theoretical work showing that rare alleles will have large

effect sizes if fitness effects and trait effects are correlated [59, 1, 278, 217, 141, 31, 242]. Yet, simulations of causal loci harboring multiple rare variants with large additive effects predict an excess of low-frequency significant markers relative to empirical findings[255, 77].

SNP-based heritability estimates have concluded that there is little missing heritability for height and BMI, and that the causal loci simply have effect sizes that are too small to reach genome-wide significance under current GWAS sample sizes[264, 263]. Further, extensions to these methods decompose genetic variance into additive and dominance components and find that dominance variance is approximately one fifth of the additive genetic variance on average across seventy-nine complex traits [275]. When taken into account together with results from GWAS, these observations can be interpreted as evidence that the genetic architecture of human traits is best-explained by a model of small additive effects. However, a recent large twin study found a substantial contribution of dominance variance for fourteen out of eighteen traits [35]. The reason for this discrepancy in results remains unclear. One possibility is a statistical artifact; for example, twin studies may be prone to mistakenly infer non-additive effects when none exist. Another possibility, which we return to later, is that this apparently contradictory results are expected under a different model of gene action.

The design, analysis, and interpretation of GWAS are heavily influenced by the "standard model" of quantitative genetics[60]. This model assigns an effect size to a mutant allele, but formally makes no concrete statement regarding the molecular nature of the allele. Early applications of this model to the problem of human complex traits include Risch's work on the power to detect causal mutations [200, 199] and Pritchard's work showing that rare alleles under purifying selection may contribute to heritable variation in complex traits [187]. When applied to molecular data, such as SNP genotypes in a GWAS, these models treat the SNPs themselves as the loci of interest. For example, influential power studies informing the design of GWAS assign effect sizes directly to SNPs and assume Risch's model of multiplicative epistasis [221]. Similarly, the single-marker logistic regression used as the

primary analysis of GWAS data typically assumes an additive or recessive model at the level of individual SNPs [252]. Finally, recent methods designed to estimate the heritability of a trait explained by genotyped markers assigns additive and dominance effects directly to SNPs [264, 265, 263, 275]. Naturally, the results of such analyses are interpreted in light of assumed model of gene action.

A weakness of the multiplicative epistasis model [200, 199] when applied to SNPs is that the concept of a gene, defined as a physical region where loss-of-function mutations have the same phenotype [15], is lost. Specifically, under the standard model, the genetic concept of a failure to complement is a property of SNPs and not "gene regions" (see [232] for a detailed discussion of this issue). We have recently introduced an alternative model of gene action, one in which risk mutations are unconditionally deleterious and fail to complement at the level of a "gene region" [232]. This model, influenced by the standard operational definition of a gene [15], gives rise to the sort of allelic heterogeneity typically observed for human Mendelian diseases [225], and to a distribution of GWAS "hit" minor allele frequencies [255, 77] consistent with empirical results [232]. In this article, we explore this "gene-based" model under more complex demographic scenarios as well as its properties with respect to the estimation of variance components using SNP-based approaches [265] and twin studies. We also compare this model to the standard models of strictly additive co-dominant effects, and multiplicative epistasis with dominance.

We further explore the power of several association tests to detect a causal gene region under each genetic and demographic model. We find significant heterogeneity in the performance of burden tests [169, 262, 232] across models of the trait and demographic history. We find that population expansion reduces the power to detect causal gene-regions due to an increase in rare variation, in agreement with work by [141, 242]. The behavior of the tests under different models provides us with insight as to the circumstances in which each test is best suited.

In total, our results show that modeling gene action is key to modeling GWAS, and thus plays an important role in both the design and interpretation of such studies. Further, the model of gene-based recessivity best explains the differences between estimates of additive and dominance variance components from SNP-based methods [275] and from twin studies [35] and is consistent with the distribution of frequencies of significant associations in GWAS [255, 77]. Further, the genetic model plays a much more important role than the demographic model, which is expected based on previous work on additive models showing that the genetic load is approximately unaffected by changes in population size over time, [217, 141]. Consistent with recent work by [242], we find that rapid population growth in the recent past increases the contribution of rare variants to total genetic variance. However, we show here that different models of gene action are qualitatively different with respect to the partitioning of genetic variance across the allele frequency spectrum. We also show that these conclusions hold under the more complex demographic models that have been proposed for human populations [229, 217].

2.6 Results and Discussion

2.6.1 The Models

As in [232],we simulate a 100 kilobase region of human genome, contributing to a complex disease phenotype and fitness. The region evolves forward in time subject to neutral and deleterious mutation, recombination, selection, and drift. To perform genetic association and heritability estimation studies *in silico*, we need to impose a trait onto simulated individuals. In doing so, we introduce strong assumptions about the molecular underpinnings of a trait and its evolutionary context.

How does the molecular genetic basis of a trait under natural selection influence population

genetic signatures in the genome? This question is very broad, and therefore it was necessary to restrict ourselves to a small subset of molecular and evolutionary scenarios. We analyzed a set of approaches to modeling a single gene region experiencing recurrent unconditionally-deleterious mutation contributing to a quantitative trait subject to Gaussian stabilizing selection. The expected fitness effect of a mutation is always deleterious because trait effects are sampled from an exponential distribution. Therefore, we do not allow for compensatory mutations that may occur in more general models of stabilizing selection. Specifically, we studied three different genetic models and two different demographic models, holding the fitness model as a constant. Parameters are briefly described in Table 2.1.

We implemented three disease-trait models of the phenotypic form $P = G + E$. G is the genetic component, and $E = N(0, \sigma_e^2)$ is the environmental noise expressed as a Gaussian random variable with mean 0 and standard deviation σ_e^2. In this context, σ_e^2 should be thought of as both the contribution from the environment and from the remaining genetic variance at loci in linkage equilibrium with the simulated 100kb region. The genetic models are named the additive co-dominant (AC) model, multiplicative recessive (Mult. recessive; MR) model and the gene-based recessive (GBR) model. The MR model has a parameter, h, that controls the degree of recessivity; we call this model the complete MR (cMR) when $h = 0$ and the incomplete MR (iMR) when $0 \leq h \leq 1$. Here, $h = 1$ corresponds to co-dominance, which is different from the typical formulation used when modeling the fitness effects of mutations directly. It is also important to note that here recessivity is being defined in terms of phenotypic effects; this may be unusual for those more accustomed to dealing directly with recessivity for fitness effects. An idealized relationship between dominance for fitness effects and trait effects of a mutation on an unaffected genetic background is shown in Fig A.15.

The critical conceptual difference between recessive models is whether dominance is a property of a locus (nucleotide/SNP) in a gene or the gene overall. Mathematically, this amounts

to whether one first determines diploid genotypes at sites (and then multiplies across sites to get a total genetic effect) or calculates a score for each haplotype (the maternal and paternal alleles). For completely co-dominant models, this distinction is irrelevant, however for a model with arbitrary dominance one needs to be more specific. As an example, imagine a compound heterozygote for two biallelic loci, i.e. genotype Ab/aB. In the case of traditional multiplicative recessivity the compound heterozygote is wild type for both loci and therefore wild-type over all; this implies that these loci are in different genes (or independent functional units of the same gene) because the mutations are complementary. However, in the case of gene-based recessivity [232], neither haplotype is wild-type and so the individual is not wild-type; the failure of mutant alleles to complement defines these loci as being in the same gene [15].

For a diploid with m_i causative mutations on the i^{th} haplotype, we may define the additive model as

$$G_{AC} = \sum_{i=1}^{2} \sum_{j=1}^{m_i} c_{i,j}, \tag{2.1}$$

where $c_{i,j}$ is the effect size of the j^{th} mutation on the i^{th} haplotype. Each $c_{i,j}$ is sampled from an exponential distribution with mean of λ, to reflect unconditionally deleterious mutation. In other words, when a new mutation arises its effect c is drawn from an exponential distribution, and remains constant throughout it's entire sojourn in the population.

The GBR model is the geometric mean of the sum of effect sizes on each haplotype [232]. We sum the causal mutation effects on each allele (paternal and maternal) to obtain a haplotype score. We then take the square root of the product of the haplotype scores to determine the

total genetic value of the diploid.

$$G_{GBR} = \sqrt{\sum_{j=1}^{m_1} c_{1,j} \times \sum_{j=1}^{m_2} c_{2,j}} \tag{2.2}$$

Finally, the MR model depends on the number of positions for which a diploid is heterozygous (m_{Aa}) or homozygous (m_{aa}) for causative mutations,

$$G_{MR} = \left(\prod_{j=1}^{m_{Aa}}(1 + hc_j)\right)\left(\prod_{j=1}^{m_{aa}}(1 + 2c_j)\right) - 1. \tag{2.3}$$

Thus, $h = 0$ is a model of multiplicative epistasis with complete recessivity (cMR), and $h = 1$ closely approximates the additive model when effect sizes are small.

Here, phenotypes are subject to Gaussian stabilizing selection with an optimum at zero and standard deviation of $\sigma_s = 1$ such that the fitness, w, of a diploid is proportional to a Gaussian function[28].

$$w = e^{-\frac{P^2}{2\sigma_s^2}} \tag{2.4}$$

The AC and MR models draw no distinction between a "mutation" and a "gene" (as discussed in [232]). The GBR is also a recessive model, but recessivity is at the level of a *haplotype* (or allele) and is not an inherent property of individual mutations (see [232] for motivation of this model). Viewed in light of the traditional AC and MR models, the recessivity of a

site in the GBR model is a function of the local genetic background on which it is found. Based on several qualitative comparisons we find that the GBR model is approximated by iMR models with $0.1 \leq h \leq 0.25$. However, no specific iMR model seems to match well in all aspects. The demographic models are that of a constant sized population (no growth) and rapid population expansion (growth).

The use of the MR model is inspired by Risch's work[200, 199], linking a classic evolutionary model of multiple loci interacting multiplicatively[86, 34] to the the genetic epidemiological parameter relative risk. Risch and Merikangas [201] used this model to calculate the power to detect causal risk variants as a function of their frequency and effect size. Pritchard extended Risch's model to consider a trait explicitly as a product of the evolutionary process[187]. Pritchard's work demonstrated that the equilibrium frequency distribution suggested an important role for rare deleterious mutations when a trait evolves in a constant sized, randomly mating population with recurrent mutation and constant effect sizes. However, multiplicative epistasis is only one model of gene action. Exploring the effect of different genotype-to-phenotype models on the population and quantitative genetic properites of complex traits is the focus of the current work.

2.6.2 Additive and dominance genetic variance in the population

The amount of narrow sense heritability, $h^2 = (V_A)/(V_P)$, explained by variants across the frequency spectrum is directly related to the effect sizes of those variants [60]. Thus, this measure is an important predictor of statistical power of GWAS and should inform decisions about study design and analysis [215]. Empirically, SNP-based estimates of heritability have inferred negligible dominance variance underlying most quantitative traits [275]. We have a particular interest in the amount of additive variance, V_A, that is due to rare alleles and how much of genetic variance, V_G, is attributable to V_A under different recessive models.

We follow the approach of[217], by calculating the cumulative percent of V_G explained by the additive effects of variants less than or equal to frequency x, $(V_{A;q\leq x})/(V_G)$. The product of this ratio and broad-sense heritability is an estimate of the narrow-sense heritability, h^2. This calculation is a population-wide equivalent to a SNP-based estimate of heritability in a population sample. In addition we calculate the same distribution for dominance effects $(V_{D;q\leq x})/(V_G)$ using the orthogonal model of [275]. Methods based on summing effect sizes [60] or the site frequency spectrum [217] would not apply to the GBR model, because the effect of a variant is not independent of other variants (*e.g.*, there is intralocus epistasis). Therefore, we resort to a regression-based approach, where we regress the genotypes of the population onto the total genetic value as defined in our disease trait models (see Material and Methods). In the limit of Hardy-Weinberg and linkage equilibrium, the regression estimates are equivalent to standard quantitative genetic estimates [60] (Fig A.14). For consistency, we applied the regression approach to all models. Overall, these distributions are substantially different across genetic models, demographic scenarios and model parameters (Fig 2.1).

Under the AC model, all of V_G is explained by additive effects if all variants are included in the calculation; in Fig 2.1 the solid variance curves reach unity in the AC panel. Low frequency and rare variants ($q < 0.01$) explain a large portion of narrow sense heritability (26% - 95%) even in models without rapid population expansion. Further, the variance explained at any given frequency threshold increases asymptotically to unity as a function of increasing λ (Fig A.4). While the total heritability of a trait in the population is generally insensitive to population size changes (Fig A.1, see also [141, 217, 241]), rapid population growth increases the fraction of additive genetic variation due to rare alleles (Fig 2.1).

Here, increasing λ corresponds to stronger selection against causative mutations, due to their increased average effect size. Recent work by Zuk et al. [278], takes a similar approach and relates the allele frequency distribution directly to design of studies for detecting the role of rare variants. However, our findings contrast with those of Zuk[278] and agree with those of

Lohmueller [141], in that we predict that population expansion will substantially increase the heritability, or portion of genetic variance, that is due to rare variants. Our results under the AC model agree with those of Simons et al. [217], in that we find that increasing strength of selection, increasing λ in our work, increases the contribution to heritability of rare variants. However, under the GBR model and the cMR model the distribution of genetic variance over risk allele frequency as function λ is non-monotonic (Fig 2.1 and Fig A.4).

For all recessive models, we find that total V_A is less than V_G (Fig 2.1). For the MR models, all additional genetic variation is explained by the dominance variance component; in Fig 2.1 the dotted variance curves reach unity in the MR panels. As expected, genetic variation under the MR model with partial recessivity ($h = 0.25$) is primarily additive [60, 93], whereas V_G under the cMR model ($h = 0$) is primarily due to dominance. The GBR model shows little dominance variance and is the only model considered here for which the total V_G explained by $V_A + V_D$ is less than the true V_G for all λ. This can be clearly seen in Fig 2.1 where the dotted curves do not reach unity in the GBR panel. These observations concerning the GBR model are consistent with the finding of [275] that dominance effects of SNPs do not contribute significantly to the heritability for complex traits.

Under the GBR model, large trait values are usually due to compound heterozygote genotypes (*e.g.*, Ab/aB, where A and B represent different sites in the same gene) [232]. Therefore, the recessivity is at the level *of the gene region* while the typical approach to estimating V_A and V_D assigns effect sizes and dominance to individual mutations. Thus, compound heterozygosity, which is commonly observed for Mendelian diseases (see [232] and references therein) would be interpreted as variation due to *interactions* (epistasis) between risk variants. Importantly, the GBR model assumes that such interactions should be local, occurring amongst causal mutations in the same locus. While the GBR model is reflective of the original definition of a gene in which recessive mutations fail to complement, we emphasize that this does not imply that mutations are necessarily exomic. The GBR model is of a general

genomic region in which mutations act locally in *cis* to disrupt the function of that region with respect to a phenotype.

The increase in the number of rare alleles due to population growth is a well established theoretical and empirical result [82, 81, 206, 72, 14, 5, 42, 73, 74, 75, 80, 109, 111, 196]. The exact relationship between rare alleles[187, 77, 255, 48, 148], and the demographic and/or selective scenarios from which they arose[198, 217, 141], and the genetic architecture of common complex diseases in humans is an active area of research. An important parameter dictating the relationships between demography, natural selection, and complex disease risk is the degree of correlation between a variant's effect on disease trait and its effect on fitness [59, 141, 217, 1]. In our simulations, we do not impose an explicit degree of correlation between the phenotypic and fitness effects of a variant. Rather, this correlation is context dependent, varying according to the current genetic burden of the population, the genetic background in which the variant is present and random environmental noise. However, if we re-parameterized our model in terms of [59], then we would have $\tau \leq 0.5$ (Gaussian function is greater than or equal to its quadratic approximation), which is consistent with recent attempts at estimating that parameter [1, 150]. Our approach is reflective of weak selection acting directly on the complex disease phenotype, but the degree to which selection acts on genotype is an outcome of the model. While the recent demographic history has little effect on key mean values such as broad-sense heritability of a trait or population genetic burden (Fig A.1 and Fig A.3), the structure of the individual components in the population which add up to those mean values varies considerably. The specific predictions with respect to the composition of the populations varies drastically across different modeling approaches. It is therefore necessary to carefully consider the structure of a genetic model in a simulation study.

The conclusions reached here also hold when we consider more complex demographic scenarios relevant to human populations. Under the demographic model for European populations

from [229], the additive and GBR models show the same behavior as in Fig 2.1 (Fig A.17). At all key time points where population size changes, $V_A = V_G$ for the additive model, and the variance explained by rare mutations depends primarily on λ (Fig A.17). For the GBR model, $V_A < V_G$ (as in Fig 2.1), and plateaus at the same ratio V_A/V_G for all time points except immediately after the bottleneck, which results in a short-lived increase in V_A/V_G that is undetectable by the time growth begins (Fig A.17). All recessive models (GBR, iMR and cMR) may show a transient increase in total V_G after the bottleneck, depending on the value of λ (A.18). However, the GBR and iMR models with $h > 0.25$ showed a return to constant population size levels by the final time point. The changes in V_A and V_G under recessive models is likely due to the transfer of non-additive variation into V_A during a bottleneck, which has been studied thoroughly in the theoretical literature[167, 9]. As in Fig 2.1, the genetic model, and not the demographic details, drive the relationship between mutation frequency and additive genetic variance. In agreement with existing literature, site based recessive models show complex dynamics during bottlenecks and population expansion (Fig A.18 and Fig A.19). However, with respect to load, the GBR model behaves more like a codominant model and is largely insensitive to changes in population size(Fig A.18 and Fig A.19). Thus, complex traits evolving under the GBR model are not expected to show large differences in load between extant human populations.

2.6.3 Estimating additive and dominance variance from population samples

The previous section shows that the relationship between genetic variance and allele frequency in the entire population strongly depends on the genetic model. Recent estimates of variance components from large population samples of unrelated individuals have inferred that dominance variance (V_D) is negligible for most traits [275]. However, a recent study of more than 10^4 Swedish twins and 18 traits obtained a contradictory result, inferring signifi-

cant non-additive variance for most traits, which was interpreted as V_D [37]. In this section, we show that this apparent inconsistency is expected under certain models of gene action.

We applied GREMLd, MAF-stratified GREMLd (MS-GREMLd), and MAF-stratified Haseman-Elston regression (see Methods for details). We found MS-GREMLd to be numerically unstable on our simulated data, and thus we present results for non-MS-stratified GREMLd. The numerical stability issues likely resulted from some combination of small number of SNPs per region ($\mathcal{O}(1000)$), low total V_G in a region, or high variance in effect sizes across causal mutations [131]. Further, for large λ, where V_G is primarily due to rare alleles (Fig 2.1), heritability in a sample may not reflect heritability in the entire population (Fig A.13).

Fig 2.2 shows the GREMLd additive and dominance heritability estimates, as compared to the respective population value, over λ. Under the cMR model ($h = 0$), the dominance component is much larger than the additive component as predicted from Fig 2.1. When GREMLd is performed on cMR model data after removing variants with $MAF \leq 0.01$, as done in [275], the total heritability estimate (AD) is quite accurate until $\lambda \geq 0.25$ where a downward bias is observed. As anticipated, GREMLd using unfiltered data yields results with a slight upward bias [132]. However, for the iMR ($h = 0.25$) model the filtered GREMLd estimates are only accurate for $\lambda < 0.1$ reflecting the preponderance of rare causal variants for larger values of λ. Unfiltered GREMLd estimates under the iMR ($h = 0.25$) model show a slight upward bias for small values of λ, but are otherwise accurate. This shows that GREMLd is performing as expected under the site-based model for which it is designed. The MS-HE regression results are generally consistent with the GREMLd results.

The GREMLd and MS-HE estimates are accurate under the GBR model when λ is small, because most heritability is additive in that case(Fig 2.1). However, under the GBR model, both filtered and unfiltered GREMLd heritability estimates show downward bias when λ is large (Fig 2.2). The MS-HE regression results reveal a similar pattern, which indicates that the downward bias for large values of λ is not strictly due to removal of rare variants in the

filtered GREMLd analysis. Instead, the bias shown for large values of λ is likely due to the presence of substantial non-additive heritability, which is not captured by the dominance effects of SNPs.

In contrast to the variance component methods, our simulated large twin studies provide approximately unbiased estimates of total heritability for large values of λ, but were biased upward for small effect sizes under the AC and GBR models (Fig 2.2). The variance in twin-study estimates was quite large, possibly because only a sinlge locus was simulated rather than the whole genome. Formally, twin studies estimate an additive and a non-additive component of variance and interpreting the non-additive component as epistatic or dominance variance is a matter of perspective. However, the GBR model is inspired by the definition of a gene as a physical region in which recessive mutations leading to the same phenotypic outcome fail to complement [15], consistent with the allelic heterogeneity observed for human Mendelian disorders (see [232] for further discussion). Thus, the model of recessivity at the level of the gene region is picked up as non-additive variance in twin studies, but missed by variance component methods (GREML and HE regression) because the dominance in the GBR model is due to Ab/aB (compound heterozygotes) genotypes rather than a/a genotypes (heterozygotes for a specific loss of function variant) assumed by variance component methods. Thus the contradictory results of applying variance component methods [275] and analysis of large twin studies [37] in order to estimate V_A and V_D may be interpreted as evidence for a model of gene action such as the GBR, which may be viewed as either recessivity at the haplotype/gene level or intralocus epistasis at the level of causative mutations in a single gene region. Both interpretations are valid. The alternative explanation is that we must assert that one of the study designs is generating artifacts.

2.6.4 The genetic model affects the outcomes of GWAS

Both demography and the model of gene action affect the degree to which rare variants contribute to the genetic architecture of a trait (Fig 2.1). However, the different mappings of genotype to phenotype from model to model make it difficult to predict *a priori* the outcomes of GWAS under each model. Therefore, we sought to explicitly examine the performance of statistical methods for GWAS under each genetic and demographic model. We assessed the power of a single marker logistic regression to detect the gene region by calculating the proportion of model replicates in which at least one variant reached genome wide significance at $\alpha \leq 10^{-8}$ (Fig 2.3A). The basic logistic regression is equivalent to testing for association under the AC model. We simulated both a perfect "genotyping chip" (all markers with $MAF \geq 0.05$) and complete re-sequencing including all markers (Fig 2.3B).

One of the most promiment feature of Fig 2.3 is the curvature of power as a function of λ. This reflects the competing forces of increasing average genetic effect and decreasing average allele frequency which occurs as λ increases (Fig A.5). As λ increases, the total genetic variance explained by the locus increases until the model enters the House-of-cards [237] regime. At which point, the genetic variance is much less dependent on λ (Fig A.1). When λ is large, however, the average allele frequency does continue to decrease (Fig A.5) which drives power down.

Across all genetic models, the single marker logistic regression has less power under population expansion (Fig 2.3A). The loss of power is attributable to a combination of rapid growth resulting in an excess of rare variants overall [82, 81, 206, 72, 14, 5, 42, 73, 74, 75, 80, 109, 111, 196], and the increasing efficacy of selection against causal variants in growing populations [217]. While complete resequencing is more powerful than a gene-chip design, the relative power gained is modest under growth (Fig 2.3A). Region-based rare variant association tests behave similarly with respect to population growth (Fig 2.3B).

There are important differences in the behavior of the examined statistical methods across genetic models. We focus first on the single marker tests (Fig 2.3A). For gene-chip strategies, power increases for "site-based" models as recessivity of risk variants increases (compare power for AC, iMR, and cMR models in Fig 2.3B). This increase in power is due to the well-known fact that recessive risk mutations are shielded from selection when rare (due to being mostly present as heterozyogtes), thus reaching higher frequencies on average (Fig A.5), and that the single-marker test is most powerful when risk variants are common [221]. Further, for the complete multiplicative-recessive model (cMR), the majority of V_G is due to common variants (Fig 2.1), explaining why resequencing does not increase power for this model (Fig 2.3A).

For single-marker tests, the GBR model predicts large gains in power under re-sequencing for intermediate λ (the mean trait-effect size of newly arising causal mutations), similar to the AC or iMR model. But, when λ is larger power may actually be less under the GBR model than under AC or iMR. For all models, causal mutations are more rare with increasing λ (Fig A.7). However, as a function of frequency, all V_G may be attributed to V_A or V_D in the site-specific models whereas there is increasing intralocus epistasis in the GBR model as a function of λ (Fig 2.1). It is well-known that the single marker test has lower power when causal mutations have low frequencies, are poorly tagged by more common SNPs, or have small main effects [221, 33].

Region-based rare variant association tests show many of the same patterns across genetic model and effect size distribution as single marker tests, but there are some interesting differences. The ESM test[232, 208] is the most powerful method tested for the AC, iMR, and GBR models (Fig 2.3b), with the c-Alpha test as a close second in some cases. For those models, the power of naive SKAT, linear kernel SKAT and SKAT-O, is always lower than the ESM and c-Alpha tests. This is peculiar since the c-Alpha test statistic is the same as the linear kernel SKAT test. The major difference between SKAT and ESM/c-

Alpha is in the evaluation of statistical significance. SKAT uses an analytical approach to determine p-values while the ESM/c-Alpha tests use an explicit permutation approach. This implies that using permutation based p-values results in greater power. Yet, under the cMR model the linear kernel SKAT is the most powerful, followed by c-alpha. The cMR model does not predict a significant burden of rare alleles and so the default beta weights of SKAT are not appropriate, and the linear kernel is superior. The ESM test does poorly on this model because there are not many marginally significant low-frequency markers. It is logical to think that these tests would all perform better if all variants were included. The massive heterogeneity in the performance of region-based rare variant tests across models strongly suggests that multiple methods should be used when prior knowledge of underlying parameters is not available. In agreement with [141, 240], we predict that population growth reduces the power to associate variants in a causal gene region with disease status (Fig 2.3) when the disease also impacts evolutionary fitness. We have recently released software to apply the ESM test to case control data [208] in order to facilitate applying this test to real data.

2.6.5 The distribution of minor allele frequencies of GWAS hits

It was noted by[255, 77], that an excess of rare significant hits, relative to empirical data, is predicted by AC models where large effect mutations contribute directly to fitness and the disease trait. We confirm that AC models are inconsistent with the empirical data (Fig 2.4), except when $\lambda \leq 0.01$. The empirical data in Fig 2.4 represent a pooled data set with the same diseases and quality filters as in [255], but updated to include more recent data. The data are described in A.1, and can be visualized alone more clearly in A.16. Close to half of the data comes from GWAS studies uploaded to the NHGRI database after 2011, yet the same qualitative pattern is observed. This contradicts the hypothesis that the initial observation of an excess of common significant hits relative to the prediction under

an AC model was simply due to small sample sizes and low marker density in early GWAS previously analyzed in [77, 255]. Yet the initial observation is in fact robust and the meta-pattern provides an appropriate point of comparison when considering the compatibility of explicit population-genetic models with existing GWAS data.

The GBR model predicts few rare significant hits and an approximately uniform distribution across the remainder of MAF domain (Fig 2.4), even for intermediate and large values of λ. For smaller values of λ, the GBR predicts an excess of common significant hits. The more uniform distribution of significant single markers seen under the GBR is consistent with the flatter distribution of genetic variance (Fig 2.1). If one considers trying to determine an approximate dominance coefficient in the GBR model, it would be found that there is a distribution of coefficients across sites. Yet, when simulating iMR model, we find that an intermediate degree of dominance, $h = 0.25$, results in distribution of significant hits which is similar to the GBR results (Fig 2.4).

Most of the models fail a KS test comparing the simulated and empirical distribution of significant hits(A.21).The cMR ($h = 0$) model shows a visual excess of intermediate frequency variants(Fig 2.4), but this does not result in rejection under the KS test (A.21) which is largely insensitive to deviations in the tails. According to the KS test, the remaining models(AC, GBR, iMR) perform best when there are fewer data points in the simulated data due to low GWAS power. This suggests that all models would be rejected with enough replicates. We note that there is no compelling reason to expect any specific value of λ to be a particularly good fit to the empirical data. The empirical data are composed of genome-wide data for multiple traits. We feel that the mutational parameters, λ and mutation rate to causal variants, are likely to vary across the genome and across traits. Thus, the empirical data reflect a mixture of different underlying models and ascertainment schemes. The reason we emphasize this feature of the data is to demonstrate that models with rare alleles of large effect do not necessarily imply a visual excess of rare significant GWAS hits.

In consideration of the rare allele of large effect hypothesis, [48] proposed a model where multiple rare alleles dominate disease risk and create synthetic associations with common SNPs. However, later it was shown that this particular model was inconsistent with GWAS theoretically and empirically [176, 255, 77]. Here, we have shown that there exist models in which rare alleles explain a substantial portion of heritability that are not inconsistent with findings from GWAS. We find that the MAF distribution of significant hits in a GWAS varies widely with choice of genetic model. In particular, we confirm the results of Wray et al. [255], that AC evolutionary models predict an excess of low frequency significant hits unless trait effect sizes are quite small. Also, the cMR model predicts an excess of intermediate and common significant hits. Utilizing a GBR model or an iMR model with $h = 0.25 - 0.5$, reconciles this inconsistency by simultaneously predicting the importance of rare alleles of large effect and the correct allele frequency distribution among statistically significant single markers.

2.6.6 Conclusion

Several empirical observations provide support for the presence of gene-based recessivity underlying variation for some complex traits in humans. The minor allele frequency distribution of significant GWAS hits is relatively flat[255, 77], which our results show is consistent with either the presence of small additive effect loci or gene-/site-based partially-recessive loci with intermediate to large effects (Fig 2.4). Models with loci of large additive effects predict an excess of rare significant hits. Oppositely, models with complete site-based recessivity predict an excess of common significant hits for all simulated mutation effect size distributions.

SNP based estimates of dominance heritability are much lower than estimates of dominance from twins [275, 37]. Of the models we explored, only the gene-based recessive model

with intermediate to large effects is consistent with difference between twin and SNP based estimates of dominance variance (Fig 2.2). Under a site-based recessive model of partial recessivity (*e.g.* h =0.25), there should be no significant difference between estimates of dominance variance from SNP and twin studies, provided that the statistical assumptions are met for both approaches (Fig 2.2). These results are complementary to the work by Zuk et. al [278], who show that twin studies can over estimate heritability under a model with gene interactions. It now appears clear that the underlying genetic model does not have the same impact on SNP-based and family based study designs; an issue which should be further explored. Our findings also support a more thorough investigation into the importance of compound heterozygosity in the genetics of complex traits. However, it may be difficult to directly observe non-additive gene-level effects through analysis of individual SNP markers.

Additionally, the genetic model appears to be important in the design and analysis of association studies. While changes in population size do affect the relationship between effect size and mutation frequency [82, 81, 206, 72, 14, 5, 42, 73, 74, 75, 80, 109, 111, 196] (Fig 2.1 and A.5), different mappings of genotype to trait value do this in radically different ways for the same demographic history (Fig 2.1). From an empirical perspective, our findings suggest that re-sequencing in large samples is likely the best way forward in the face of the allelic heterogeneity imposed by the presence of rare alleles of large effect. Re-sequencing of candidate genes [101, 123, 153, 207] and exomes [98, 254, 189, 229, 171, 143, 40] in case-control panels have observed an abundance of rare variants associated with case status. Here we show that under a model of mutation-selection balance on the genic level, neither current single-marker nor popular multi-marker tests are especially powerful at detecting large genomic regions harboring multiple risk variants (Fig 2.3). However, we show that using permutations to derive p-values improves the power of SKAT[131] with a linear kernel (equivalent test statistic to c-Alpha [169]). Similarly, another permutation based test, the ESM test [208], has more robust power across demographic and genetic models (Fig 2.3).

Conceptually, *cis*-effects arise naturally from the original definition of a gene in which mutant recessive alleles fail to complement [15]. We show that *cis*-effects within a locus, represented by the GBR model, can have an important impact on the population level architecture of a complex trait. This conclusion is important for future simulation studies as well as the interpretation of empirical data. It is important to note that despite our use of the term "gene-based" this model may apply to any functional genomic element in which there are multiple mutable sites affecting a trait in *cis*, not just to genes. From a theoretical perspective, our work motivates the development of a more generalized gene-based model to include arbitrary dominance and arbitrary locus size. Empirically, we find that the GBR model is broadly consistent with a variety of observations from the human statistical genetics literature. Thus, there is an evident need for improved region-based association tests and the development of genetic variance component methods for haplotypes.

2.7 Materials and Methods

2.7.1 Forward simulation

Using the fwdpp template library v0.2.8 [231], we implemented a forward in time individual-based simulation of a Wright-Fisher population with mutation under the infinitely many sites model[113], recombination, and selection occurring each generation. We simulated populations of size $N = 2e4$ individuals for a time of $8N$ generations with a neutral mutation rate of $\mu = 0.00125$ per gamete per generation and a per diploid per generation recombination rate of $r = 0.00125$. Deleterious mutations occurred at a rate of $\mu_d = 0.1\mu$ per gamete per generation. These parameters correspond to $\theta = 4N\mu = \rho = 4Nr = 100$ and thus our simulation approximates a 100Kb region of the human genome. For simulations with growth, we simulated an additional 500 generations of exponential growth from $N_i = 2e4$

to $N_{final} = 1e6$. This demographic model is much simpler than current models fit to empirical data[80]. However, this simple model allows us to more easily get a sense of the impact of population expansion[141, 217]. 250 simulation trials were performed for each parameter/model combination unless specified otherwise.

2.7.2 Exploring the gene region's contribution to heritability

Broad-sense heritability can be calculated directly from our simulated data as $H^2 = \frac{V_G}{V_P}$. We explored broad-sense heritability as a function of mean causative effect size λ under each model; $\lambda \in \{0.01, 0.025, 0.05, 0.1, 0.125, 0.25, 0.5\}$. We compare our simulation results to $V_G \sim 4\mu_d\sigma_s^2$ for additive models and $V_G \sim 2\mu_d\sigma_s^2$ for recessive models [237, 219]. In our simulations, $\sigma_s^2 = 1$, and we tuned the environmental standard deviation σ_e to generate simulations for which $E[H^2] \sim 0.04$ or ~ 0.08. For $E[H^2] \sim 0.04$, we set $\sigma_e = 0.11$ for the additive codominant model, $\sigma_e = 0.075$ for the gene based and complete multiplicative recessive models and $\sigma_e = 0.098$ for the incomplete mutliplicative recessive model ($h = 0.25$). For $E[H^2] \sim 0.08$, we set $\sigma_e = 0.075$ for the additive codominant model, $\sigma_e = 0.053$ for the gene based and complete multiplicative recessive models and $\sigma_e = 0.068$ for the incomplete mutliplicative recessive model ($h = 0.25$).

2.7.3 Determining the genetic load of the population

Genetic load is defined as the relative deviation in a populations fitness from the fitness optimum, $L = (w_{max} - \bar{w})/(w_{max})$. We set the phenotypic optimum to be zero; $P_{opt} = 0$. When determining fitness for the SBR models, we subtract one from all phenotypes. This implies that $w_{max} = e^{-\frac{P_{opt}^2}{2\sigma_s^2}} = 1$ and that load is a simple function of the phenotypes of the population, $L = 1 - e^{-\frac{P^2}{2\sigma_s^2}}$. We also used the mean number of mutations per individual, and the mean frequency and effect sizes of segregating risk variants as proxies for the genetic

load [217, 79]. Lastly, we calculated Burden Ratios (B_r) [8] as the ratio of load between an equilibrium and non-equilibrium population. We calculated B_r using both the true load and the number of mutations per individual.

2.7.4 Additive and dominance genetic variance over allele frequency

We used an approach based sequential (type-1) regression sums of squares to estimate the contribution of the additive and dominance effects of variants to the total genetic variation due to a locus. Given a genotype matrix (rows are individuals and columns are risk variants) of (0,1, or 2) copies of a risk allele (*e.g.* all mutations affecting phenotype), we sort the columns by decreasing risk mutation frequency. Then, within frequency classes, columns were sorted by decreasing effect sizes. For each variant a dominance component was also coded as 0, 2q, or 4q-2 according to the orthogonal model of [275], where q is the frequency of the variant in the population. We then used the `R` package `biglm` [144] to regress the individual genetic values (G in the previous section) onto this matrix. The variance explained by the additive and dominance effects of the m markers with $q \leq x$ is then approximately $r^2 = (\sum_{i=1}^{m} \Sigma SS_{reg,i})/(SS_{tot})$. Averaging results across replicates, this procedure results in a Monte-Carlo estimate of the fraction of V_G that is due to additive and dominance effects of variants with population frequency less than or equal to x is $(V_{A;q\leq x} + V_{D;q\leq x})/(V_{G;q\leq 1})$ [217]. This fraction can be easily partitioned into strictly additive and dominance components.

2.7.5 Additive and dominance heritability in random population samples

We employed three different SNP-based approaches to estimating heritability from population samples: GREMLd, minor allele frequency stratified GREMLd (MS-GREMLd)[275], and MS-Haseman-Elston (HE) regression [88, 214]. For comparison, we calculated the true total heritability in the sample as $H^2_{sample} = (V_{G;sample})/(V_{P;sample})$. Unfortunately, due to the nature of our simulated data MS-GREMLd did not result in sufficiently reliable results. Under MS-GREMLd, many replicates resulted in numerical errors in GCTA. These problems were present at a rate of less than 1/100 replicates using non-MS GREMLd, but were increased by splitting the data into multiple GRMs.

Using raw individual phenotypic values as quantitative trait values, random samples from simulated populations (n=6000) were converted to .bed format using PLINK 1.90a[188]. PLINK was also used to test for HWE ($p < 1e-6$) and filter on minor allele frequency. GCTA 1.24.4 [265] was used to make genetic relatedness matrices (GRM) for both additive and dominance components with the flags –autosome and –make-grm(-d).

For non-MS runs, we tested the effect of filtering on MAF by performing the analysis on unfiltered datasets and with markers with $MAF < 0.01$ removed. For MS estimates we stratified the additive and dominance GRM's into two bins $MAF \leq 0.01$ and $MAF > 0.01$. GREMLd analysis was performed in GCTA with Fisher scoring, no variance component constraint and a max of 200 iterations. MS-HE regression was carried out by regressing the off diagonal elements of each GRM onto the cross product of the scaled and centered phenotypes in a multiple linear regression setting in R [193].

2.7.6 Twin studies

To simulate twin studies we sampled 2000 monozygotic (MZ) and 2000 dizygotic (DZ) twins pairs from the final generation of the simulations. Parents were sampled randomly without replacement. MZ twin pairs were formed by sampling a single gamete pair, one recombinant from each parent, and two environmental random deviates. DZ twin pairs were formed by sampling two gamete pairs, two recombinant gametes from each parent, and two environmental random deviates. Our simulated studies are ideal in that there are no correlated environmental effects, but potentially problematic due to low total heritability. We explored the use of structural equation modeling (SEM) using the package OpenMx [170], but chose to rely strictly on estimates of twin correlation obtained directly from the data. For monozygotic (MZ) twins, we used only a single child gamete pair with two unique environmental deviates. For dizygotic (DZ) twins we used two child gamete pairs, each with a unique environmental deviate. Broad sense heritability is the correlation between MZ twin pairs; $H^2 = r_{MZ}$. Under a purely additive model, the DZ twin correlation should be half of the MZ twin correlation. Non-additive genetic components of phenotypic variance reduce the DZ twin correlation. If all non-additive heritability is due to dominance, then the dominance heritability can be calculated as twice the difference between the MZ twin correlation and two-times the DZ twin correlation: $\delta^2 = 2 * (r_{MZ} - 2 * r_{DZ})$. The additive heritability can then be calculated as the difference between the broad-sense and non-additive component: $h^2 = H^2 - \delta^2 = 4 * r_{DZ} - r_{MZ}$ [60].

These direct estimates of MZ and DZ twin correlations in our simulations are reliable as we have no measurement error, shared environmental effects, gene-by-environment effects, or gene-by-gene interactions. Additionally, we only simulate a single genomic region contributing $H^2 \sim 0.04$, which made use of SEM difficult numerically. This creates a limitation in that we can not discuss when a model with dominance is a better fit to the data than the additive only model. But, the benefit of using direct estimates is that we can clearly see

what signals are present in the data. To further clarify the data visualization, we pooled our 512 twin-study replicates into groups of 8, creating 64 sets of MZ-DZ twin phenotypes. This did not have an effect on the central tendencies of our estimates, but it reduced the variance. The twin study error bars in Fig 2.2 are based on 64 sets of 64,000 individuals, which is larger than a typical twin study. However, one reason our results have high variance is because we only simulate a single locus, rather than a whole trait.

2.7.7 Case-control studies

Following [232], we sampled 3000 cases and 3000 controls from each simulated population. Cases were randomly sampled from the upper 15% of phenotypic values in the population, and controls were randomly sampled from within 0.5 standard deviations of the population mean(as in [232]). This is the liability scale model (see [60]). We define a "GWAS" to be a study including all markers with MAF $\geq$ 5% and a re-sequencing study to include all markers. In all cases we used a minor allele count logistic regression as the single marker test. For single marker tests, the p-value cut off for significance is $p \leq 1e-08$ which is common in current GWAS [48, 157]. Power is determined by the percentage of simulation replicates in which at least one marker reaches genome wide significance.

2.7.8 Region-based tests of association due to rare alleles

We applied multiple region-based tests to our simulated data, ESM_K [232], several variations of SKAT [262] and c-Alpha[169]. We used the `R` package from the SKAT authors to implement their test (`http://cran.r-project.org/web/packages/SKAT/index.html`). The remaining tests were implemented in a custom `R` package (see SOFTWARE AVAILABILITY below). For the ESM_K and c-Alpha we performed up to $2e6$ permutations of case-control labels to determine empirical p-values. Common variants ($q \geq 0.05$) were removed prior to

performing region-based rare variant association tests.

2.7.9 Distribution of Significant GWAS Hits

Following[255, 77], we calculated the distribution of the minor allele frequency (MAF) of the most significant SNPs in a GWAS in empirical and simulated data. The empirical data was obtained from the NHGRI-EBI GWAS database (`http://www.ebi.ac.uk/gwas/`) on 02/05/2015. We considered the same diseases and applied the same filters as in Table 3 of [255]. Specific information regarding the empirical data can be obtained in Table A.1.

In order to mimic ascertained SNP data, we sampled markers from our case/control panels according to their minor allele frequencies [154], as done in [232]. Additionally, we removed all markers with MAF < 0.01 to reflect common quality controls used in GWAS. The simulated data were grouped by genetic model, demographic scenario, heritability level, and mutation effect distribution. We then plotted the minor allele frequency of the most significant marker with a single-marker score $-log_{10}(p) \geq 8$, for all replicates where significant markers were present.Finally, we performed a two-sample KS test in `R` between each group of simulated GWAS hit allele frequencies and the empirical data.

2.7.10 Human demography

We simulated a demographic model for Europeans based on [229] as described in [217]. For simplicity, we ignored migration between the European (EA) and African American (AA) populations. The model was implemented using the Python package `fwdpy` version 0.0.4, which uses `fwdpp` [231] version 0.5.1 as a C++ back-end. During the evolution of the EA population, we recorded the genetic variance in the population, V_G, and the number of deleterious mutations per diploid (a measure of genetic load [217]) every 50 generations. In a

separate set of simulations, we applied the regression method described above to calculate cumulative additive genetic variance as a function of allele frequency. Because the regressions are computationally demanding, we applied the method in the generation immediately before, and at the start of, any changes in population size.

These simulations were run with no neutral mutations, and the recombination rate and mutation rate to causative mutations were the same as in the simulations described above.

The Python scripts for these simulations and iPython/Jupyter notebooks used for generating figures are available online (see Software availability section below).

2.7.11 Software availability

Our simulation code and code for downstream analyses are freely available at

- `http://github.com/ThorntonLab/disease_sims`
- `http://github.com/molpopgen/buRden`
- `http://github.com/molpopgen/fwdpy`
- `http://github.com/molpopgen/TennessenEAonly`

2.8 Acknowledgements

We are thankful to Joseph Farran, Harry Mangalam, Adam Brenner, Garr Updegraff, and Edward Xia for administering the University of California, Irvine High Performance Computing cluster. We are thankful to Kirk Lohmueller for helpful detailed comments throughout

this project. We would like to thank Peter Andolfatto, Bogdan Pasaniuc and Nick Mancuso for helpful discussion.

2.9 Chapter 2 Tables

Parameter	**Description**
N	Population size
P	Phenotype
P_{opt}	Optimum phenotype
G	Genetic contribution to phenotype
E	Environmental contribution to phenotype
λ	Mean and standard deviation of trait effects
c_i	Specific trait effect of site i
h	Dominance coefficient for trait effects
w	Fitness, based on gaussian function
σ_s^2	Environmental variance
σ_e^2	The total inverse selection intensity
V_A	Additive genetic variance
V_D	Dominance genetic variance
V_G	Genetic variance
$V_{A;q\leq x}$	Additive variance explained by variants below frequency q

Table 2.1: Description of parameters used in the models

2.10 Chapter 2 Figures

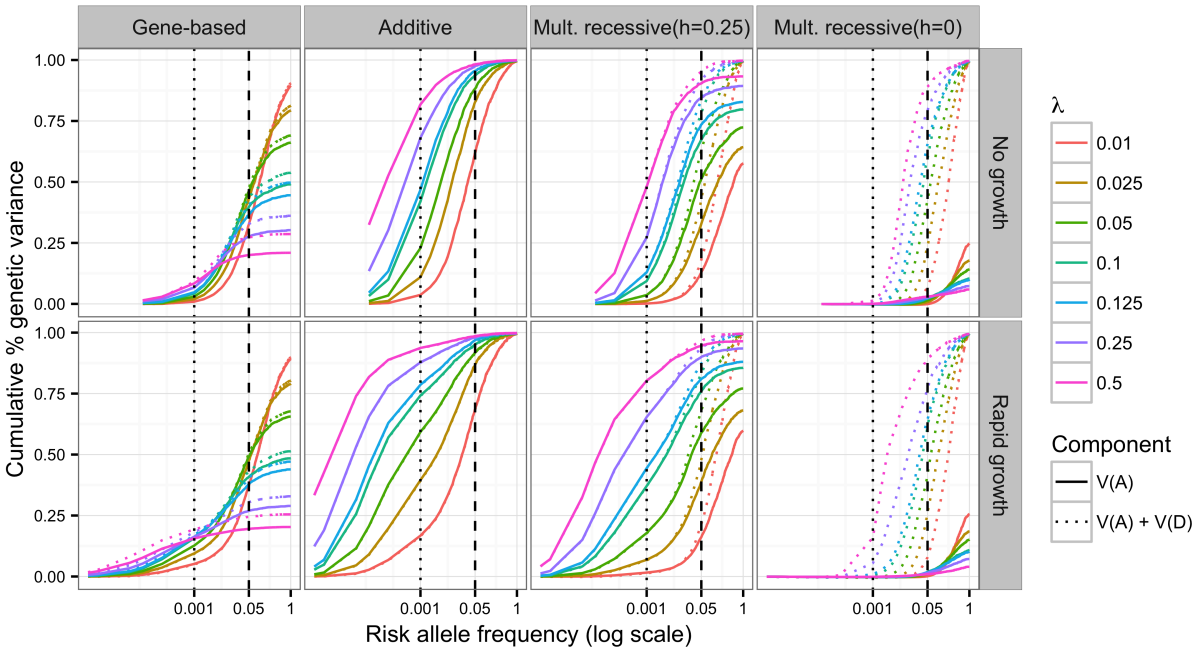

Figure 2.1: Variance explained over allele frequency. The cumulative additive and dominance genetic variance which can be explained by markers whose frequencies, q, are $\leq x$. Each color represents a different value of λ: the mean effects size of a new deleterious mutation. Shown here are the gene-based (GBR), additive co-dominant (AC), incomplete multiplicative recessive (Mult. recessive ($h = 0.25$); iMR) and complete multiplicative recessive (Mult. recessive ($h = 0$);cMR) models. Solid lines show the additive variance alone and dotted lines show the combined additive and dominance variance. All data shown are for models where $H^2 \sim 0.08$. These particular results are robust to changes H^2 when V_G is not changed, as is the case here. The additive and dominance genetic variance is estimated by the adjusted r^2 of the regression of all markers (and their corresponding dominance encoding) with $q \leq x$ onto total genotypic value (see methods for details); data are displayed as the mean of 250 simulation replicates. The vertical dotted and dashed lines correspond to the $q = 0.001$ and $q = 0.05$, respectively. The curves under no growth appear to be truncated with respect to rapid growth because the range of the x-axis differs between growth and no growth (minimum $q = 1/2N$).

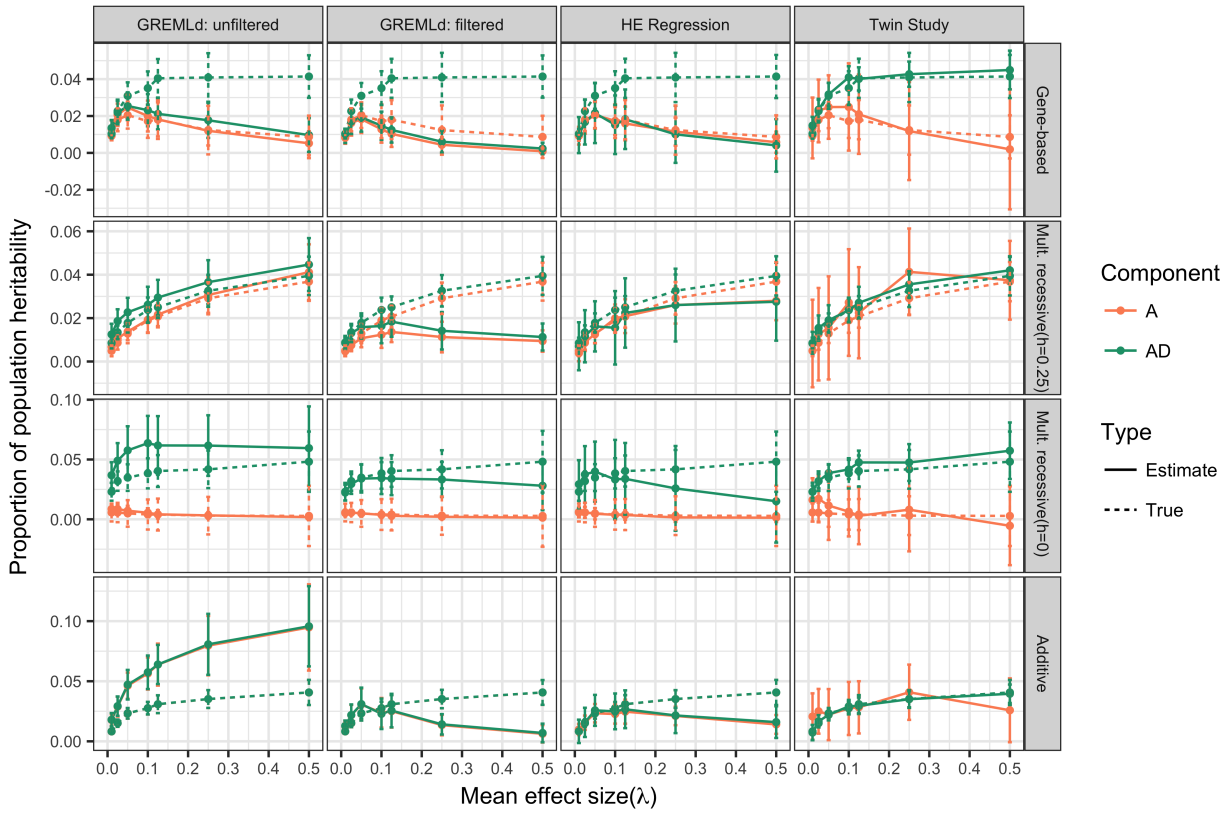

Figure 2.2: Heritability estimates compared to population heritability. Heritability estimates and population heritability as a function of λ: the mean effect size of a new deleterious mutation. Additive (A; orange) component of true heritability is calculated by multiplying the end point($q = 1$) of the variance curves in Fig 2.1 by the broad-sense heritability values summarized in Fig A.1. HE-regression and GREMLd estimates were obtained from random population samples (n = 6000). GREMLd analysis was performed in GCTA using genotype data that was either unfiltered or filtered to remove variants with MAF<0.01. Twin study estimates are directly calculated using MZ and DZ twin correlations from 64 sets of twin studies. Each study consisted of pooling 2000 MZ twin pairs and 2000 DZ twin pairs from each of 8 model replicates for a total of 64,000 individual phenotypes. Data are plotted as the median across replicate sets ± half the interquartile range. Shown are the additive co-dominant (AC), gene-based (GBR) incomplete multiplicative recessive (Mult. recessive ($h = 0.25$); iMR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

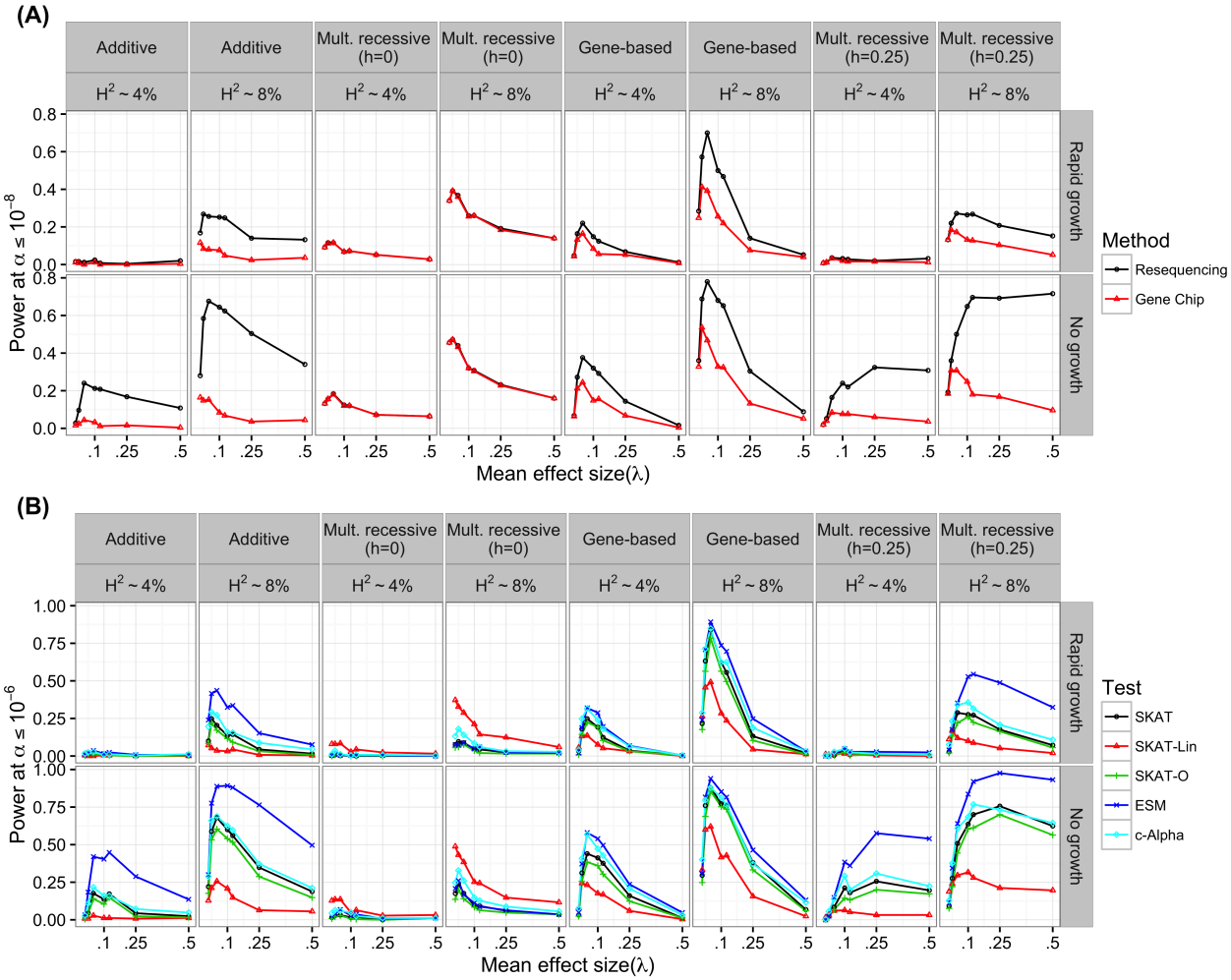

Figure 2.3: Power of association tests. (A) The power of a single marker logistic regression, at significance threshold of $\alpha \leq 10^{-8}$, as a function of λ: the mean effect size of a new deleterious mutation. For single marker tests we define power as the number of simulation replicates in which any single marker reaches genome wide significance. Two study designs were emulated. For the gene chip design only markers with $MAF > 0.05$ were considered and all markers were considered for the resequencing design. Genetic models shown here are the additive co-dominant (AC), gene-based (GBR), complete multiplicative site-based recessive (Mult. recessive ($h = 0$); cMR) and incomplete multiplicative site-based recessive models (Mult. recessive ($h = 0.25$); iMR) (B) The power of region-based rare variant association tests to detect association with the simulated causal gene region at significance threshold of $\alpha \leq 10^{-6}$. For region-based tests, we define power as the percent of simulation replicates in which the p-value of the test was less than α. The p-values for the ESM, c-Alpha were evaluated using 2×10^6 permutations. SKAT p-values were determined by the SKAT R package and represent numerical approximations to the presumed analytical p-value.

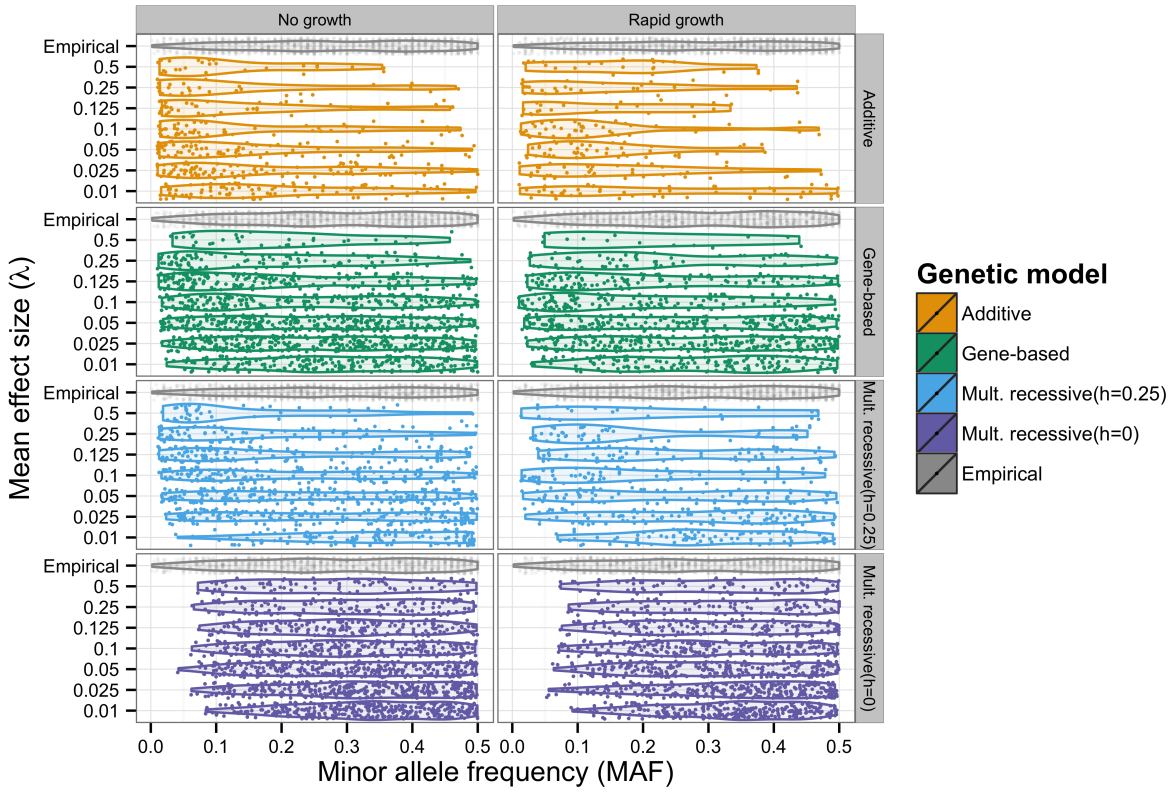

Figure 2.4: Distribution of significant GWAS hits. Horizontal violin plots depict the distribution of minor allele frequencies (MAF) of the most strongly associated single marker in a GWAS. Individual hits are plotted as translucent points and jittered to provide a sense of the total number and density of hits. Each panel contains simulated data pooled across model replicates for each value of λ with empirical data for comparison. Empirical data are described in Materials and Methods. In cases where more than one marker was tied for the lowest p-value, one was chosen at random. Shown here are the additive co-dominant (AC), gene-based (GBR), incomplete multiplicative recessive (Mult. recessive ($h = 0.25$); iMR) and complete multiplicative recessive (Mult. recessive ($h = 0$);cMR) models. All data shown are for models where $H^2 \sim 0.08$, because single marker test power was too low under $H^2 \sim 0.04$ to make informative density plots. To further increase the number simulated data points, we perfromed n=1,250 replicates at each level for this figure. Simulated data were subjected to ascertainment sampling such that the MAF distribution of all markers on the simulated genotyping chip was uniform. Specific information regarding the empirical data can be obtained in Table A.1.

Chapter 3

Efficient software for multi-marker, region-based analysis of GWAS data

3.1 Article

Efficient software for multi-marker, region-based analysis of GWAS Data

Jaleal S. Sanjak, Anthony D. Long, Kevin R. Thornton

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, 92697
Center for Complex Biological Systems, University of California Irvine, Irvine, CA 92697, USA

Corresponding Author: Jaleal S. Sanjak

Email: jsanjak@uci.edu

3.2 Preface

This chapter was originally published in *G3: Genes, Genomes, Genetics* under the title "Efficient software for multi-marker, region-based analysis of GWAS Data " [208]. It is reprinted here in its original form. The software was written by myself and Kevin Thornton with input from Anthony Long. I performed the statistical analysis of the dataset, drew the primary conclusions, and wrote the text of the paper.

3.3 Abstract

Genome-wide association studies (GWAS) have associated many single variants with complex disease, yet the better part heritable complex disease risk remains unexplained. Analytical tools designed to work under specific population genetic models are needed. Rare variants are increasingly shown to be important in human complex disease, but most existing GWAS data do not cover rare variants. Explicit population genetic models predict that genes contributing to complex traits and experiencing recurrent, unconditionally-deleterious mutation will harbor multiple rare, causative mutations of subtle effect. It is difficult to identify genes harboring rare variants of large effect that contribute to complex disease risk via the single marker association tests typically used in GWAS. Gene/region-based association tests may have the power detect associations by combining information from multiple markers, but have yielded limited success in practice. This is partially because many methods have not been widely applied. Here we empirically demonstrate the utility of a procedure based on the rank truncated product (RTP) method, filtered to reduce the effects of LD. We apply the procedure to the Wellcome Trust Case Control Consortium (WTCCC) data set and uncover previously-unidentified associations, some of which have been replicated in much larger studies. We show that, in the absence of significant rare variant coverage, RTP based

methods still have the power to detect associated genes. We recommend that the RTP based methods be applied to all existing GWAS data to maximize the usefulness of those data. For this, we provide efficient software implementing our procedure.

3.4 Introduction

Revealing the genetic basis of common human diseases, such as diabetes and heart disease, remains a central challenge in human genetics. Family-based and twin-based studies estimate that the genetic component of disease risk is typically large. Genome-wide association studies (GWAS) have identified many genetic variants associated with complex human diseases [253], yet the heritability explained by specific statistically significant variants remains small in comparison to the total heritability estimates [151, 244]. Various hypotheses explaining the "missing heritability problem" exist [151, 244, 77, 205]. Gene-by-gene, gene-by-environment, and other complex epistatic interactions might create statistical challenges for the detection of causal variants [56, 251] or might inflate total heritability estimates [277]. The missing heritability could be attributable to many common well-tagged variants which do not reach statistical significance because of their miniscule effect sizes [64, 246]. Rare variants with large effects (RALE) might drive heritability and escape detection because they are not well-tagged by current genotyping methods [158, 39]. Quantifying the roles of these non-mutually-exclusive hypotheses is important for the design of future studies and the development of new analytical tools [245]. We still do not know exactly how mutational effect sizes underlying specific diseases map onto the human site-frequency spectrum. However, it is becoming increasingly clear that rare variants are an important contributor to the genetic basis of complex diseases [6, 182, 254, 189, 45, 98, 171, 100].

The RALE hypothesis is particularly appealing to some because it is a prediction that arises naturally from population-genetic models of mutation-selection balance [84]. Specifically,

it arises from a model in which equilibrium allele frequencies and phenotypic effect sizes both reflect a balance between two things: recurrent unconditionally deleterious mutations occurring in a disease gene, and their elimination by natural selection [187]. A previous simulation study [232] investigated a novel model where standing quantitative genetic variation in complex disease genes of large effect is maintained via partially non-complementing mutations. An important prediction of this model is that a gene region can harbor several, individually rare, variants which all contribute to a complex disease phenotype. Such allelic heterogeneity is predicted to pose complications for genome wide association studies [158]. In particular, we know that single-marker association tests do not have sufficient statistical power in these cases [104, 215, 221]. Further, associations under this model are a mixture of two different types [232]. First, associations may be due to tagging a causal marker whose effect size is small, implying a sufficiently small effect on fitness, allowing the mutation to reach intermediate frequency ($> 5\%$ in the population). The second class of association is due to non-causative mutations in linkage disequilibrium (LD) with causal markers. These "tagged" associations tend to be rare, and of relatively large effect [232]. Under this model, "missing heritability" arises from a combination of allelic heterogeneity and a lack of power to identify risk variants.

Under the model of non-complementing mutations, regions harboring risk alleles show a statistical signature of a large number of markers with single-marker p-values approaching, but still below, a genome-wide significance threshold [232]. They further showed that, under this model, the excess of significant markers (ESM) test, a permutation-based regional association test, had more power to detect a causal gene region in typical GWAS data than single marker methods and many popular region-based tests [232], even for GWAS containing only common markers ($MAF > 0.05$). Although the test statistic of the ESM test is inspired by order statistics, under the permutation procedure for evaluating statistical significance it is equivalent to the rank truncated product (RTP) of p-values [51]. This equivalence was not initially recognized by [232]. Multiple variations on the RTP exist to address issues related

to correlation between p-values [47] and the need to specify a truncation threshold [269]. Although the RTP test has been used recently to obtain pathway- or gene-level associations in GWAS and other genomic applications [162, 20, 2, 137, 131, 4, 124], it is not widely-used. Here we demonstrate the utility of mining existing data sets with an RTP approach, which we call the ESM test from here on, and provide an efficient implementation which can perform genome-wide scans without the need to restrict only to coding regions.

GWAS data do not have sufficient coverage of rare variants for direct analysis, but the ESM test is a powerful tool for extracting useful information despite this fact. Here we perform an empirical analysis of the performance of the ESM test on the Wellcome Trust Case Control Consortium (WTCCC) GWAS data set [252]. We chose this dataset to determine the empirical efficacy of the ESM test because the dataset is well-characterized and easy to obtain. In addition, the choice of a dataset without substantial rare variant coverage, allows us to show that the ESM test has the power to detect the slight differences in allele frequencies between cases and control at common neutral markers, which is predicted by RALE models. We discover four novel gene regions that contribute to complex disease variation not detected in the original study, and propose that the ESM test is even better-suited to data sets that employ more modern denser SNP chips.

3.5 Materials and Methods

3.5.1 Dataset

Data were obtained from the Wellcome Trust Case Control Consortium and are as decribed in [252]. Briefly, we obtained $\sim$ 2000 cases for each of 7 diseases and a set of $\sim$ 3000 shared controls typed on an Affymetrix 500K SNP chip. Diseases included in the dataset are Bipolar Disorder (BD), Coronary Artery Disease (CAD), Hypertension (HT), Chron's

disease (IBD), Rheumatoid Arthritis (RA), Type 1 Diabetes(T1D), and Type 2 Diabetes (T2D). Case and control samples are obtained from across Great Britain. Control samples contain two subgroups; $\sim$ 1500 individuals come from the 1958 British Birth Cohort(1958BC) and $\sim$ 1500 belong to the national UK Blood Services donor pool (NBS).

3.5.2 Data Preprocessing

The raw WTCCC data were formatted for use in PLINK 1.90a[188]. SNP's listed in the WTCCC genotype file by their Affymetrix identification were translated into RefSNP (rsID) with the Affymetrix chip annotations. The SNP identifications and chromosome positions were updated to the most recent dbSNP Build 144. The SNP and individual exclusions lists provided were applied and only genotyping calls with quality score over 0.9 were included.

3.5.3 Basic Association and Permutation

The basic single marker association test is executed with the PLINK 1.90a command *–assoc*. A total of N permuted single marker p-values are obtained from PLINK!1.90a by specifying –mperm=N. We take $N = 2 \times 10^6$ permutations such that the resolution of our permutation p-value is $\frac{1}{N} = 0.5 \times 10^{-6}$, which can allow us to establish a region as genome-wide significant below a marginal p-value threshold of $\alpha \leq 1e-6$. We stored the observed association p-values, the permuted association p-values, and the R^2 between each marker (from plink *–ld* command) into HDF5 file format for use in the ESM test.

3.5.4 Excess of Significant Markers Test

We implement the ESM test as described in [232]. The test is a permutation based variation of rank truncated Fisher's combined p-value method using a null hypothesis based on order statistics. The test statistic is the sum of the differences between the observed and expected $-log_{10}(p)$. However the expected value under the null is the same for each permutation and thus the statistic is equivalent to the sum of observed $-log_{10}(p)$, i.e. the RTP. For a set of m markers the expected p-value, under the null model of no association, of the i^{th} most significant marker is $\frac{i}{m}$. Let $\mathbf{Y}$ be a vector of length m containing the observed $-log_{10}(p)$, sorted in order of decreasing significance, from the single marker association test. Then the ESM test statistic is defined to be:

$$ESM = \sum_{i=1}^{m} (Y_i + log_{10}(\frac{i}{m}))$$

For each region, we calculate the ESM test statistic based for the observed data and for each permutation of the data. For a given region, let the set of ESM test statistics be $\text{ESM}_j : j = 0, ..., N$, such that ESM_0 is the observed value and the rest are calculated from permuted data. Then the p-value for that region is:

$$p = \frac{\sum_{j=1}^{N} I(j)}{N}$$

where,

$$I(j) = \begin{cases} 1 & ESM_j \geq ESM_0 \\ 0 & ESM_j < ESM_0 \end{cases}$$

We performed the ESM test using a two stage sliding window approach. Using 100 kilobase windows we performed a genome scan with a jump size of 50 kilobases, with m =25. The effect of changing m was explored in [232] and the choice of 25 was based on average SNP density in the WTCCC data. Within each region we filtered markers based on LD, taking only SNPs whose R^2 was less than 0.2; always removing the SNP with the greater chromosomal position. While choosing this particular LD pruning rule is arbitrary, it prevents the introduction of bias due to selecting SNPs based on association significance. Regions which contained a marginally significant hit, ESM p-value less than 1e-04, were re-scanned using a finer (1 kilobase) jump size. The code for implementing the test can obtained at from github: `https://github.com/ThorntonLab/ESMtest`. Contiguous genomic regions which contain windows reaching genome-wide significance at α $\leq$1e-6 were taken and explored for functional annotations. This significance threshold results in a predicted genome-wide Type-1 error rate of approximately 0.06; the mean (across diseases)number of total windows analyzed is 58,724 and thus the idealized type-1 error rate is $58,724 * 1e-6 = 0.0587$. However, this estimate is quite conservative because the windows are spatially auto-correlated across the genome, making the effective number of tests performed much lower than the number of windows analyzed.

3.5.5 Intersection with other GWAS data

Significant regions were initially queried against the NHGRI GWAS database. Regions were classified as being potentially novel if there were no significant SNPs in the NHGRI GWAS

database for the specific disease whose genomic position fell within the boundaries of the region. Regions which contained significant SNPs in the NHGRI GWAS database that were not contributed by the Wellcome Trust were also taken for further analysis. The regions were queried against gene and transcript annotations in human reference genome GRCh38 using the R package biomaRt [54, 55]. The resulting gene and transcript annotations were manually curated for novelty and functional relevance.

3.6 Results and Discussion

We implement the ESM test as a sliding-window genome-wide scan for significant regions; we use 100Kb windows and 2 million permutations to reach genome-wide significance at an empirical $p \leq$1e-6. Region-/set-based methods result in far fewer tests than single-marker methods. By analyzing 60,000 windows with a marginal $\alpha \leq$1e-6, our genome-wide type 1 error rate is roughly 0.06; this estimate is conservative because the windows are not independent, and thus we effectively perform fewer tests than is suggested by the number of windows analyzed. Permutation procedures on genomic datasets are notoriously computationally expensive and are thus typically avoided despite their appealing statistical properties. With this in mind, we developed an efficient and freely-available computational pipeline to implement the ESM test which relies on new software and PLINK 1.90a [188] (see Methods). The pipeline leverages PLINK's fast permutation procedures for single marker association tests, stores the data in I/O optimized HDF5 file format, and performs the test. Our analysis recapitulates most, but not all, of the associations established in the standard analysis of [252] and finds new associations demonstrating that the ESM test is an excellent candidate for application in addition to standard methods.

3.6.1 Overlap between the ESM test and standard analysis

The majority of the regions found in [252] that show strong associations with case-control status are also significant under the ESM test. In [252], the standard 1-df χ^2 test resulted in 21 regions showing strong association signals ($p \leq$5e-7). Table A.2 shows that 18 of these regions also have an ESM test $p \leq$1e-6. Of the three regions which do not reach genome-wide significance under the ESM test, two have p-values between 1e-4 and 1e-6 (Table A.3). In particular, multiple windows containing rs2542151, the main SNP reported for region chr18:12.77-12.92(Mb) in association with inflammatory bowel disease, reach ESM $p =$9e-6. A third SNP, rs420259, in region chr16:23.38-23.7(Mb) reported in association with bipolar disorder by [252] did not replicate in other studies [236] and the region does not show strong association via the ESM test. Applying the SKAT [262, 133] test to the same genomic windows results in less overlap with the WTCCC results (Tables A.5 and A.6). Some of the regions not deemed significant by SKAT have been validated in other studies and can be viewed as false negatives. The ESM test has fewer false negatives. Because SKAT is not a permutation based test, it is orders of magnitude faster computationally. However, our concern should primarily be on getting better answers within the constraints of what is tractable. The ESM test is computationally feasible (Figure A.25) and is shown here to give useful results. When we look at the overlaps and differences between the results of the ESM test and single marker test, we make two important observations. First, the ESM test has the power to detect genomic regions in association with disease status. Second, because there are regions which are only identified by either the ESM test or the single marker test we should view these methods as complementary. The second point is conceptually important, but computationally trivial because one has to do a single marker test to serve as the input to the ESM test. The suggested workflow is essentially: run the single marker test, run the ESM test, analyze both results separately and then observe their union and intersection.

3.6.2 Strong associations replicated in independent datasets

Table 3.1 shows that the ESM test identifies four genomic regions that were not significant in the original WTCCC single-marker based analysis. Three out of four of these regions have since been associated with disease statuses in independent studies published in the years following the introduction of the WTCCC analysis (Table 3.1). These subsequent independent studies all leveraged datasets employing larger case/control panels and/or more densely genotyped SNPs than were originally used in [252]. Published simulations suggest that the ESM test should accrue additional benefits when used on datasets with improved genotyping (See Figure 3 in [232]). In contrast, applying SKAT [262, 133] to these same data and genomic windows was less promising. Although SKAT finds three significant regions which are not significant with a single marker test (Table A.4), only two have support in studies and it finds no completely novel candidate genes. The number of new results is not significantly different between the ESM test and SKAT, but there does appear to be a qualitative difference in the level of plausibility. However, at present we can not rule out differences in optimal approach to partitioning the genome or differences in the type of signal detected in explaining the observed differences in ESM and SKAT results. Overall three of the four novel associations identified using the ESM test are replicated, providing empirical support that the ESM test can detect novel true positive associations, even in relatively small data sets. We briefly describe the known biological significance of these three genomic regions below.

The region chr7:129.99-130.12(Mb) is strongly associated with coronary artery disease (CAD) (Figure 3.1 and Table 3.1). This region overlaps two genes: *ZC3HC1* and *KLHDC10*. A missense mutation in ZC3HC1, which is also an *cis*-eQTL for *KLHDC10* has been previously associated with CAD [58]. Neither gene currently has a clearly understood role in the etiology of CAD. The region chr22:37.09-37.21(Mb), containing *IL2RB*, is associated with type-1 diabetes (T1D) (Figure 3.1 and Table 3.1). This region was nominally associated with

rheumatoid arthritis (RA) by the WTCCC, but not with T1D. *IL2RB* has been associated with both diseases in multiple studies [179, 57, 174, 38]. Epidemiological associations with immune related genes like *IL2RB* have motivated many important basic and clinical research studies [180]. Finally, we find an intergenic region, chr1:172.87-172.99(Mb), which contains SNPs previously associated with inflammatory bowel disease (IBD) [66, 105] and Celiac Disease [50], to be associated with IBD (Figure 3.1 and Table 3.1). Both nearby genes, *TNFSF18* and *FASLG*, are part of the immunologically important TNF superfamily. The presence of putatively active regulatory elements within this associated region (Figure A.24), supports the association between variation in regulatory sequences and common diseases [156, 155].

3.6.3 Novel association: SEMA3C

The ESM test finds one additional novel region, not shown to be genome-wide significant in any study to date, showing strong association with CAD: chr7:8.08-8.09(Mb) (Table 3.1 and Figure 3.1). The only known protein-coding gene in this region is *SEMA3C* (Figure 3.2). A single SNP (rs4236644) in *SEMA3C* reached marginal significance (p =2e-6) in a meta-analysis of GWAS for total serum bilirubin levels [102]. *SEMA3C* is a secreted neurovascular guiding molecule which has a number of developmental functions and plays a role in cardiovascular development during embyrogenesis [190, 61]. Certain congenital heart diseases are attributed to disregulation of *SEMA3C* and its associated receptor *PLXNA2* [120]. *SEMA3C* is also an adipokine indicated in extracellular changes during white adipose tissue hypertrophy in human obesity [160]. In total, *SEMA3C* is a plausible candidate gene driving the observed ESM signal. However, we should note that the nearby (0.5Mb away) gene *CD36* is associated with heart-disease-related traits including response to blood lipid drugs [68], platelet count, and HDL cholesterol in African Americans [191, 41]. Although Figure 3.2 demonstrates lower support for *CD36*, its presence could be driving the association

with *SEMA3C* through long range linkage disequilibrium. Alternatively, the presence of *CD36* might reflect the typical spatial clustering of functionally related genes found in many organisms [97]. Overall, the association of *SEMA3C* with CAD is consistent with its known physiological function in the development of the heart, and thus makes it an intriguing candidate for future studies.

3.6.4 Discussion

The power of the ESM test is highlighted by the fact that it can identify novel, biologically plausible associations in an approximately 10 year old data set that has been highly studied. We provide open-source software implementing the test which can be applied to GWAS data in PLINK .ped/.bed file format. As a caveat, although the test is simple, performing the millions of permutations on GWAS data sets is computationally intensive. Individual-level genotype data is a requirement of the ESM test. The test cannot be applied to summary statistics from case/control studies. If it is applied to data with greater SNP coverage across the genome, a finer-scale sliding window may be desirable, requiring more permutations to keep Type-1 errors low. Nevertheless, simulations suggest that the power of ESM test will increase significantly when the test is applied to data sets that have employed more modern higher density SNP chips [232]. False positives due to LD between markers is often a concern for region based analysis. While it has been shown that using permutation does adequately address the impact of LD on variations of Fisher's combined p-value [164, 3]. However, when SNP pruning is applied , as it is here, to reduce the maximum pairwise correlation to 0.2 the effect is predicted to be quite insignificant [3]. This agrees with the observation from [232], that the ESM test did not result in any false positives under neutral simulations.

We find that using rank truncated product methods in conjunction with single-marker analysis yields an approximately twenty percent gain in power over single-marker analysis alone,

as illustrated by the finding of 4 new results on top of the pre-existing 21 results from the standard method. It is clear to see the potential benefit of applying the ESM test in this way to all of the existing GWAS data. Given the extent of GWAS data currently in existence, it is conceivable that a broad application of the ESM test would establish thousands of new associations. An additional benefit of a broad application the ESM test is the opportunity to validate hits in new datasets with older ones, as we demonstrated here.

A key limitation of region/SNP-set based tests in general, including rank truncated product methods, is that one cannot simply validate a single or small set of markers in a second panel. It is instead necessary to do deep genotyping of a candidate region in an independent panel in order to gain a perspective on the genetic variation present in the associated region. A corollary is that the lack of simple single SNP markers makes the estimation of effect sizes and variance explained by a detected gene region difficult; this problem should be a focus of future studies. Using existing data, rank truncated product methods have power to detect new associations between genomic regions and disease. Notably, the development of more powerful region-based tests seems likely. The ESM test was designed to detect an association signal in case/control panels under a particular gene action model and a small range of population genetic scenarios. Recent work [166] demonstrates that predictions from simulation studies regarding performance of region based tests are impacted by various model details. Thus, future research should focus on the behavior of association tests under various models of gene action and demography.

3.7 Chapter 3 Tables

Disease	Chr	Position(Mb)	Gene Region	Source
CAD	7	80.78-80.88	SEMA3C	*This analysis*
CAD	7	129.993-130.123	ZC3HC1/KLHDC10	[58]
T1D, RA	22	37.096-37.203	IL2RB	[179, 57, 174, 38]
IBD	1	172.872-172.983	FASLG/TNFSF18	[66, 105, 50]

Table 3.1: New Associations. Regions with ESM test $p \leq$1e-6 with no corresponding hit from [252] are reported below. Three out of four regions contain corresponding hits in NHGRI GWAS database not due to [252] or were otherwise previously indicated in the particular disease as cited in the source column below. One region is novel based on our analysis and overlaps with a biologically plausible gene SEMA3C.

3.8 Chapter 3 Figures

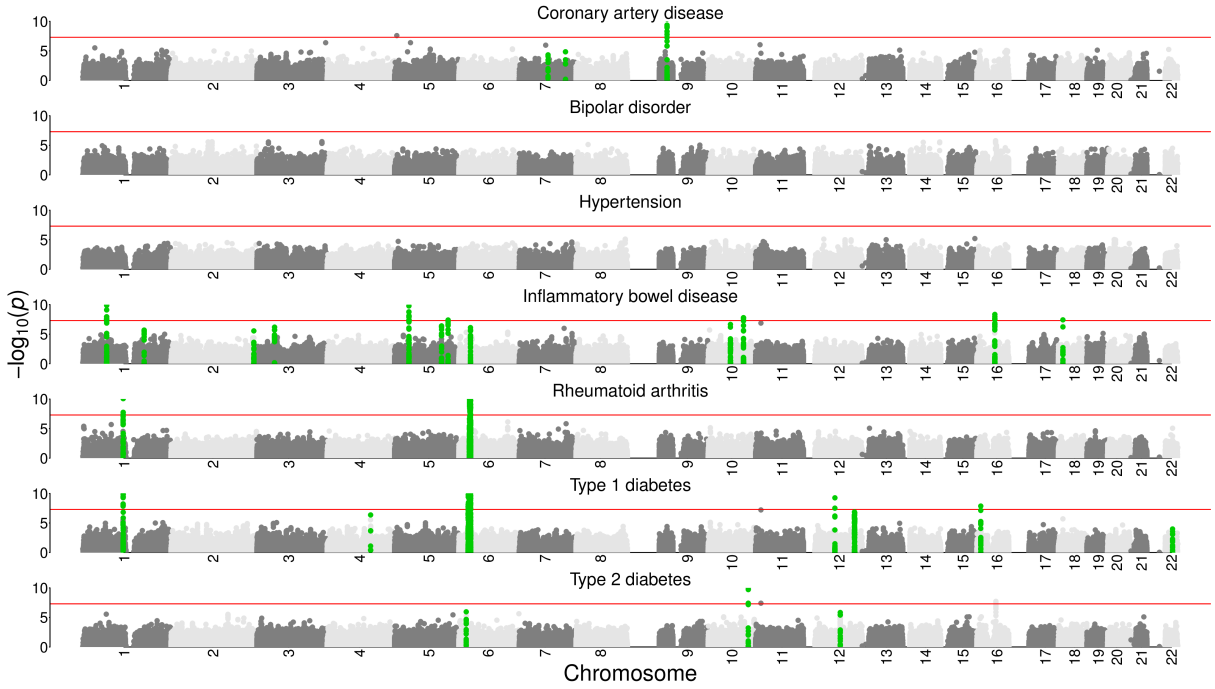

Figure 3.1: Manhattan plots with ESM significant regions highlighted. Single marker $-log_{10}(p)$p-values versus chromosomal position (BP) for all seven diseases analyzed, with SNPs in ESM significant (ESM $p \leq$ 1e-6) regions highlighted in green. Horizontal lines are placed at $-log_{10}(p) = 8$ to illustrate the typical single marker genome-wide significance threshold. SNP clusters which are highlighted in green, but do not contain a single genome-wide significant SNP are reported as novel.

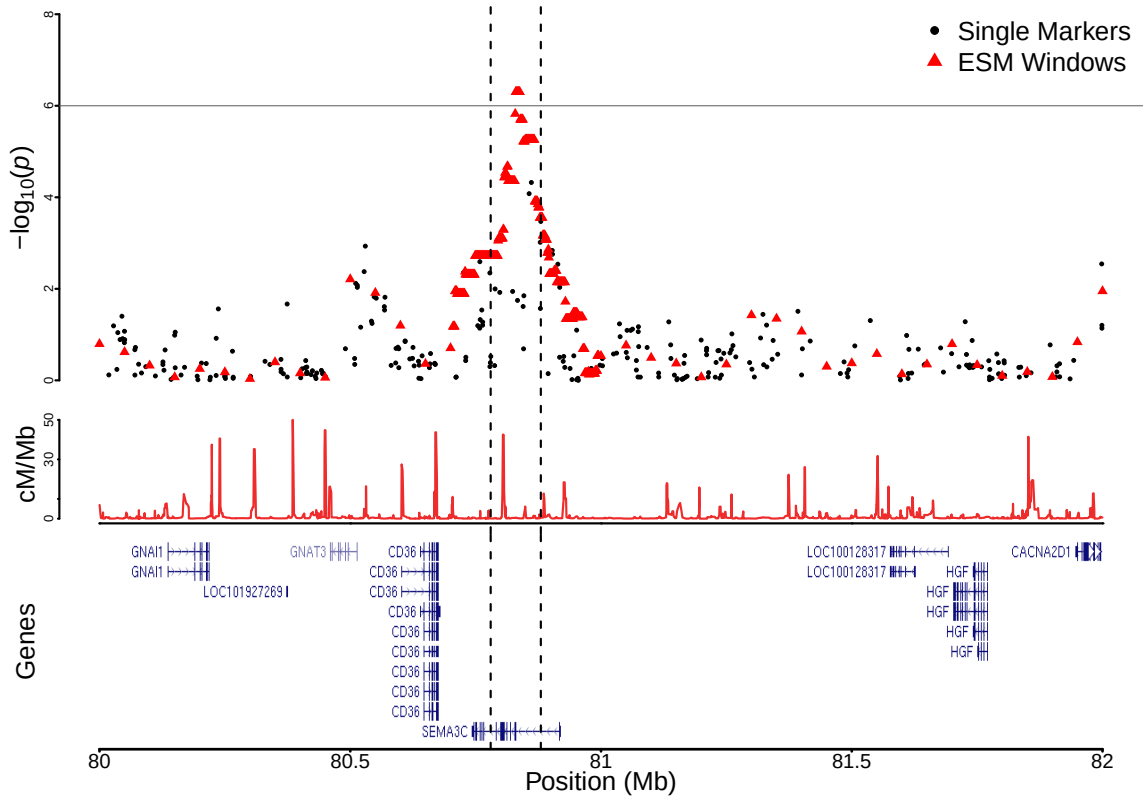

Figure 3.2: Region plot for SEMA3C hit. The top panel contains single marker (black points) and ESM test (red triangles) $-log_{10}(p)$-values for coronary artery disease versus chromosomal position in the region chr7:80-82 (Mb). Each ESM test point is plotted at the midpoint of a genomic window to which that $-log_{10}(p)$-values corresponds. The single 100Kb ESM significant (ESM p $\leq$1e-6) region chr7:80.78-80.88 (Mb) is demarcated by vertical dashed lines, and the horizontal lines are placed at $-log_{10}(p) = 6$ to indicate the ESM test significance threshold. The middle panel contains the recombination rate in cM/Mb obtained from HapMap througout the same region. The lower panel shows the refseq gene UCSC genome browser track for the region.

Chapter 4

Evidence of directional and stabilizing selection in contemporary humans

4.1 Article

Evidence of directional and stabilizing selection in contemporary humans

Jaleal S. Sanjak, Julia Sidorenko, Matthew R. Robinson, Kevin R. Thornton, Peter M. Visscher

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, 92697 Center for Complex Biological Systems, University of California Irvine, Irvine, CA 92697, USA Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia Institute for Molecular Biosciences, The University of Queensland, Brisbane, Queensland, Australia Department of Computational Biology, University of Lausanne, Lausanne, 1010, Switzerland

Corresponding Author: Peter M. Visscher

Email: peter.visscher@uq.edu.au

4.2 Preface

This chapter was originally published in *Proceedings of the National Academy of Sciences of the United States of America* under the title "Evidence of directional and stabilizing selection in contemporary humans " . It is reprinted here in its original form. I permformed the initial statistical analysis of the data, which lead to the key observations in the paper. In the writing of the paper, small corrections were applied to all analyses and they were redone by Julia Sidorenko. I wrote the text of the paper with the help of Peter Visscher and the advice of Matt Robinsion and Kevin Thornton.

4.3 Significance statement

Combining high throughput molecular genetic data with extensive phenotyping enables the direct study of natural selection in humans. We see first hand how and at what rates contemporary human populations are evolving. Here we demonstrate that the genetic variants associated with several traits including age at first birth in females and body-mass index in males are also associated with reproductive success. In addition, for several traits, we demonstrate that individuals at either extreme of the phenotypic range have reduced fitness–the hallmark of stabilizing selection. Overall, the data are indicative of a moving optimum model for contemporary evolution of human quantitative traits.

4.4 Abstract

Modern molecular genetic datasets, primarily collected to study the biology of human health and disease, can be used to directly measure the action of natural selection and reveal important features of contemporary human evolution. Here we leverage the UK Biobank data to test for the presence of linear and non-linear natural selection in a contemporary population of United Kingdom. We obtain phenotypic and genetic evidence consistent with the action of linear/directional selection. Phenotypic evidence suggests that stabilizing selection, which acts to reduce variance in the population without necessarily modifying the population mean, is widespread and relatively weak in comparison to estimates from other species.

4.5 Introduction

Natural selection can strongly affect patterns of phenotypic variation. This fact has lead to considerable interest in understanding how natural selection and other evolutionary forces combined to shape the alellic spectrum underlying variation within and between populations. Most of this work has focused on searching the genome for signatures of past selective events [192]. Yet selection fundamentally acts on phenotypes, not genotypes. Therefore, the relationships between phenotypes and fitness must be studied in contemporary populations to observe natural selection directly. In doing so, we can gain insights about the direction and magnitude of phenotypic evolution. Theoretically, such observations allow one to predict future evolutionary change, and they can serve as points of comparison with inferences of selection obtained from other sources of data. Here we report observational evidence that is consistent with the action of natural selection in a contemporary human population.

Directional selection results in a covariance between the trait and fitness, and can lead to changes in the mean value of a trait in a population [204, 185, 186]. Further, if phenotypic

variation for the trait is caused by genetic factors, then directional selection can result in changes in the genetic composition of a population. Phenotypes may also be subject to stabilizing or disruptive selection, which are both non-linear forms of selection. The key distinction between stabilizing and disruptive selection is whether the relationship between fitness and a phenotype is concave down or up, respectively. Stabilizing selection, which is commonly invoked in theoretical studies of quantitative traits [85, 203, 112, 125, 237], will tend to reduce phenotypic variation while disruptive selection will tend to increase it. In a seminal paper on the direct study of natural selection, Lande and Arnold [126] put forth a statistical framework by which the magnitude of both directional and non-linear selection could be estimated from observational data via regression of fitness onto phenotypes and their squared values.

Application of the Lande and Arnold framework to human populations has yielded evidence consistent with the action of directional selection on physiology, life-history and body-size traits in both pre- and post-industrial societies [222]. While important differences between the studied populations exist [212, 226], a few interesting trends have emerged. Multiple studies have suggested that directional selection has acted to lower the age at first birth in females [118, 90, 249, 30, 223, 234], increase the age at menopause [118, 30], increase weight in females [7, 30, 223], and decrease height in females [7, 172, 30, 223, 226] in contemporary post-industrial populations.

Direct evidence for the action of stabilizing selection in humans is more scarce. Birth weight is one reported example of a human trait under stabilizing selection [107], although the intensity of selection has decreased in post-industrial societies [239]. A twin study of female reproductive life history traits showed evidence for a phenotypic optimum for age at menarche [118]. Additionally, phenotypic evidence has been presented that is indicative of the simultaneous action of directional and stabilizing selection on height in the Dutch [227]. However, a recent study in the contemporary United States found no evidence for any

non-linear selection [13]–although sample size may have limited the power to detect such effects. While selection acts on phenotypes, evolution requires genetic variation. The genetic covariance between a phenotype and fitness determines the expected evolutionary change [204, 185, 186] of that phenotype in a population. Genetic covariances between traits can be estimated from pedigree information or directly from molecular genetic data [133].

The use of molecular genetic data has multiple advantages over traditional sources of data for the study contemporary selection [234, 13, 122]. The most obvious advantage is the availability of data; genetic data from large samples of unrelated individuals are increasingly accessible to many researchers. Another advantage comes in the ability to control for possible cultural transmission of traits, which is generally confounded with genetics in observational studies because parents pass both on to their offspring [222]. This issue can be partially mitigated by accounting for population structure [184] and geography in samples of unrelated individuals.

In the first attempt to use SNP-array data to study contemporary natural selection on complex traits, Tropf, et al. [234] found a negative genetic correlation between relative lifetime reproductive success (rLRS)–the individual life time reproductive success divided by the mean–and age at first birth using a bivariate linear mixed modeling approach [230, 133]. However, Beauchamp [13] noted that the bivariate analyses are under-powered with modest sample sizes, and chose to analyze genetic predictors derived from the results of independent large genome-wide association studies (GWAS). Significant negative correlations between polygenic prediction scores for female educational attainment and rLRS have been found in the populations of the contemporary United States [13] and Iceland [122]. But, reliance on external GWAS summary statistics limits analyses to traits which have already been thoroughly characterized at the genetic level.

Here, we analyze the phenotypic and genetic correlates of rLRS in the UK Biobank (UKB). The UKB is a large population-based prospective study of the genetic and environmental

determinants of aging related disease [228]. The dataset consists of over 500,000 individuals from the United Kingdom who have been genotyped at common SNPs and clinically phenotyped for many different traits. These data provide paired genotype and phenotype samples large enough to accurately measure additive genetic correlation between many heritable complex traits [248].

First, we apply the Lande and Arnold framework through regression analyses of the relationship between a suite of phenotypes and a proxy for fitness, rLRS, in 217,728 females and 158,638 males. Then, the genetic data available from 157,807 female and 115,902 male unrelated samples is used to estimate genetic correlations between the phenotypes and rLRS through LD-score regression analysis [22, 23]. This analysis was supported by the observation that rLRS had a low, but measurable heritability. Our analyses replicate the main results of other recent studies [234, 13, 122], and uncover a host of other significant genetic correlations with rLRS. We also report estimates of quadratic relationships with rLRS, which may be interpreted as evidence consistent with stabilizing or disruptive selection, informing efforts to model the processes that maintain heritable variation in human complex traits [187, 59, 1, 217, 141, 278, 173, 166, 242, 209]. Our observations are consistent with the action of weak directional and stabilizing selection, and limited disruptive selection the UK Biobank population.

4.6 Phenotypic observations

We estimate linear (β) and quadratic (γ) selection gradients by regressing rLRS onto phenotypes and squared phenotypes [126]. Because of possible heterogeneity in selection pressures and rLRS measurement precision–documented number of live births in females versus self-reported number of children fathered in males–all analyses were performed on a sex-specific basis. In total, we analyzed 37 traits in females and 33 traits in males; the traits and results

are listed in SI Appendix, Dataset A.10. The histogram of $\hat{\beta}$ (SI Appendix, Fig. A.26A) shows that the observed signals of directional selection are weaker than what has been found in other species [117]. Such weak selection gradients are unlikely to lead to large changes in phenotypic distributions over clinically or socially relevant timescales [13, 122]. Yet, it is important to note that the measured rLRS may be biased because it is conditional on survival to post-reproductive age and may not be completed rLRS for males. Despite the weak signal, we find that 23 female traits and 21 male traits have significant non-zero directional selection gradients ($\hat{\beta}$) at a family-wise error rate (FWER) ≤ 0.05. However, many of these traits are highly correlated (SI Appendix, Fig. A.29 and Fig. A.30) and should not be viewed as separable axes of selection.

The $\hat{\beta}$ estimates for traits with a significant estimate in at least one sex are shown in Fig. 4.1A. Overall, the $\hat{\beta}$ estimates were not highly correlated between sexes. This implies that there is some sex-specific selection acting on these phenotypes, consistent with recent work on the genetic and phenotypic correlates of viability [165]. In many instances, the difference between sexes is driven by a large difference in the magnitude, not the sign, of $\hat{\beta}$. For example, the estimate for educational attainment in females is $\hat{\beta}_{EA,F} = -0.0612 \pm 0.0022$ ($p < 10^{-172}$) while the estimate in males is $\hat{\beta}_{EA,M} = -0.0086 \pm 0.003$ ($p \approx 10^{-2.3}$). Conversely, the estimate for birth weight in males, $\hat{\beta}_{BW,M} = 0.021 \pm 0.0038$ ($p < 10^{-7}$), is much larger than the estimate in females of $\hat{\beta}_{BW,F} = 0.0047 \pm 0.0027$ ($p = 0.084$). Height is the only trait we studied for which the data indicate sexually antagonistic selection. In females $\hat{\beta}_{HT,F} = -0.028 \pm 0.0021$ ($p < 10^{-39}$), while in males $\hat{\beta}_{HT,M} = 0.022 \pm 0.0025$ ($p < 10^{-18}$). Fig. 4.2 further illustrates that the predicted phenotypic optimum is above and below the population mean height for males and females respectively, consistent with multiple previous studies showing a difference in contemporary selection pressures on height between males and females [226]. Further, the empirical relationships between LRS and height, as illustrated in SI Appendix, Fig. A.27, is very similar to that which is predicted by a Gaussian stabilizing selection model (SI Appendix, Fig. A.28).

In contrast to a recent study [13], 12 traits in females and 14 traits in males have a significant non-linear selection gradient estimate ($\hat{\gamma}$). It is important to note that the sample size available in [13] was nearly two orders of magnitude smaller than that of the present study. The histogram of $\hat{\gamma}$ values in SI Appendix, Fig. A.26B shows a skew towards negative values. Specifically, 47 of the 64 sex-trait combinations examined show a negative quadratic selection gradient (median $\hat{\gamma} = -0.0059$), of which 26 were significant. 24 sex-trait pairs had a non-zero $\hat{\beta}$ and a significant negative $\hat{\gamma}$, which is indicative of the simultaneous action of directional and stabilizing selection.

Fig. 4.1B shows that, unlike many of the $\hat{\beta}$ estimates, the estimates of $\hat{\gamma}$ were quite similar in both sexes. For example, the estimates for height are $\hat{\gamma}_{HT,F} = -0.0189 \pm 0.0014$ ($p < 10^{-37}$) in females and $\hat{\gamma}_{HT,M} = -0.015 \pm 0.0017$ ($p < 10^{-17}$) in males respectively. Fig. 4.1B shows that among traits with significant $\hat{\gamma}$ in both sexes the male estimate tends to be further from zero (with height being an exception). We find no traits with significant $\hat{\gamma}$ in both sexes with opposite signs.

Fig. 4.1B shows that age at menopause, fluid intelligence score and age at first birth (AFB) all have a positive $\hat{\gamma}$ in females. In addition, the $\hat{\gamma}$ for educational attainment is positive in both sexes. A positive value of γ can be interpreted as evidence of disruptive selection. However, our results for AFB are more indicative of a plateauing of directional selection towards the upper phenotypic extreme rather than true disruptive selection (SI Appendix, Fig. A.31). The situation is somewhat less clear for the other phenotypes with a significant positive $\hat{\gamma}$ (SI Appendix, Fig. A.32, Fig. A.35 and Fig. A.34) and these results should be followed up more closely in future work.

A multiple regression analysis provided a more conservative perspective on the phenotypic correlates of rLRS. Due to multi-collinearity (SI Appendix, A.29 and A.30) and non-overlapping missing data we had to choose only a subset of traits for the multiple regression. The full multiple regression results are included in SI Appendix, Dataset A.10 and are sum-

marized in SI Appendix, Table A.7. In males, the estimates of β for hand grip strength, pulse rate, BMI, and systolic blood pressure are significant in the multiple regression and retain their direction of association from the single-trait regression models. In females, the estimates of β for educational attainment (EA), AFB, age at menarche, bone mineral density, systolic blood pressure, and waist-to-hip ratio are significant in the multiple regression model. However, the direction of the association between EA and rLRS in females is positive in the multiple regression setting. This stands in sharp contrast to the separate regression results and strongly points against a simple linear relationship between EA and increased rLRS. Rather, it appears that correlated factors, such as AFB, drive the apparent selection [103] on EA.

To further explore the relationship between AFB, EA and rLRS we fit a reduced multiple regression model with EA, AFB and their interaction. In the reduced model, all three terms (two linear and one interaction) were highly significant (SI Appendix, Table A.7). As in the initial multiple regression, the direction of association for EA is positive in the reduced model. In addition, the interaction term between EA and AFB is strongly positive ($\hat{\beta_{AFB:EA}} = 0.03 \pm 0.0016 (p < 10^{-112})$). One hypothesis consistent with this observation is that the effect of EA on rLRS becomes more positive as AFB increases and that the negative regression coefficient in the EA alone model can be fully explained by the strong negative association between AFB and rLRS. In simpler terms these results suggest that among females who have children later in life, those individuals with higher EA will tend to have more children. This is despite the fact that people with higher EA tend to have fewer children overall and is consistent with prior work in the Icelandic population [122].

The estimates of γ were much less significant in the multiple regression compared to the separate regressions. For females, the estimates of γ for age at first live birth and BMI were significant–with the BMI estimate reversing direction to be positive. In males, the estimates of γ for EA and BMI were significant–with both retaining their direction of association.

4.7 Genetic correlations with rLRS

The phenotypic results are consistent with the action of natural selection, but for adaptation to occur there must be effects on the genetic level. To this end, we analyzed genetic data from 157,807 female and 115,902 male unrelated samples. Estimates of the genetic correlations between several traits and rLRS, $r_{g,rLRS}$, were obtained from the data using LD-score regression [22, 23]. LD-score regression uses the regression of the cross-products of z-statistics onto a measure of linkage disequilibrium in a genomic window (the LD-score), assuming a polygenic architecture, to estimate genetic covariance components from GWAS results. We also analyzed the UKB interim data release using a linear mixed modeling (LMM) approach. This approach was not computationally feasible on the full dataset; we report the results on the full dataset using LD-score regression in the main text, but see the SI Appendix for a discussion of the LMM results. All genetic variance and covariance estimates are contained in SI Appendix, Dataset A.11.

Theory predicts that traits highly correlated with fitness will have low heritability [161]. As expected, rLRS has a low but significant SNP heritability in the UKB dataset, which means that we have power to detect strong genetic correlations. Specifically, the LD-score regression estimates of $h^2_{SNP,rLRS}$ were 0.056 and 0.033 in females and males respectively with respective standard errors of 0.0046 and 0.0054. Fig. 4.3 shows $\hat{r}_{g,rLRS}$ for the subset of traits for which an estimate was marginally significant ($p \leq 10^{-3}$) in at least one sex.

The estimated genetic correlation with rLRS was significant for several anthropometric traits. For example, the estimates of $\hat{r}_{g,rLRS}$ for Height are $-0.1278 \pm 0.0274 (p < 10^{-5})$ in females and $-0.0074 \pm 0.0412 (p = 0.18)$ in males. Recall that we estimated a significant negative selection gradient in females with a small but significant positive selection gradient in males. The phenotypic results are in agreement with prior studies in western populations [226] suggesting that selection on reproductive success favors shorter females and taller males.

However, because we see no evidence for a genetic correlation between height and rLRS in males we do not predict that the observed phenotypic selection in males would induce a response to selection (in a single generation).

BMI provides another important example of evidence for directional selection on an anthropometric trait. We estimate that the $\hat{r}_{g,rLRS}$ for BMI is $0.104 \pm 0.0344 (p = 10^{-2.6})$ in females and $0.31 \pm 0.046 (p < 10^{-10})$ in males. These results are qualitatively similar to our phenotypic results which indicated positive directional selection in both sexes with a larger estimate in males. Although the genetic result for females did not pass our study-wise significance threshold, the results are consistent with the hypothesis that contemporary selection on reproduction favors higher BMI in males and support exploration of the same hypothesis in females.

The genetic correlation estimate for age at first birth (AFB) in females was the strongest observed in our study. We estimate that the $\hat{r}_{g,rLRS}$ for AFB is $-0.593 \pm 0.035 (p < 10^{-16})$. This result is consistent both with our phenotypic observations and prior pedigree based results [30]. Educational attainment (EA) is also strongly negatively correlated with rLRS; we estimate the $\hat{r}_{g,rLRS}$ for EA to be $-0.316 \pm 0.037 (p < 10^{-16})$ in females and $-0.2539 \pm 0.052 (p < 10^{-5})$ in males. However, the most likely explanation for these genetic results is something very similar to what we observed on the phenotypic level for these two traits, which would agree with work on contemporary selection in an Icelandic population [122].

Another interesting aspect of the observed negative directional selection on AFB is that it would suggest selection for increased female reproductive lifespan. However, the evidence is less clear when we compare the results on AFB to other female reproductive life-history traits such as the ages at menarche (AAM) and menopause (AMP). In fact, we estimate that the genetic correlation with rLRS is positive for AAM ($\hat{r}_{g,rLRS} = 0.133 \pm 0.032 (p < 10^{-4.4})$) and negative for AMP ($\hat{r}_{g,rLRS} = -0.168 \pm 0.045 (p < 10^{-3.7})$). The genetic result for AMP is particularly unexpected because it is inconsistent with both our phenotypic result (even

though the phenotypic correlation is very small, 0.02) and a prior result obtained in a pedigree study [30]. Further, we estimate that the genetic correlation between AAM and AFB is strongly positive despite the fact that signs of the the $\hat{r}_{g,rLRS}$ estimates for the two traits are opposite. We intuitively expect a positive relationship between AAM and AFB because the latter requires the former. However, the positive genetic correlation between rLRS and AAM is less explicable.

Estimation of the genetic evidence of non-linear selection was not performed because of lack of statistical power. Theory predicts that the additive genetic variance for a squared phenotype is likely to be very small and, when present, is confounded with genetic control of phenotypic variability. In addition, the empirical heritability estimates for squared phenotypes are small (SI Appendix, Fig. A.44). Despite the lack of power, a polygenic predictor for height, constructed from a meta analysis of the GIANT-UKB joint dataset, did show a marginally significant negative quadratic regression coefficient in females (see SI Appendix for details).

4.8 Discussion

Estimates of linear and quadratic selection gradients were obtained via simple linear regression of a broad set of phenotypes onto a proxy for fitness. The results suggest that many traits measured in the UKB are under the influence of directional and stabilizing selection. However, many of the selection gradient estimates were not significant in a multiple regression setting, implicating apparent selection [103]. Yet, the population genetic architecture of a trait may still be modified by apparent selection.

For example, the direction of association between female educational attainment and rLRS is positive in the multiple regression, which opposes results from our single trait regressions, genetic correlation analyses and multiple other published results [249, 62, 96, 218, 13, 122].

Our findings lead to the prediction that variants with a positive effect on female educational attainment would decrease in frequency over time even if variance in educational attainment itself does not directly cause variance in reproductive success. Consistent with this prediction, recent work demonstrated that the mean polygenic score for educational attainment has declined over time in the Icelandic population [122], but also suggest that this trend may be explained by factors like female age at first birth.

Consistent with previous studies, our results support a hypothesis of strong negative selection on female age at first birth [106, 118, 91, 90, 249, 30, 163, 223, 16, 234]. We also observed a small but positive relationship between age at menopause in females and rLRS on the phenotypic level, which agrees with previous results [118, 30, 222, 16]. But, we find support for a negative genetic relationship between rLRS and age at menopause. Further, both genetic and phenotypic data suggest a positive correlation between age at menarche and rLRS. Thus, it is unclear whether the total reproductive lifespan is positively or negatively correlated with rLRS in our data. As larger samples from diverse populations become available, we may gain a more clear view of the selective forces acting on reproductive traits in contemporary humans.

There is clear evidence for correlation between rLRS and several anthropometric traits. Our findings are consistent with previous reports of selection for increased BMI [30, 13]. Additionally, the data suggest that the relationship between rLRS and height is more negative in females than in males, which agrees with other results in the literature [223, 227, 13].

Our estimates are conditional on survival to post-reproductive ages, so the intensity of selection could be different for traits that strictly influence survival. Birth weight is a classic example of a trait under strong stabilizing selection, where high and low birth weights are correlated with reduced survival in both males an females [107]. Yet we find no evidence for stabilizing selection on Birth weight in males and only a marginally significant estimate of $\hat{\gamma} = -0.0057 \pm 0.0019$ ($p < 10^{-2}$) in females.

There are a few other important caveats and limitations to our present analyses. All of our results are conditional on the suite of phenotypes that we have measured; there is a real possibility that there are unmeasured phenotypes that drive or confound some of our results. This issue is related to the phenomenon of apparent selection and should always be kept in mind when studying phenotypic selection [103]. In addition, the genetic correlations are estimated using common SNP markers (MAF>0.01), which may be a source of bias because the genetic variants with deleterious effects on fitness are likely to be rare and thus absent from our analyses. Yet, this should simply reduce the power of our analyses. Further, there is evidence that the population of the UKB may not be perfectly representative of the whole population of United Kingdom [70]. The potential ascertainment bias (heathy particpant bias) in the UKB is important to consider and may have a quantitative effect on our estimates, but the bias is not likely to be large enough [70] to disrupt the conclusions of our work in a qualitative way.

The distributions of $\hat{\beta}$ and $\hat{\gamma}$ provide useful context for considering the types and strengths of selective forces at play in contemporary human populations [117]. These insights support ongoing efforts to use theoretical evolutionary models to understand the maintenance of heritable variation for complex traits in humans [85, 203, 112, 125, 237, 187, 59, 1, 217, 141, 278, 173, 166, 242, 209]. The estimates of β and γ are qualitatively consistent with estimates from other species [117], but the quantitative range is an order of magnitude smaller. However, the selection gradients from our study are estimated with much more precision that those in other species, where the sampling variance may have inflated the range of estimated coefficients. So, while the signal of selection appears to be statistically significant we do not expect that selection can explain observed secular trends in the phenotypes we studied.

Stabilizing selection appears to be more common form of non-linear selection. The most common model of stabilizing selection used in evolutionary quantitative genetics is the Gaussian

stabilizing selection model [28]. One of the most important parameters of the Gaussian stabilizing selection model is the inverse selection intensity normalized by the phenotypic variance, $\frac{V_S}{V_P}$. This ratio quantifies how fast fitness, modeled by a Gaussian function, decreases as a function of distance to the theoretical fitness optimum. $\frac{V_S}{V_P}$ can be estimated by the negative reciprocal of the quadratic selection gradient [103]. Based on the first and third quartile values of $\hat{\gamma}$, we estimate that a reasonable range for human phenotypes is $\frac{V_S}{V_P} \sim 28 - 173$ with a median of 65, which would be considered weak, but non-trivial in a theoretical context[237]. Theoretical arguments suggest that a thorough characterization of the effects of stabilizing at the genetic level will require larger sample sizes and/or methods of interrogating non-additive genetic variance to directly observe stabilizing selection acting on genetic variation in a population sample like the UKB.

We have shown the power of combining high throughput molecular genetic data with extensive phenotyping to study the ongoing dynamics of human evolution [222]. Our work supports further study of a dynamic moving-optimum model for the evolution of complex traits in humans. Presently, we do not know if the genetic architectures of complex traits are commensurate with equilibrium models parameterized by their contemporary selection gradients. If they are not, further research is needed to better understand how contemporary evolutionary forces differ from the ones that shaped the genetic architecture of the trait.

4.9 Materials and methods

Phenotypic and Genetic data were obtained from the UK Biobank and may be accessed by all bonafide researchers from the UK Biobank Access Management System. Only data from samples of self-reported white-british ancestry over the ages of 45 years for females and 50 years for men were used in all analyses, unless otherwise noted. Phenotypic analyses were performed using linear regression in R [193]. Genetic correlations were calculated

using LD-score regression software according to the protocol developed in [22]. Statistical significance was determined using bonferroni corrected p-values at family wise error rate of 0.05. The Northwest Multicentre Research Ethics Committee (MREC) approved the study and all participants in the UK Biobank study provided written informed consent. For detailed descriptions of the data preparation and analyses, see the SI Appendix, Supporting Materials and Methods.

4.10 Acknowledgements

This research was conducted using the UK Biobank Resource under project 12505. The University of Queensland authors are supported by the Australian Research Council (Discovery Project 160103860) and the Australian National Health and Medical Research Council (grants 1078037 and 1113400). This work was supported by NIH grant R01-GM115564 to KRT. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1321846. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.11 Chapter 4 Figures

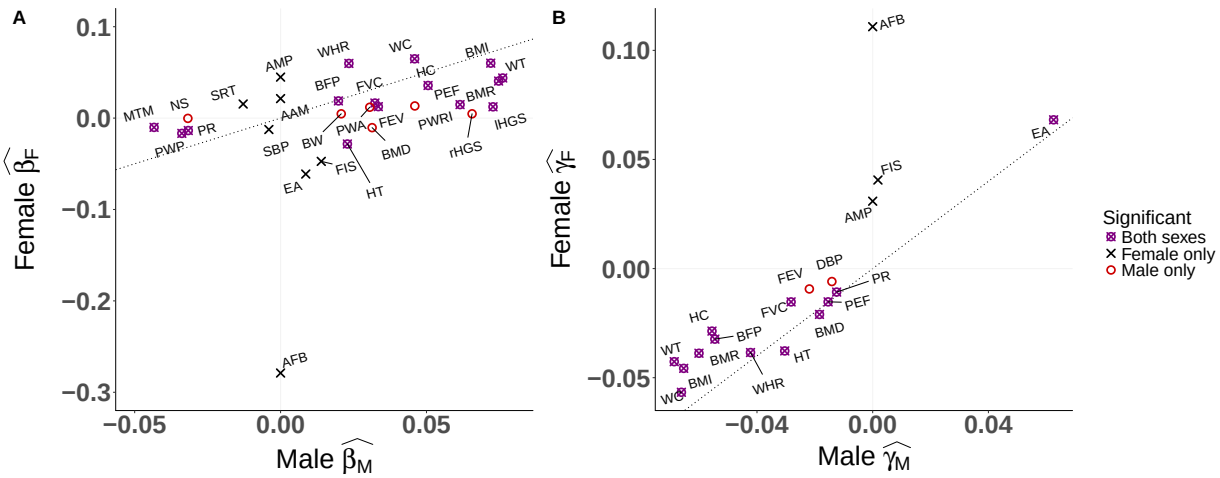

Figure 4.1: Scatter plot showing the magnitude of (A) linear selection gradients $\hat{\beta}$ and (B) quadratic selection gradients $\hat{\gamma}$ for a selection of traits in Females and Males. Traits were selected on the basis of being significant (FWER$\leq$ 0.05) in at least one sex. Estimates are on the z-score scale for theoretical interpretation and consistency across traits. Points are labeled with the following abbreviated trait descriptions: age at first birth (AFB), age at menarche (AAM), age at menopause (AMP), basal metabolic rate (BMR), birth weight (BW), body-fat percentage (BFP), body-mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (SBP), educational attainment (EA), fluid intelligence score (FIS), forced expiratory volume (FEV), forced vital capacity (FVC), hand grip strength (HGS), height (HT), hip circumference (HC), mean time to correctly identify matches (MTM), neuroticism score (NS), peak expiratory flow (PEF), pulse rate (PR), pulse-wave arterial stiffness index (PWA), pulse-wave peak-to-peak time (PWP), pulse-wave reflection index (PWRI), SRT vision estimate (SRT), bone mineral density (BMD), waist circumference (WC), waist-to-hip ratio (WHR) and weight (WT). Note that data on AFB, AMP and AAM are not available for males and their regression values were set to zero.

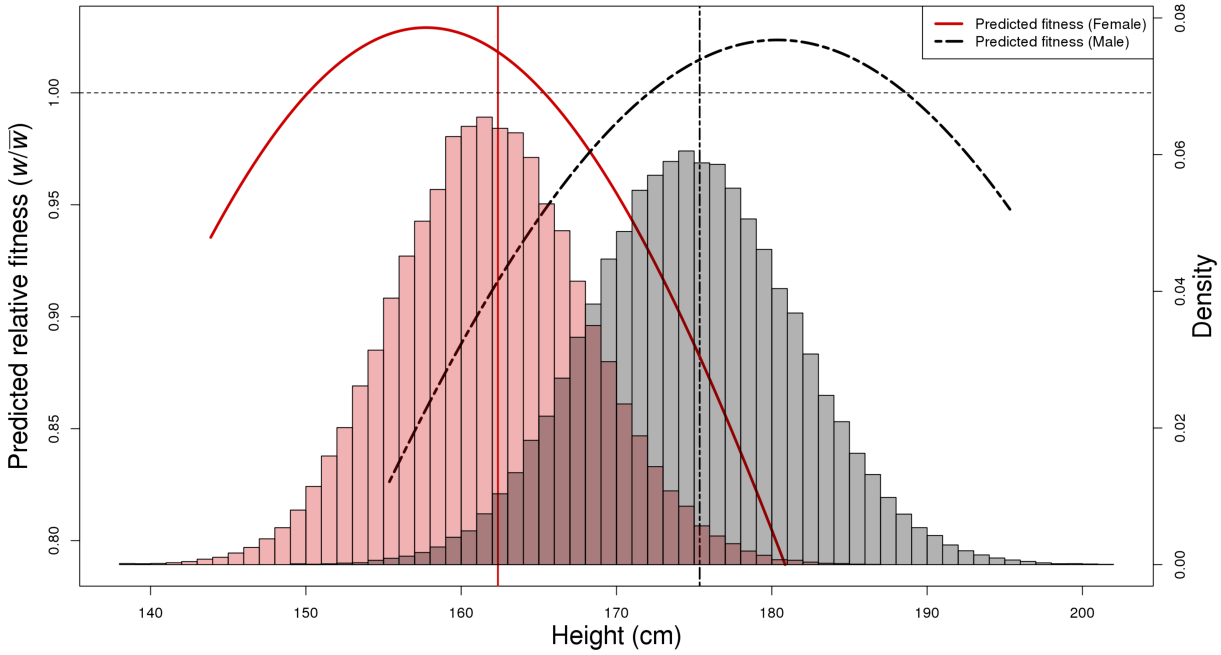

Figure 4.2: Predicted relative fitness as a function of Height. Predicted relative fitness as a function of Height. Linear and quadratic selection gradients were converted into parameters of a Gaussian fitness function. Using the parameterized Gaussian fitness function, relative fitness values across the observed phenotypic range were predicted and shown by solid red (female) and dashed black (male) lines. The population means are indicated by vertical solid red (female) and dashed black (male) lines. Histograms of female (red) and male (gray) phenotypes are overlaid with an axis on the right hand side. The horizontal dashed line indicates a relative predicted fitness of 1.

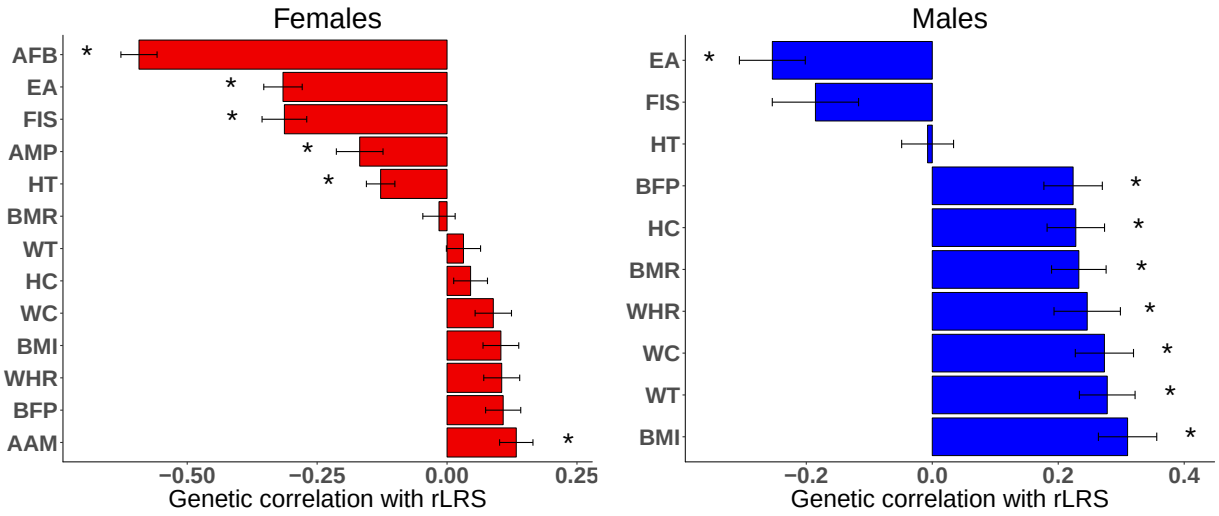

Figure 4.3: Bar plots showing genetic correlations between a selection of traits and rLRS for Females (red) and Males (blue). Traits were selected on the basis of being marginally significant ($p \leq 0.001$) in at least one sex, and were sorted in ascending order of the estimate for each sex. Data are displayed as the correlation estimate plus or minus the standard error ($\sim p \leq 0.001$, *FWER≤ 0.05). Bars are labeled with the following abbreviated trait descriptions: age at first birth (AFB), educational attainment (EA), age at menarche (AAM), age at menopause (AMP), fluid intelligence score (FIS), height (HT), basal metabolic rate (BMR), weight (WT), hip circumference (HC), waist circumference (WC), waist-to-hip ratio (WHR), body-fat percentage (BFP) and body-mass index (BMI).

Chapter 5

Estimating the number of effective alleles at a QTL

5.1 Chapter description

Estimating the number of effective alleles at a QTL

5.2 Preface

This chapter will not be published in its current form outside of this dissertation. I performed all data generation and analysis steps presented here. R Code for implementing the CaSANOVA method were obtained from Bondell et al.(2009)[18]. Additional R code for replicating simulations from King et al.(2014)[115] were provided to me directly by Elizabeth King.

5.3 Abstract

Standard methods of QTL mapping assume that causal loci are biallelic. However, there are both theoretical and empirical reasons to believe that there are several segregating alleles at many QTL. How much functional allelic diversity is typically present at a QTL? Can we obtain reliable estimates of the number of alleles present at a specific QTL? Few methods exist for estimating the true number of segregating alleles at a detected QTL. CaSANOVA/GFLasso is one promising method that penalizes a linear regression model based on the pairwise differences between estimated haplotype effects. Here I apply the CaSANOVA/GFLasso method to several different QTL mapping panel designs. I assess both the power to detect a QTL and the efficacy of CaSANOVA/GFLasso to estimate the number of effective alleles. I find that CaSANOVA/GFLasso can provide unbiased estimates of the number of effective alleles at large effect QTL in very large random population samples. Further, I show that performing an intercross of phased outbred lines (POLs) does not increase the ability to estimate the number of alleles, but does provide very good power to detect QTL.

5.4 Introduction

Most analyses of genetic association assume the presence of a biallelic quantitative trait locus (QTL) in linkage with available genetic markers[60, 146]. However, King et al.(2014) showed evidence that it is more typical to find multiple segregating alleles at expression QTL[115]. They posed the following question: is it possible to directly estimate the number of segregating alleles at QTL?

King et al.(2014) addressed this question by comparing the goodness-of-fit of a series of linear models with haplotypes split into allelic groupings. They presented a simulation study

and found that their model selection method was accurate when the number of true alleles present was low. King et al.(2014) also showed that their method was biased downward in the presence of many alleles. This left the door open for the development of new approaches to the multiple alleles problem.

The creation of large multi-parent populations[29] from which samples can be taken for use in quantitative trait mapping experiments [87] presents an opportunity to develop and test new methods for estimating the number of alleles at a QTL. It might be possible to sequence samples of diverse of haploid yeast strains derived from known founder populations[29, 87]. Haploid strains can be crossed to create diploids allowing for the estimation of both additive and dominance effects in a multi-allelic context. Such an experimental design could elucidate more complex aspects of the genetic architecture of quantitative traits such as allelic heterogeneity[115] and gene-based dominance[209].

The problem of multiple alleles has already been considered in the statistical genetics literature. Bondell et al.(2009)[18, 238] and Kim et al.(2009)[110] both proposed similar penalized linear regression approaches to deal with multiple alleles. The methods are called Collapsing and Shrinkage in ANOVA (CaSANOVA)[18] and the Graph Guided Fused Lasso(GFLasso)[110] respectively. In the context of the multiple alleles problem, both methods estimate haplotype effects while shrinking the effect estimates toward one another. The shrinkage is achieved by placing a penalty on the pairwise differences between haplotype effects.

Here I explore several simulated multi-parental diploid QTL mapping panels. I estimate the power to detect additive and dominance QTL under each design and then apply the CaSANOVA/GFLasso method to estimate the number of segregating alleles. Three types of panels are explored. First, I explore a simple random sample of diploid individuals that is amenable to analysis with a standard genetic association test. Second, I simulate a blocked phased outbred line (POL) design[87] that requires the use linear mixed modeling[213]. In

the proposed POL designs, several small full POL intercrosses, e.g. Hallin et al.(2016)[87], are constructed and analyzed jointly. Lastly, I test CaSANOVA/GFLasso method on the design used in King et al.(2014) with the simulation approach developed therein.

I show that the CaSANOVA/GFLasso method can provide unbiased estimates of the number of alleles in random population sample, provided the sample sizes are quite large compared to the effect size of the QTL. Under POL designs, the CaSANOVA/GFLasso method fails to effectively collapse haplotype effects. This failure is potentially an algorithmic one, as I employed a FaST-LMM approximation[138, 194] due to computational limitations. Alternative direct optimization methods, such as those based on proximal gradients[270], might provide superior performance. Despite the poor performance on the multiple alleles problem, POL designs represent a powerful class of experiments for detecting additive and dominance QTL. Finally, I show that the CaSANOVA/GFLasso method behaves very similarly to the model selection method employed in King et al.(2014), when applied to the same simulations. This suggests that perhaps the study design, not the statistical method, is the main factor limiting our ability to estimate the number of segregating alleles at a QTL.

5.5 Results and Discussion

5.5.1 The GWAS design

I assume the existence of an outcrossing population founded from 18 inbred lines that has been freely recombining for many generations with no selection. From this population, pairs of haploid individuals can be sampled and crossed to create diploids. In the simplest experimental design, one could randomly generate a panel of diploid individuals whose genotype at a particular locus will correspond to one of the 18 founder haplotypes at that locus.

This type of random sample is amenable to single locus analysis, i.e. GWAS, and I call this approach the GWAS design. I simulated multi-haplotype single locus genotypes under the GWAS design, performed ordinary least squares linear regression to estimate the haplotype effects[115] and then applied the CaSANOVA/GFLasso method to estimate the number of effective alleles present. Figure 5.1 shows an example of how the haplotype effect estimates go from being scattered (Figure 5.1a) to aligned in clear groups after (Figure 5.1b) the application of the CaSANOVA/GFLasso method.

As expected, the GWAS design provides good power to detect additive QTL (Figure 5.2). Figure 5.2 shows that with multiple replicate experiments and the sample size of N=2000, there is nearly complete power to detect a QTL that explains 1% of phenotypic variance. I explored several genetic models of dominance, ranging from pure dominance (max) through co-dominance (add) to pure recessivity (min). Figure 5.3 shows that the GWAS design has good power to detect a dominance QTL under the purely recessive (min) model, while showing no indication of false positives under co-dominance (add). One explanation for the difference in power under the min and max models (Figure 5.3) is that I included a putative "wild-type" allele in all my simulations. The "wild-type" allele was given a value of zero, while all other alleles were given positive effects; this modeling approach means that the min function creates a more pronounced deviation from additivity.

The CaSANOVA/GFLasso can shrink and collapse the haplotype effect estimates in an unbiased way under the GWAS design (Figure 5.4). Yet, Figure 5.4 also shows that there is a large amount of variance around the true value, even for large sample sizes. While the variance does decrease as sample size increases, it may still be prohibitively expensive to sequence sample sizes large enough to reliably estimate the number of alleles at any particular QTL.

5.5.2 POL designs

The POL design can be powerful method for mapping QTL[87]. In a complete POL cross, two distinct sets of haploid lines from an out crossing population are intercrossed in an all by all fashion. Here I tried to strike a balance between the number of unique haploid genomes sequenced and the total number of phenotypic measurements made. Therefore, I constructed many small complete POL intercrosses and analyzed them jointly; Figure 5.5 illustrates the POL designs used here.

The POL design has more power than the GWAS design to detect additive QTL, when the number of individuals sequenced is similar. Figure 5.2 shows that with N=750 under a GWAS design, power is quite low, but under the POL design with N=768 Figure 5.6 shows that power is high. The POL design similarly shows more power to detect dominance QTL than the GWAS design, as shown by Figure 5.7. However, the POL design also shows a tendency to produce false positives. Figure 5.7 shows that when the heritability explained by the focal locus is low, the LMM will sometimes find a significant dominance QTL under an additive model. The false positives disappear when there is more power, either with increased repetition or QTL effect size (Figure 5.7).

It appears extremely difficult to correctly estimate the number of effective alleles in the POL design. The CaSANOVA/GFLasso method regularly fails to shrink and collapse any haplotype effects under the FaST-LMM framework (Figure 5.8). I next considered whether it was possible to ignore the correlation structure in the POL design and simply perform single marker GWAS.

I found that the average linkage disequilibrium (LD) in the panel decayed extremely rapidly as a function of lines used (Figure 5.9). The lack of LD suggested that a single marker analysis might not be heavily biased by the structure of the cross. However, using the standard regression approach (used in the previous GWAS section) did not improve the

ability to estimate the number of alleles (Figure 5.12).

5.5.3 Comparison to King et al.(2014)

King et al.(2014) proposed a model selection method to determine the best grouping of haplotypes. They then used the number of groups as an estimate of the number of effective alleles at a QTL. In a simulation based on real genotypes from their RIL crosses, they showed that the model selection method was accurate when the number of alleles was low, but was conservative when the number of alleles was high.

I used scripts provided by King et al.(2014) to generate similar simulated data. Then I applied the basic linear model version of CaSANOVA/GFLasso method to those data. Figure 5.13 shows that the CaSANOVA/GFLasso method produces a very similar behavior on the King et al.(2014) simulated data as the model selection method. I tested whether expanding the size of the RIL cross would improve the behavior, but Figure 5.13 showed that the conservative behavior persists even with 10,000 RIL crosses.

5.5.4 Discussion

Estimating the number of alleles at a locus remains an open problem. It appears that it is much easier to estimate this number when the true number of alleles is much smaller than the number of known founder haplotypes (Figures 5.4 and 5.13). This could be because it is necessary to have an allele represented several times in the linear model coefficient vector in order to properly assign it to a new allele group. Of course, with increasing sample size the standard error of the coefficients become small and the model will converge on the truth.

The POL design deserves serious consideration as an approach to mapping QTL. It has high power to detect both additive and dominance QTL, with lower sequencing cost as compared

to the GWAS design. However, it performed extremely poorly on the allele estimation problem.

I believe that there are potentially several problems with how the POL design was analyzed in this work. First, the FaST-LMM approach only approximately solves the model and is a source of error. Direct methods to optimize the LMM with CaSANOVA/GFLasso penalty, such as proximal gradient methods[270], might provide superior performance on this problem. In addition, I did not explicitly account for the polygenic background in my analysis of the POL design. Therefore, including a genome wide empirical kinship matrix[87] is an obvious next step.

Interestingly, the CaSANOVA/GFLasso method performed conservatively on the King et al.(2014) simulated data. This behavior was qualitatively similar to the model selection method employed in King et al.(2014). This indicates that the major limitation to estimating the number of alleles is the study design, not the statistical method. I would argue that the CaSANOVA/GFLasso method has received more attention in the statistical literature [18, 238, 110, 270] and is possibly more well understood. This makes the CaSANOVA/GFLasso method a good candidate for further exploration.

5.6 Methods

5.6.1 Genetic association test in a random sample

Single locus genotypes were simulated based on a randomly mating population at linkage equilibrium which was constructed from a set of 18 founder lines. Each founder haplotype was assumed to be present at equal frequency in the population. The founder haplotype effect sizes were assigned based on the number of assumed effective alleles, which ranged

from 2 to 16.

In all cases, a base allele was assumed to have an effect of zero. The remaining alleles were assigned effects based on a draw from a gamma distribution with shape=1 and scale=0.1. The effect size distribution was chosen to mimic the simulations from Thornton et al.(2013) [232, 209], in which gamete effects were based on a sum of point mutations each with an effect drawn from an exponential distribution. Therefore, the distribution used in these simulations reflects the assumption of approximately one causal variant per effective allele and an average effect size of 0.1 with variance 0.01.

Founder haplotypes were then randomly assigned to an effective allele. Population samples of size N = 750, 1,000, 1,500, 2,000, or 2,500 were taken at random. Phenotypes for each diploid individual were determined by the standard quantitative genetic model, which assumes independent genetic and environmental contributions[60, 146].

$$P = G + E$$

The environmental contribution was drawn from a Gaussian distribution with mean zero and variance σ_e^2. The environmental variance used was tuned to make the heritability due to the focal locus equal to H^2=0.25%-50%; tuning was done in each replicate by first determining the actual variance of G and then back calculating the necessary σ_e^2.

Genetic values were determined by one of five distinct genetic models. All of the models can be described by a single power mean function

$$G_p = (\frac{1}{2}(g_1^p + g_2^p))^{\frac{1}{p}}$$

The parameter p is set to either negative infinity, -1, 0,1, or positive infinity. These values of p correspond to the minimum, harmonic mean, geometric mean, arithmetic mean (additive),

and maximum functions. From a biological perspective the minimum, harmonic mean, and geometric mean all reflect recessive models of decreasing completeness, the arithmetic mean is an additive co-dominant model and the maximum function is a model of complete dominance. Importantly, the geometric mean function (p=0) is equivalent to the gene-based recessive model from Thornton et al.(2013) and chapter 2 of this document.

To simulate replicate experiments, several environmental effects (E) are drawn for each individual in the sample. These environmental effects are averaged and added to the the genetic effect to determine a line/individual mean value(y). The line/individual means are regressed against genotype to estimate haplotype effects.

Finally a penalized linear regression model was fit based on the CaSANOVA/GFLasso methods[18, 110]. We regress phenotype (y) onto the model matrix (X) of haplotype encodings. Specifically, the basic model matrix (X) contains no intercept column and 18 columns corresponding to count (0,1 or 2) of the corresponding founder haplotype. To estimate the founder effects vector β, we minimize the sum of squared error subject to penalty on the sum of pairwise differences between haplotype coefficients.

$$\hat{\beta} = argmin_{\beta}(||y - X\beta||^2 + \lambda \sum_{i} \sum_{i<j} |\beta_i - \beta_j|)$$

Subject to

$$\sum_{i} \beta_i = 0$$

The parameter λ determines the strength of the penalty applied to the pairwise differences and is tuned by Bayesian Information Criterion (BIC). Code to implement this method was adapted from [18, 238]. The final haplotype coefficient vector values were rounded to the parameter ϵ = 1e-6, which was used as the convergence criteria CaSANOVA algorithm. The number of unique values in the haplotype coefficient vector was used to determine the

number of effective alleles in the population.

In addition to the number of effective alleles, it was also possible to calculate the power to detect QTL under this study design. To do so, an unpenalized linear regression model is fit and significance is evaluated with an F-test. The F-test is performed with $\alpha = \frac{0.05}{1200}$, because there can be assumed to be approximately 1,200 10kb loci in a yeast genome(Anthony Long, personal communication). The unpenalized linear regression model is fit with the basic model matrix(X), as outlined above, as well as a dominance effect model matrix(D) that is encoded according to the multi-allelic haplotype model of Da et al. 2015 [46]. With these two matrices it is possible to evaluate the power to detect additive and dominance QTL.

5.6.2 Phased outbred line intercross (POL)

In the genetic association procedure described above, each individual represented a cross between two randomly chosen haploid lines. Therefore, to perform a GWAS with N diploids one must sample and sequence 2*N haploid lines, to be crossed at random without replacement. An alternative approach which reduces the number of sequenced lines and increases the number of phenotypic measurements is an intercross of haploid lines. The diploids constructed from the cross of haploid lines are called phased outbred lines (POLs)[87].

In a POL intercross, two distinct sets of lines are fully crossed to each line in the other set. This is distinct from a diallel because in a diallel there is only one set of lines. Haploid lines extracted from a randomly mating population constructed from inbred founders are similar to recombinant inbred lines. Therefore a haploid line intercross is very similar conceptually to the recombinant inbred line cross design[115].

To simulate a POL design, two sets of N=768 haploid genomes were generated with 1,200 loci per genome. A founder haplotype is assigned to each locus in each haploid genome at

random, assuming equal founder frequencies in the population. Both the total heritability of the trait and heritability explained a focal major effect QTL are varied in the simulations. Remaining heritability not explained by the focal major effect QTL is distributed across a polygenic background, assumed to be composed of 50% of the genome, i.e. 600 loci.

Three POL designs were simulated. All three are highly sparse POL designs in that a full intercross is only performed within small blocks of the total 768 by 768 design matrix. Design 1 consisted of 96 sets of complete 8 by 8 POL blocks, Design 2 was 64 sets of 12 by 12 POL blocks and Design 3 alternates between 8 by 8 and 12 by 12 POL blocks. Within each POL block, all haploids in set a are crossed to all haploids in set α. Single locus haplotype effects are drawn in the same manner as described in the previous section. The polygenic background is assumed to be purely additive.

In these POL designs the diploid individuals within row/column of a POL block share an entire genome-wide haplotype. Therefore phenotypic measurements within a POL block are far from independent and a linear mixed effect model (LMM) must be used to properly account for the study design[87]. Specifically, the fixed effect additive X and dominance D matrices, as described above, are fit along with a random effect matrix Z corresponding to each haploid line. The Z matrix has length equal to the total number of phenotypic measurements and width $2N = 1536$. The full LMM is outlined below:

$$y = X\beta + D\delta + Zu + \epsilon$$

In order to estimate the number of effective alleles, we first remove the dominance matrix, as was done in the previous section. Then the same CaSANOVA/GFLasso penalty is applied to the fixed effects vector (β). Previous authors have described how to apply penalization terms to fixed effects in an LMM setting [17, 210, 194]. We modify the approach outlined in Schelldorfer(2011)[210], by applying the CaSANOVA/GFLasso penalty to the objective

function of the LMM; this function is the standard negative log-likelihood function of the LMM[213] without the constant term.

$$Q = \frac{1}{2}log|V| + \frac{1}{2}(y - X\beta)V^{-1}(y - X\beta) + \lambda \sum_{i} \sum_{i<j} |\beta_i - \beta_j|$$

Where V is the covariance matrix of the random effects. Because this function is neither convex nor separable in terms of the parameters(β), most optimization procedures such as gradient descent will fail[210, 235]. Therefore, the factored spectral transformation(FaST)[138] is applied to the LMM, to reduce it to the standard penalized regression problem[194]. The FaST-LMM transformation is outlined in both Lippert et al.(2011)[138] and Rakitsch et el.(2013)[194]. It essentially involves fitting the LMM without fixed effects, then performing an eigendecomposition on V, rotating the model matrix(X) and scaling the phenotypic observations(y) to turn remove the random effects from the model equation. The transformed model matrix and phenotype vector can then be used in the standard CaSANOVA/GFLasso method.

5.6.3 Testing CaSANOVA/GFLasso on the King et al simulations

In addition to the random population sample and the RIX design, CaSANOVA/GFLasso was applied to the simulations from King et al 2014[115]. King et al 2014 [115] analyzed a set of over 596 RILs of Drosophila melanogaster. They simulated genetic data by sampling from their empirical RIL genotype calls. Allele effects were sampled from a Gaussian distribution or took on fixed values from 1 to the number of simulated effective alleles. Here, genotype data was generated with scripts from King et al.(2014). These data were fed into the CaSANOVA/GFLasso method described above. In addition, genotype data of the same structure was generated for a much larger hypothetical RIL panel of 10,000 RILs.

5.7 Chapter 5 Figures

a b

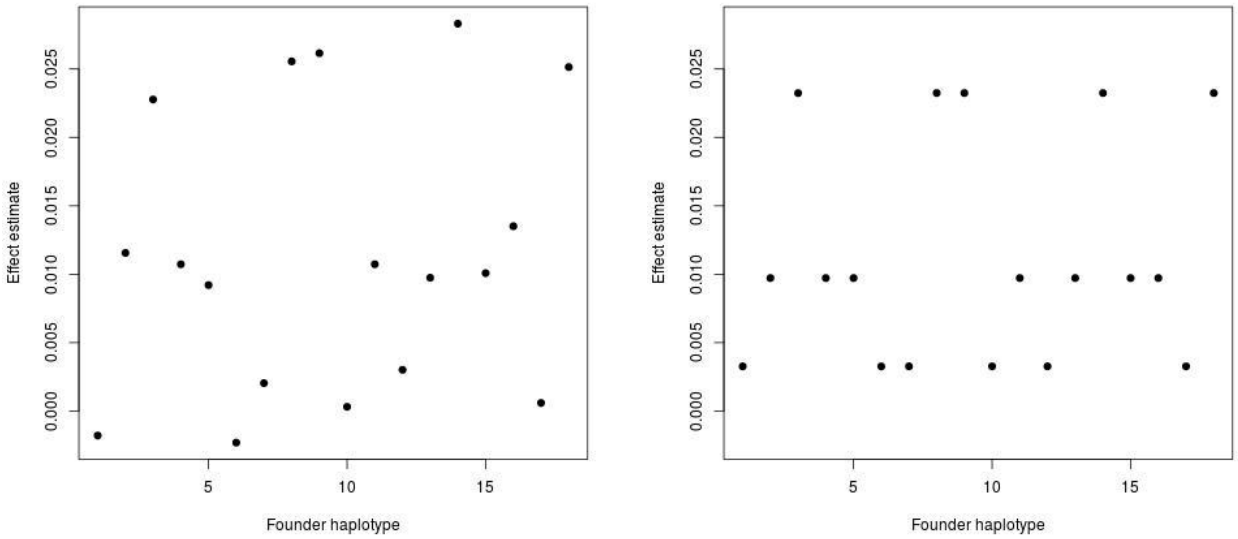

Figure 5.1: Example of CaSANOVA/GFLasso method. Estimates of founder haplotype effects are shown based on the (a) original linear model and (b) the CaSANOVA/GFLasso method. In this example, a sample of N=1,500 diploid individual is analyzed under the GWAS design. There are 3 functional alleles at an additive QTL that explains 10% of phenotypic variance. Three replicate experiments were performed and the the CaSANOVA/GFLasso method was applied to the line means.

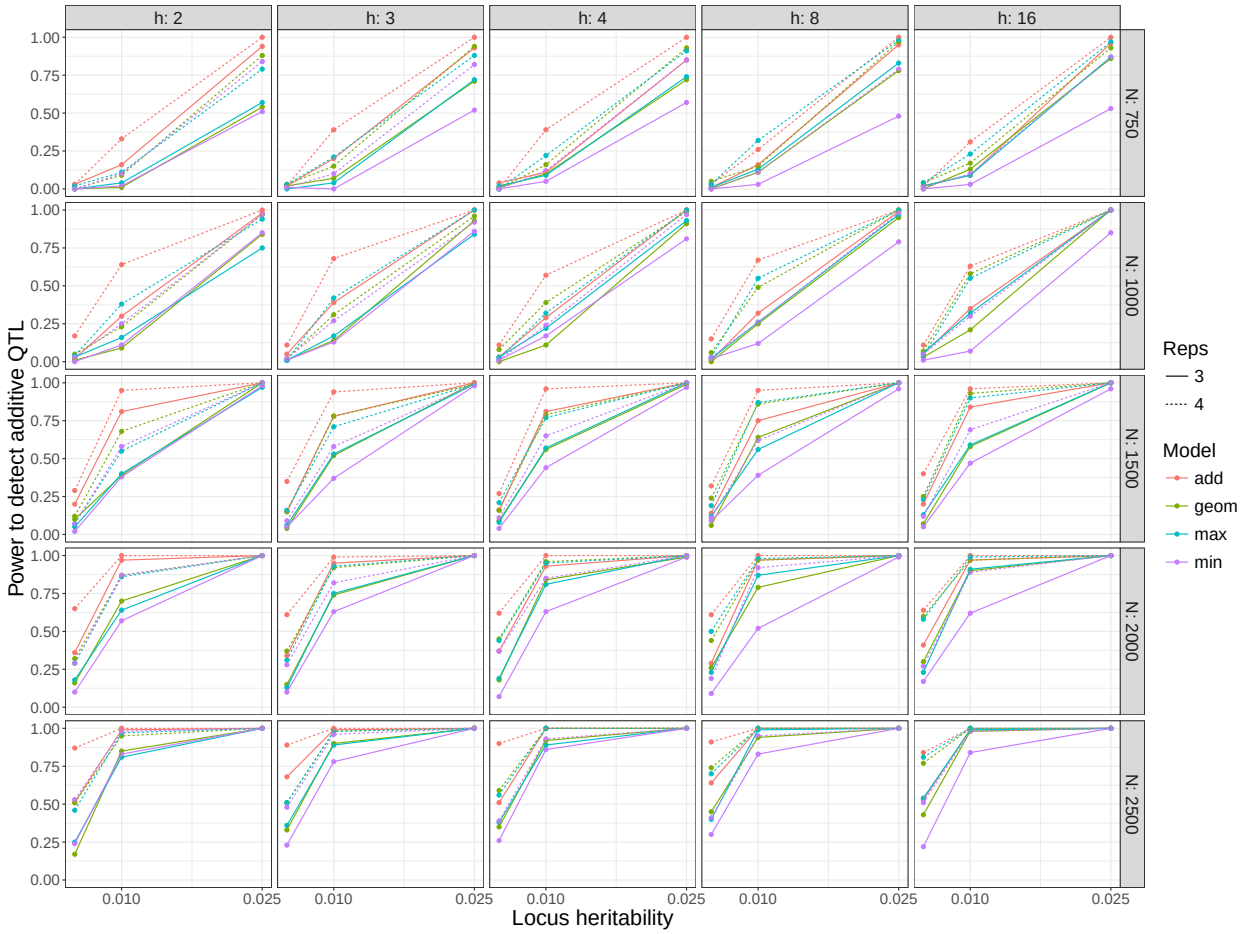

Figure 5.2: Power of to detect an additive effect QTL with GWAS design. These panels show power at $\alpha = \frac{0.05}{1200}$ as function of the heritability explained by the QTL. The horizontal facet shows the results when there are h=2 to h=16 functional alleles at the QTL. The color of the lines represent the genetic model and the line shape corresponds to the number of replicate experiments simulated

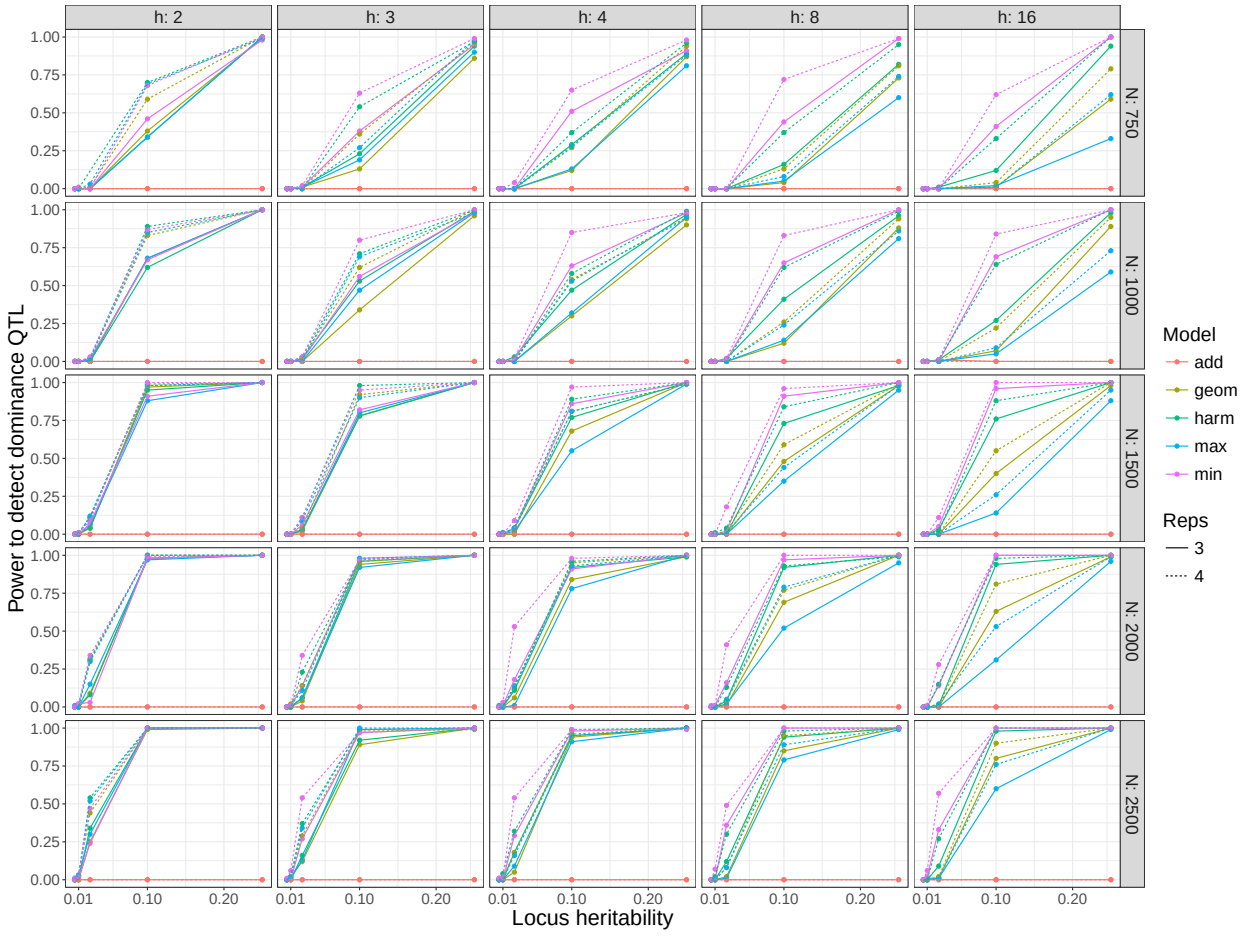

Figure 5.3: Power of to detect a dominance effect QTL with GWAS design. These panels show power at $\alpha = \frac{0.05}{1200}$ as function of the heritability explained by the QTL. The horizontal facet shows the results when there are h=2 to h=16 functional alleles at the QTL. The vertical facet shows how the results change as a function of sample size. The color of the lines represent the genetic model and the line shape corresponds to the number of replicate experiments simulated

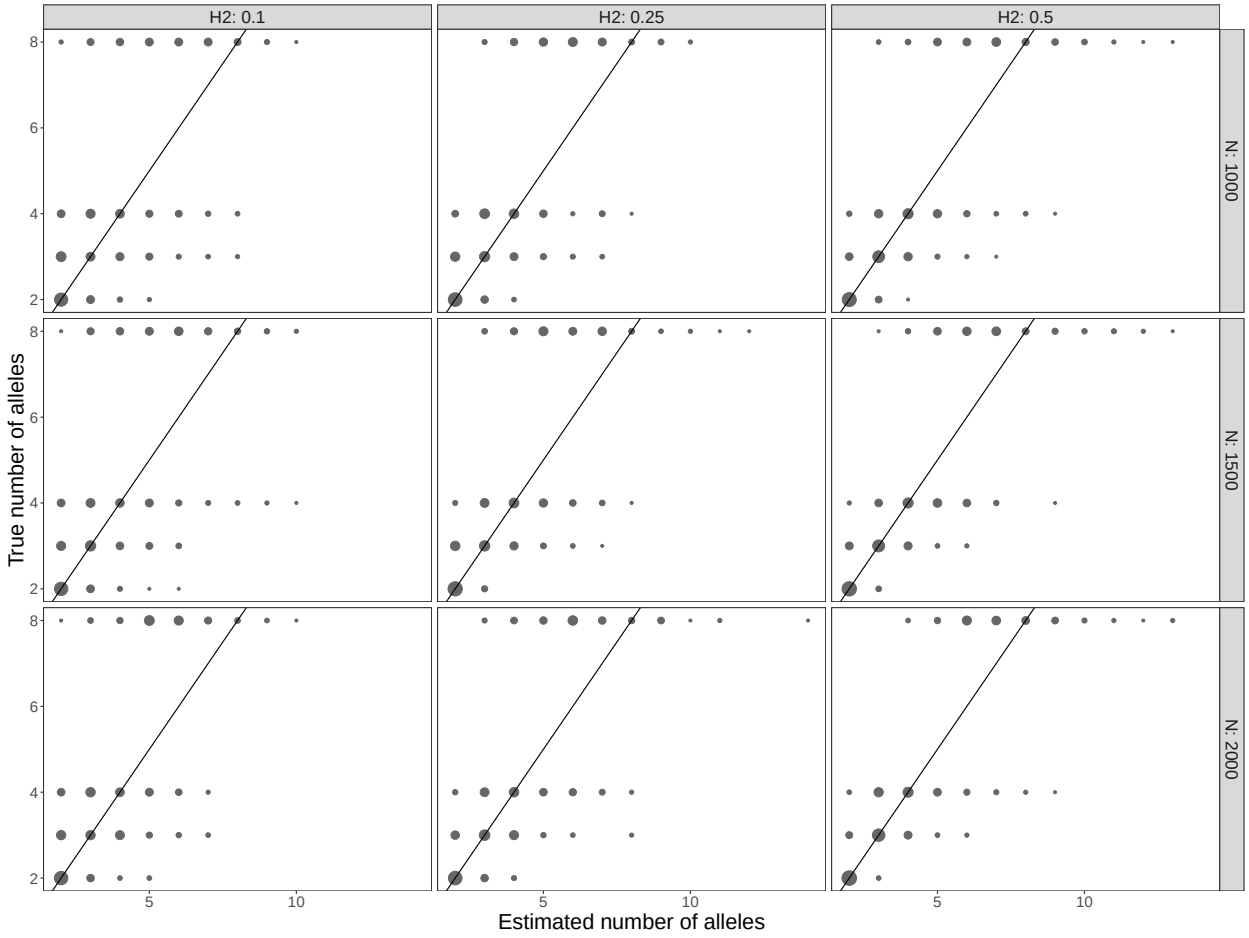

Figure 5.4: Estimated number of effective alleles using CaSANOVA/GFLasso with GWAS design. These panels show the relationship between the true number of alleles and the estimated number of alleles at an additive QTL. These data reflect 4 replicate simulations. The horizontal facet shows how the results change as a function of heritability explained by the QTL. The vertical facet shows how the results change as a function of sample size. The size of the gray circles correspond to the number of simulation replicates taking on that value. The line $y = x$ is illustrated in black.

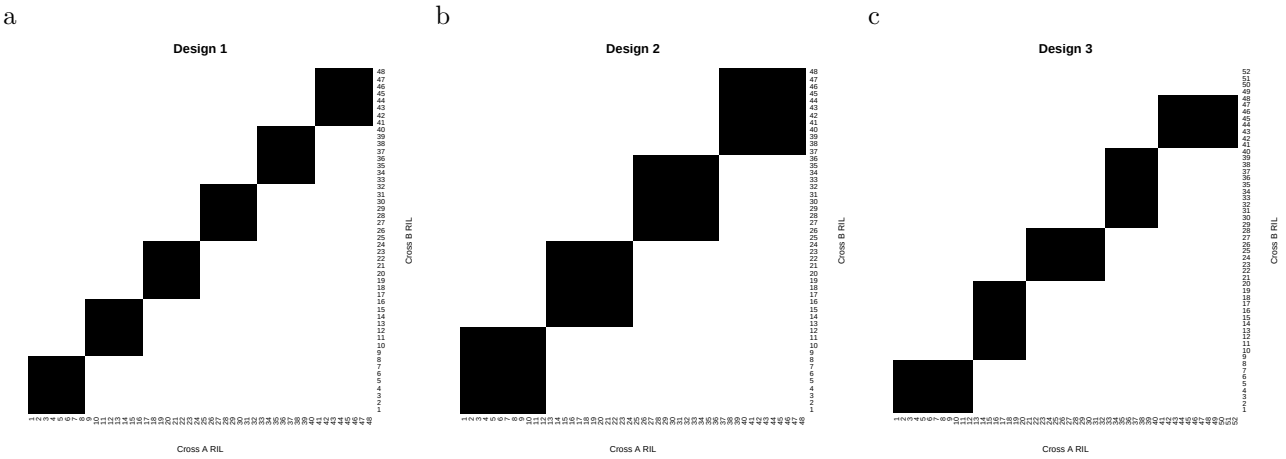

Figure 5.5: POL study designs. I tested three different mapping panel designs based on phased outbred lines (POLs). All three designs involved two sets of 768 haploids (RIL), with either (a) 96 sets of 8 by 8, (b) 64 sets of 12 by 12, or (c) 38 sets of 12 by 12 and 39 sets of 8 by 8 alternating. This figure only shows up to 52 of the 768 haploids (RIL) used in each cross.

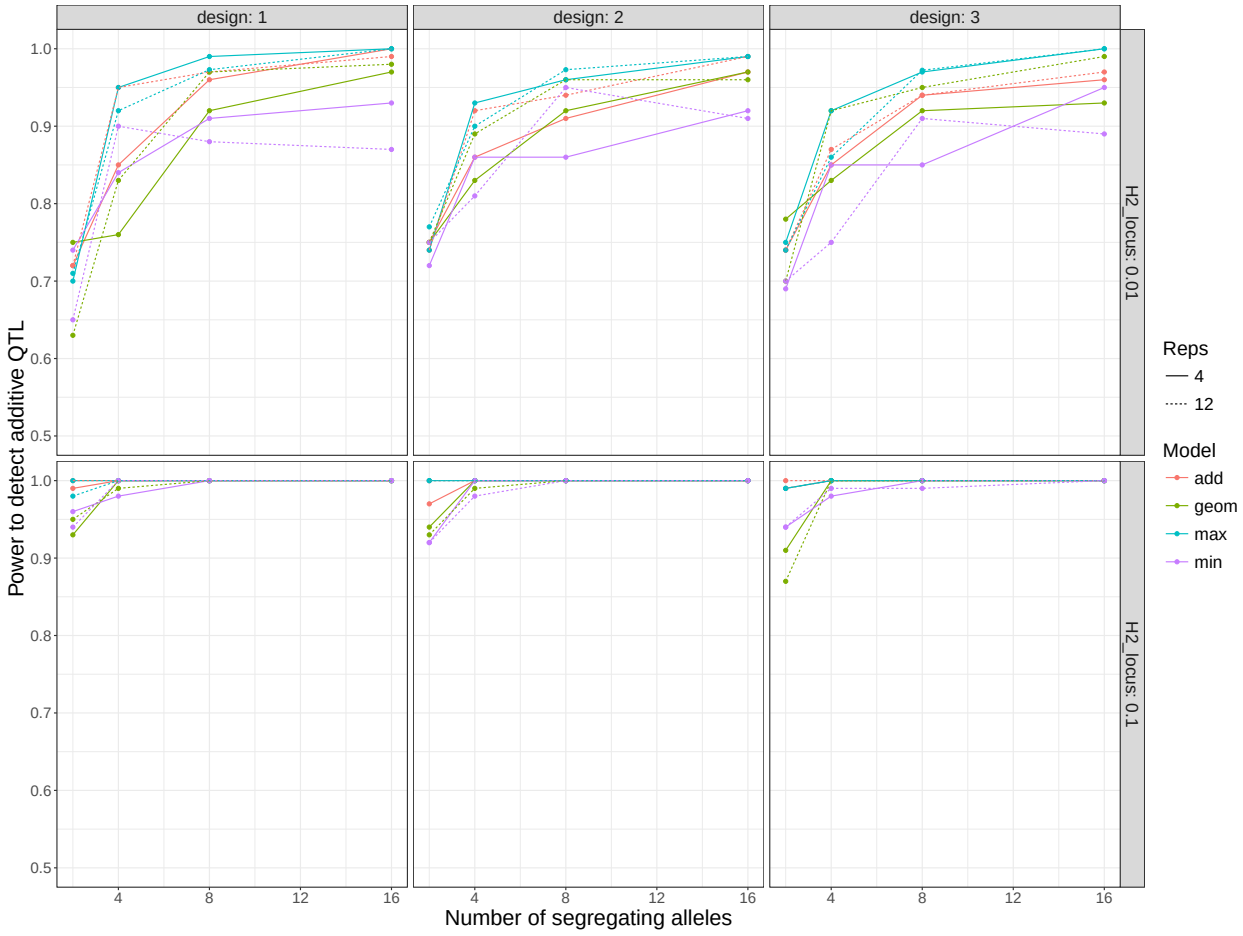

Figure 5.6: Power of LMM detect an additive effect QTL with POL design. These panels show power at $\alpha = \frac{0.05}{1200}$ as function of the number of alleles segregating at a QTL. The horizontal facet shows how the results change as a function of experimental design. The vertical facet shows how the results change as a function of heritability explained by the QTL. The color of the lines represent the genetic model and the line shape corresponds to the number of replicate experiments simulated.

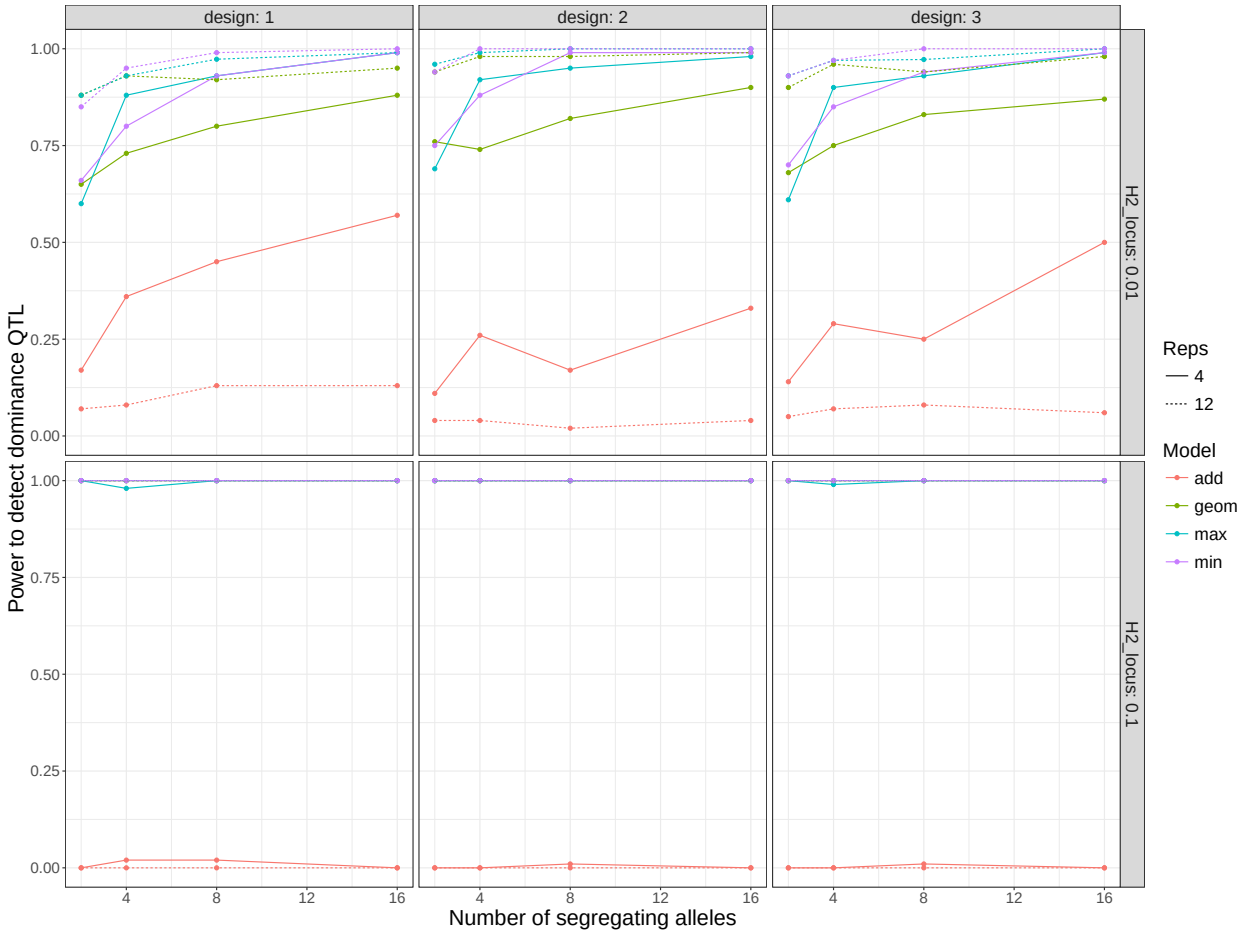

Figure 5.7: Power of LMM to detect a dominance effect QTL with POL design. These panels show power at $\alpha = \frac{0.05}{1200}$ as function of the number of alleles segregating at a QTL. The horizontal facet shows how the results change as a function of experimental design. The vertical facet shows how the results change as a function of heritability explained by the QTL. represent the genetic model and the line shape corresponds to the number of replicate experiments simulated.

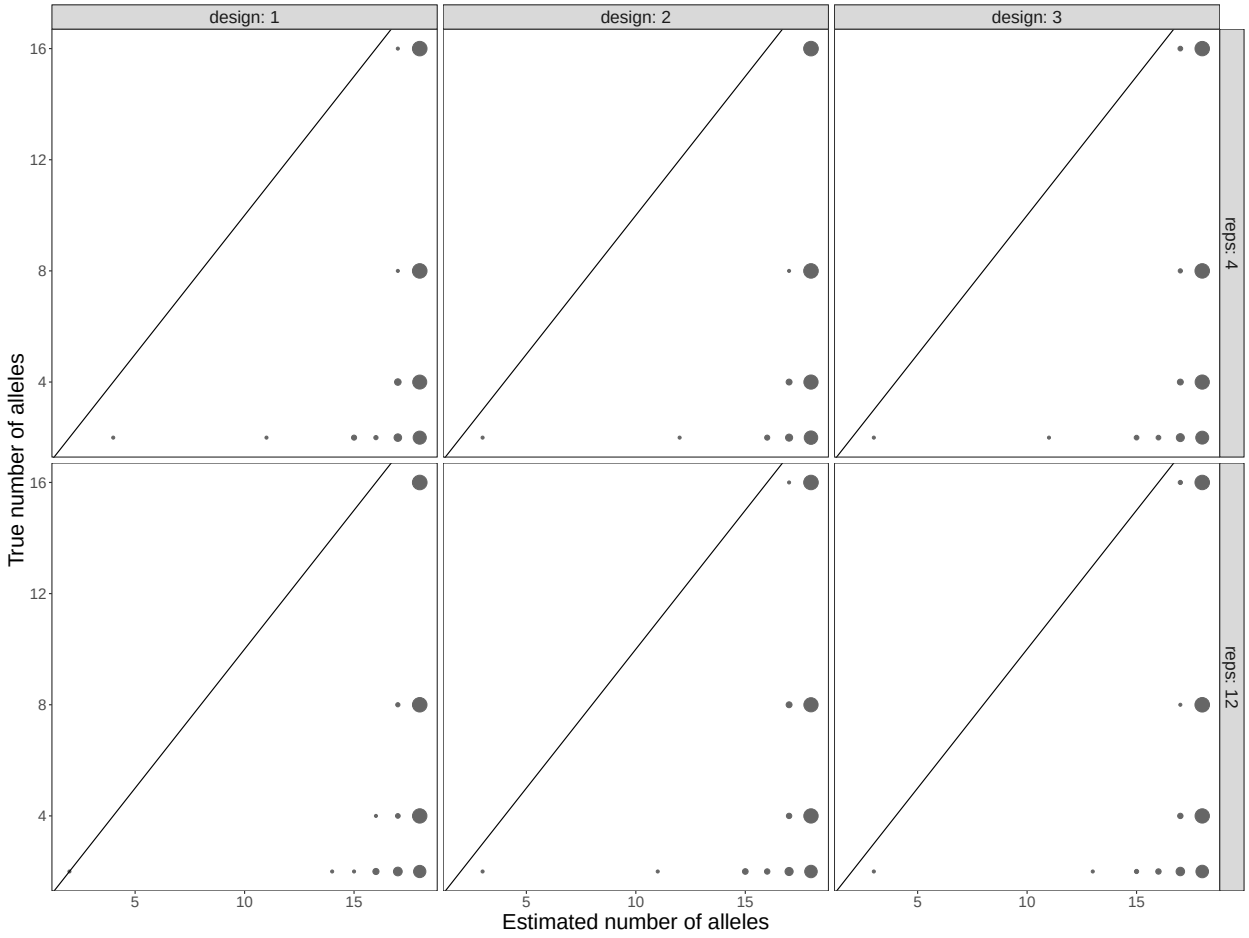

Figure 5.8: Estimated number of effective alleles using LMM-CaSANOVA with POL design. These panels show the relationship between the true number of alleles and the estimated number of alleles at an additive QTL that explains 50% of phenotypic variance. The horizontal facet shows how the results change as a function of experimental design. The vertical facet shows how the results change as a function of experimental replicates. The size of the gray circles correspond to the number of simulation replicates taking on that value. The line $y = x$ is illustrated in black.

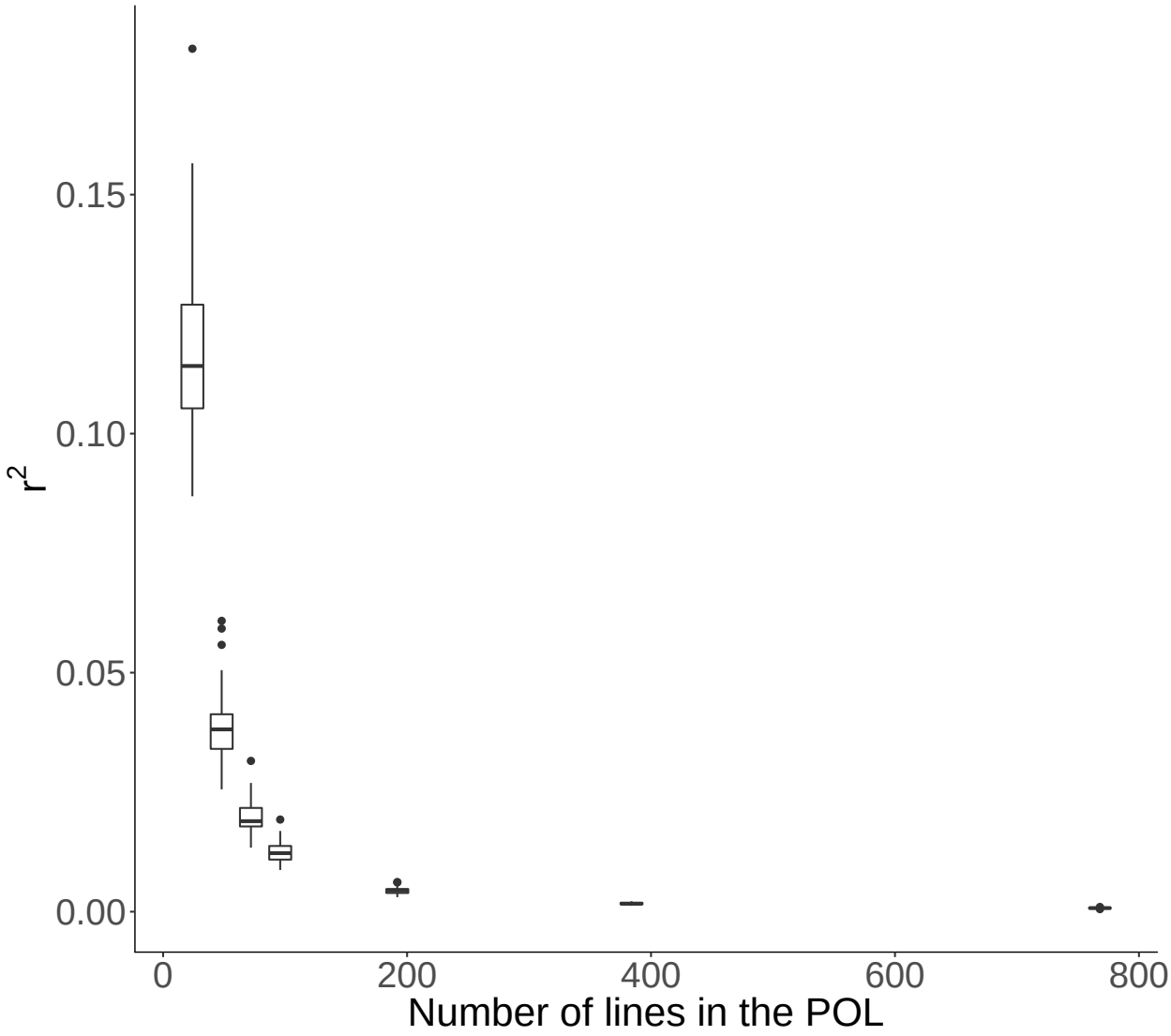

Figure 5.9: Decay of overall LD[271] between neighboring loci in a POL design as a function of number of haploid lines used in the POL cross.

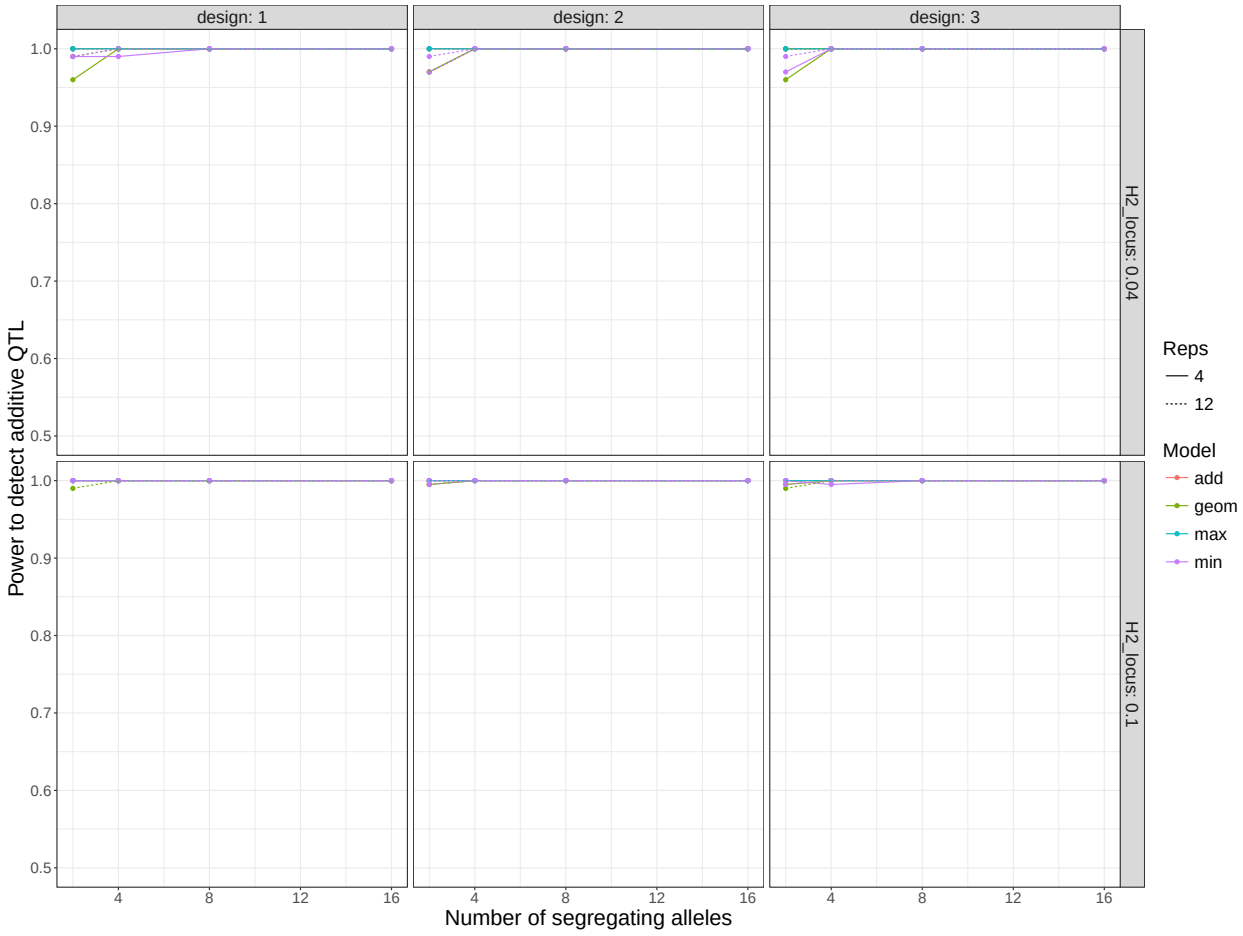

Figure 5.10: Power of linear regression to detect an additive effect QTL with POL design. These panels show power at $\alpha = \frac{0.05}{1200}$ as function of the number of alleles segregating at a QTL. The horizontal facet shows how the results change as a function of experimental design. The vertical facet shows how the results change as a function of heritability explained by the QTL. The color of the lines represent the genetic model and the line shape corresponds to the number of replicate experiments simulated.

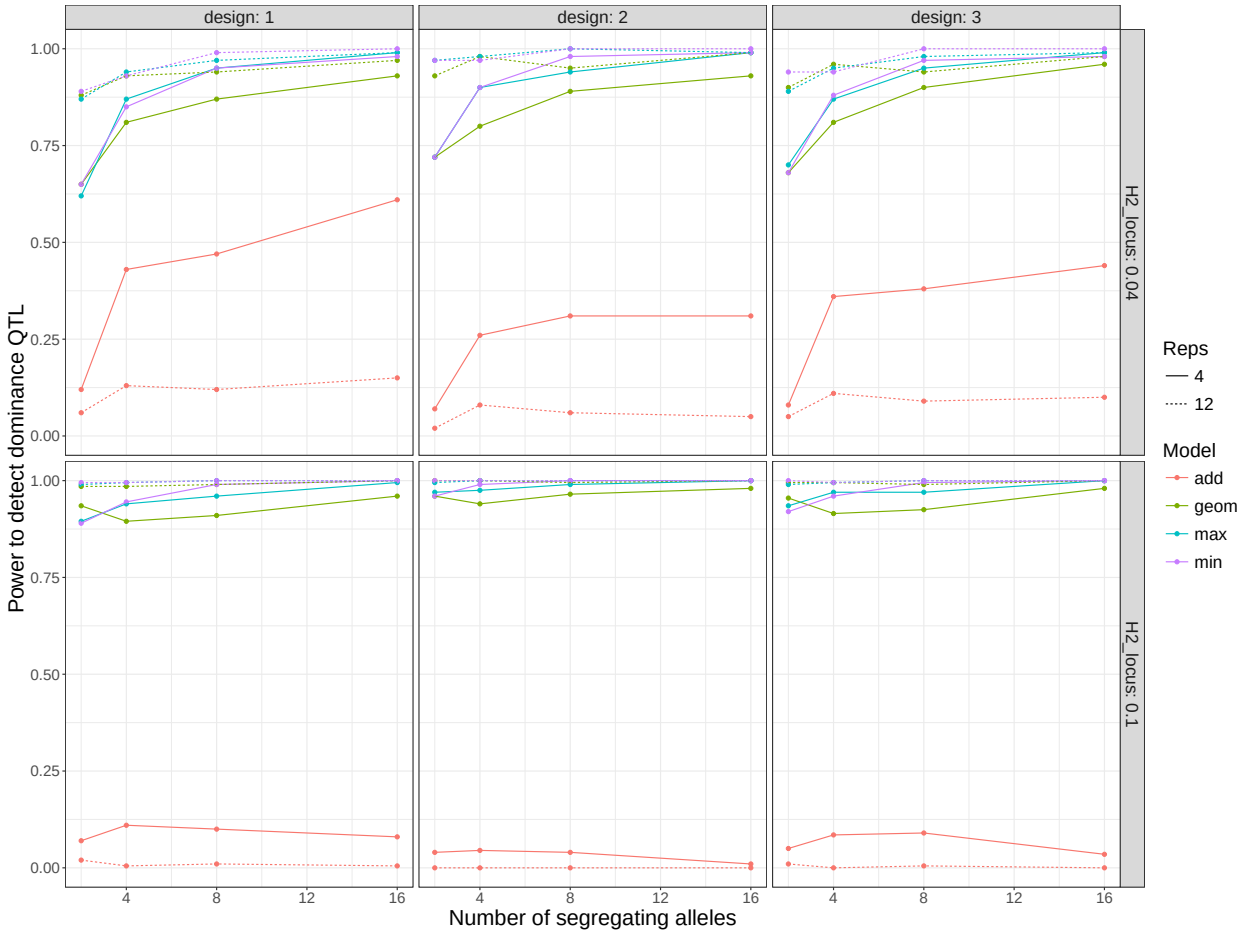

Figure 5.11: Power of linear regression to detect a dominance effect QTL with POL design. These panels show power at $\alpha = \frac{0.05}{1200}$ as function of the number of alleles segregating at a QTL. The horizontal facet shows how the results change as a function of experimental design. The vertical facet shows how the results change as a function of heritability explained by the QTL. The color of the lines represent the genetic model and the line shape corresponds to the number of replicate experiments simulated.

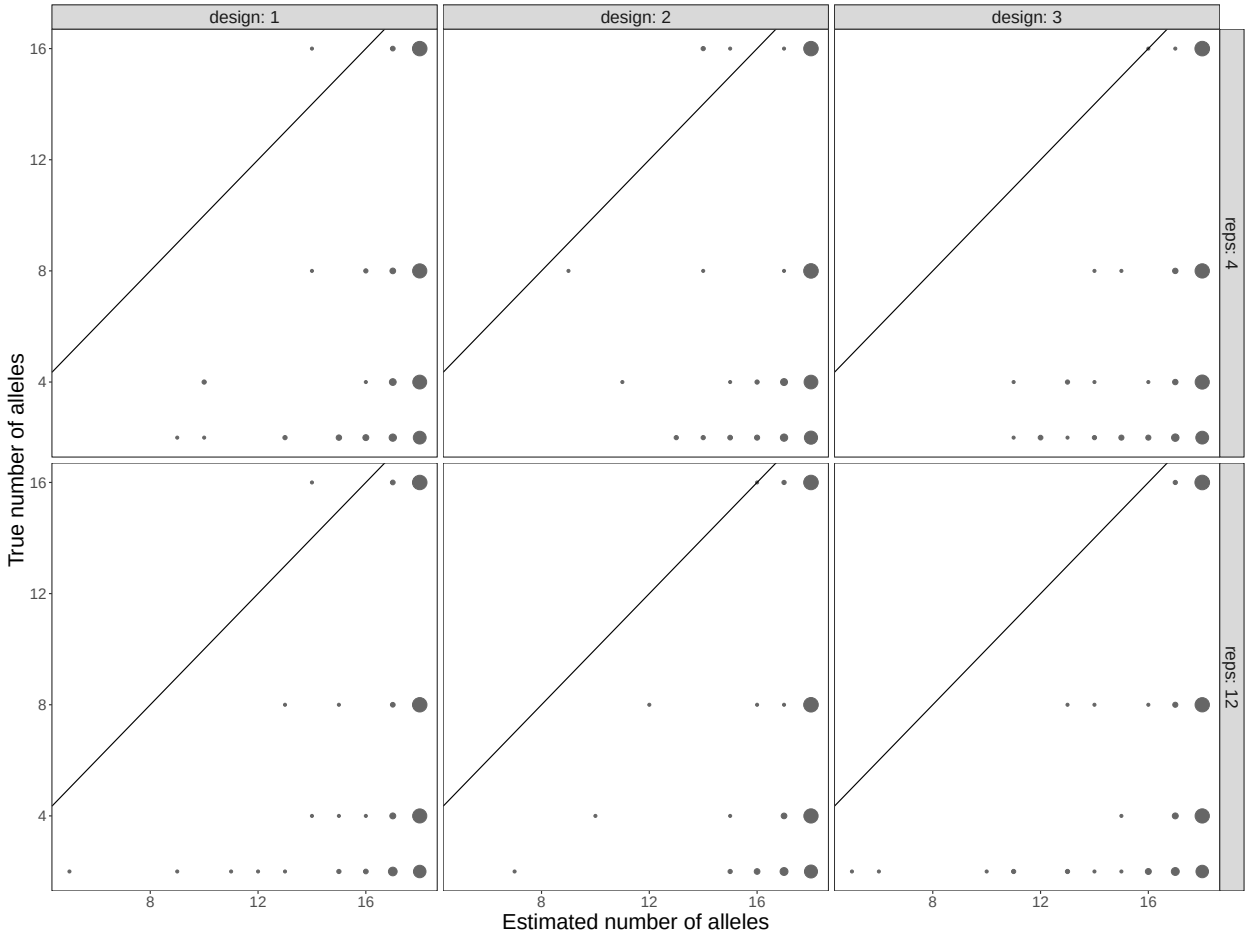

Figure 5.12: Estimated number of effective alleles using CaSANOVA with POL design. These panels show the relationship between the true number of alleles and the estimated number of alleles at an additive QTL that explains 50% of phenotypic variance. The horizontal facet shows how the results change as a function of experimental design. The vertical facet shows how the results change as a function of experimental replicates. The size of the gray circles correspond to the number of simulation replicates taking on that value. The line $y = x$ is illustrated in black.

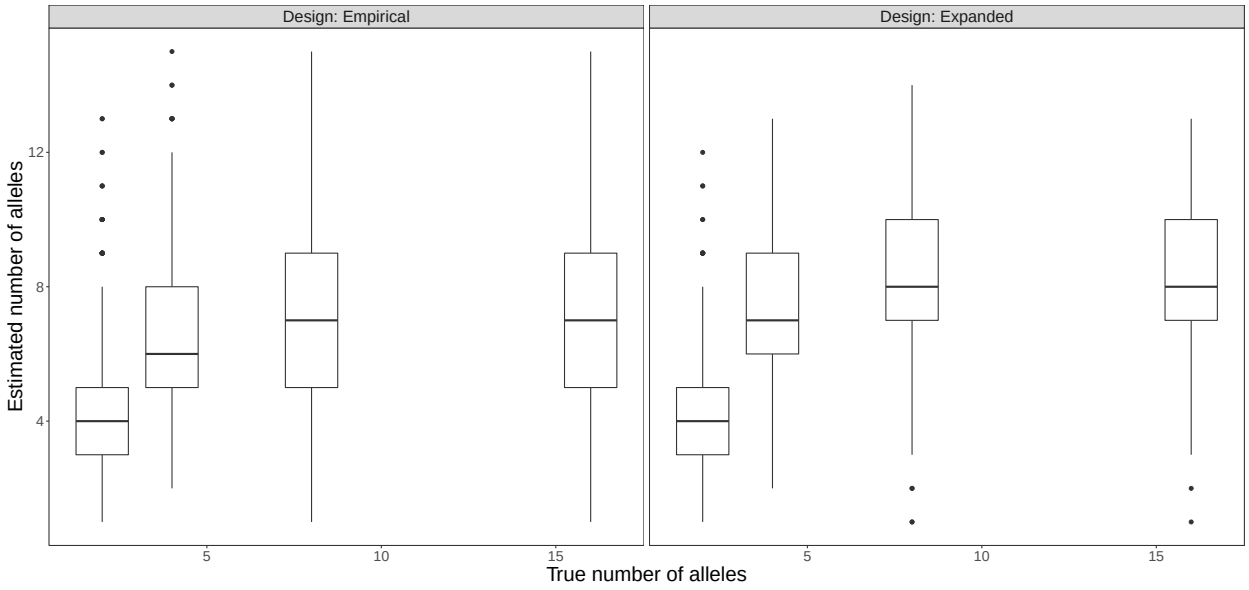

Figure 5.13: Estimated number of effective alleles using CaSANOVA with King et al., 2014 design.

Chapter 6

Conclusion

6.1 Chapter description

Conclusion

6.2 Understanding the evolution of human complex traits and its implications in statistical genetics

Throughout this document I have considered several issues in understanding the genetic architecture of complex traits. First, I performed a detailed simulation study using realistic models of human demography and explored different concepts of dominance. This work contributed to the missing heritability problem by providing a solution to one major weakness of the RALE hypothesis. Specifically I showed that with gene-based dominance it is possible to have a RALE model without the existence of many low-frequency statistical significant markers. I also showed that molecular genetic methods for estimating dominance variance

would fail under gene-based dominance in such a way that is predictable and consistent with empirical results. Using these simulation studies as a goal, Kevin R. Thornton and I were forced to push the boundary on what was possible to simulate explicitly forward in time. The bulk of the software engineering was done by Kevin. R. Thornton and I am tremendously grateful for that.

There are several ways in which this type of simulation study could be expanded upon. First, it would be useful to simulate larger populations in order to maintain pace with the sample sizes being collected today in human genetics. Understanding how our observations change as a function of sample size could be a very powerful statistical signature from which to draw inferences. Another potentially profitable avenue of research includes the expansion to genome-wide simulation. My work to date focused on simulating single 100 kilobase regions, but it might be possible soon to perform multi-locus simulations. Multi-locus simulations would enable a more general discussion of the structure of polygenic genetic architectures and their statistical properties. Lastly, I would encourage the development of multivariate trait simulations, as the study of genetic covariance in human complex traits has expanded considerably in recent years[23, 139, 140].

In the second chapter of this document I described an efficient implementation of the ESM test. I demonstrated empirically that it had more power to detect trait associated genomic regions than existing methods. This test is fundamentally based on order statistics and thus may present a profitable avenue for further development of new statistical tests. There are currently very few order statistic based tests in the genetics literature[52]. The implementation itself could be made more efficient by using an adaptive permutation approach rather than the current brute force method. Further, the software could be made more flexible by allowing for tests of arbitrary sets of markers, as opposed to the current sliding window approach. Testing on arbitrary sets of markers could enable genetic or biochemical pathway based analysis, potentially broadening the scope of the test.

Chapter 3 focused on an empirical analysis of contemporary selection in a population sample from the United Kingdom. This work was important for two major reasons. First it represented the first inference of widespread stabilizing selection in humans. This allowed me to estimate a reasonable range for V_s, which is critical in theoretical models of mutation selection balance on quantitative traits. Second, I demonstrated genetic evidence of direction selection for several traits contributing to growing body of evidence[30, 234, 13, 122] supporting a dynamic model of contemporary human evolution.

Several interesting questions are brought up by the work presented in chapter 3. Are the genetic architectures of specific complex traits commensurate with the relevant parameter estimates obtained in contemporary populations? Answering this question will give us a clue as to how contemporary evolutionary forces might differ from those which have shaped the genetic basis of a complex trait. I argue that we can turn to simulation studies such as those presented in chapter 1 to begin addressing this issue. Another question is whether we can provide genetic evidence of stabilizing selection. Theoretical analyses presented in chapter 3 suggest that doing so will be very difficult statistically and that having 500,000 samples is still not enough. However, finding this genetic evidence is very important it has been shown both in my own work and by[195] that phenotypic evidence alone is not completely reliable. Thus, while the work presented in chapter 3 is very promising, it awaits validation at the genetic level.

In conclusion, we are still a very long way from fully understanding the genetic architecture of complex traits. And perhaps even further from understanding how these architectures came to exist and how they will evolve over time. However, the scientific community seems committed to collecting the type of population scale genome sequencing data necessary to continue advancing in this area. More attention needs to be paid to large scale population sequencing in model organisms, where statistical results can be more easily followed up with experimental manipulations. In fact, both the RALE and infinitesimal models of genetic

architecture will require that we sequence extremely large samples before we can see all the variation. The difference is that under the RALE we expect to see more and more associated variants within the same loci, while the infinitesimal model we expect to find more and more loci. The advances in the scale of data collection will be met with continuous development of theory[216]. There will not be a single unified model for all traits[28], but we can make it our goal to know what the key measurements are and how to properly make them. Then we will be well on our way to understanding the genetics of complex traits.

Bibliography

[1] V. Agarwala, J. Flannick, S. Sunyaev, and D. Altshuler. Evaluating empirical bounds on complex disease genetic architecture. *Nature genetics*, 45(12):1418–27, 2013.

[2] H. Ahsan, J. Halpern, M. G. Kibriya, B. L. Pierce, L. Tong, E. Gamazon, V. McGuire, A. Felberg, J. Shi, F. Jasmine, S. Roy, R. Brutus, M. Argos, S. Melkonian, J. Chang-Claude, I. Andrulis, J. L. Hopper, E. M. John, K. Malone, G. Ursin, M. D. Gammon, D. C. Thomas, D. Seminara, G. Casey, J. A. Knight, M. C. Southey, G. G. Giles, R. M. Santella, E. Lee, D. Conti, D. Duggan, S. Gallinger, R. Haile, M. Jenkins, N. M. Lindor, P. Newcomb, K. Michailidou, C. Apicella, D. J. Park, J. Peto, O. Fletcher, I. dos Santos Silva, M. Lathrop, D. J. Hunter, S. J. Chanock, A. Meindl, R. K. Schmutzler, B. Müller-Myhsok, M. Lochmann, L. Beckmann, R. Hein, E. Makalic, D. F. Schmidt, Q. M. Bui, J. Stone, D. Flesch-Janys, N. Dahmen, H. Nevanlinna, K. Aittomäki, C. Blomqvist, P. Hall, K. Czene, A. Irwanto, J. Liu, N. Rahman, C. Turnbull, A. M. Dunning, P. Pharoah, Q. Waisfisz, H. Meijers-Heijboer, A. G. Uitterlinden, F. Rivadeneira, D. Nicolae, D. F. Easton, N. J. Cox, and A. S. Whittemore. A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 23(4):658–69, apr 2014.

[3] G. Alves and Y.-K. Yu. Accuracy evaluation of the unified P-value from combining correlated P-values. *PloS one*, 9(3):e91225, jan 2014.

[4] H. Arem, K. Yu, X. Xiong, K. Moy, N. D. Freedman, S. T. Mayne, D. Albanes, A. A. Arslan, M. Austin, W. R. Bamlet, L. Beane-Freeman, P. Bracci, F. Canzian, M. Cotterchio, E. J. Duell, S. Gallinger, G. G. Giles, M. Goggins, P. J. Goodman, P. Hartge, M. Hassan, K. Helzlsouer, B. Henderson, E. A. Holly, R. Hoover, E. J. Jacobs, A. Kamineni, A. Klein, E. Klein, L. N. Kolonel, D. Li, N. Malats, S. Männistö, M. L. McCullough, S. H. Olson, I. Orlow, U. Peters, G. M. Petersen, M. Porta, G. Severi, X.-O. Shu, K. Visvanathan, E. White, H. Yu, A. Zeleniuch-Jacquotte, W. Zheng, G. S. Tobias, D. Maeder, M. Brotzman, H. Risch, J. N. Sampson, and R. Z. Stolzenberg-Solomon. Vitamin D metabolic pathway genes and pancreatic cancer risk. *PloS one*, 10(3):e0117574, jan 2015.

[5] S. Aris-Brosou and L. Excoffier. The impact of population expansion and mutation

rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution*, 13(3):494–504, mar 1996.

[6] P. L. Auer, M. Nalls, J. F. Meschia, B. B. Worrall, W. T. Longstreth, S. Seshadri, C. Kooperberg, K. M. Burger, C. S. Carlson, C. L. Carty, W.-M. Chen, L. A. Cupples, A. L. DeStefano, M. Fornage, J. Hardy, L. Hsu, R. D. Jackson, G. P. Jarvik, D. S. Kim, K. Lakshminarayan, L. A. Lange, A. Manichaikul, A. R. Quinlan, A. B. Singleton, T. A. Thornton, D. A. Nickerson, U. Peters, and S. S. Rich. Rare and Coding Region Genetic Variants Associated With Risk of Ischemic Stroke: The NHLBI Exome Sequence Project. *JAMA neurology*, 72(7):781–8, jul 2015.

[7] S. M. Bailey and S. M. Garn. Socioeconomic interactions with physique and fertility. *Human Biology*, 51(3):317–33, sep 1979.

[8] D. J. Balick, R. Do, C. A. Cassa, D. Reich, and S. R. Sunyaev. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. *PLoS genetics*, 11(8):e1005436, aug 2015.

[9] N. H. Barton, M. Turelli, N. H. Barton ", and M. Turelli3. Effects of Genetic Drift on Variance Components under a General Model of Epistasis. *Source: Evolution INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION PUBLISHED BY THE SOCIETY FOR THE STUDY OF EVOLUTION Evolution*, 58(5810):2111–2132, 2004.

[10] N. N. Barton, A. A. Etheridge, A. Veber, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, jul 2017.

[11] W. A. Baseler, D. Thapa, R. Jagannathan, E. R. Dabkowski, T. L. Croston, and J. M. Hollander. miR-141 as a regulator of the mitochondrial phosphate carrier (Slc25a3) in the type 1 diabetic heart. *American journal of physiology. Cell physiology*, 303(12):C1244–51, dec 2012.

[12] S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genetic epidemiology*, 35(7):606–19, nov 2011.

[13] J. P. Beauchamp. Genetic evidence for natural selection in humans in the contemporary United States. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):7774–7779, jul 2016.

[14] M. A. Beaumont. Detecting Population Expansion and Decline Using Microsatellites. *Genetics*, 153(4):2013–2029, dec 1999.

[15] S. Benzer. Fine Structure of a Genetic Region in Bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America*, 41(6):344–354, jun 1955.

[16] E. Bolund, S. Bouwhuis, J. E. Pettay, and V. Lummaa. Divergent selection on, but no genetic conflict over, female and male timing and rate of reproduction in a human population. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1772), 2013.

[17] H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, 66(4):1069–1077, dec 2010.

[18] H. D. Bondell and B. J. Reich. Simultaneous Factor Selection and Collapsing Levels in ANOVA. *Biometrics*, 65(1):169–177, mar 2009.

[19] E. A. Boyle, Y. I. Li, and J. K. Pritchard. No Title. *Cell*, 169(7):1177–1186, jun 2017.

[20] A. V. Brenner, G. Neta, E. M. Sturgis, R. M. Pfeiffer, A. Hutchinson, M. Yeager, L. Xu, C. Zhou, W. Wheeler, M. A. Tucker, S. J. Chanock, and A. J. Sigurdson. Common single nucleotide polymorphisms in genes related to immune function and risk of papillary thyroid cancer. *PloS one*, 8(3):e57243, jan 2013.

[21] S. R. Browning and E. A. Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–31, apr 2012.

[22] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, L. Duncan, J. R. B. Perry, N. Patterson, E. B. Robinson, M. J. Daly, A. L. Price, B. M. Neale, M. J. Daly, A. L. Price, and B. M. Neale. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, sep 2015.

[23] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, and B. M. Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, feb 2015.

[24] J. J. Bull. Evolution of Phenotypic Variance. *Evolution*, 41(2):303, mar 1987.

[25] M. G. Bulmer. The genetic variability of polygenic characters under optimizing selection, mutation and drift. *Genetical Research*, 19(01):17, feb 1972.

[26] M. G. Bulmer. *The mathematical theory of quantitative genetics*. Clarendon, 1985.

[27] R. Burger and J. Hofbauer. Mutation load and mutation-selection-balance in quantitative genetic traits. *Journal of Mathematical Biology*, 32(3):193–218, feb 1994.

[28] R. R. Burger, R. R. Burger, R. R. Burger, and R. R. Burger. *The mathematical theory of selection, recombination, and mutation*. Wiley, 2000.

[29] M. K. Burke, G. Liti, and A. D. Long. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of Saccharomyces cerevisiae. *Molecular biology and evolution*, 31(12):3228–39, dec 2014.

[30] S. G. Byars, D. Ewbank, D. R. Govindaraju, and S. C. Stearns. Colloquium papers: Natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1787–92, jan 2010.

[31] A. Caballero, A. Tenesa, and P. D. Keightley. The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses. *Genetics*, 201(4):1601–13, oct 2015.

[32] M. Chakraborty, R. Zhao, X. Zhang, S. Kalsow, and J. Emerson. Extensive hidden genetic variation shapes the structure of functional elements in Drosophila. *doi.org*, page 114967, mar 2017.

[33] J. M. Chapman, J. D. Cooper, J. A. Todd, and D. G. Clayton. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human heredity*, 56(1-3):18–31, 2003.

[34] B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, aug 1993.

[35] H.-S. Chen, C. M. Hutter, L. E. Mechanic, C. I. Amos, V. Bafna, E. R. Hauser, R. D. Hernandez, C. Li, D. A. Liberles, K. McAllister, J. H. Moore, D. N. Paltoo, G. J. Papanicolaou, B. Peng, M. D. Ritchie, G. Rosenfeld, J. S. Witte, E. M. Gillanders, and E. J. Feuer. Genetic simulation tools for post-genome wide association studies of complex diseases. *Genetic epidemiology*, 39(1):11–9, jan 2015.

[36] L. Chen, Q. Gong, J. P. Stice, and A. A. Knowlton. Mitochondrial OPA1, apoptosis, and heart failure. *Cardiovascular research*, 84(1):91–9, oct 2009.

[37] X. Chen, R. Kuja-Halkola, I. Rahman, J. Arpegård, A. Viktorin, R. Karlsson, S. Hägg, P. Svensson, N. Pedersen, P. Magnusson, A. Tenesa, C. Haley, M. Neale, H. Maes, P. Visscher, S. Medland, M. Ferreira, K. Morley, G. Zhu, B. Cornes, G. Montgomery, N. Martin, J. Yang, S. Lee, M. Goddard, P. Visscher, J. van Dongen, P. Slagboom, H. Draisma, N. Martin, D. Boomsma, B. Maher, J. Yang, B. Benyamin, B. McEvoy, S. Gordon, A. Henders, D. Nyholt, P. Madden, A. Heath, N. Martin, G. Montgomery, et Al., Z. Zhu, A. Bakshi, A. Vinkhuyzen, G. Hemani, S. Lee, I. Nolte, J. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, L. C. Study, et Al., P. Magnusson, C. Almqvist, I. Rahman, A. Ganna, A. Viktorin, H. Walum, L. Halldner, S. Lundström, F. Ullén, N. Långström, et Al., J. DeLeeuw, M. Neale, M. Hunter, J. Pritikin, M. Zahery, T. Brick, R. Kirkpatrick, R. Estabrook, T. Bates, H. Maes, S. Boker, J. Arpegård, A. Viktorin, Z. Chang, U. de Faire, P. Magnusson, P. Svensson, J. van Dongen, G. Willemsen, W. Chen, E. de Geus, D. Boomsma, T. Polderman, B. Benyamin, C. de Leeuw, P. Sullivan, A. van Bochoven, P. Visscher, D. Posthuma, K. Clément, T. Sørensen, P. Magnusson, F. Rasmussen, O. Zuk, S. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. Daly, B. Neale, S. Sunyaev, E. Lander, D. Golan, E. Lander, S. Rosset, O. Zuk, E. Hechter, S. Sunyaev, E. Lander, M. Ritchie, A. Mäki-Tanila, W. Hill, D. Rettew, I. Rebollo-Mesa, J. Hudziak, G. Willemsen, D. Boomsma, M. Keller, W. Coventry, A. Heath, N. Martin, M. Keller, S. Medland, and L. Duncan. Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *The American Journal of Human Genetics*, 97(5):708–714, nov 2015.

[38] E. R. Chimusa, N. Zaitlen, M. Daya, M. Möller, P. D. van Helden, N. J. Mulder, A. L. Price, and E. G. Hoal. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Human molecular genetics*, 23(3):796–809, feb 2014.

[39] E. T. Cirulli and D. B. Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics*, 11(6):415–25, jun 2010.

[40] E. T. Cirulli, B. N. Lasseigne, S. Petrovski, P. C. Sapp, P. A. Dion, C. S. Leblond, J. Couthouis, Y.-F. Lu, Q. Wang, B. J. Krueger, Z. Ren, J. Keebler, Y. Han, S. E. Levy, B. E. Boone, J. R. Wimbish, L. L. Waite, A. L. Jones, J. P. Carulli, A. G. Day-Williams, J. F. Staropoli, W. W. Xin, A. Chesi, A. R. Raphael, D. McKenna-Yasek, J. Cady, J. M. B. Vianney de Jong, K. P. Kenna, B. N. Smith, S. Topp, J. Miller, A. Gkazi, A. Al-Chalabi, L. H. van den Berg, J. Veldink, V. Silani, N. Ticozzi, C. E. Shaw, R. H. Baloh, S. Appel, E. Simpson, C. Lagier-Tourenne, S. M. Pulst, S. Gibson, J. Q. Trojanowski, L. Elman, L. McCluskey, M. Grossman, N. A. Shneider, W. K. Chung, J. M. Ravits, J. D. Glass, K. B. Sims, V. M. Van Deerlin, T. Maniatis, S. D. Hayes, A. Ordureau, S. Swarup, J. Landers, F. Baas, A. S. Allen, R. S. Bedlack, J. W. Harper, A. D. Gitler, G. A. Rouleau, R. Brown, M. B. Harms, G. M. Cooper, T. Harris, R. M. Myers, and D. B. Goldstein. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*, 347(6229):1436–1441, feb 2015.

[41] M. A. Coram, Q. Duan, T. J. Hoffmann, T. Thornton, J. W. Knowles, N. A. Johnson, H. M. Ochs-Balcom, T. A. Donlon, L. W. Martin, C. B. Eaton, J. G. Robinson, N. J. Risch, X. Zhu, C. Kooperberg, Y. Li, A. P. Reiner, and H. Tang. Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *American journal of human genetics*, 92(6):904–16, jun 2013.

[42] A. Coventry, L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell, J. Crosby, J. E. Hixson, T. J. Rea, D. M. Muzny, L. R. Lewis, D. A. Wheeler, A. Sabo, C. Lusk, K. G. Weiss, H. Akbar, A. Cree, A. C. Hawes, I. Newsham, R. T. Varghese, D. Villasana, S. Gross, V. Joshi, J. Santibanez, M. Morgan, K. Chang, W. H. Iv, A. R. Templeton, E. Boerwinkle, R. Gibbs, and C. F. Sing. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications*, 1:131, jan 2010.

[43] J. F. Crow and M. Kimura. The theory of genetic loads. *Proc. XI Int. Congr. Genet.*, pages 495–505, 1964.

[44] J. F. J. F. Crow and M. Kimura. *An introduction to population genetics theory*. Blackburn Press, 2009.

[45] C. Cruchaga, C. M. Karch, S. C. Jin, B. A. Benitez, Y. Cai, R. Guerreiro, O. Harari, J. Norton, J. Budde, S. Bertelsen, A. T. Jeng, B. Cooper, T. Skorupa, D. Carrell,

D. Levitch, S. Hsu, J. Choi, M. Ryten, J. Hardy, D. Trabzuni, M. E. Weale, A. Ramasamy, C. Smith, C. Sassi, J. Bras, J. R. Gibbs, D. G. Hernandez, M. K. Lupton, J. Powell, P. Forabosco, P. G. Ridge, C. D. Corcoran, J. T. Tschanz, M. C. Norton, R. G. Munger, C. Schmutz, M. Leary, F. Y. Demirci, M. N. Bamne, X. Wang, O. L. Lopez, M. Ganguli, C. Medway, J. Turton, J. Lord, A. Braae, I. Barber, K. Brown, P. Passmore, D. Craig, J. Johnston, B. McGuinness, S. Todd, R. Heun, H. Kölsch, P. G. Kehoe, N. M. Hooper, E. R. L. C. Vardy, D. M. Mann, S. Pickering-Brown, N. Kalsheker, J. Lowe, K. Morgan, A. David Smith, G. Wilcock, D. Warden, C. Holmes, P. Pastor, O. Lorenzo-Betancor, Z. Brkanac, E. Scott, E. Topol, E. Rogaeva, A. B. Singleton, M. I. Kamboh, P. St George-Hyslop, N. Cairns, J. C. Morris, J. S. K. Kauwe, and A. M. Goate. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*, 505(7484):550–4, jan 2014.

[46] Y. Da, C. Henderson, W. Fikse, J. Philipsson, R. Powell, P. VanRaden, P. VanRaden, G. Wiggans, I. Misztal, L. Vleck, P. VanRaden, H. Patterson, R. Thompson, Y. Da, C. Wang, S. Wang, G. Hu, C. Wang, Y. Da, B. Hayes, M. Goddard, J. Yang, B. Benyamin, B. McEvoy, S. Gordon, A. Henders, D. Nyholt, R. Fisher, R. Fisher, C. Cockerham, O. Kempthorne, M. Lynch, B. Walsh, O. Kempthorne, D. Falconer, T. Mackay, J. Álvarez-Castro, R.-C. Yang, D. Balding, S. Garnier, V. Truong, J. Brocheton, T. Zeller, M. Rovital, P. Wild, R. Morris, N. Kaplan, J. Barrett, B. Fry, J. Maller, M. Daly, P. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, B. Browning, S. Browning, P. Scheet, M. Stephens, B. Holdt, J. Pollinger, K. Lohmueller, E. Han, H. Parker, P. Quignon, M. Calus, A. Roos, R. Veerkamp, T. Villumsen, L. Janss, M. Lund, B. Cuyabano, G. Su, M. Lund, H. Mulder, M. Calus, R. Veerkamp, T. Meuwissen, M. Goddard, M. Erbe, B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, R. Brøndum, G. Su, M. Lund, P. Bowman, M. Goddard, B. Hayes, C. Henderson, C. Wang, D. Prakapenka, S. Wang, S. Pulugurta, H. Runesha, and Y. Da. Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC Genetics*, 16(1):144, dec 2015.

[47] O. De la Cruz, X. Wen, B. Ke, M. Song, and D. L. Nicolae. Gene, region and pathway level analyses in whole-genome studies. *Genetic epidemiology*, 34(3):222–31, apr 2010.

[48] S. P. Dickson, K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein. Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biology*, 8(1), 2010.

[49] R. Do, D. Balick, H. Li, I. Adzhubei, S. Sunyaev, and D. Reich. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics*, 47(2):126–131, jan 2015.

[50] P. C. A. Dubois, G. Trynka, L. Franke, K. A. Hunt, J. Romanos, A. Curtotti, A. Zhernakova, G. A. R. Heap, R. Adány, A. Aromaa, M. T. Bardella, L. H. van den Berg, N. A. Bockett, E. G. de la Concha, B. Dema, R. S. N. Fehrmann, M. Fernández-Arquero, S. Fiatal, E. Grandone, P. M. Green, H. J. M. Groen, R. Gwilliam, R. H. J. Houwen, S. E. Hunt, K. Kaukinen, D. Kelleher, I. Korponay-Szabo, K. Kurppa, P. MacMathuna,

M. Mäki, M. C. Mazzilli, O. T. McCann, M. L. Mearin, C. A. Mein, M. M. Mirza, V. Mistry, B. Mora, K. I. Morley, C. J. Mulder, J. A. Murray, C. Núñez, E. Oosterom, R. A. Ophoff, I. Polanco, L. Peltonen, M. Platteel, A. Rybak, V. Salomaa, J. J. Schweizer, M. P. Sperandeo, G. J. Tack, G. Turner, J. H. Veldink, W. H. M. Verbeek, R. K. Weersma, V. M. Wolters, E. Urcelay, B. Cukrowska, L. Greco, S. L. Neuhausen, R. McManus, D. Barisani, P. Deloukas, J. C. Barrett, P. Saavalainen, C. Wijmenga, and D. A. van Heel. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295–302, apr 2010.

[51] F. Dudbridge and B. P. C. Koeleman. Rank truncated product of P-values, with application to genomewide association scans. *Genetic epidemiology*, 25(4):360–6, dec 2003.

[52] F. Dudbridge and B. P. C. Koeleman. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American journal of human genetics*, 75(3):424–35, sep 2004.

[53] J. Dupuis, C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo, A. U. Jackson, E. Wheeler, N. L. Glazer, N. Bouatia-Naji, A. L. Gloyn, C. M. Lindgren, R. Mägi, A. P. Morris, J. Randall, T. Johnson, P. Elliott, D. Rybin, G. Thorleifsson, V. Steinthorsdottir, P. Henneman, H. Grallert, A. Dehghan, J. J. Hottenga, C. S. Franklin, P. Navarro, K. Song, A. Goel, J. R. B. Perry, J. M. Egan, T. Lajunen, N. Grarup, T. Sparsø, A. Doney, B. F. Voight, H. M. Stringham, M. Li, S. Kanoni, P. Shrader, C. Cavalcanti-Proença, M. Kumari, L. Qi, N. J. Timpson, C. Gieger, C. Zabena, G. Rocheleau, E. Ingelsson, P. An, J. O'Connell, J. Luan, A. Elliott, S. A. McCarroll, F. Payne, R. M. Roccasecca, F. Pattou, P. Sethupathy, K. Ardlie, Y. Ariyurek, B. Balkau, P. Barter, J. P. Beilby, Y. Ben-Shlomo, R. Benediktsson, A. J. Bennett, S. Bergmann, M. Bochud, E. Boerwinkle, A. Bonnefond, L. L. Bonnycastle, K. Borch-Johnsen, Y. Böttcher, E. Brunner, S. J. Bumpstead, G. Charpentier, Y.-D. I. Chen, P. Chines, R. Clarke, L. J. M. Coin, M. N. Cooper, M. Cornelis, G. Crawford, L. Crisponi, I. N. M. Day, E. J. C. de Geus, J. Delplanque, C. Dina, M. R. Erdos, A. C. Fedson, A. Fischer-Rosinsky, N. G. Forouhi, C. S. Fox, R. Frants, M. G. Franzosi, P. Galan, M. O. Goodarzi, J. Graessler, C. J. Groves, S. Grundy, R. Gwilliam, U. Gyllensten, S. Hadjadj, G. Hallmans, N. Hammond, X. Han, A.-L. Hartikainen, N. Hassanali, C. Hayward, S. C. Heath, S. Hercberg, C. Herder, A. A. Hicks, D. R. Hillman, A. D. Hingorani, A. Hofman, J. Hui, J. Hung, B. Isomaa, P. R. V. Johnson, T. Jørgensen, A. Jula, M. Kaakinen, J. Kaprio, Y. A. Kesaniemi, M. Kivimaki, B. Knight, S. Koskinen, P. Kovacs, K. O. Kyvik, G. M. Lathrop, D. A. Lawlor, O. Le Bacquer, C. Lecoeur, Y. Li, V. Lyssenko, R. Mahley, M. Mangino, A. K. Manning, M. T. Martínez-Larrad, J. B. McAteer, L. J. McCulloch, R. McPherson, C. Meisinger, D. Melzer, D. Meyre, B. D. Mitchell, M. A. Morken, S. Mukherjee, S. Naitza, N. Narisu, M. J. Neville, B. A. Oostra, M. Orrù, R. Pakyz, C. N. A. Palmer, G. Paolisso, C. Pattaro, D. Pearson, J. F. Peden, N. L. Pedersen, M. Perola, A. F. H. Pfeiffer, I. Pichler, O. Polasek, D. Posthuma, S. C. Potter, A. Pouta, M. A. Province, B. M. Psaty, W. Rathmann, N. W. Rayner, K. Rice, S. Ripatti, F. Rivadeneira, M. Roden, O. Rolandsson, A. Sandbaek, M. Sandhu, S. Sanna, A. A. Sayer, P. Scheet, L. J. Scott, U. Seedorf, S. J. Sharp,

B. Shields, G. Sigurethsson, E. J. G. Sijbrands, A. Silveira, L. Simpson, A. Singleton, N. L. Smith, U. Sovio, A. Swift, H. Syddall, A.-C. Syvänen, T. Tanaka, B. Thorand, J. Tichet, A. Tönjes, T. Tuomi, A. G. Uitterlinden, K. W. van Dijk, M. van Hoek, D. Varma, S. Visvikis-Siest, V. Vitart, N. Vogelzangs, G. Waeber, P. J. Wagner, A. Walley, G. B. Walters, K. L. Ward, H. Watkins, M. N. Weedon, S. H. Wild, G. Willemsen, J. C. M. Witteman, J. W. G. Yarnell, E. Zeggini, D. Zelenika, B. Zethelius, G. Zhai, J. H. Zhao, M. C. Zillikens, I. B. Borecki, R. J. F. Loos, P. Meneton, P. K. E. Magnusson, D. M. Nathan, G. H. Williams, A. T. Hattersley, K. Silander, V. Salomaa, G. D. Smith, S. R. Bornstein, P. Schwarz, J. Spranger, F. Karpe, A. R. Shuldiner, C. Cooper, G. V. Dedoussis, M. Serrano-Ríos, A. D. Morris, L. Lind, L. J. Palmer, F. B. Hu, P. W. Franks, S. Ebrahim, M. Marmot, W. H. L. Kao, J. S. Pankow, M. J. Sampson, J. Kuusisto, M. Laakso, T. Hansen, O. Pedersen, P. P. Pramstaller, H. E. Wichmann, T. Illig, I. Rudan, A. F. Wright, M. Stumvoll, H. Campbell, J. F. Wilson, R. N. Bergman, T. A. Buchanan, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. T. Valle, D. Altshuler, J. I. Rotter, D. S. Siscovick, B. W. J. H. Penninx, D. I. Boomsma, P. Deloukas, T. D. Spector, T. M. Frayling, L. Ferrucci, A. Kong, U. Thorsteinsdottir, K. Stefansson, C. M. van Duijn, Y. S. Aulchenko, A. Cao, A. Scuteri, D. Schlessinger, M. Uda, A. Ruokonen, M.-R. Jarvelin, D. M. Waterworth, P. Vollenweider, L. Peltonen, V. Mooser, G. R. Abecasis, N. J. Wareham, R. Sladek, P. Froguel, R. M. Watanabe, J. B. Meigs, L. Groop, M. Boehnke, M. I. McCarthy, J. C. Florez, and I. Barroso. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–16, feb 2010.

[54] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*, 21(16):3439–40, aug 2005.

[55] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 4(8):1184–91, jan 2009.

[56] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11(6):446–50, jun 2010.

[57] H. Eleftherohorinou, C. J. Hoggart, V. J. Wright, M. Levin, and L. J. M. Coin. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Human Molecular Genetics*, 20(17):3494–3506, jun 2011.

[58] A. Erbilgin, M. Civelek, C. E. Romanoski, C. Pan, R. Hagopian, J. A. Berliner, and A. J. Lusis. Identification of CAD candidate genes in GWAS loci and their expression in vascular cells. *Journal of lipid research*, 54(7):1894–905, jul 2013.

[59] A. Eyre-Walker. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association

studies. *Proceedings of the National Academy of Sciences of the United States of America*, 107 Suppl:1752–1756, 2010.

[60] D. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics, Ed. 2.* Longmans Green, Harlow, Essex, UK, 1996.

[61] L. Feiner, A. L. Webber, C. B. Brown, M. M. Lu, L. Jia, P. Feinstein, P. Mombaerts, J. A. Epstein, and J. A. Raper. Targeted disruption of semaphorin 3C leads to persistent truncus arteriosus and aortic arch interruption. *Development*, 128(16):3061–3070, aug 2001.

[62] M. Fieder and S. Huber. The effects of sex and childlessness on the association between status and reproductive output in modern society. *Evolution and Human Behavior*, 28(6):392–398, nov 2007.

[63] R. A. Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1918.

[64] R. A. Fisher. *The Genetical Theory Of Natural Selection : Fisher, R. A : Free Download & Streaming : Internet Archive*. Oxford Univ. Press, Oxford, UK, 1930.

[65] W. H. Fleming. Equilibrium Distributions of Continuous Polygenic Traits. *SIAM Journal on Applied Mathematics*, jul 1979.

[66] A. Franke, D. P. B. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, C. A. Anderson, J. C. Bis, S. Bumpstead, D. Ellinghaus, E. M. Festen, M. Georges, T. Green, T. Haritunians, L. Jostins, A. Latiano, C. G. Mathew, G. W. Montgomery, N. J. Prescott, S. Raychaudhuri, J. I. Rotter, P. Schumm, Y. Sharma, L. A. Simms, K. D. Taylor, D. Whiteman, C. Wijmenga, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Büning, A. Cohen, J.-F. Colombel, M. Cottone, L. Stronati, T. Denson, M. De Vos, R. D'Inca, M. Dubinsky, C. Edwards, T. Florin, D. Franchimont, R. Gearry, J. Glas, A. Van Gossum, S. L. Guthery, J. Halfvarson, H. W. Verspaget, J.-P. Hugot, A. Karban, D. Laukens, I. Lawrance, M. Lemann, A. Levine, C. Libioulle, E. Louis, C. Mowat, W. Newman, J. Panés, A. Phillips, D. D. Proctor, M. Regueiro, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, F. Seibold, A. H. Steinhart, P. C. F. Stokkers, L. Torkvist, G. Kullak-Ublick, D. Wilson, T. Walters, S. R. Targan, S. R. Brant, J. D. Rioux, M. D'Amato, R. K. Weersma, S. Kugathasan, A. M. Griffiths, J. C. Mansfield, S. Vermeire, R. H. Duerr, M. S. Silverberg, J. Satsangi, S. Schreiber, J. H. Cho, V. Annese, H. Hakonarson, M. J. Daly, and M. Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–25, dec 2010.

[67] K. Fransen, M. C. Visschedijk, S. van Sommeren, J. Y. Fu, L. Franke, E. A. M. Festen, P. C. F. Stokkers, A. A. van Bodegraven, J. B. A. Crusius, D. W. Hommes, P. Zanen, D. J. de Jong, C. Wijmenga, C. C. van Diemen, and R. K. Weersma. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. *Human molecular genetics*, 19(17):3482–8, sep 2010.

[68] A. C. Frazier-Wood, S. Aslibekyan, I. B. Borecki, P. N. Hopkins, C.-Q. Lai, J. M. Ordovas, R. J. Straka, H. K. Tiwari, and D. K. Arnett. Genome-wide association study indicates variants associated with insulin signaling and inflammation mediate lipoprotein responses to fenofibrate. *Pharmacogenetics and genomics*, 22(10):750–7, oct 2012.

[69] R. M. Freathy, N. J. Timpson, D. A. Lawlor, A. Pouta, Y. Ben-Shlomo, A. Ruokonen, S. Ebrahim, B. Shields, E. Zeggini, M. N. Weedon, C. M. Lindgren, H. Lango, D. Melzer, L. Ferrucci, G. Paolisso, M. J. Neville, F. Karpe, C. N. A. Palmer, A. D. Morris, P. Elliott, M.-R. Jarvelin, G. D. Smith, M. I. McCarthy, A. T. Hattersley, and T. M. Frayling. Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes*, 57(5):1419–26, may 2008.

[70] A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, and N. E. Allen. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*, pages 1–9, aug 2017.

[71] W. Fu, R. Gittelman, M. Bamshad, and J. Akey. Characteristics of Neutral and Deleterious Protein-Coding Variation among Individuals and Populations. *The American Journal of Human Genetics*, 95(4):421–436, 2014.

[72] Y. X. Fu. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *Genetics*, 147(2):915–925, oct 1997.

[73] F. Gao and A. Keinan. High burden of private mutations due to explosive human population growth and purifying selection. *BMC genomics*, 15 Suppl 4(Suppl 4):S3, 2014.

[74] E. Gazave, D. Chang, A. G. Clark, and A. Keinan. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics*, 195(3):969–978, 2013.

[75] E. Gazave, L. Ma, D. Chang, A. Coventry, F. Gao, D. Muzny, E. Boerwinkle, R. a. Gibbs, C. F. Sing, A. G. Clark, and A. Keinan. Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences of the United States of America*, 111(2):757–62, 2014.

[76] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, P. K.-H. Tam, L.-C. Tsui, M. M. Y. Waye, J. T.-F. Wong, C. Zeng, Q. Zhang, M. S. Chee, L. M. Galver, S. Kruglyak, S. S. Murray, A. R. Oliphant, A. Montpetit, T. J. Hudson, F. Chagnon, V. Ferretti, M. Leboeuf, M. S. Phillips, A. Verner, P.-Y. Kwok, S. Duan, D. L. Lind, R. D. Miller, J. P. Rice, N. L. Saccone, P. Taillon-Miller, M. Xiao, Y. Nakamura, A. Sekine, K. Sorimachi, T. Tanaka, Y. Tanaka, T. Tsunoda, E. Yoshino, D. R. Bentley, P. Deloukas, S. Hunt, D. Powell, D. Altshuler, S. B. Gabriel, H. Zhang, C. Zeng, I. Matsuda, Y. Fukushima, D. R. Macer,

E. Suda, C. N. Rotimi, C. A. Adebamowo, T. Aniagwu, P. A. Marshall, O. Matthew, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, L. D. Stein, F. Cunningham, A. Kanani, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, P. Donnelly, J. Marchini, G. A. T. McVean, S. R. Myers, L. R. Cardon, G. R. Abecasis, A. Morris, B. S. Weir, J. C. Mullikin, S. T. Sherry, M. Feolo, D. Altshuler, M. J. Daly, S. F. Schaffner, R. Qiu, A. Kent, G. M. Dunston, K. Kato, N. Niikawa, B. M. Knoppers, M. W. Foster, E. W. Clayton, V. O. Wang, J. Watkin, R. A. Gibbs, J. W. Belmont, E. Sodergren, G. M. Weinstock, R. K. Wilson, L. L. Fulton, J. Rogers, B. W. Birren, H. Han, H. Wang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Todani, T. Fujita, S. Tanaka, A. L. Holden, E. H. Lai, F. S. Collins, L. D. Brooks, J. E. McEwen, M. S. Guyer, E. Jordan, J. L. Peterson, J. Spiegel, L. M. Sung, L. F. Zacharia, K. Kennedy, M. G. Dunn, R. Seabrook, M. Shillito, B. Skene, J. G. Stewart, D. L. Valle (chair), E. W. Clayton (co-chair), L. B. Jorde (co-chair), J. W. Belmont, A. Chakravarti, M. K. Cho, T. Duster, M. W. Foster, M. Jasperse, B. M. Knoppers, P.-Y. Kwok, J. Licinio, J. C. Long, P. A. Marshall, P. N. Ossorio, V. O. Wang, C. N. Rotimi, C. D. M. Royal, P. Spallone, S. F. Terry, E. S. Lander (chair), E. H. Lai (co-chair), D. A. Nickerson (co-chair), G. R. Abecasis, D. Altshuler, D. R. Bentley, M. Boehnke, L. R. Cardon, M. J. Daly, P. Deloukas, J. A. Douglas, S. B. Gabriel, R. R. Hudson, T. J. Hudson, L. Kruglyak, P.-Y. Kwok, Y. Nakamura, R. L. Nussbaum, C. D. M. Royal, S. F. Schaffner, S. T. Sherry, L. D. Stein, and T. Tanaka. The International HapMap Project. *Nature*, 426(6968):789–796, dec 2003.

[77] G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.

[78] D. Golan, E. S. Lander, and S. Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[79] S. Gravel. When Is Selection Effective? *Genetics*, 203(1):451–62, may 2016.

[80] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, C. D. Bustamante, D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, D. B. Jaffe, E. Shefler, C. L. Sougnez, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani,

D. Riches, W. Song, C. Turcotte, S. Wang, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, M. Bainbridge, D. Challis, A. Sabo, J. Yu, X. Fang, X. Guo, Y. Li, R. Luo, S. Tai, H. Wu, X. Zheng, Y. Zhou, E. P. Garrison, W. Huang, A. R. Indap, D. Kural, W.-P. Lee, W. F. Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, A. D. Ball, E. Banks, B. L. Browning, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, A. Boyko, J. Degenhardt, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stutz, M. Bauer, R. K. Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, R. Agarwala, H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. V. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. C. Melton, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. M. Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. Snyder, X. Zhan, S. Zollner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. C. Sahinalp, P. H. Sudmant, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, J. W. Wallis, M. C. Wendl, Q. Zhang, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Q. Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, C. Coafra, H. Dinh, C. Kovar, S. Lee, L. Nazareth, J. Wilkinson, C. Scott, N. Gharani, J. S. Kaye, A. Kent, T. Li, A. L. McGuire, P. N. Ossorio, C. N. Rotimi, Y. Su, L. H. Toji, C. TylerSmith, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, A. Abdallah, C. R. Juenger, N. C. Clemm, A. Duncanson, E. D. Green, M. S. Guyer, J. L. Peterson, 1000 Genomes Project, C. D. Bustamante, D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. Li, M. Jian, G. Li,

R. Li, H. Liang, G. Tian, B. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, D. B. Jaffe, E. Shefler, C. L. Sougnez, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, M. Bainbridge, D. Challis, A. Sabo, J. Yu, X. Fang, X. Guo, Y. Li, R. Luo, S. Tai, H. Wu, X. Zheng, Y. Zhou, E. P. Garrison, W. Huang, A. R. Indap, D. Kural, W.-P. Lee, W. F. Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, A. D. Ball, E. Banks, B. L. Browning, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, A. Boyko, J. Degenhardt, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stutz, M. Bauer, R. K. Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, R. Agarwala, H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. V. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. C. Melton, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. M. Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. Snyder, X. Zhan, S. Zollner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. C. Sahinalp, P. H. Sudmant, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, J. W. Wallis, M. C. Wendl, Q. Zhang, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Q. Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, C. Coafra,

H. Dinh, C. Kovar, S. Lee, L. Nazareth, J. Wilkinson, C. Scott, N. Gharani, J. S. Kaye, A. Kent, T. Li, A. L. McGuire, P. N. Ossorio, C. N. Rotimi, Y. Su, L. H. Toji, C. TylerSmith, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, A. Abdallah, C. R. Juenger, N. C. Clemm, A. Duncanson, E. D. Green, M. S. Guyer, J. L. Peterson, 1000 Genomes Project, and C. D. Bustamante. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, jul 2011.

[81] R. C. Griffiths and S. Tavare. Ancestral Inference in Population Genetics. *Statistical Science*, 9(3):307–319, aug 1994.

[82] R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 344(1310):403–10, jun 1994.

[83] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, 5(10):e1000695, oct 2009.

[84] J. B. S. Haldane. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(07):838—-844, 1927.

[85] J. B. S. Haldane. The measurement of natural selection. *Proc. 9th Int. Congr. Genet.*, 1:480–487, 1954.

[86] J. B. S. Haldane. The cost of natural selection, 1957.

[87] J. Hallin, K. Märtens, A. I. Young, M. Zackrisson, F. Salinas, L. Parts, J. Warringer, G. Liti, P. M. Visscher, M. A. Brown, M. I. McCarthy, J. Yang, E. E. Eichler, J. Yang, B. Lehner, O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, M. Abney, M. S. McPeek, C. Ober, B. Lehner, G. Liti, J. Schacherer, J. S. Bloom, J. S. Bloom, I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, L. Kruglyak, A. I. Young, R. Durbin, D. W. Threadgill, K. W. Hunter, R. W. Williams, S.-W. Tsaih, L. Lu, D. C. Airey, R. W. Williams, G. A. Churchill, F. Zou, L. Parts, F. A. Cubillos, C. J. R. Illingworth, L. Parts, A. Bergström, G. Liti, V. Mustonen, K. Märtens, J. Hallin, J. Warringer, G. Liti, L. Parts, M. Zackrisson, J. Warringer, S. Ibstedt, J. Warringer, D. Anevski, B. Liu, A. Blomberg, P. K. Joshi, P. M. Magwene, Q.-M. Wang, W.-Q. Liu, G. Liti, S.-A. Wang, F.-Y. Bai, A. Bergström, F. A. Cubillos, M. Plech, J. A. G. M. de Visser, R. Korona, E. Zörgö, R. Shapira, T. Levy, S. Shaked, E. Fridman, L. David, I. M. Ehrenreich, K. Lorenz, B. A. Cohen, E. Zörgö, A. C. Gerstein, J. Berman, A. C. Gerstein, S. P. Otto, M. J. P. Chaisson, R. K. Wilson, E. E. Eichler, E. O. Perlstein, D. M. Ruderfer, D. C. Roberts, S. L. Schreiber, L. Kruglyak, M. Mülleder, J. Gerke, K. Lorenz, B. Cohen, M. B. Taylor, I. M. Ehrenreich, M. Nordborg, D. Weigel, T. F. C. Mackay, A. E. Melchinger, J. Tang, J. Hua, R. Mott, G. Liti, J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, H. Wickham, and H. Wickham. Powerful decomposition of complex traits in a diploid model. *Nature Communications*, 7:13311, nov 2016.

[88] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1):3–19, mar 1972.

[89] P. C. Havugimana, G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V.-u.-N. Dar, A. Bezginov, G. W. Clark, G. C. Wu, S. J. Wodak, E. R. M. Tillier, A. Paccanaro, E. M. Marcotte, and A. Emili. A census of human soluble protein complexes. *Cell*, 150(5):1068–81, aug 2012.

[90] S. Helle, N. Allal, R. Sear, A. Prentice, R. Mace, O. Basso, E. Nohr, K. Christensen, J. Olson, J. Brommer, L. Gustafsson, H. Pietiäinen, J. Merilä, G. Brush, A. Boyce, G. Harrison, R. Bukowski, G. G. Smith, F. Malone, R. Ball, D. Nyberg, C. Comstock, f. t. F. R. Consortium, et Al., K. Burnham, D. Anderson, M. Devi, J. Kumari, C. Srikumari, P. Ellison, A. Frisancho, J. Sanchez, D. Pallardel, L. Yanez, E. Georgiadis, D. Gigante, B. Horta, R. Lima, F. Barros, C. Victora, D. Gigante, K. Rasmussen, C. Victora, S. Helle, V. Lummaa, J. Jokela, P. Hindmarsh, M. Geary, C. Rodeck, J. Kingdom, T. Cole, L. Hurt, C. Ronsmans, S. Thomas, B. Jacobsen, I. Heuch, G. Kvåle, G. Jasienska, A. Kemkes-Grottenthaler, E. Ketterson, V. Nolan, K. Kirk, S. Blomberg, D. Duffy, A. Heath, I. Owens, N. Martin, E. L. Bourg, J. Liljestrand, S. Bergström, S. Westman, R. Loos, C. Derom, R. Eeckels, R. Derom, R. Vlietinck, V. Lummaa, K. Marsál, P. Persson, T. Larsen, H. Lilja, A. Selbing, B. Sultan, R. Martorell, H. Delgado, V. Valverde, R. Klein, W. Mueller, A. Must, S. Phillips, E. Naumova, M. Blum, S. Harris, B. Dawson-Hughes, et Al., D. Nettle, M. Okasha, P. McCarron, J. McEwen, G. G. Smith, N. Onland-Moret, P. Peeters, C. van Gils, F. Clavel-Chapelon, T. Key, A. Tjønneland, et Al., G. Parker, N. Royle, I. Hartley, J. Pettay, S. Helle, J. Jokela, V. Lummaa, M. Ritamies, D. Roff, D. Roff, D. Fairbairn, T. Scholl, M. Hediger, R. Sear, R. Sear, N. Allal, R. Mace, I. McGregor, R. Sear, R. Mace, I. McGregor, K. Silventoinen, A. Vetta, and T. Williams. A tradeoff between reproduction and growth in contemporary Finnish women. *Evolution and Human Behavior*, 29(3):189–195, may 2008.

[91] S. Helle, V. Lummaa, and J. Jokela. Are reproductive and somatic senescence coupled in humans? Late, but not early, reproduction correlated with longevity in historical Sami women. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1558), 2005.

[92] B. M. Henn, L. R. Botigué, S. Peischl, I. Dupanloup, M. Lipatov, B. K. Maples, A. R. Martin, S. Musharoff, H. Cann, M. P. Snyder, L. Excoffier, J. M. Kidd, and C. D. Bustamante. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences*, 113(4):E440–E449, jan 2016.

[93] W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics*, 4(2):e1000008, feb 2008.

[94] W. G. Hill and H. A. Mulder. Genetic analysis of environmental variation. *Genetics Research*, 92(5-6):381–395, dec 2010.

[95] D. Houle. Comparing evolvability and variability of quantitative traits. *Genetics*, 130(1):195 LP – 204, jan 1992.

[96] S. Huber, F. L. Bookstein, and M. Fieder. Socioeconomic status, education, and reproduction in modern women: An evolutionary perspective. *American Journal of Human Biology*, 22(5):578–587, sep 2010.

[97] L. D. Hurst, C. Pál, and M. J. Lercher. The evolutionary dynamics of eukaryotic gene order. *Nature reviews. Genetics*, 5(4):299–310, apr 2004.

[98] J. R. Huyghe, A. U. Jackson, M. P. Fogarty, M. L. Buchkovich, A. Stančáková, H. M. Stringham, X. Sim, L. Yang, C. Fuchsberger, H. Cederberg, P. S. Chines, T. M. Teslovich, J. M. Romm, H. Ling, I. McMullen, R. Ingersoll, E. W. Pugh, K. F. Doheny, B. M. Neale, M. J. Daly, J. Kuusisto, L. J. Scott, H. M. Kang, F. S. Collins, G. R. Abecasis, R. M. Watanabe, M. Boehnke, M. Laakso, and K. L. Mohlke. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature genetics*, 45(2):197–201, 2013.

[99] D. International HapMap 3 Consortium, D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, L. Peltonen, E. Dermitzakis, P. E. Bonnen, D. M. Altshuler, R. A. Gibbs, P. I. W. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, F. Yu, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, R. A. Gibbs, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, E. Dermitzakis, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, P. E. Bonnen, R. A. Gibbs, C. Gonzaga-Jauregui, A. Keinan, A. L. Price, F. Yu, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, S. F. Schaffner, Q. Zhang, M. J. R. Ghori, R. McGinnis, W. McLaren, S. Pollack, A. L. Price, S. F. Schaffner, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O. Wang, D. Reddy, C. N. Rotimi, C. D. Royal, R. R. Sharp, C. Zeng, L. D. Brooks, and J. E. McEwen. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, sep 2010.

[100] C. T. Johansen, J. Wang, M. B. Lanktree, H. Cao, D. Adam, M. R. Ban, R. a. Martins, B. a. Kennedy, R. G. Hassell, M. E. Visser, S. M. Schwartz, B. F. Voight, R. Elosua, C. J. O. Donnell, G. M. Dallinga-thie, S. S. Anand, M. W. Huff, S. Kathiresan, and R. a. Hegele. Mutation skew in genes identified by genome-wide association study of hypertriglyceridemia. *Nature genetics*, 42(8):684–687, 2011.

[101] C. T. Johansen, J. Wang, M. B. Lanktree, H. Cao, A. D. McIntyre, M. R. Ban, R. A. Martins, B. A. Kennedy, R. G. Hassell, M. E. Visser, S. M. Schwartz, B. F. Voight, R. Elosua, V. Salomaa, C. J. O'Donnell, G. M. Dallinga-Thie, S. S. Anand, S. Yusuf, M. W. Huff, S. Kathiresan, and R. A. Hegele. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature genetics*, 42(8):684–7, aug 2010.

[102] A. D. Johnson, M. Kavousi, A. V. Smith, M. H. Chen, A. Dehghan, T. Aspelund, J. P. Lin, C. M. van Duijn, T. B. Harris, L. A. Cupples, A. G. Uitterlinden, L. Launer, A. Hofman, F. Rivadeneira, B. Stricker, Q. Yang, C. J. O'Donnell, V. Gudnason, and J. C. Witteman. Genome-wide association meta-analysis for total serum bilirubin levels. *Human Molecular Genetics*, 18(14):2700–2710, 2009.

[103] T. Johnson and N. Barton. Theoretical models of selection and mutation on quantitative traits. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1459):1411–25, jul 2005.

[104] H. R. Johnston, Y. Hu, and D. J. Cutler. Population genetics identifies challenges in analyzing rare variants. *Genetic epidemiology*, 39(3):145–8, mar 2015.

[105] L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J.-P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Büning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, and J. H. Cho. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–24, nov 2012.

[106] P. Kaar, J. Jokela, T. Helle, and I. Kojola. Direct and Correlative Phenotypic Selection on Life-History Traits in Three Pre-Industrial Human Populations. *Proceedings of the Royal Society of London B: Biological Sciences*, 263(1376), 1996.

[107] M. N. Karn and P. L. S. Birth Weight and Gestation Time in Relation to Maternal Age, Parity and Infant Survival. *Annals of Eugenics*, 16(1):147–164, jan 1951.

[108] P. D. Keightley and W. G. Hill. Variation Maintained in Quantitative Traits with Mutation-Selection Balance: Pleiotropic Side-Effects on Fitness Traits. *Proceedings of the Royal Society B: Biological Sciences*, 242(1304):95–100, nov 1990.

[109] a. Keinan and a. G. Clark. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science*, 336(6082):740–743, 2012.

[110] S. Kim and E. P. Xing. Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *PLoS Genetics*, 5(8):e1000587, aug 2009.

[111] M. Kimmel, R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins, and L. B. Jorde. Signatures of Population Expansion in Microsatellite Repeat Data. *Genetics*, 148(4):1921–1930, apr 1998.

[112] M. Kimura. A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proceedings of the National Academy of Sciences of the United States of America*, 54(3):731–6, sep 1965.

[113] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(694):893–903, 1969.

[114] C. R. King, P. J. Rathouz, and D. L. Nicolae. An evolutionary framework for association testing in resequencing studies. *PLoS genetics*, 6(11):e1001202, nov 2010.

[115] E. G. King, B. J. Sanderson, C. L. McNeil, A. D. Long, and S. J. Macdonald. Genetic Dissection of the Drosophila melanogaster Female Head Transcriptome Reveals Widespread Allelic Heterogeneity. *PLoS Genetics*, 10(5):e1004322, may 2014.

[116] J. F. C. Kingman. A Simple Model for the Balance between Selection and Mutation. *Journal of Applied Probability*, 15(1):1–12, mar 1978.

[117] J. G. Kingsolver and S. E. Diamond. Phenotypic selection in natural populations: what limits directional selection? *The American Naturalist*, 177(3):346–57, mar 2011.

[118] K. M. Kirk, S. P. Blomberg, D. L. Duffy, a. C. Heath, I. P. Owens, and N. G. Martin. Natural selection and quantitative genetics of life-history traits in Western women: a twin study. *Evolution*, 55(2):423–435, 2001.

[119] M. Kirkpatrick and P. Jarne. The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. *The American Naturalist*, 155(2):154–167, feb 2000.

[120] K. Kodo, T. Nishizawa, M. Furutani, S. Arai, E. Yamamura, K. Joo, T. Takahashi, R. Matsuoka, and H. Yamagishi. GATA6 mutations cause human cardiac outflow tract defects by disrupting semaphorin-plexin signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33):13933–8, aug 2009.

[121] K. Kojima. Role of Epistasis and Overdominance in Stability of Equilibriea with Selection. *Proceedings of the National Academy of Sciences of the United States of America*, 45(7):984–9, jul 1959.

[122] A. Kong, M. L. Frigge, G. Thorleifsson, H. Stefansson, A. I. Young, F. Zink, G. A. Jonsdottir, A. Okbay, P. Sulem, G. Masson, D. F. Gudbjartsson, A. Helgason, G. Bjornsdottir, U. Thorsteinsdottir, and K. Stefansson. Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences of the United States of America*, page 201612113, jan 2017.

[123] I. K. Kotowski, A. Pertsemlidis, A. Luke, R. S. Cooper, G. L. Vega, J. C. Cohen, and H. H. Hobbs. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *American journal of human genetics*, 78(3):410–422, mar 2006.

[124] Y.-C. Lai, C.-F. Kao, M.-L. Lu, H.-C. Chen, P.-Y. Chen, C.-H. Chen, W. W. Shen, J.-Y. Wu, R.-B. Lu, and P.-H. Kuo. Investigation of associations between NR1D1, RORA and RORB genes and bipolar disorder. *PloS one*, 10(3):e0121245, jan 2015.

[125] R. Lande. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genetical research*, 26(3):221–35, dec 1975.

[126] R. Lande and S. Arnold. The measurement of selection on correlated characters. *Evolution*, 37(6):1210–1226, 1983.

[127] E. S. Lander. The New Genomics: Global Views of Biology. *Science*, 274(5287):536–539, oct 1996.

[128] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki,

D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001.

[129] E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–48, sep 1994.

[130] B. D. H. Latter. Natural selection for an intermediate optimum. *Australian Journal of Biological Sciences*, 13:30–35, 1960.

[131] S. Lee, with contributions from Larisa Miropolsky, and M. Wu. *SKAT: SNP-set (Sequence) Kernel Association Test*, 2014.

[132] S. H. Lee, J. Yang, G.-B. Chen, S. Ripke, E. A. Stahl, C. M. Hultman, P. Sklar, P. M. Visscher, P. F. Sullivan, M. E. Goddard, and N. R. Wray. Estimation of SNP heritability from dense genotype data. *American journal of human genetics*, 93(6):1151–5, dec 2013.

[133] S. H. Lee, J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics (Oxford, England)*, 28(19):2540–2, oct 2012.

[134] R. D. Lewontin. *The Genetic Basis of Evolutionary Change* . New York and London: Columbia University Press, 1974.

[135] B. Li and S. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83:311–321, 2008.

[136] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, jul 2011.

[137] W.-Q. Li, R. M. Pfeiffer, P. L. Hyland, J. Shi, F. Gu, Z. Wang, S. Bhattacharjee, J. Luo, X. Xiong, M. Yeager, X. Deng, N. Hu, P. R. Taylor, D. Albanes, N. E. Caporaso, S. M. Gapstur, L. Amundadottir, S. J. Chanock, N. Chatterjee, M. T. Landi, M. A. Tucker, A. M. Goldstein, and X. R. Yang. Genetic polymorphisms in the 9p21 region associated with risk of multiple cancers. *Carcinogenesis*, 35(12):2698–705, dec 2014.

[138] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, sep 2011.

[139] P.-R. Loh, G. Bhatia, A. Gusev, H. K. Finucane, B. K. Bulik-Sullivan, S. J. Pollack, T. R. de Candia, S. H. Lee, N. R. Wray, K. S. Kendler, M. C. O'Donovan, B. M. Neale, N. Patterson, A. L. Price, and A. L. Price. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):1385–1392, nov 2015.

[140] P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, feb 2015.

[141] K. E. Lohmueller. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS genetics*, 10(5):e1004379, may 2014.

[142] K. E. Lohmueller, A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clark, and C. D. Bustamante. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451(7181):994–997, 2008.

[143] M. Luciano, V. Svinti, A. Campbell, R. E. Marioni, C. Hayward, A. F. Wright, M. S. Taylor, D. J. Porteous, P. Thomson, J. G. D. Prendergast, N. D. Hastie, S. M. Farrington, G. Scotland, M. G. Dunlop, and I. J. Deary. Exome Sequencing to Detect Rare Variants Associated With General Cognitive Ability: A Pilot Study. *Twin research and human genetics : the official journal of the International Society for Twin Studies*, pages 1–9, mar 2015.

[144] T. Lumley. *biglm: bounded memory linear and generalized linear models*, 2013.

[145] L. Luo, E. Boerwinkle, and M. Xiong. Association studies for next-generation sequencing. *Genome research*, 21(7):1099–108, jul 2011.

[146] M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits*. Sinauer, 1998.

[147] B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2), 2009.

[148] M. C. Maher, L. H. Uricchio, D. G. Torgerson, and R. D. Hernandez. Population genetics of rare variants and complex diseases. *Human heredity*, 74(3-4):118–28, jan 2012.

[149] A. Mäki-Tanila and W. G. Hill. Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–67, sep 2014.

[150] N. Mancuso, N. Rohland, K. A. Rand, A. Tandon, A. Allen, D. Quinque, S. Mallick, H. Li, A. Stram, X. Sheng, Z. Kote-Jarai, D. F. Easton, R. A. Eeles, L. Le Marchand, A. Lubwama, D. Stram, S. Watya, D. V. Conti, B. Henderson, C. A. Haiman, B. Pasaniuc, and D. Reich. The contribution of rare variation to prostate cancer heritability. *Nature genetics*, advance on, nov 2015.

[151] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, oct 2009.

[152] J. Marchini, L. R. Cardon, M. S. Phillips, and P. Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, may 2004.

[153] N. J. Marini, J. Gin, J. Ziegle, K. H. Keho, D. Ginzinger, D. a. Gilbert, and J. Rine. The prevalence of folate-remedial MTHFR enzyme variants in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(23):8055–8060, jun 2008.

[154] G. T. Marth, E. Czabarka, J. Murvai, and S. T. Sherry. The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. *Genetics*, 166(1):351–372, 2004.

[155] A. Mathelier, W. Shi, and W. W. Wasserman. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics*, 31(2):67–76, feb 2015.

[156] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, 337(6099):1190–5, sep 2012.

[157] M. I. McCarthy and J. N. Hirschhorn. Genome-wide association studies: Potential next steps on a genetic journey. *Human Molecular Genetics*, 17(2):156–165, 2008.

[158] J. McClellan and M.-C. King. Genetic heterogeneity in human disease. *Cell*, 141(2):210–7, apr 2010.

[159] G. A. McVean, D. M. Altshuler (Co-Chair), R. M. Durbin (Co-Chair), G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A.

Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs (Principal Investigator), H. Dinh, C. Kovar, S. Lee, L. Lewis, D. Muzny, J. Reid, M. Wang, J. Wang (Principal Investigator), X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, G. Li, J. Li, Y. Li, Z. Li, X. Liu, Y. Lu, X. Ma, Z. Su, S. Tai, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, Y. Yin, W. Zhang, J. Zhao, M. Zhao, X. Zheng, Y. Zhou, E. S. Lander (Principal Investigator), D. M. Altshuler, S. B. Gabriel (Co-Chair), N. Gupta, P. Flicek (Principal Investigator), L. Clarke, R. Leinonen, R. E. Smith, X. Zheng-Bradley, D. R. Bentley (Principal Investigator), R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach (Principal Investigator), R. Sudbrak (Project Leader), M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, S. T. Sherry (Principal Investigator), G. A. McVean (Principal Investigator), E. R. Mardis (Co-Principal Investigator) (Co-Chair), R. K. Wilson (Co-Principal Investigator), L. Fulton, R. Fulton, G. M. Weinstock, R. M. Durbin (Principal Investigator), S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt (Principal Investigator), C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton (Principal Investigator), R. A. Gibbs (Principal Investigator), F. Yu (Project Leader), M. Bainbridge, D. Challis, U. S. Evani, J. Lu, D. Muzny, U. Nagaswamy, J. Reid, A. Sabo, Y. Wang, J. Yu, J. Wang (Principal Investigator), L. J. M. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, N. Qin, H. Shao, B. Wang, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, G. T. Marth (Principal Investigator), E. P. Garrison, D. Kural, W.-P. Lee, W. Fung Leong, A. N. Ward, J. Wu, M. Zhang, C. Lee (Principal Investigator), L. Griffin, C.-H. Hsieh, R. E. Mills, X. Shi, M. von Grotthuss, C. Zhang, M. J. Daly (Principal Investigator), M. A. DePristo (Project Leader), D. M. Altshuler, E. Banks, G. Bhatia, M. O. Carneiro, G. del Angel, S. B. Gabriel, G. Genovese, N. Gupta, R. E. Handsaker, C. Hartl, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. F. Schaffner, K. Shakir, S. C. Yoon (Principal Investigator), J. Lihm, V. Makarov, H. Jin (Principal Investigator), W. Kim, K. Cheol Kim, J. O. Korbel (Principal Investigator), T. Rausch, P. Flicek (Principal Investigator), K. Beal, L. Clarke, F. Cunningham, J. Herrero, W. M. McLaren, G. R. S. Ritchie, R. E. Smith, X. Zheng-Bradley, A. G. Clark (Principal Investigator), S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, P. C. Sabeti (Principal Investigator), S. R. Grossman, S. Tabrizi, R. Tariyal, D. N. Cooper (Principal Investigator), E. V. Ball, P. D. Stenson, D. R. Bentley (Principal Investigator), B. Barnes, M. Bauer, R. Keira Cheetham, T. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, K. Ye (Principal Investigator), M. A. Batzer (Principal Investigator), M. K. Konkel, J. A. Walker, D. G. MacArthur (Principal Investigator), M. Lek, Sudbrak (Project Leader), V. S. Amstislavskiy, R. Herwig, M. D. Shriver (Principal Investigator), C. D. Bustamante (Principal Investigator), J. K. Byrnes, F. M. De La Vega, S. Gravel, E. E. Kenny, J. M. Kidd, P. Lacroute, B. K. Maples, A. Moreno-Estrada, F. Zakharia, E. Halperin (Principal Investigator), Y. Baran, D. W. Craig (Principal Investigator), A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry (Principal Investigator), C. Xiao, J. Sebat (Principal Investigator), V. Bafna, K. Ye, E. G. Burchard (Principal Investigator), R. D. Hernandez (Principal Investigator), C. R. Gignoux, D. Haussler (Principal Investigator), S. J.

Katzman, W. James Kent, B. Howie, A. Ruiz-Linares (Principal Investigator), E. T. Dermitzakis (Principal Investigator), T. Lappalainen, S. E. Devine (Principal Investigator), X. Liu, A. Maroo, L. J. Tallon, J. A. Rosenfeld (Principal Investigator), L. P. Michelson, G. R. Abecasis (Principal Investigator) (Co-Chair), H. Min Kang (Project Leader), P. Anderson, A. Angius, A. Bigham, T. Blackwell, F. Busonero, F. Cucca, C. Fuchsberger, C. Jones, G. Jun, Y. Li, R. Lyons, A. Maschio, E. Porcu, F. Reinier, S. Sanna, D. Schlessinger, C. Sidore, A. Tan, M. Kate Trost, P. Awadalla (Principal Investigator), A. Hodgkinson, G. Lunter (Principal Investigator), G. A. McVean (Principal Investigator) (Co-Chair), J. L. Marchini (Principal Investigator), S. Myers (Principal Investigator), C. Churchhouse, O. Delaneau, A. Gupta-Hinch, Z. Iqbal, I. Mathieson, A. Rimmer, D. K. Xifara, T. K. Oleksyk (Principal Investigator), Y. Fu (Principal Investigator), X. Liu, M. Xiong, L. Jorde (Principal Investigator), D. Witherspoon, J. Xing, E. E. Eichler (Principal Investigator), B. L. Browning (Principal Investigator), C. Alkan, I. Hajirasouliha, F. Hormozdiari, A. Ko, P. H. Sudmant, E. R. Mardis (Co-Principal Investigator), K. Chen, A. Chinwalla, L. Ding, D. Dooling, D. C. Koboldt, M. D. McLellan, J. W. Wallis, M. C. Wendl, Q. Zhang, R. M. Durbin (Principal Investigator), M. E. Hurles (Principal Investigator), C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, A. J. Coffey, V. Colonna, P. Danecek, N. Huang, L. Jostins, T. M. Keane, H. Li, S. McCarthy, A. Scally, J. Stalker, K. Walter, Y. Xue, Y. Zhang, M. B. Gerstein (Principal Investigator), A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, L. Habegger, A. O. Harmanci, M. Jin, E. Khurana, X. Jasmine Mu, C. Sisu, Y. Li, R. Luo, H. Zhu, C. Lee (Principal Investigator) (Co-Chair), L. Griffin, C.-H. Hsieh, R. E. Mills, X. Shi, M. von Grotthuss, C. Zhang, G. T. Marth (Principal Investigator), E. P. Garrison, D. Kural, W.-P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll (Project Leader), D. M. Altshuler, E. Banks, G. del Angel, G. Genovese, R. E. Handsaker, C. Hartl, J. C. Nemesh, K. Shakir, S. C. Yoon (Principal Investigator), J. Lihm, V. Makarov, J. Degenhardt, P. Flicek (Principal Investigator), L. Clarke, R. E. Smith, X. Zheng-Bradley, J. O. Korbel (Principal Investigator) (Co-Chair), T. Rausch, A. M. Stütz, D. R. Bentley (Principal Investigator), B. Barnes, R. Keira Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, K. Ye (Principal Investigator), M. A. Batzer (Principal Investigator), M. K. Konkel, J. A. Walker, P. Lacroute, D. W. Craig (Principal Investigator), N. Homer, D. Church, C. Xiao, J. Sebat (Principal Investigator), V. Bafna, J. J. Michaelson, K. Ye, S. E. Devine (Principal Investigator), X. Liu, A. Maroo, L. J. Tallon, G. Lunter (Principal Investigator), G. A. McVean (Principal Investigator), Z. Iqbal, D. Witherspoon, J. Xing, E. E. Eichler (Principal Investigator) (Co-Chair), C. Alkan, I. Hajirasouliha, F. Hormozdiari, A. Ko, P. H. Sudmant, K. Chen, A. Chinwalla, L. Ding, M. D. McLellan, J. W. Wallis, M. E. Hurles (Principal Investigator) (Co-Chair), B. Blackburne, H. Li, S. J. Lindsay, Z. Ning, A. Scally, K. Walter, Y. Zhang, M. B. Gerstein (Principal Investigator), A. Abyzov, J. Chen, D. Clarke, E. Khurana, X. Jasmine Mu, C. Sisu, R. A. Gibbs (Principal Investigator) (Co-Chair), F. Yu (Project Leader), M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, L. Lewis, J. Lu, D. Muzny, U. Nagaswamy, J. Reid, A. Sabo, J. Yu, X. Guo, Y. Li, R. Wu, G. T. Marth (Principal Investigator) (Co-Chair), E. P. Garrison, W. Fung Leong, A. N. Ward, G. del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G. Clark (Principal In-

vestigator), J. L. Rodriguez-Flores, P. Flicek (Principal Investigator), L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur (Principal Investigator), C. D. Bustamante (Principal Investigator), S. Gravel, D. W. Craig (Principal Investigator), A. Christoforides, N. Homer, T. Izatt, S. T. Sherry (Principal Investigator), C. Xiao, E. T. Dermitzakis (Principal Investigator), G. R. Abecasis (Principal Investigator), H. Min Kang, G. A. McVean (Principal Investigator), E. R. Mardis (Principal Investigator), D. Dooling, L. Fulton, R. Fulton, D. C. Koboldt, R. M. Durbin (Principal Investigator), S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, M. B. Gerstein (Principal Investigator), S. Balasubramanian, L. Habegger, E. P. Garrison, R. A. Gibbs (Principal Investigator), M. Bainbridge, D. Muzny, F. Yu, J. Yu, G. del Angel, R. E. Handsaker, V. Makarov, J. L. Rodriguez-Flores, H. Jin (Principal Investigator), W. Kim, K. Cheol Kim, P. Flicek (Principal Investigator), K. Beal, L. Clarke, F. Cunningham, J. Herrero, W. M. McLaren, G. R. S. Ritchie, X. Zheng-Bradley, S. Tabrizi, D. G. MacArthur (Principal Investigator), M. Lek, C. D. Bustamante (Principal Investigator), F. M. De La Vega, D. W. Craig (Principal Investigator), A. A. Kurdoglu, T. Lappalainen, J. A. Rosenfeld (Principal Investigator), L. P. Michelson, P. Awadalla (Principal Investigator), A. Hodgkinson, G. A. McVean (Principal Investigator), K. Chen, Y. Chen, V. Colonna, A. Frankish, J. Harrow, Y. Xue, M. B. Gerstein (Principal Investigator) (Co-Chair), A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, E. Khurana, X. Jasmine Mu, C. Sisu, R. A. Gibbs (Principal Investigator), G. Fowler, W. Hale, D. Kalra, C. Kovar, D. Muzny, J. Reid, J. Wang (Principal Investigator), X. Guo, G. Li, Y. Li, X. Zheng, D. M. Altshuler, P. Flicek (Principal Investigator) (Co-Chair), L. Clarke (Project Leader), J. Barker, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley (Principal Investigator), T. Cox, S. Humphray, S. Kahn, R. Sudbrak (Project Leader), M. W. Albrecht, M. Lienhard, D. W. Craig (Principal Investigator), T. Izatt, A. A. Kurdoglu, S. T. Sherry (Principal Investigator) (Co-Chair), V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, C. Xiao, H. Zhang, D. Haussler (Principal Investigator), G. R. Abecasis (Principal Investigator), G. A. McVean (Principal Investigator), C. Alkan, A. Ko, D. Dooling, R. M. Durbin (Principal Investigator), S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti (Co-Chair), B. M. Knoppers (Co-Chair), G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, D. Reich, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, B. Timmermann, S. Tishkoff, L. H. Toji, C. Tyler Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, C. Zhi Ming, G. Yang, C. Jia You, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, N. C. Clemm, A. Duncanson, M. Dunn, E. D. Green, M. S. Guyer, J. L. Peterson, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin,

R. E. Handsaker, H. Min Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, oct 2012.

[160] N. Mejhert, F. Wilfling, D. Esteve, J. Galitzky, V. Pellegrinelli, C.-I. Kolditz, N. Viguerie, J. Tordjman, E. Näslund, P. Trayhurn, D. Lacasa, I. Dahlman, V. Stich, P. Lång, D. Langin, A. Bouloumié, K. Clément, and M. Rydén. Semaphorin 3C is a novel adipokine linked to extracellular matrix composition. *Diabetologia*, 56(8):1792–801, aug 2013.

[161] J. Merila and B. C. Sheldon. Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. *Heredity*, 83(2):103, aug 1999.

[162] T. E. Meyer, L. W. Chu, Q. Li, K. Yu, P. S. Rosenberg, I. Menashe, A. P. Chokkalingam, S. M. Quraishi, W.-Y. Huang, J. M. Weiss, R. Kaaks, R. B. Hayes, S. J. Chanock, and A. W. Hsing. The association between inflammation-related genes and serum androgen levels in men: the prostate, lung, colorectal, and ovarian study. *The Prostate*, 72(1):65–71, jan 2012.

[163] E. Milot, F. M. Mayer, D. H. Nussey, M. Boisvert, F. Pelletier, and D. Réale. Evidence for evolution in response to natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences of the United States of America*, 108(41):17040–5, oct 2011.

[164] V. Moskvina, K. M. Schmidt, A. Vedernikov, M. J. Owen, N. Craddock, P. Holmans, and M. C. O'Donovan. Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. *European journal of human genetics : EJHG*, 20(8):890–6, aug 2012.

[165] H. Mostafavi, T. Berisa, F. R. Day, J. R. B. Perry, M. Przeworski, and J. K. Pickrell. Identifying genetic variants that affect viability in large cohorts. *bioRxiv*, 15(9):e2002458, sep 2017.

[166] L. Moutsianas, V. Agarwala, C. Fuchsberger, J. Flannick, M. A. Rivas, K. J. Gaulton, P. K. Albers, G. McVean, M. Boehnke, D. Altshuler, and M. I. McCarthy. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS genetics*, 11(4):e1005165, apr 2015.

[167] Y. Naciri-Graven and J. Goudet2. THE ADDITIVE GENETIC VARIANCE AFTER BOTTLENECKS IS AFFECTED BY THE NUMBER OF LOCI INVOLVED IN EPISTATIC INTERACTIONS. *Evolution*, 57(574):706–716, 2003.

[168] S. A. Naser, M. Arce, A. Khaja, M. Fernandez, N. Naser, S. Elwasila, and S. Thanigachalam. Role of ATG16L, NOD2 and IL23R in Crohn's disease pathogenesis. *World journal of gastroenterology : WJG*, 18(5):412–24, feb 2012.

[169] B. M. Neale, M. a. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3), 2011.

[170] M. C. Neale, M. D. Hunter, J. N. Pritikin, M. Zahery, T. R. Brick, R. M. Kickpatrick, R. Estabrook, T. C. Bates, H. H. Maes, and S. M. Boker. Open{M}x 2.0: {E}xtended structural equation and statistical modeling. *Psychometrika*, 2015.

[171] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, N.Y.)*, 337(6090):100–4, jul 2012.

[172] D. Nettle. Women's height, reproductive success and the evolution of sexual dimorphism in modern humans. *Proceedings. Biological sciences*, 269(1503):1919–23, sep 2002.

[173] T.-L. North and M. A. Beaumont. Complex trait architecture: the pleiotropic model revisited. *Scientific reports*, 5:9351, jan 2015.

[174] Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. A. Pappas, L. Jiang, J. Yin, L. Ye, D.-F. Su, J. Yang, G. Xie, E. Keystone, H.-J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. A. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. H. Coenen, P. L. C. M. van Riel, M. A. F. J. van de Laar, H.-J. Guchelaar, T. W. J. Huizinga, P. Dieudé, X. Mariette, S. L. Bridges, A. Zhernakova, R. E. M. Toes, P. P. Tak, C. Miceli-Richard, S.-Y. Bang, H.-S. Lee, J. Martin, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapää-Dahlqvist, L. Arlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, S. Eyre, J. Bowes, A. Barton, N. de Vries, L. W. Moreland, L. A. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S.-C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. A. Brown, H. Yamanaka, T. Mimori, A. Takahashi, H. Xu, T. W. Behrens, K. A. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, and R. M. Plenge. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–81, feb 2014.

[175] A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G.-B. Chen, V. Emilsson, S. F. W. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma, M. Pina Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E. Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J. Lahti, S. J. van der Lee, C. DeLeeuw, P. A. Lind, K.-O. Lindgren, T. Liu, M. Mangino, J. Marten, E. Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot, Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt,

K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot, B. St Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H.-J. Westra, J. Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K. Bønnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J.-E. De Neve, P. Deloukas, I. Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J. Forstner, I. Gandin, B. Gunnarsson, B. V. Halldórsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G. Holliday, G. Homuth, M. A. Horan, J.-J. Hottenga, P. L. de Jager, P. K. Joshi, A. Jugessur, M. A. Kaakinen, M. Kähönen, S. Kanoni, L. Keltigangas-Järvinen, L. A. L. M. Kiemeney, I. Kolcic, S. Koskinen, A. T. Kraja, M. Kroh, Z. Kutalik, A. Latvala, L. J. Launer, M. P. Lebreton, D. F. Levinson, P. Lichtenstein, P. Lichtner, D. C. M. Liewald, L. Cohort Study, A. Loukola, P. A. Madden, R. Mägi, T. Mäki-Opas, R. E. Marioni, P. Marques-Vidal, G. A. Meddens, G. McMahon, C. Meisinger, T. Meitinger, Y. Milaneschi, L. Milani, G. W. Montgomery, R. Myhre, C. P. Nelson, D. R. Nyholt, W. E. R. Ollier, A. Palotie, L. Paternoster, N. L. Pedersen, K. E. Petrovic, D. J. Porteous, K. Räikkönen, S. M. Ring, A. Robino, O. Rostapshova, I. Rudan, A. Rustichini, V. Salomaa, A. R. Sanders, A.-P. Sarin, H. Schmidt, R. J. Scott, B. H. Smith, J. A. Smith, J. A. Staessen, E. Steinhagen-Thiessen, K. Strauch, A. Terracciano, M. D. Tobin, S. Ulivi, S. Vaccargiu, L. Quaye, F. J. A. van Rooij, C. Venturini, A. A. E. Vinkhuyzen, U. Völker, H. Völzke, J. M. Vonk, D. Vozzi, J. Waage, E. B. Ware, G. Willemsen, J. R. Attia, D. A. Bennett, K. Berger, L. Bertram, H. Bisgaard, D. I. Boomsma, I. B. Borecki, U. Bültmann, C. F. Chabris, F. Cucca, D. Cusi, I. J. Deary, G. V. Dedoussis, C. M. van Duijn, J. G. Eriksson, B. Franke, L. Franke, P. Gasparini, P. V. Gejman, C. Gieger, H.-J. Grabe, J. Gratten, P. J. F. Groenen, V. Gudnason, P. van der Harst, C. Hayward, D. A. Hinds, W. Hoffmann, E. Hyppönen, W. G. Iacono, B. Jacobsson, M.-R. Järvelin, K.-H. Jöckel, J. Kaprio, S. L. R. Kardia, T. Lehtimäki, S. F. Lehrer, P. K. E. Magnusson, N. G. Martin, M. McGue, A. Metspalu, N. Pendleton, B. W. J. H. Penninx, M. Perola, N. Pirastu, M. Pirastu, O. Polasek, D. Posthuma, C. Power, M. A. Province, N. J. Samani, D. Schlessinger, R. Schmidt, T. I. A. Sørensen, T. D. Spector, K. Stefansson, U. Thorsteinsdottir, A. R. Thurik, N. J. Timpson, H. Tiemeier, J. Y. Tung, A. G. Uitterlinden, V. Vitart, P. Vollenweider, D. R. Weir, J. F. Wilson, A. F. Wright, D. C. Conley, R. F. Krueger, G. Davey Smith, A. Hofman, D. I. Laibson, S. E. Medland, M. N. Meyer, J. Yang, M. Johannesson, P. M. Visscher, T. Esko, P. D. Koellinger, D. Cesarini, and D. J. Benjamin. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604):539–542, may 2016.

[176] G. Orozco, J. C. Barrett, and E. Zeggini. Synthetic associations in the context of genome-wide association scan signals. *Human molecular genetics*, 19(R2):R137–44, oct 2010.

[177] J.-H. Park, S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics*, 42(7):570–5, jul 2010.

[178] S. Peischl and L. Excoffier. Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology*, 24(9):2084–2094, may 2015.

[179] V. Plagnol, J. M. M. Howson, D. J. Smyth, N. Walker, J. P. Hafler, C. Wallace, H. Stevens, L. Jackson, M. J. Simmonds, P. J. Bingley, S. C. Gough, and J. A. Todd. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS genetics*, 7(8):e1002216, aug 2011.

[180] P. Pozzilli, E. Maddaloni, and R. Buzzetti. Combination immunotherapies for type 1 diabetes mellitus. *Nature reviews. Endocrinology*, 11(5):289–97, may 2015.

[181] N. J. Prescott, K. M. Dominy, M. Kubo, C. M. Lewis, S. A. Fisher, R. Redon, N. Huang, B. E. Stranger, K. Blaszczyk, B. Hudspith, G. Parkes, N. Hosono, K. Yamazaki, C. M. Onnie, A. Forbes, E. T. Dermitzakis, Y. Nakamura, J. C. Mansfield, J. Sanderson, M. E. Hurles, R. G. Roberts, and C. G. Mathew. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Human molecular genetics*, 19(9):1828–39, may 2010.

[182] N. J. Prescott, B. Lehne, K. Stone, J. C. Lee, K. Taylor, J. Knight, E. Papouli, M. M. Mirza, M. A. Simpson, S. L. Spain, G. Lu, F. Fraternali, S. J. Bumpstead, E. Gray, A. Amar, H. Bye, P. Green, G. Chung-Faye, B. Hayee, R. Pollok, J. Satsangi, M. Parkes, J. C. Barrett, J. C. Mansfield, J. Sanderson, C. M. Lewis, M. E. Weale, T. Schlitt, and C. G. Mathew. Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in BTNL2 and implicates other immune related genes. *PLoS genetics*, 11(2):e1004955, feb 2015.

[183] A. L. Price, G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L.-J. Wei, and S. R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *American journal of human genetics*, 86(6):832–8, jun 2010.

[184] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, aug 2006.

[185] G. R. Price. Selection and Covariance. *Nature*, 227(5257):520–521, aug 1970.

[186] G. R. Price. Extension of covariance selection mathematics. *Annals of Human Genetics*, 35(4):485–490, apr 1972.

[187] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69(1):124–37, jul 2001.

[188] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75, sep 2007.

[189] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kähler, L. Duncan, E. Stahl, G. Genovese, E. Fernández, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. E. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. N.

Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. a. McCarroll, and P. Sklar. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–90, 2014.

[190] A. W. Püschel, R. H. Adams, and H. Betz. Murine semaphorin D/collapsin is a member of a diverse gene family and creates domains inhibitory for axonal extension. *Neuron*, 14(5):941–948, may 1995.

[191] R. Qayyum, B. M. Snively, E. Ziv, M. A. Nalls, Y. Liu, W. Tang, L. R. Yanek, L. Lange, M. K. Evans, S. Ganesh, M. A. Austin, G. Lettre, D. M. Becker, A. B. Zonderman, A. B. Singleton, T. B. Harris, E. R. Mohler, B. A. Logsdon, C. Kooperberg, A. R. Folsom, J. G. Wilson, L. C. Becker, and A. P. Reiner. A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. *PLoS genetics*, 8(3):e1002491, jan 2012.

[192] L. Quintana-Murci, J. Pritchard, J. Pritchard, N. Cox, T. Manolio, F. Collins, N. Cox, D. Goldstein, L. Hindorff, D. Hunter, M. McCarthy, G. Abecasis, L. Cardon, D. Goldstein, J. Little, J. Ioannidis, D. Reich, E. Lander, M. Zwick, D. Cutler, A. Chakravarti, N. Schork, S. Murray, K. Frazer, E. Topol, W. Bodmer, C. Bonilla, D. Goldstein, Q. Zhu, D. Ge, J. Maia, M. Zhu, S. Petrovski, S. Dickson, Y. Lu, D. Goldstein, M. Angrist, G. Cavalleri, A. Rienzo, B. Crespi, G. Abecasis, A. Auton, L. Brooks, M. DePristo, R. Durbin, R. Handsaker, F. Racimo, S. Sankararaman, R. Nielsen, E. Huerta-Sanchez, J. Kelso, K. Prufer, K. Veeramah, M. Hammer, J. Novembre, S. Ramachandran, B. Henn, L. Cavalli-Sforza, M. Feldman, V. Sousa, S. Peischl, L. Excoffier, K. Lohmueller, R. Nielsen, I. Hellmann, M. M. Hubisz, C. Bustamante, A. Clark, P. Sabeti, S. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, C. Jeong, A. Rienzo, J. Vitti, S. Grossman, P. Sabeti, F. Key, J. Teixeira, C. Filippo, A. Andres, S. Grossman, K. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, L. Barreiro, L. Quintana-Murci, J. Brinkworth, L. Barreiro, E. Karlsson, D. Kwiatkowski, P. Sabeti, R. Blekhman, O. Man, L. Herrmann, A. Boyko, A. Indap, C. Kosiol, A. Eyre-Walker, P. Keightley, G. Kryukov, L. Pennacchio, S. Sunyaev, A. Boyko, S. Williamson, A. Indap, J. Degenhardt, R. Hernandez, K. Lohmueller, A. Eyre-Walker, P. Keightley, C. Bustamante, A. Fledel-Alon, S. Williamson, R. Nielsen, M. M. Hubisz, S. Glanowski, M. Kimura, T. Maruyama, J. Crow, T. Ohta, H. Akashi, N. Osada, T. Ohta, A. Coventry, L. Bull-Otterson, X. Liu, A. Clark, T. Maxwell, J. Crosby, G. Marth, F. Yu, A. Indap, K. Garimella, S. Gravel, W. Leong, A. Keinan, A. Clark, M. Nelson, D. Wegmann, M. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Tennessen, A. Bigham, T. O'Connor, W. Fu, E. Kenny, S. Gravel, W. Fu, T. O'Connor, G. Jun, H. Kang, G. Abecasis, S. Leal, M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, T. Fennell, V. Agarwala, J. Flannick, S. Sunyaev, T. Go, D. Altshuler, G. Gibson, M. Maher, L. Uricchio, D. Torgerson, R. Hernandez, S. Gravel, B. Henn, R. Gutenkunst, A. Indap, G. Marth, A. Clark, K. Lohmueller, A. Indap, S. Schmidt, A. Boyko, R. Hernandez, M. M. Hubisz, S. Peischl, I. Dupanloup, M. Kirkpatrick, L. Excoffier, A. Eyre-Walker, I. Adzhubei, S. Schmidt, L. Peshkin, V. Ramensky, A. Gerasimova, P. Bork, G. Cooper, E. Stone, G. Asimenos, N. Program, E. Green, S. Batzoglou, P. Kumar, S. Henikoff, P. Ng, C. Dong, P. Wei, X. Jian, R. Gibbs, E. Boerwinkle, K. Wang, M. Kircher,

D. Witten, P. Jain, B. O'Roak, G. Cooper, J. Shendure, Y. Itan, L. Shang, B. Boisson, M. Ciancanelli, J. Markle, R. Martinez-Barricarte, R. Gutenkunst, R. Hernandez, S. Williamson, C. Bustamante, R. Do, D. Balick, H. Li, I. Adzhubei, S. Sunyaev, D. Reich, W. Fu, R. Gittelman, M. Bamshad, J. Akey, Y. Simons, M. Turchin, J. Pritchard, G. Sella, B. Henn, L. Botigue, C. Bustamante, A. Clark, S. Gravel, F. Casals, A. Hodgkinson, J. Hussin, Y. Idaghdour, V. Bruat, T. Maillard, E. Lim, P. Wurtz, A. Havulinna, P. Palta, T. Tukiainen, K. Rehnstrom, B. Henn, L. Botigue, S. Peischl, I. Dupanloup, M. Lipatov, B. Maples, S. Klopfstein, M. Currat, L. Excoffier, K. Lohmueller, L. Segurel, L. Quintana-Murci, L. Scheinfeldt, S. Tishkoff, J. Pritchard, A. Rienzo, J. Pritchard, J. Pickrell, G. Coop, L. Quintana-Murci, L. Barreiro, K. Siddle, L. Quintana-Murci, T. Bersaglieri, P. Sabeti, N. Patterson, T. Vanderploeg, S. Schaffner, J. Drake, S. Tishkoff, F. Reed, A. Ranciaro, B. Voight, C. Babbitt, J. Silverman, N. Enattah, T. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen, H. Rasinpera, Y. Itan, A. Powell, M. Beaumont, J. Burger, M. Thomas, A. Ranciaro, M. Campbell, J. Hirbo, W. Ko, A. Froment, P. Anagnostou, S. Beleza, A. Santos, B. McEvoy, I. Alves, C. Martinho, E. Cameron, C. Miller, S. Beleza, A. Pollen, D. Schluter, R. Kittles, M. Shriver, H. Norton, R. Kittles, E. Parra, P. McKeigue, X. Mao, K. Cheng, R. Lamason, M. Mohideen, J. Mest, A. Wong, H. Norton, M. Aros, A. Hancock, D. Witonsky, G. Alkorta-Aranburu, C. Beall, A. Gebremedhin, R. Sukernik, X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. Cuo, J. Pool, A. Bigham, M. Bauchet, D. Pinto, X. Mao, J. Akey, R. Mei, T. Simonson, Y. Yang, C. Huff, H. Yun, G. Qin, D. Witherspoon, A. Hancock, D. Witonsky, A. Gordon, G. Eshel, J. Pritchard, G. Coop, G. Alkorta-Aranburu, C. Beall, D. Witonsky, A. Gebremedhin, J. Pritchard, A. Rienzo, G. Coop, J. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, L. Barreiro, M. Ben-Ali, H. Quach, G. Laval, E. Patin, J. Pickrell, M. Deschamps, G. Laval, M. Fagny, Y. Itan, L. Abel, J. Casanova, M. Fumagalli, M. Sironi, E. Karlsson, J. Harris, S. Tabrizi, A. Rahman, I. Shlyakhter, N. Patterson, D. Kwiatkowski, H. Laayouni, M. Oosting, P. Luisi, M. Ioana, S. Alonso, I. Ricano-Ponce, C. Louicharoen, E. Patin, R. Paul, I. Nuchprayoon, B. Witoonpanich, C. Peerapittayamongkol, J. Manry, G. Laval, E. Patin, S. Fornarino, Y. Itan, M. Fumagalli, S. Mukherjee, N. Sarkar-Roy, D. Wagener, P. Majumder, L. Quintana-Murci, A. Clark, P. Sabeti, D. Reich, J. Higgins, H. Levine, D. Richter, S. Schaffner, M. Sironi, M. Clerici, E. Vasseur, M. Boniotto, E. Patin, G. Laval, H. Quach, J. Manry, G. Wlasiuk, M. Nachman, C. Jeong, G. Alkorta-Aranburu, B. Basnyat, M. Neupane, D. Witonsky, J. Pritchard, J. Pickrell, G. Coop, J. Novembre, S. Kudaravalli, J. Li, D. Absher, P. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, K. Tang, K. Thornton, M. Stoneking, B. Voight, S. Kudaravalli, X. Wen, J. Pritchard, C. Carlson, D. Thomas, M. Eberle, J. Swanson, R. Livingston, M. Rieder, J. Kelley, J. Madeoy, J. Calhoun, W. Swanson, J. Akey, L. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, H. Chen, N. Patterson, D. Reich, W. Jin, S. Xu, H. Wang, Y. Yu, Y. Shen, B. Wu, B. Weir, L. Cardon, A. Anderson, D. Nielsen, W. Hill, J. Akey, G. Zhang, K. Zhang, L. Jin, M. Shriver, J. Akey, S. Williamson, M. M. Hubisz, A. Clark, B. Payseur, C. Bustamante, R. Nielsen, M. Fagny, E. Patin, D. Enard, L. Barreiro, L. Quintana-Murci, G. Laval, R. Hernandez, J. Kelley, E. Elyashiv, S. Melton, A. Auton, G. McVean, J. Granka, B. Henn, C. Gignoux, J. J. Kidd, C. Bustamante,

M. Feldman, B. Vernot, A. Stergachis, M. Maurano, J. Vierstra, S. Neph, R. Thurman, H. Fraser, J. Pickrell, M. Schaub, A. Boyle, A. Kundaje, S. Batzoglou, M. Snyder, S. Nakagome, G. Alkorta-Aranburu, R. Amato, B. Howie, B. Peter, R. Hudson, B. Peter, E. Huerta-Sanchez, R. Nielsen, M. Allentoft, M. Sikora, K. Sjogren, S. Rasmussen, M. Rasmussen, J. Stenderup, I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. Roodenberg, J. Berg, G. Coop, M. Turchin, C. Chiang, C. Palmer, S. Sankararaman, D. Reich, J. Hirschhorn, P. Messer, D. Petrov, D. Charlesworth, J. Klein, A. Sato, S. Nagl, C. O'HUigin, A. Allison, J. Klein, Y. Satta, C. O'HUigin, N. Takahata, A. Hughes, M. Nei, F. Prugnolle, A. Manica, M. Charpentier, J. Guegan, V. Guernier, F. Balloux, L. Segurel, E. Thompson, T. Flutre, J. Lovstad, A. Venkat, S. Margulis, R. Cagliani, F. Guerini, M. Fumagalli, S. Riva, C. Agliardi, D. Galimberti, E. Leffler, Z. Gao, S. Pfeifer, L. Segurel, A. Auton, O. Venn, J. Teixeira, C. Filippo, A. Weihmann, J. Meneu, F. Racimo, M. Dannemann, R. Single, M. Martin, X. Gao, D. Meyer, M. Yeager, J. J. Kidd, A. Andres, M. M. Hubisz, A. Indap, D. Torgerson, J. Degenhardt, A. Boyko, M. DeGiorgio, K. Lohmueller, R. Nielsen, M. Rasmussen, M. M. Hubisz, I. Gronau, A. Siepel, A. Ferrer-Admetlla, E. Bosch, M. Sikora, T. Marques-Bonet, A. Ramirez-Soriano, A. Muntasell, P. Bronson, S. Mack, H. Erlich, M. Slatkin, A. Andres, M. Dennis, W. Kretzschmar, J. Cannons, S. Lee-Lin, B. Hurle, P. Norman, L. Abi-Rached, K. Gendzekhadze, D. Korbel, M. Gleimer, D. Rowley, M. Fumagalli, M. Fracassetti, R. Cagliani, D. Forni, U. Pozzoli, G. Comi, E. Hollox, J. Armour, M. Haber, M. Mezzavilla, Y. Xue, C. Tyler-Smith, S. Vattathil, J. Akey, S. Wong, S. Gochhait, D. Malhotra, F. Pettersson, Y. Teo, C. Khor, P. Uciechowski, H. Imhoff, C. Lange, C. Meyer, E. Browne, D. Kirsten, F. Broushaki, M. Thomas, V. Link, S. Lopez, L. Dorp, K. Kirsanow, Z. Hofmanova, S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, D. Diez-Del-Molino, R. Nielsen, M. M. Hubisz, I. Hellmann, D. Torgerson, A. Andres, A. Albrechtsen, B. Georgi, B. Voight, M. Bucan, A. Battle, S. Mostafavi, X. Zhu, J. Potash, M. Weissman, C. McCormick, M. Gerstein, A. Kundaje, M. Hariharan, S. Landt, K. Yan, C. Cheng, H. Fraser, A. Hirsh, L. Steinmetz, C. Scharfe, M. Feldman, I. Jordan, L. Marino-Ramirez, Y. Wolf, E. Koonin, D. Torgerson, A. Boyko, R. Hernandez, A. Indap, X. Hu, T. White, S. Katzman, A. Kern, G. Bejerano, G. Fewell, L. Fulton, R. Wilson, J. Drake, C. Bird, J. Nemesh, D. Thomas, C. Newton-Cheh, A. Reymond, J. Casanova, L. Abel, L. Quintana-Murci, A. Alcais, L. Quintana-Murci, D. Thaler, E. Schurr, L. Abel, J. Casanova, S. Boisson-Dupuis, X. Kong, S. Okada, S. Cypowyj, A. Puel, L. Abel, R. P. de Diego, V. Sancho-Shimizu, L. Lorenzo, A. Puel, S. Plancoulaine, C. Picard, J. Casanova, L. Abel, L. Quintana-Murci, V. Colonna, Q. Ayub, Y. Chen, L. Pagani, P. Luisi, M. Pybus, E. Corona, R. Chen, M. Sikora, A. Morgan, C. Patel, A. Ramesh, J. Young, Y. Chang, J. Kim, J. Chretien, M. Klag, M. Levine, R. Chen, E. Corona, M. Sikora, J. Dudley, A. Morgan, A. Moreno-Estrada, K. Andersen, I. Shylakhter, S. Tabrizi, S. Grossman, C. Happi, P. Sabeti, F. Key, B. Peter, M. Dennis, E. Huerta-Sanchez, W. Tang, L. Prokunina-Olsson, M. Fumagalli, M. Sironi, U. Pozzoli, A. Ferrer-Admetlla, L. Pattini, R. Nielsen, J. Dudley, Y. Kim, L. Liu, G. Markov, K. Gerold, R. Chen, J. Neel, M. Fumagalli, U. Pozzoli, R. Cagliani, G. Comi, S. Riva, M. Clerici, T. Raj, M. Kuchroo, J. Replogle, S. Raychaudhuri, B. Stranger, P. Jager, A. Zhernakova, C. Elbers, B. Ferwerda, J. Romanos, G. Trynka, P. Dubois, L. Uricchio, N. Zaitlen, C. Ye, J. Witte,

R. Hernandez, K. Prufer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, M. Meyer, M. Kircher, M. Gansauge, H. Li, F. Racimo, S. Mallick, R. Green, J. Krause, A. Briggs, T. Maricic, U. Stenzel, M. Kircher, S. Sankararaman, S. Mallick, M. Dannemann, K. Prufer, J. Kelso, S. Paabo, D. Reich, R. Green, M. Kircher, J. Krause, N. Patterson, E. Durand, D. Reich, N. Patterson, M. Kircher, F. Delfin, M. Nandineni, I. Pugach, B. Vernot, J. Akey, B. Vernot, J. Akey, S. Sankararaman, S. Mallick, N. Patterson, D. Reich, C. Simonti, B. Vernot, L. Bastarache, E. Bottinger, D. Carrell, R. Chisholm, E. Huerta-Sanchez, X. Jin, null Asan, Z. Bianba, B. Peter, N. Vinckenbosch, L. Abi-Rached, M. Jobin, S. Kulkarni, A. McWhinnie, K. Dalva, L. Gragert, F. Mendez, J. Watkins, M. Hammer, F. Mendez, J. Watkins, M. Hammer, F. Mendez, J. Watkins, M. Hammer, M. Dannemann, A. Andres, and J. Kelso. Understanding rare and common diseases in the context of human evolution. *Genome Biology*, 17(1):225, dec 2016.

[193] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[194] B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt. A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, jan 2013.

[195] M. D. Rausher. The Measurement of Selection on Quantitative Traits: Biases Due to Environmental Covariances between Traits and Fitness. *Evolution*, 46(3):616, jun 1992.

[196] D. E. Reich and D. B. Goldstein. Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences*, 95(14):8119–8123, jul 1998.

[197] D. E. Reich and D. B. Goldstein. Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, 20(1):4–16, jan 2001.

[198] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502–510, sep 2001.

[199] N. Risch. Linkage strategies for genetically complex traits. I. Multilocus models. *American journal of human genetics*, 46:222–228, 1990.

[200] N. Risch. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *American journal of human genetics*, 46:242–253, 1990.

[201] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)*, 273(5281):1516–7, sep 1996.

[202] N. J. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–856, jun 2000.

[203] A. Robertson. The effect of selection against extreme deviants based on deviation or on homozygosis. *Journal of Genetics*, 54(2):236–248, may 1956.

[204] A. Robertson. A mathematical model of the culling process in dairy cattle. *Animal Production*, 8(01):95–108, feb 1966.

[205] M. R. Robinson, N. R. Wray, and P. M. Visscher. Explaining additional genetic variation in complex traits. *Trends in Genetics*, 30:124–132, 2014.

[206] A. Rogers and H. Harpending. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.*, 9(3):552–569, may 1992.

[207] S. Romeo, L. A. Pennacchio, Y. Fu, E. Boerwinkle, A. Tybjaerg-Hansen, H. H. Hobbs, and J. C. Cohen. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature genetics*, 39(4):513–6, apr 2007.

[208] J. S. Sanjak, A. D. Long, and K. R. Thornton. Efficient Software for Multi-marker, Region-Based Analysis of GWAS Data. *G3 (Bethesda, Md.)*, 6(4):1023–30, jan 2016.

[209] J. S. Sanjak, A. D. Long, K. R. Thornton, S. Sherry, T. Brick, and R. Kickpatrick. A Model of Compound Heterozygous, Loss-of-Function Alleles Is Broadly Consistent with Observations from Complex-Disease GWAS Datasets. *PLoS Genetics*, 13(1):e1006573, jan 2017.

[210] J. SCHELLDORFER, P. BÜHLMANN, and S. V. DE GEER. Estimation for High-Dimensional Linear Mixed-Effects Models Using 1-Penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, jun 2011.

[211] N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–9, jun 2009.

[212] R. Sear, N. Allal, I. A. Mcgregor, and R. Mace. Height, Marriage and Reproductive Success in Gambian Women. *Research in Economic Anthropology*, 23:203–224, 2004.

[213] S. R. S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. Wiley, 2006.

[214] P. C. Sham and S. Purcell. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *American journal of human genetics*, 68(6):1527–32, jun 2001.

[215] P. C. Sham and S. M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature reviews. Genetics*, 15(5):335–46, may 2014.

[216] Y. B. Simons, K. Bullaughey, R. R. Hudson, and G. Sella. A model for the genetic architecture of quantitative traits under stabilizing selection. apr 2017.

[217] Y. B. Simons, M. C. Turchin, J. K. Pritchard, and G. Sella. The deleterious mutation load is insensitive to recent population history. *Nature Genetics*, 46(August 2013):220–4, 2014.

[218] V. Skirbekk. Fertility trends by social status. *Demographic Research*, 18:145–180, mar 2008.

[219] M. Slatkin. Heritable variation and heterozygosity under a balance between mutations and stabilizing selection. *Genetical research*, 50:53–62, 1987.

[220] M. Slatkin and R. Lande. Niche Width in a Fluctuating Environment-Density Independent Model. *The American Naturalist*, 110(971):31–55, jan 1976.

[221] C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genetics*, 5(5):e1000477, may 2009.

[222] S. C. Stearns, S. G. Byars, D. R. Govindaraju, and D. Ewbank. Measuring selection in contemporary human populations. *Nature Reviews Genetics*, 11(9):611, aug 2010.

[223] S. C. Stearns, D. R. Govindaraju, D. Ewbank, and S. G. Byars. Constraints on the coevolution of contemporary human males and females. *Proceedings. Biological sciences*, 279(1748):4836–44, dec 2012.

[224] J. R. Stinchcombe, A. F. Agrawal, P. A. Hohenlohe, S. J. Arnold, and M. W. Blows. Estimating nonlinear selection gradients using quadratic regression coefficients: double or nothing? *Evolution*, 62(9):2435–40, sep 2008.

[225] A. Strachan, T and Read. *Human Molecular Genetics* . (New York: Garland Science, Taylor \& Francis Group), 2011.

[226] G. Stulp and L. Barrett. Evolutionary perspectives on human height variation. *Biological Reviews*, 91(1):206–234, feb 2016.

[227] G. Stulp, L. Barrett, F. C. Tropf, and M. Mills. Does natural selection favour taller stature among the tallest people on earth? *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1806), 2015.

[228] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3), mar 2015.

[229] J. a. Tennessen, a. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. a. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, and J. M. Akey. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, 337(6090):64–69, 2012.

[230] R. Thompson. The Estimation of Variance and Covariance Components with an Application when Records are Subject to Culling. *Biometrics*, 29(3):527, sep 1973.

[231] K. R. Thornton. A C++ Template Library for Efficient Forward-Time Population Genetic Simulation of Large Populations. *Genetics*, 198(1):1–21, 2014.

[232] K. R. Thornton, A. J. Foran, and A. D. Long. Properties and Modeling of GWAS when Complex Disease Risk Is Due to Non-Complementing, Deleterious Mutations in Genes of Large Effect. *PLoS Genetics*, 9(2), 2013.

[233] J. A. Todd, N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, R. Bailey, S. Nejentsev, S. F. Field, F. Payne, C. E. Lowe, J. S. Szeszko, J. P. Hafler, L. Zeitels, J. H. M. Yang, A. Vella, S. Nutland, H. E. Stevens, H. Schuilenburg, G. Coleman, M. Maisuria, W. Meadows, L. J. Smink, B. Healy, O. S. Burren, A. A. C. Lam, N. R. Ovington, J. Allen, E. Adlem, H.-T. Leung, C. Wallace, J. M. M. Howson, C. Guja, C. Ionescu-Tîrgovite, M. J. Simmonds, J. M. Heward, S. C. L. Gough, D. B. Dunger, L. S. Wicker, and D. G. Clayton. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics*, 39(7):857–64, jul 2007.

[234] F. C. Tropf, G. Stulp, N. Barban, P. M. Visscher, J. Yang, H. Snieder, and M. C. Mills. Human fertility, molecular genetics, and natural selection in modern societies. *PLoS One*, 10(6):e0126821, 2015.

[235] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, mar 2009.

[236] J. Y. Tung, C. B. Do, D. A. Hinds, A. K. Kiefer, J. M. Macpherson, A. B. Chowdry, U. Francke, B. T. Naughton, J. L. Mountain, A. Wojcicki, and N. Eriksson. Efficient replication of over 180 genetic associations with self-reported medical data. *PloS one*, 6(8):e23473, jan 2011.

[237] M. Turelli. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology*, 25(2):138–193, apr 1984.

[238] J.-Y. Tzeng and H. D. Bondell. A comprehensive approach to haplotype-specific analysis by penalized likelihood. *European Journal of Human Genetics*, 18(1):95–103, jan 2010.

[239] L. Ulizzi, Terrenato, and L. Natural selection associated with birth weight. VI. Towards the end of the stabilizing component. *Annals of Human Genetics*, 56(2):113–118, may 1992.

[240] L. H. Uricchio, R. Torres, J. S. Witte, and R. D. Hernandez. Population genetic simulations of complex phenotypes with implications for rare variant association tests. *Genetic epidemiology*, 39(1):35–44, jan 2015.

[241] L. H. Uricchio, J. S. Witte, and R. D. Hernandez. Selection and explosive growth may hamper the performance of rare variant association tests. Technical report, mar 2015.

[242] L. H. Uricchio, N. A. Zaitlen, C. J. Ye, J. S. Witte, and R. D. Hernandez. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Research*, may 2016.

[243] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51, feb 2001.

[244] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS

discovery. *American journal of human genetics*, 90(1):7–24, jan 2012.

[245] P. M. Visscher, M. E. Goddard, E. M. Derks, and N. R. Wray. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Molecular psychiatry*, 17(5):474–85, may 2012.

[246] P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era–concepts and misconceptions. *Nature reviews. Genetics*, 9(4):255–66, apr 2008.

[247] P. M. Visscher and D. Posthuma. Statistical Power to Detect Genetic Loci Affecting Environmental Sensitivity. *Behavior Genetics*, 40(5):728–733, sep 2010.

[248] P. M. P. Visscher, G. Hemani, A. A. A. E. Vinkhuyzen, G.-B. Chen, S. S. H. Lee, N. R. N. Wray, M. M. E. Goddard, J. Yang, L. Hindorff, P. Sethupathy, H. Junkins, E. Ramos, J. Mehta, T. Manolio, F. Collins, N. Cox, D. Goldstein, L. Hindorff, J. Yang, B. Benyamin, B. McEvoy, S. Gordon, A. Henders, S. S. H. Lee, N. R. N. Wray, M. M. E. Goddard, P. M. P. Visscher, I. Deary, J. Yang, G. Davies, S. Harris, A. Tenesa, S. S. H. Lee, J. Yang, M. M. E. Goddard, P. M. P. Visscher, N. R. N. Wray, S. S. H. Lee, S. Ripke, B. Neale, S. Faraone, S. Purcell, H. So, M. Li, P. Sham, F. Dudbridge, E. Stahl, D. Wegmann, G. Trynka, J. Gutierrez-Achury, R. Do, D. Benjamin, D. Cesarini, M. van der Loos, C. Dawes, P. Koellinger, T. Meuwissen, B. Hayes, M. M. E. Goddard, J. Yang, S. S. H. Lee, M. M. E. Goddard, P. M. P. Visscher, H. Patterson, R. Thompson, J. Williams, J. Blangero, P. Sham, S. Cherny, S. Purcell, J. Hewitt, F. Rijsdijk, J. Hewitt, P. Sham, P. Sham, S. Purcell, P. M. P. Visscher, J. Hopper, A. A. A. E. Vinkhuyzen, N. R. N. Wray, J. Yang, M. M. E. Goddard, P. M. P. Visscher, J. Yang, T. Manolio, L. Pasquale, E. Boerwinkle, N. Caporaso, E. Reeve, A. Robertson, K. Koots, J. Gibson, W. Chen, G. Abecasis, M. M. E. Goddard, D. Speed, G. Hemani, M. Johnson, D. Balding, P. M. P. Visscher, J. Yang, M. M. E. Goddard, N. R. N. Wray, J. Yang, B. Hayes, A. Price, M. M. E. Goddard, B. Devlin, K. Roeder, P. M. P. Visscher, S. S. H. Lee, D. Harold, D. Nyholt, M. M. E. Goddard, and K. Zondervan. Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genetics*, 10(4):e1004269, apr 2014.

[249] J. Weeden, M. J. Abrams, M. C. Green, and J. Sabini. Do high-status people really have fewer children? *Human Nature*, 17(4):377–392, dec 2006.

[250] R. K. Weersma, P. C. F. Stokkers, I. Cleynen, S. C. S. Wolfkamp, L. Henckaerts, S. Schreiber, G. Dijkstra, A. Franke, I. M. Nolte, P. Rutgeerts, C. Wijmenga, and S. Vermeire. Confirmation of multiple Crohn's disease susceptibility loci in a large Dutch-Belgian cohort. *The American journal of gastroenterology*, 104(3):630–8, mar 2009.

[251] W.-H. Wei, G. Hemani, and C. S. Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733, 2014.

[252] T. Wellcome, T. Case, and C. Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(June):661–678, 2007.

[253] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(December 2013):1001–1006, 2014.

[254] J. Wessel and M. O. Goodarzi. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature Communications*, 6:5897, 2015.

[255] N. R. Wray, S. M. Purcell, and P. M. Visscher. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biology*, 9(1):e1000579, 2011.

[256] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.

[257] S. Wright. Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics*, 6(2):111–23, mar 1921.

[258] S. Wright. Systems of Mating. II. the Effects of Inbreeding on the Genetic Composition of a Population. *Genetics*, 6(2):124–43, mar 1921.

[259] S. Wright. Systems of Mating. III. Assortative Mating Based on Somatic Resemblance. *Genetics*, 6(2):144–61, mar 1921.

[260] S. Wright. Systems of Mating. IV. the Effects of Selection. *Genetics*, 6(2):162–6, mar 1921.

[261] S. Wright. Systems of Mating. V. General Considerations. *Genetics*, 6(2):167–78, mar 1921.

[262] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, 2011.

[263] J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, and P. M. Visscher. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, advance on, aug 2015.

[264] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–9, jul 2010.

[265] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, jan 2011.

[266] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, jan 2011.

[267] J. Yang, R. J. F. Loos, J. E. Powell, S. E. Medland, E. K. Speliotes, D. I. Chasman, L. M. Rose, G. Thorleifsson, V. Steinthorsdottir, R. Mägi, L. Waite, A. Vernon Smith, L. M. Yerges-Armstrong, K. L. Monda, D. Hadley, A. Mahajan, G. Li, K. Kapur, V. Vitart, J. E. Huffman, S. R. Wang, C. Palmer, T. Esko, K. Fischer, J. Hua Zhao, A. Demirkan, A. Isaacs, M. F. Feitosa, J. Luan, N. L. Heard-Costa, C. White, A. U. Jackson, M. Preuss, A. Ziegler, J. Eriksson, Z. Kutalik, F. Frau, I. M. Nolte, J. V. Van Vliet-Ostaptchouk, J.-J. Hottenga, K. B. Jacobs, N. Verweij, A. Goel, C. Medina-Gomez, K. Estrada, J. Lynn Bragg-Gresham, S. Sanna, C. Sidore, J. Tyrer, A. Teumer, I. Prokopenko, M. Mangino, C. M. Lindgren, T. L. Assimes, A. R. Shuldiner, J. Hui, J. P. Beilby, W. L. McArdle, P. Hall, T. Haritunians, L. Zgaga, I. Kolcic, O. Polasek, T. Zemunik, B. A. Oostra, M. Juhani Junttila, H. Grönberg, S. Schreiber, A. Peters, A. A. Hicks, J. Stephens, N. S. Foad, J. Laitinen, A. Pouta, M. Kaakinen, G. Willemsen, J. M. Vink, S. H. Wild, G. Navis, F. W. Asselbergs, G. Homuth, U. John, C. Iribarren, T. Harris, L. Launer, V. Gudnason, J. R. O'Connell, E. Boerwinkle, G. Cadby, L. J. Palmer, A. L. James, A. W. Musk, E. Ingelsson, B. M. Psaty, J. S. Beckmann, G. Waeber, P. Vollenweider, C. Hayward, A. F. Wright, I. Rudan, L. C. Groop, A. Metspalu, K. Tee Khaw, C. M. van Duijn, I. B. Borecki, M. A. Province, N. J. Wareham, J.-C. Tardif, H. V. Huikuri, L. Adrienne Cupples, L. D. Atwood, C. S. Fox, M. Boehnke, F. S. Collins, K. L. Mohlke, J. Erdmann, H. Schunkert, C. Hengstenberg, K. Stark, M. Lorentzon, C. Ohlsson, D. Cusi, J. A. Staessen, M. M. Van der Klauw, P. P. Pramstaller, S. Kathiresan, J. D. Jolley, S. Ripatti, M.-R. Jarvelin, E. J. C. de Geus, D. I. Boomsma, B. Penninx, J. F. Wilson, H. Campbell, S. J. Chanock, P. van der Harst, A. Hamsten, H. Watkins, A. Hofman, J. C. Witteman, M. C. Zillikens, A. G. Uitterlinden, F. Rivadeneira, M. Carola Zillikens, L. A. Kiemeney, S. H. Vermeulen, G. R. Abecasis, D. Schlessinger, S. Schipf, M. Stumvoll, A. Tönjes, T. D. Spector, K. E. North, G. Lettre, M. I. McCarthy, S. I. Berndt, A. C. Heath, P. A. F. Madden, D. R. Nyholt, G. W. Montgomery, N. G. Martin, B. McKnight, D. P. Strachan, W. G. Hill, H. Snieder, P. M. Ridker, U. Thorsteinsdottir, K. Stefansson, T. M. Frayling, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher. FTO genotype is associated with phenotypic variability of body mass index. *Nature*, 490(7419):267–272, sep 2012.

[268] J. Yang, M. N. Weedon, S. Purcell, G. Lettre, K. Estrada, C. J. Willer, A. V. Smith, E. Ingelsson, J. R. O'Connell, M. Mangino, R. Mägi, P. A. Madden, A. C. Heath, D. R. Nyholt, N. G. Martin, G. W. Montgomery, T. M. Frayling, J. N. Hirschhorn, M. I. McCarthy, M. E. Goddard, P. M. Visscher, and GIANT Consortium. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, jul 2011.

[269] K. Yu, Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg, N. Caporaso, P. Kraft, and N. Chatterjee. Pathway analysis by adaptive combination of P-values. *Genetic epidemiology*, 33(8):700–9, dec 2009.

[270] Y. Yu, X. Zheng, M. Marchetti-Bowick, and E. P. Xing. Minimizing Nonconvex Non-Separable Functions. *18th International Con- ference on Artificial Intelligence and Statistics (AISTATS)*, 38:9–12, 2015.

[271] D. V. Zaykin, A. Pudovkin, and B. S. Weir. Correlation-Based Inference for Linkage Disequilibrium With Multiple Alleles. 2008.

[272] X. Zhang and W. G. Hill. Mutation-Selection Balance for Environmental Variance. *The American Naturalist*, 171(3):394–399, mar 2008.

[273] X.-S. Zhang and W. G. Hill. Joint Effects of Pleiotropic Selection and Stabilizing Selection on the Maintenance of Quantitative Genetic Variation at Mutation-Selection Balance. *Genetics*, 162(1):459–471, sep 2002.

[274] X.-S. Zhang and W. G. Hill. Evolution of the environmental component of the phenotypic variance: stabilizing selection in changing environments and the cost of homogeneity. *Evolution*, 59(6):1237–44, jun 2005.

[275] Z. Zhu, A. Bakshi, A. Vinkhuyzen, G. Hemani, S. Lee, I. Nolte, J. vanVliet Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, W. Hill, B. Weir, M. Goddard, P. Visscher, and J. Yang. Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *The American Journal of Human Genetics*, pages 377–385, 2015.

[276] Z. Zhu, Z. Zheng, F. Zhang, Y. Wu, M. Trzaskowski, R. Maier, M. Robinson, J. McGrath, P. Visscher, N. Wray, and J. Yang. Causal associations between risk factors and common diseases inferred from GWAS summary data. *doi.org*, page 168674, jul 2017.

[277] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[278] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111:E455–64, 2014.

Appendix A

Supplementary information for all chapters

A.1 Chapter 2 supplementary texts

A.1.1 Population genetic modeling of complex traits

Forward-in-time simulations, as in [1, 141, 240, 31] and [183, 114, 21, 173] , explicitly model the allele frequencies and effect sizes of mutations in a genomic region with neutral sites, selected sites and recombination. Within the forward simulation framework, methods differ in their approach to assigning phenotypes to particular genotypes. One approach, based on the work of [59], models a mutation's effect on fitness as a pleiotropic consequence of its effect on trait values. In that model, by specifying a parameter τ, the user establishes the shape of relationship between the fitness effects and expected trait effects of variants [1, 141] and [166, 173]. There is another term, ϵ, which adds noise to the fitness-trait relationship; together τ and ϵ determine the correlation between fitness effects and trait effects. A related

approach that builds off of much earlier work by [108], models trait and fitness effects as coming from a bivariate gamma distribution with a specified correlation parameter ρ [31]. In both approaches the disease-trait itself is not a component of fitness and thus standing variation may be dominated by occasional large trait effect mutations with small fitness effects that can reach intermediate frequency. The extent to which this occurs is dependent on the degree of correlation between fitness and trait effects. Furthermore, both approaches indicate that an intermediate degree of correlation between complex disease traits and fitness is most plausible [1, 31, 150](although [150] is not a population genetic simulation based implementation of [59], a similar conclusion was reached.)

The approach in [232] is similar to typical models [85, 130, 43, 116, 65] and [237] of selection on quantitative traits where phenotype is the sum of genetic and environmental components and is subjected to Gaussian stabilizing selection. A key difference between [232] and the typical quantitative trait models is that all causal mutations are unconditionally deleterious and the gene action model exhibits gene-based recessivity, i.e, allelic non-complementation. While the work in [232] presented a new genetic model, it was limited because only that model was explored under a single demographic scenario (constant sized population). To our knowledge, there has not been a joint analysis of the effect of genetic model and demography on the predicted outcomes of GWAS. Thus, we extend the approach of [232] by including a model of recent population expansion and a set of genetic models.

A.1.2 Heritability and genetic load under population growth

Before deeply exploring the predicted genetic architecture of a trait under each model, we looked at two key mean values: total genetic variance and load. The genetic variance underlying a trait is, in part, determined by the outcome of mutation-selection balance. Approximations for the expected genetic variance under models of stabilizing selection with

Gaussian mutational effect sizes and additive gene action have been derived previously[112] and [237].

According to the house-of-cards(HOC) approximation, when the variance in mutational effect sizes is large compared to total genetic variance, the genetic variance will be dependent only on the mutation rate, μ, and the intensity of stabilizing selection, $1/\sigma_s^2$. Here, μ refers to the total mutation rate in the "gene region" (per gamete, per generation), with mutations arising according to an infinitely-many sites scheme [113]. In a diploid species, $V_G \approx 4\mu\sigma_s^2$ for an additive trait, and $V_G \approx 2\mu\sigma_s^2$ for a recessive trait [237, 219]. By keeping σ_s^2 and μ constant, we can modulate the broad sense heritability ($H^2 = (V_G)/(V_G + V_E)$) by changing the environmental variance, $V_E = \sigma_e^2$. These approximations are expected to hold for arbitrary probability distributions of mutational effect sizes [27]; however all distributions discussed in [27] are reflected about zero. Here, as in [232], we draw the effect sizes of causal mutation from a standard exponential distribution, modeling unconditionally deleterious mutations. As previously shown in [232], heritability approaches the value expected under the HOC approximation when the variance in effect sizes (λ^2) is large (A.1).

Previous work, under additive genetic models, on the impact of population growth on the genetic architecture of complex traits suggests that mean heritability is constant under growth [141, 217]. We confirm this in A.1, showing that mean broad-sense heritability, $H^2 = (V_G)/(V_P)$, initially increases as λ, the mean effect of a new deleterious mutation, increases and then approximately reaches the same level as models with constant population size. This general trend is observed under each genetic model, but the MR model is qualitatively different in its behavior under population growth. The MR model predicts a broad sense heritability under growth of about 90% of constant sized population levels when $\lambda = 0.01$, and 50% when $\lambda = 0.5$ (A.2).

The degree to which recent demographic history has impacted the distribution of genetic variance over risk allele frequency in human populations is still unclear. One line of evidence

may come from the study of genetic load in humans. If fitness effects and trait effects of variants are correlated, then the composition of the genetic load of deleterious mutations in the population is highly relevant to the genetic architecture of that trait. Comparisons between populations with different demographic histories provide insight into the impact of demography on genetic load. The influential study of [142] found there to be both more non-synonymous relative to synonymous variants and a higher average number of homozygous non-synonymous sites in European Americans than African Americans. Later studies showed empirically and through simulations that the mean allele frequency of deleterious mutations is not impacted by recent demographic history [217, 71, 49]. Simulations presented by [178] suggest that load is expected to increase during a range expansion, without an increase in mean frequency of deleterious alleles, due to an increase in homozygosity at deleterious recessive sites. By invoking expansion load theory and empirical data from multiple human populations, [92] argue that load is increased in non-African populations due to serial bottlenecks during range expansion after the out of Africa event. While arguments about genetic load are sensitive to choice of metric [79], and the empirical evidence supporting one view or another is still lacking, it does appear that any differences between current human populations due to past population bottlenecks is likely to be small.

Our results show that genetic burden (load), as measured by the average relative deviation from optimum fitness, of the population is also unaffected by recent population history under the AC model (A.3), as shown in [217, 141, 8]. We find this same behavior under the GBR model, but not under the MR model. Under rapid population expansion, the load decreases slightly (at most 2%). As λ increases the load increases under the additive model in both demographic scenarios. Load is effectively constant over the range of λ under the GBR model. Increasing the heritability of the trait decreased the magnitude of the genetic load, but had no interaction with the effects of demography or increasing mean effect size.

We also find that the dynamics of load under more complex demographic models involving

multiple bottlenecks and recent growths behaves as expected from previous literature[119] and [217, 8]. The Tennessen[229] demographic model is characterized by an ancient expansion, two recent bottlenecks and subsequent rapid population expansion. In models with strong recessive selection the load should increase immediately after the bottleneck, then decay due to the purging of deleterious alleles in homozygotes [119]. Upon population rexpansion, in models with strong recessive selection, we expect the load to further drop below equilibrium levels [217, 8]. We observe this pattern most clearly in the multiplicative recessive models (cMR or iMR($h = 0.1$)) when λ is large A.18. In A.19 we also show the Burden Ratio (B_r) [8] calculated relative to a model with no bottleneck or growth, which provides a clear visualization of the aformentioned dynamics. Further, in agreement with Simons et al [217], the number of deleterious alleles per individual decreases following the bottleneck under strong recessive selection. This results in an increase in the B_r calculated using the number of alleles A.20 following the bottleneck. We note that because B_r is calculated by comparing two sets of simulations, it may not be exactly equal to one when comparing time points at which the two simulations share identical demographic histories. This is especially the case for the B_r calculated using fixed load which, being small, has high relative variance.

From these results we can conclude that the AC and GBR models are fairly comparable with respect to total genetic variance and genetic load. Therefore, the remaining differences, which we highlight in the main text, between the AC and GBR model can be attributed to the fine scale composition of the genetic variance in the population. In other words, the AC and GBR model differ in how the genetic variance and load are accounted for, despite the total amounts being roughly equivalent. However, the MR model is qualitatively different in its behavior under population growth. This makes comparison to the MR model somewhat difficult. However, there are important reasons to explore it further. Based on first principles, the application of the MR model in simulation of a single gene region is inappropriate. However, it would be appropriate for simulating each mutation as a variant of a distinct functional genomic unit. It is also the most analytically tractable model of

recessivity in population genetics, and as such it is our best reference point for comparison to the GBR model.

A.1.3 The approximate distribution of fitness effects

Here, fitness is a function of phenotype and so the distribution of fitness effects of newly arising mutations is dependent on the state of the population. However, we can achieve an approximate result by assuming that large effect mutations are rare and considering the effect of a new mutation on an unaffected genetic background [273]. This approximation is likely to be most accurate for large values of λ. In this case, we can find the exact distribution of fitness effects given the distribution of trait effects by a simple change of variables. We will assume a fitness model where fitness is 1, $1 - sh$ and $1 - s$ for 0,1, or 2 copies of the deleterious mutation. Although, excepting complete recessive selection, the expected allele frequency trajectories are determined by sh we focus on the distribution of s and emphasize that selection is still recessive ($h < 0.5$) under the additive phenotypic model(A.15).

Let $f_z(z)$ describe the density of mutant phenotypic effects and $s(z)$ describe the fitness of a homozygote for a deleterious allele. We can find the density ($f_s(s)$) and cumulative distribution ($F_s(s)$) of s by change of variables.

$$f_z(z) \sim Exp(\frac{1}{\lambda})$$

$$s(z) = 1 - e^{-\frac{(2*z)^2}{2}}$$

$$f_s(s) = f_z(s^{-1}(s))\frac{d}{ds}s^{-1}(s)$$

$$s^{-1}(s) = \frac{\sqrt{-2\log(1-s)}}{2}$$

$$f_s(s) = \frac{\frac{e^{-\frac{\sqrt{-\log(1-s)}}{\sqrt{2}\lambda}}}{\lambda}}{2\sqrt{2}(1-s)\sqrt{-\log(1-s)}}$$

$$F_s(s) = 1 - e^{\frac{\sqrt{-log(1-s)}}{\sqrt{(2)}\lambda}}$$

We checked this result via simple sampling in R [193] (A.22), using the population size scaled parameter $2Ns$. Across the range of λ simulated the distribution of fitness effects spans multiple selective regimes. In all cases there will be some weakly and strongly selected mutations. When $\lambda < 0.1$ there will also be a considerable proportion of nearly neutral mutations. Again, we emphasize that the degree of recessivity can have an important effect here, as the distribution of $2Nsh$ will be shifted to the left in A.22. It is also important to observe the appearance of a mass of lethal mutations $s \approx 1$ as λ gets larger in A.23. These approximate distributions of fitness effects reveal the relative impact of mutations in different selective regimes in each model. In general, the simulated frequency spectra (see A.7, A.8, and A.9) and genetic loads (see A.3 and A.18) are in agreement with the expecations under the approximate distribution of fitness effects.

A.1.4 Choice of genetic model effects key population genetic signatures

In this section, we explore how the choice of genetic model impacts the site frequency spectrum for risk variants. Mean genetic load decreases with degree of dominance (A.3). Recessive deleterious mutations segregate to higher frequencies in the population without increasing genetic load (A.5). The MR model demonstrates a slight decrease in load under growth (A.3). Similarly, the additive model has greater skew and kurtosis for both the number of mutations and the genetic value of a gamete over the range of λ and demographic models (A.11 and A.12).The increase skew and kurtosis implies that total genetic load in the additive models is dominated by rare large excursions from the population mean.

Population expansion has been shown to impact the site frequency spectrum[109, 154, 42, 73]. In general, we expect to find an increase in rare private mutations under a rapid population expansion scenario. We find all of our models showcase the expected pattern (A.7), but there are consistent and important differences between models. In A.7, the site frequency spectrum of risk variants from a population sample (n=100) shows a dependency on population growth, mean effect size λ, and genetic model. Population growth increases the proportion of singletons for all genetic models and values of λ. Increasing the value λ increases the proportion of singletons in each genetic model and demographic scenario, but the increase is qualitatively dependent on genetic model and independent of demography. The recessive models show the strongest dependence on λ. When λ is small the recessive models show fewer singletons as compared to the additive model, but as λ increases the relative proportion of singletons between recessive and additive genetic model increases. When the value of λ is large, the recessive models show more singletons than the additive model. The GBR model shows more singletons than the MR model in all cases. A.8 shows the site frequency spectrum for non-risk variants, which demonstrates a dependence on population growth, shifting towards low-frequency sites, but shows no dependence on genetic

model or λ. Since the neutral variant site frequency spectrum is consistent across models we determine that the difference in linked selection between models is not important in the relevant recombination rate regime.

A.1.5 Regression based estimates of genetic variance

For Fig 2.1 and A.4, we performed linear regression of the genetic component of phenotype onto genotypes. This provided an estimate of distribution of genetic variance over risk allele frequency. Under an additive model and Hardy-Weinberg linkage equilibrium (HWLE), these estimates are identical to the classic result $V_G = 2pq\alpha^2$[64]. To demonstrate this we simulated genotype data at 1000 independent markers for 5000 individuals in R[193]. Population frequencies were drawn from the constant population size neutral allele frequency distribution. Allele counts for individuals were binomial samples of size two with probability of success equal to the population frequency of the minor allele. Effects sizes were sampled randomly from an exponential distribution with mean $\lambda = 0.1$, to mimic our simulations in the main text. Regressions were performed only on markers which were not fixed in the sample. A.14 shows the regression estimate as a function of its expected value $2pq\alpha^2$. There is some noise for markers with low total variance explained, which is due to random deviations from HWLE.

A.2 Chapter 2 supplementary figures

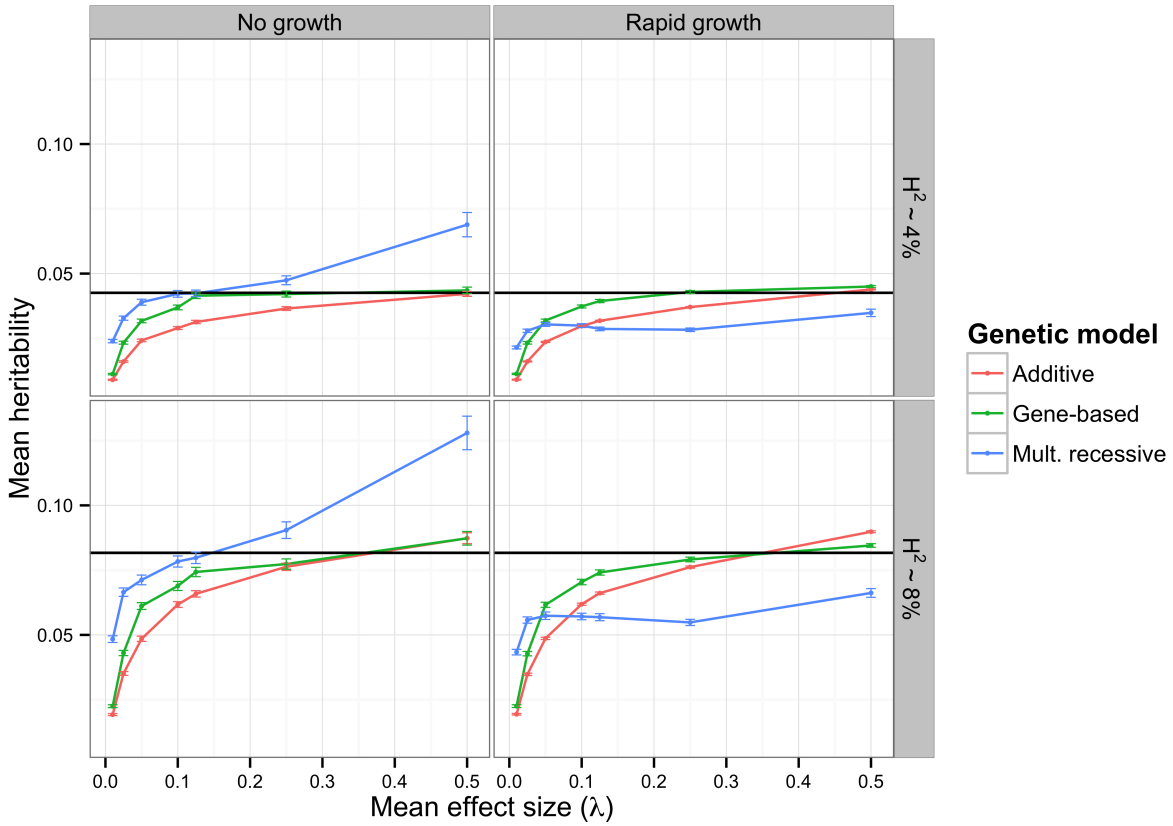

Figure A.1: Broad-sense heritability, $H^2 = (V_G)/(V_P)$, as a function of λ: the mean effect size of a new deleterious mutation, as calculated explicitly from our simulated populations. Data are plotted as the mean across model replicates $\pm$ the standard error of the mean. The solid black horizontal line shows the predicted H^2 under the respective house of cards approximation. The data is grouped by expected level of heritability and demographic scenario. For the additive model model, $H^2 \sim 8\%$ and $H^2 \sim 4\%$ imply environmental standard deviations of $\sigma_e = 0.075$ and $\sigma_e = 0.011$ respectively. For recessive models, $H^2 \sim 8\%$ and $H^2 \sim 4\%$ imply environmental standard deviations of $\sigma_e = 0.053$ and $\sigma_e = 0.075$ respectively. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

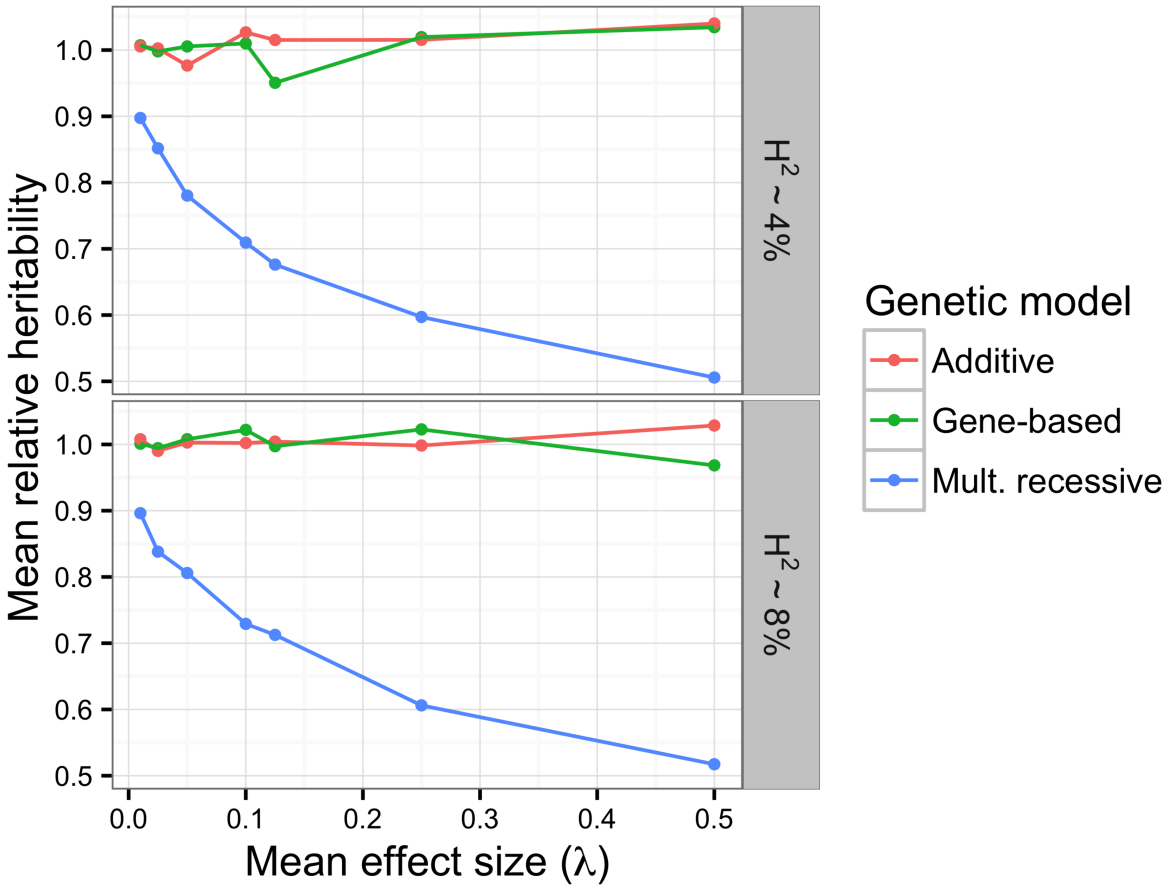

Figure A.2: The y-axis is the ratio of mean broad-sense heritability under recent rapid growth to mean broad sense heritability for a constant-sized population, *e.g.* $\text{Mean}[H^2]_{growth}/\text{Mean}[H^2]_{constant}$. This ratio is plotted as a function of the mean effect size of causative mutations (λ). For co-dominant models, $H^2 \sim 8\%$ and $H^2 \sim 4\%$ imply environmental standard deviations of $\sigma_e = 0.075$ and $\sigma_e = 0.011$ respectively. For recessive models, $H^2 \sim 8\%$ and $H^2 \sim 4\%$ imply environmental standard deviations of $\sigma_e = 0.053$ and $\sigma_e = 0.075$ respectively. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

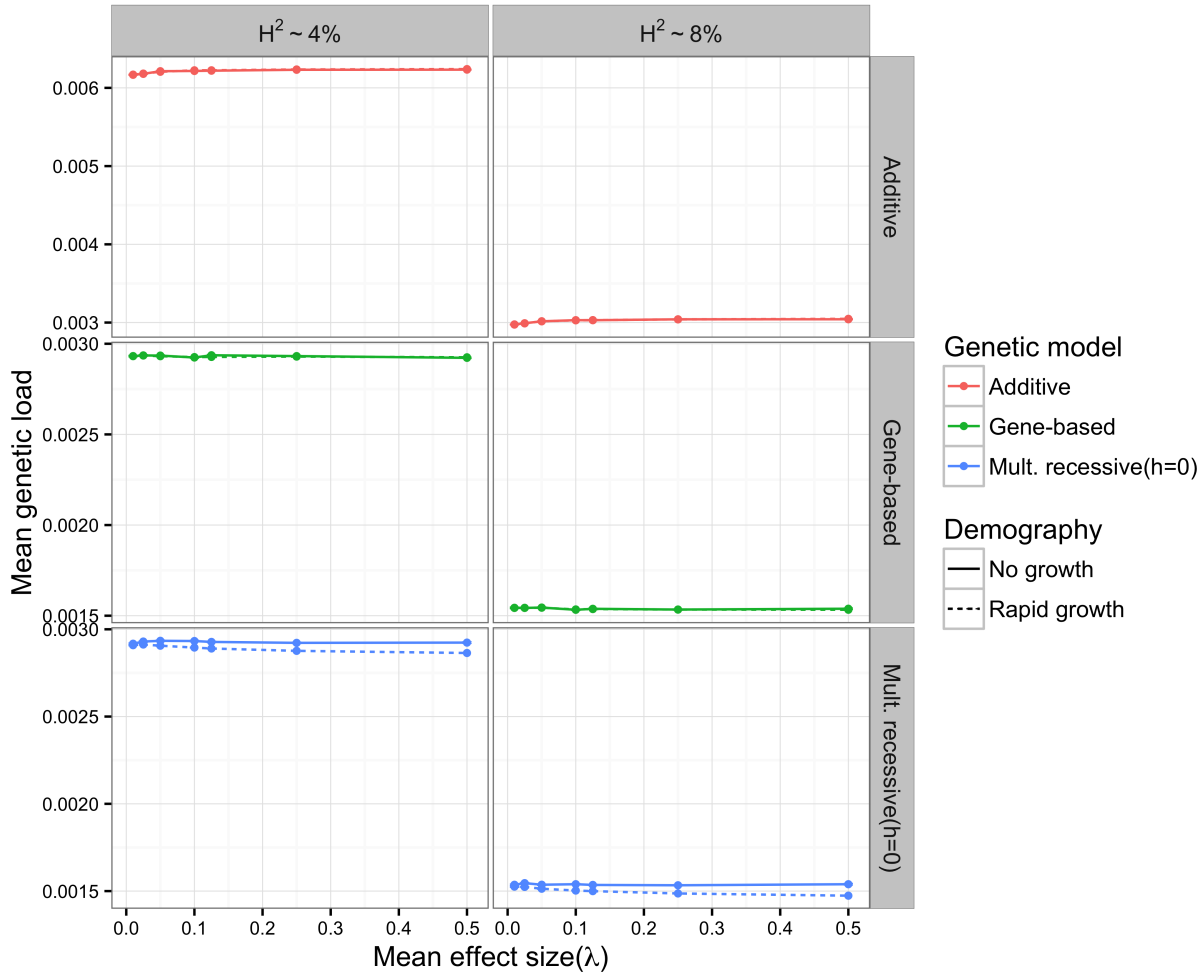

Figure A.3: Genetic load(burden), $L = \frac{w_{opt} - \bar{w}}{w_{opt}}$, as a function of λ: the mean effect size of a new deleterious mutation. Data are plotted as the mean across model replicates $\pm$ the standard error of the mean. Solid curves show values for constant sized population simulations and dashed curves show values for rapid population expansion simulations. The data is grouped by expected level of heritability and genetic model. For the additive model, $H^2 \sim 8\%$ and $H^2 \sim 4\%$ imply environmental standard deviations of $\sigma_e = 0.075$ and $\sigma_e = 0.011$ respectively. For recessive models, $H^2 \sim 8\%$ and $H^2 \sim 4\%$ imply environmental standard deviations of $\sigma_e = 0.053$ and $\sigma_e = 0.075$ respectively. Note the scales of y-axis for each plot. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

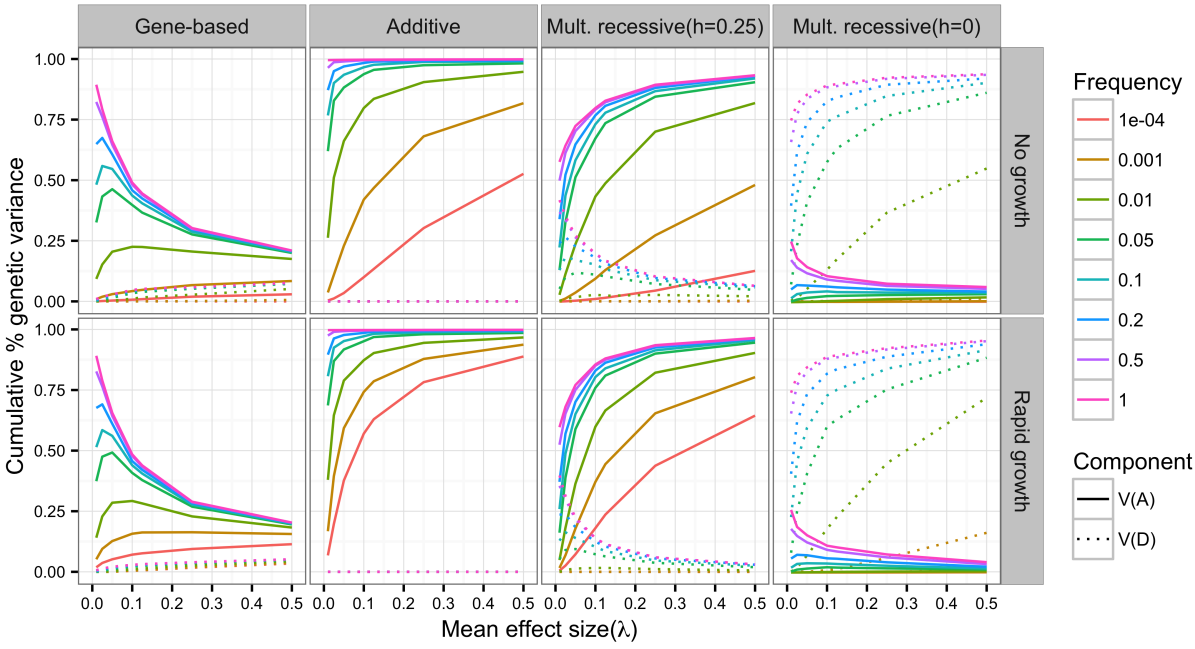

Figure A.4: The percent of cumulative genetic variance explained by additive and dominance effects of variants with frequency less than or equal to a series of frequency values over λ. Shown here are the gene-based (GBR), additive co-dominant (AC), incomplete multiplicative recessive (Mult. recessive ($h = 0.25$); iMR) and complete multiplicative recessive (Mult. recessive ($h = 0$);cMR) models. Solid lines show the additive variance alone and dotted lines show the combined additive and dominance variance. All data shown are for models where $H^2 \sim 0.08$. These particular results are robust to changes H^2 when V_G is not changed, as is the case here. The additive and dominance genetic variance is estimated by the adjusted r^2 of the regression of all markers (and their corresponding dominance encoding) with $MAF \leq x$ onto total genotypic value (see methods for details); data are displayed as the mean of 250 simulation replicates. For each frequency level we calculated the r^2 of a linear regression of genotypes of markers with frequency below that level on to total genetic value and plot it against λ: the mean effects size of a new deleterious mutation. The data are displayed as a mean across model replicates.

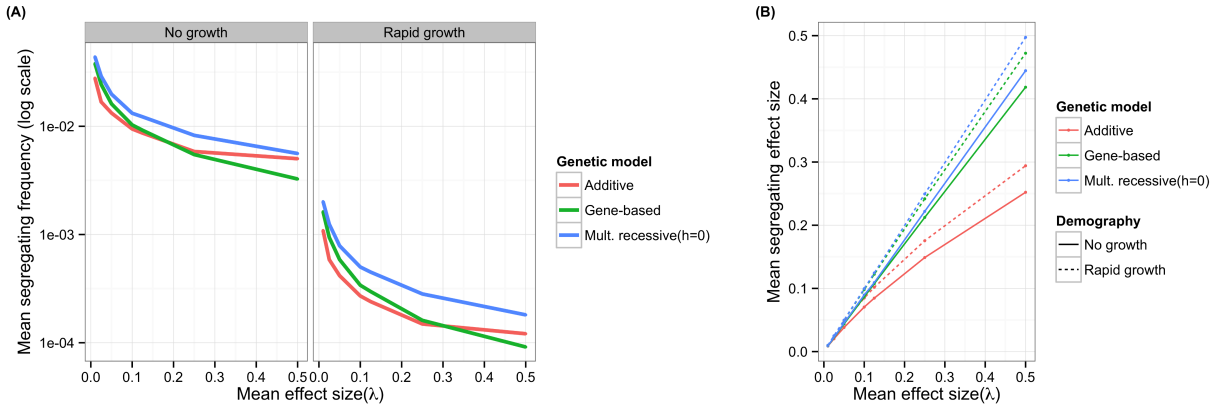

Figure A.5: The mean frequency and B) mean effect size of a segregating risk variant over λ. Note they log10 y-axis scale in A. The mean effect size is the value pulled from the exponential distribution with mean λ, not the fitness effect or the quantitative genetic effect size. The data are calculated for all risk mutations segregating in the simulated populations. Data are plotted as the mean across model replicates. For visual clarity, standard errors are not shown. In panel A, the standard error bars overlap zero under rapid population growth. The data for mean frequency are grouped by demographic scenario; the left panel shows values for constant sized population, the right panel shows values for the rapidly expanded populations. For mean effect size plots the solid curves show the constant sized population data and the dashed curves show the data for the rapidly expanded populations. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

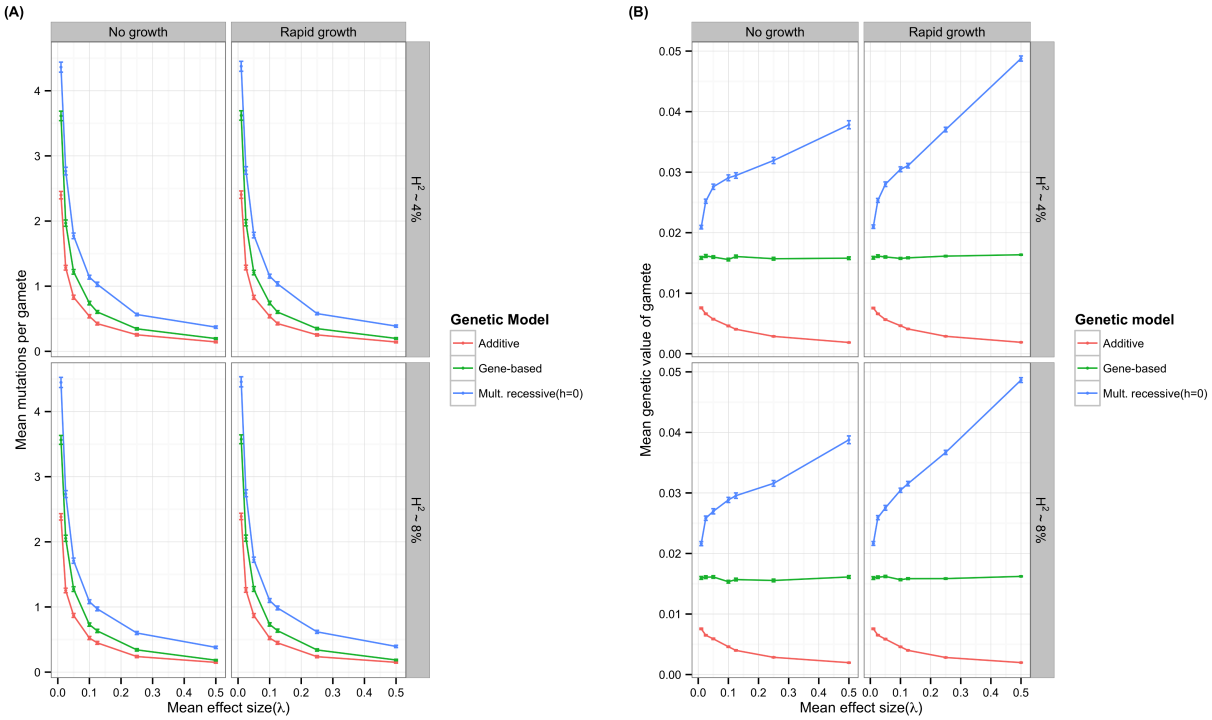

Figure A.6: A) The mean number of deleterious mutations per gamete in the population as a function of λ: the mean effect size of new causative mutation. The data plotted as mean over simulation replicates $\pm se$. The data are calculated for the entire simulated population. B) The mean genetic value of a gamete, i.e. the average sum of mutational effect sizes on a gamete as a function of λ. Data are plotted as the mean across model replicates $\pm$ the standard error of the mean. In the case the gene-based recessive model, this value is also the expected value of the mean phenotype and is accurate within the sampling variance of the mean environmental variate and random pairing of gametes in diploid. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

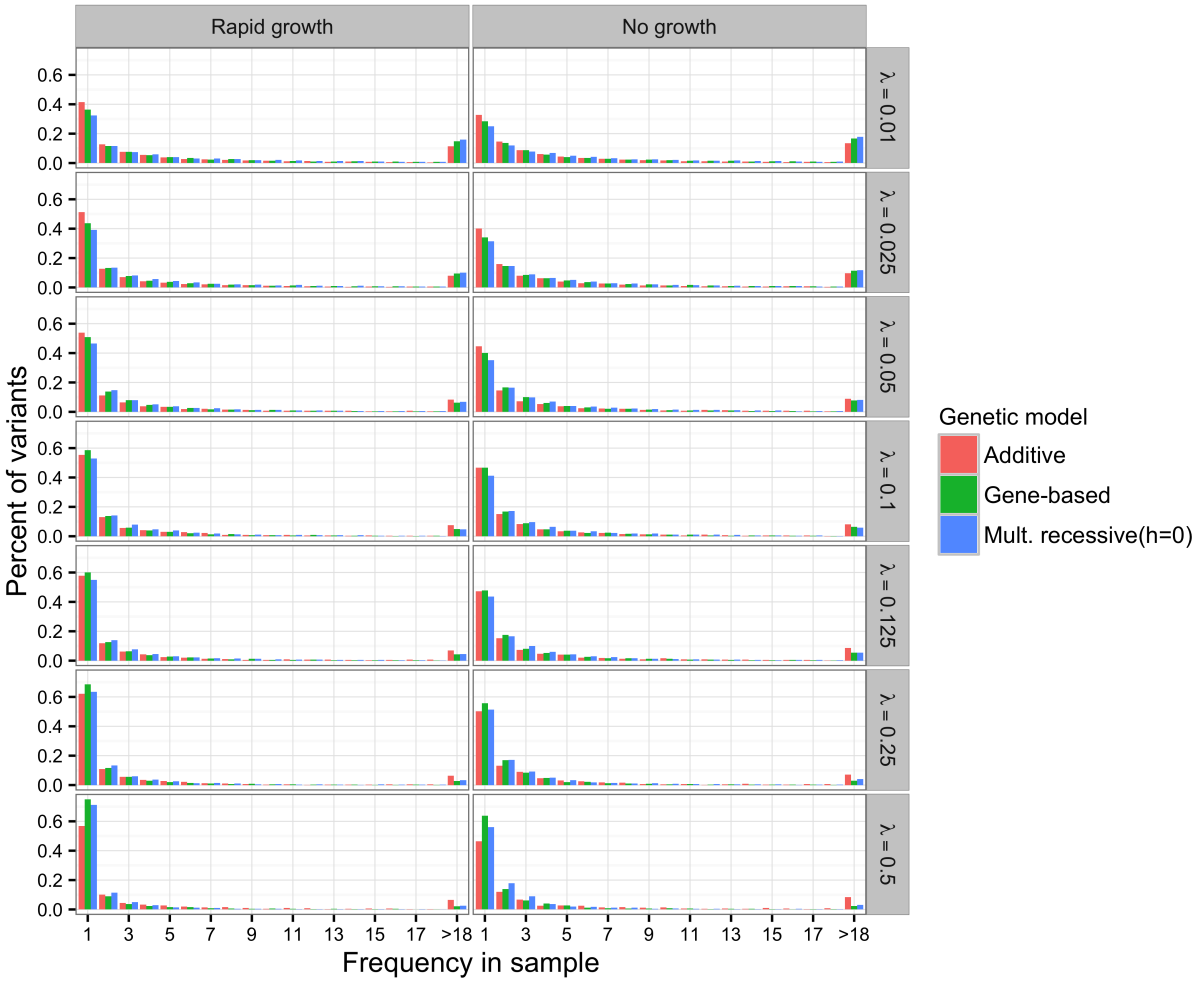

Figure A.7: For a sample $n = 100$ individuals, the relative site frequency spectrum is calculated as the proportion (y-axis) of all polymorphic sites which belong to each frequency class (x-axis). Sites with frequency was above 18 were grouped, into one category to improve visualization. The data are grouped by λ, the mean effect size of a new risk mutation, and the demographic scenario. Data shown are for simulations in which the predicted broad sense heritability is $H^2 \sim 8\%$. Plotted values are the mean proportion across simulation replicates. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

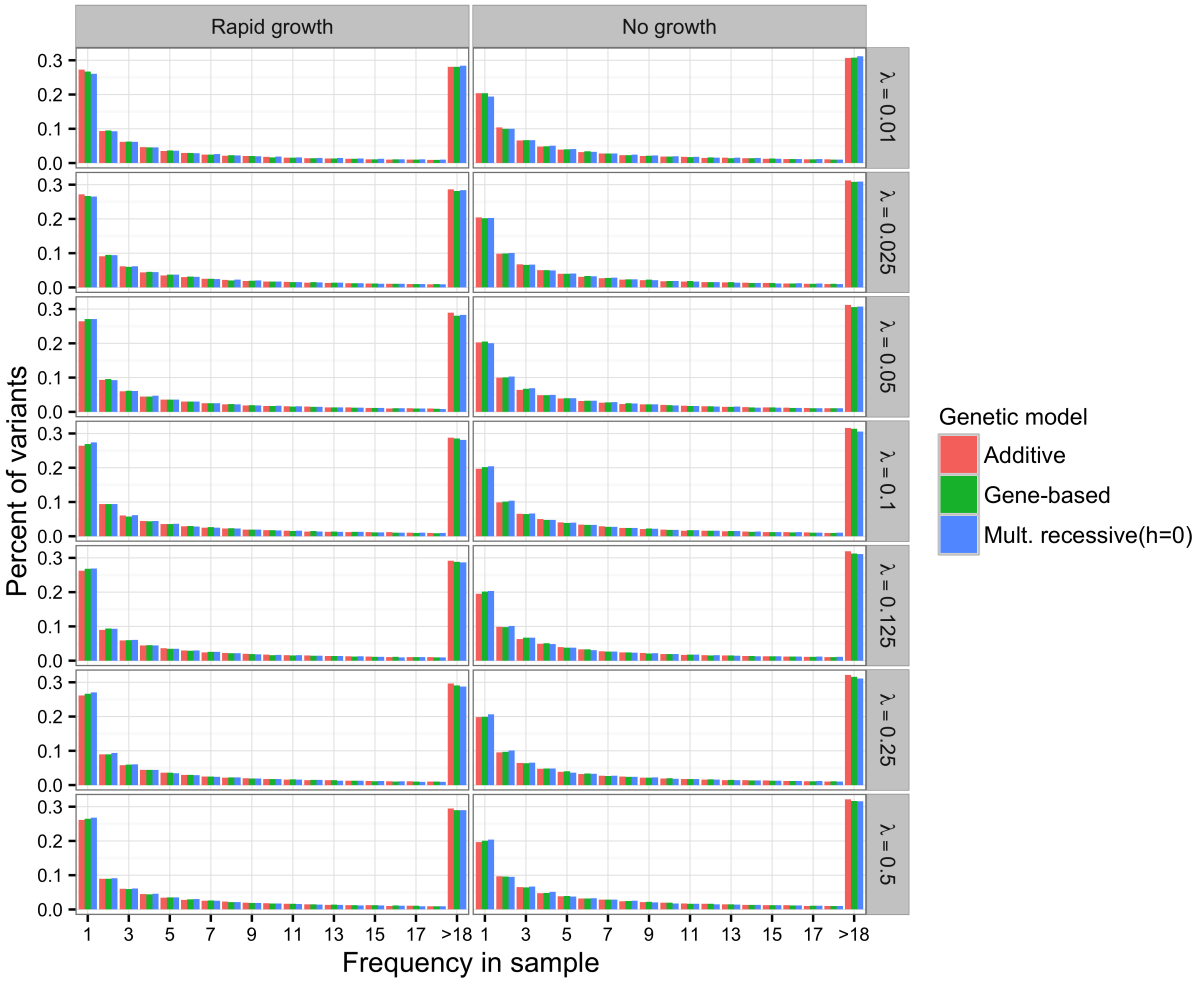

Figure A.8: For a sample $n = 100$ individuals, the relative site frequency spectrum is calculated as the proportion (y-axis) of all polymorphic sites which belong to each frequency class (x-axis). Sites with frequency was above 18 were grouped into one category to improve visualization. The data are grouped by λ, the mean effect size of a new risk mutation, and the demographic scenario. Data shown are for simulations in which the predicted broad sense heritability is $H^2 \sim 8\%$. Plotted values are the mean proportion across simulation replicates. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

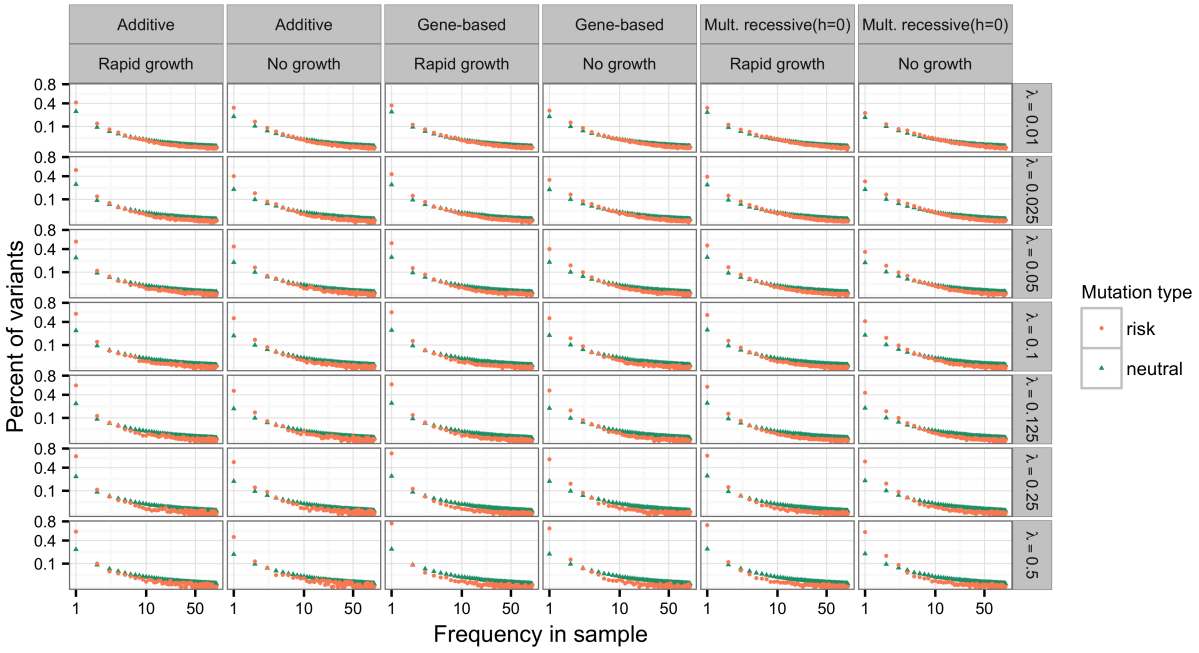

Figure A.9: For a sample $n = 100$ individuals, the relative site frequency spectrum is calculated as the proportion (y-axis) of all polymorphic sites which belong to each frequency class (x-axis). Neutral variants are in orange and risk variants are shown in green. Y-axis is on a square-root scale and X-axis is on a log10 scale to improve visualization. The data are grouped by λ, the mean effect size of a new risk mutation, the demographic scenario and genetic model. Data shown are for simulations in which the predicted broad sense heritability is $H^2 \sim 8\%$. Plotted values are the mean proportion across simulation replicates. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

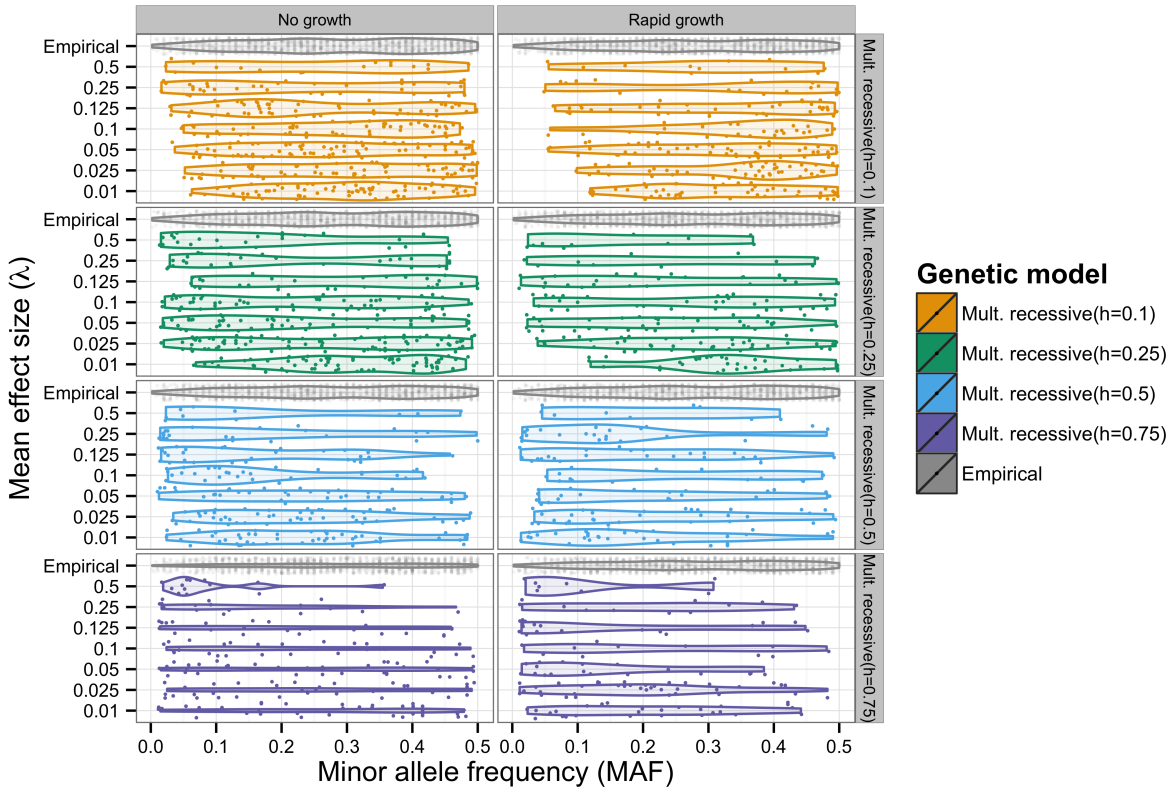

Figure A.10: Horizontal violin plots depict the distribution of minor allele frequencies (MAF) of the most strongly associated single marker in a GWAS. Individual hits are plotted as translucent points and jittered to provide a sense of the total number and density of hits. Each panel contains simulated data pooled across model replicates for each value of λ, with empirical data for comparison. The degree of dominance h was varied from 0.1 to 0.75; perfect co-dominance here is h=1. Empirical data were downloaded from the NHGRI-EBI GWAS database (http://www.ebi.ac.uk/gwas/) on 02/03/2015, diseases and inclusion criteria are as in [255]. In cases where more than one marker was tied for the lowest p-value, one was chosen at random. Simulated data were subjected to ascertainment sampling such that the MAF distribution of all markers on the simulated genotyping chip was uniform. Specific information regarding the empirical data can be obtained in A.1.

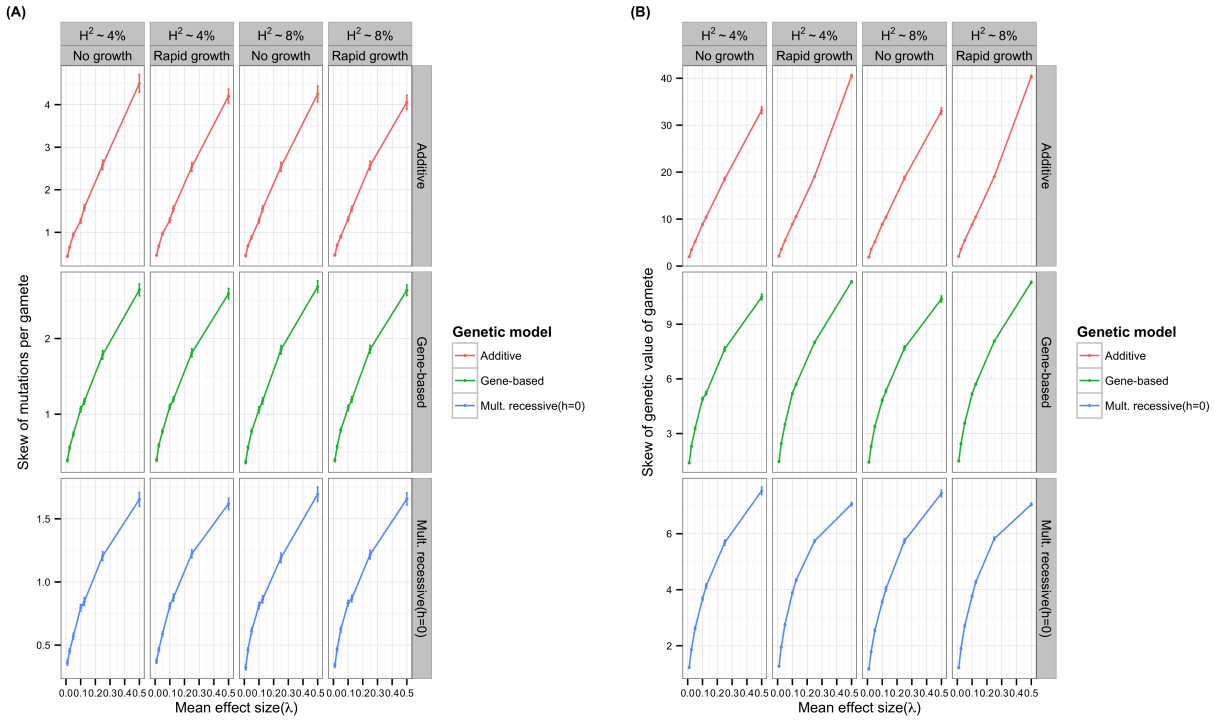

Figure A.11: The skewness A) the number of mutations per gamete and B) the genetic value (sum of mutational effects) of a gamete over λ. The data are calculated for all risk mutations segregating in the simulated populations. Moments were calculated using the boost C++ statistical accumulators library. Data are plotted as the mean across model replicates $\pm$ the standard error of the mean. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

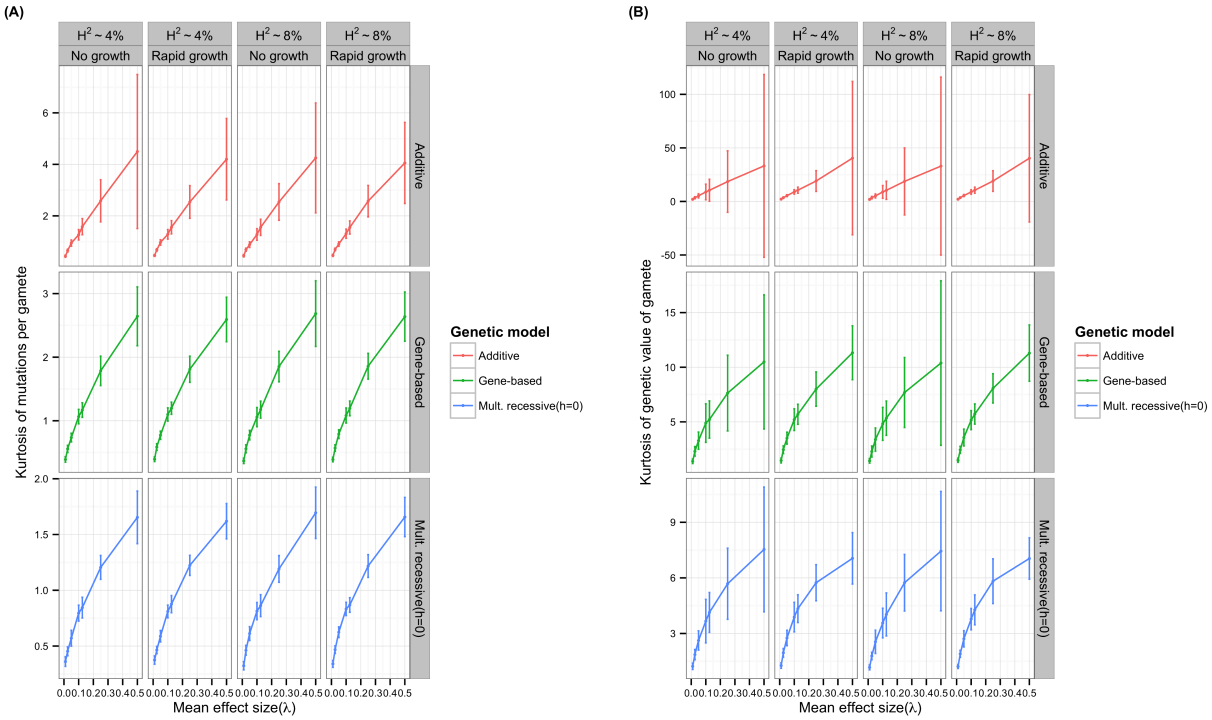

Figure A.12: The kurtosis A) the number of mutations per gamete and B) the genetic value (sum of mutational effects) of a gamete over λ. The data are calculated for all risk mutations segregating in the simulated populations. Moments were calculated using the boost C++ statistical accumulators library. Data are plotted as the mean across model replicates $\pm$ the standard error of the mean. Shown are the additive co-dominant (AC), gene-based (GBR) and complete multiplicative recessive (Mult. recessive ($h = 0$); cMR) models.

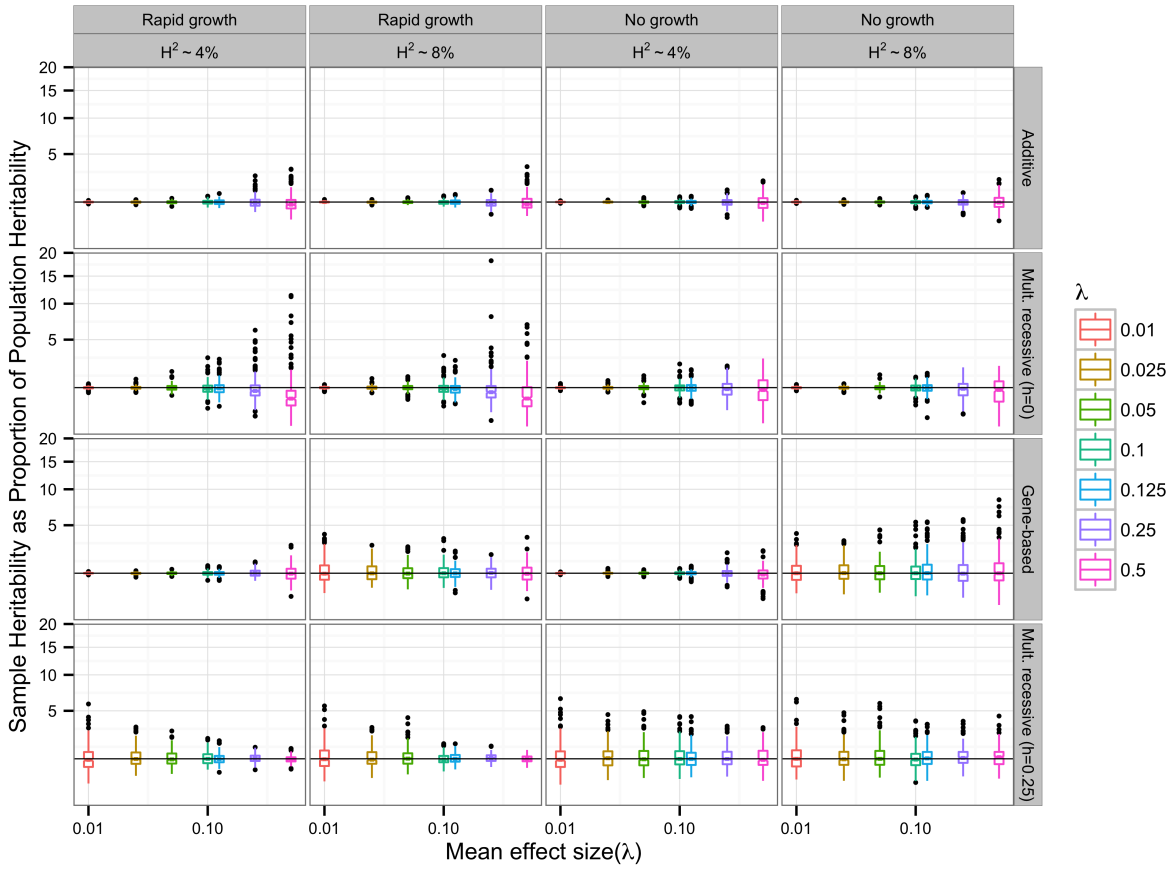

Figure A.13: Broad-sense heritability in a population sample of 6000 as a proportion of population wide broad-sense heritability. Data are grouped by demographic scenario, model and λ. The arbitrary dominance coefficient is parameterized such that $h = 0$ is complete recessivity, $h = 1$ would be exact co-dominance and $h = 2$ would be complete dominance. Multiplicative recessive (MR) models shown are only for $h = 0.25$.

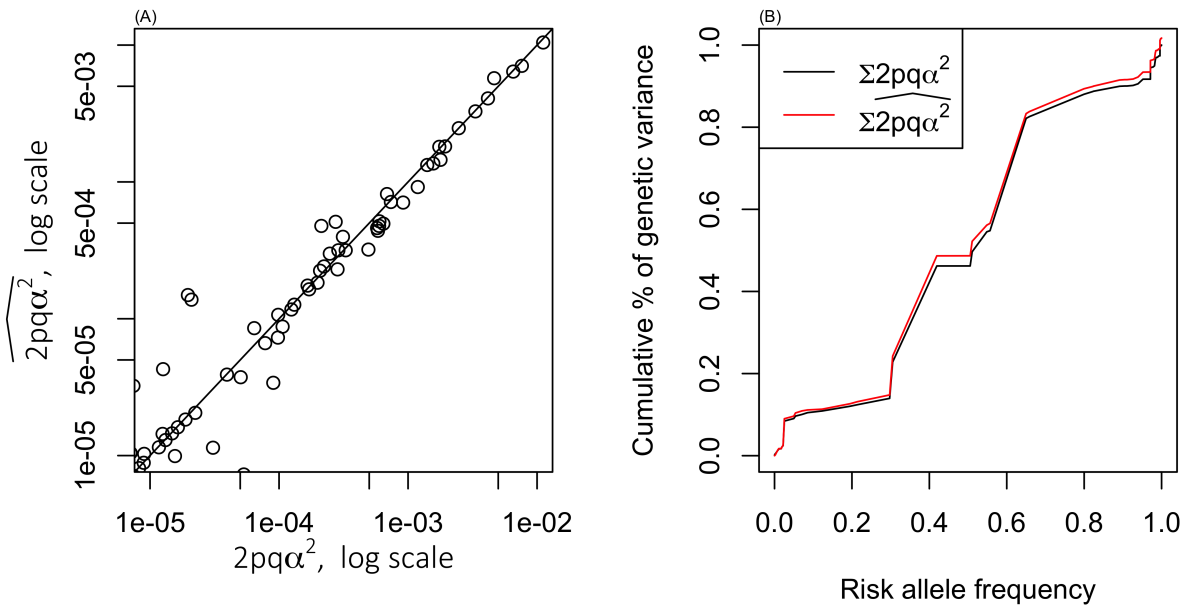

Figure A.14: (A) Regression estimates of variance explained by markers versus the classical formula $2pq\alpha^2$. (B) Cumulative percent of variance explained across the risk allele frequency, based on regression estimates and classical formula. 1000 unlinked markers were simulated with effects drawn from an exponential distribution with mean 0.1 and population frequencies drawn from the neutral Wright-Fisher allele frequency distribution. Sample data for 5000 individuals were then generated by sampling genotypes at each marker based on its allele frequency. We plot the regression estimate of variance explained by each marker against Fisher's classic result [64]: $V_G = 2pq\alpha^2$.

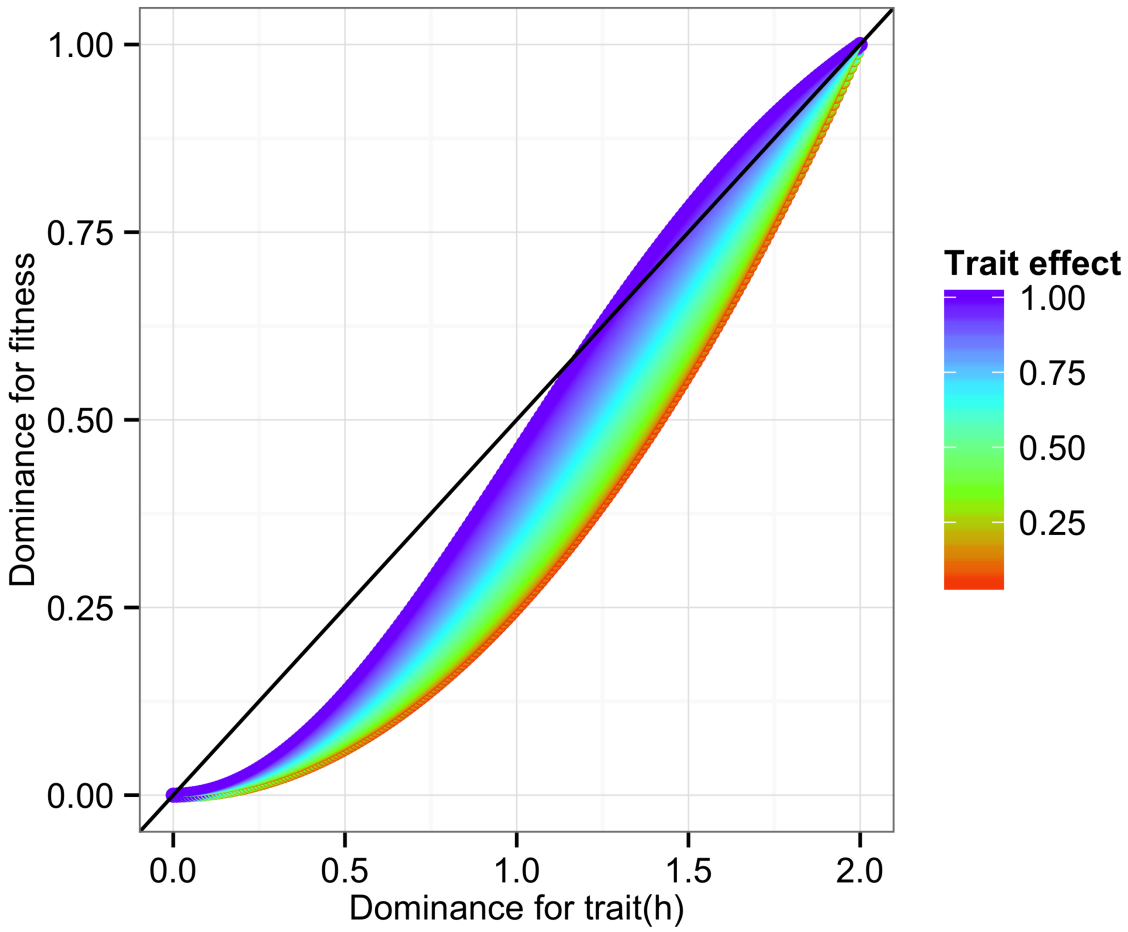

Figure A.15: The dominance of fitness effects, $\frac{s_{het}}{s_{hom}}$, as a function of the dominance for trait effects, h. Values are based on idealized fitness effects of a mutation on a previously unaffected genetic background. The the relationship between fitness and trait dominance is influenced by the trait effect size. We varied trait effect sizes from 0.01 to 1, and values are colored based on the trait effect.

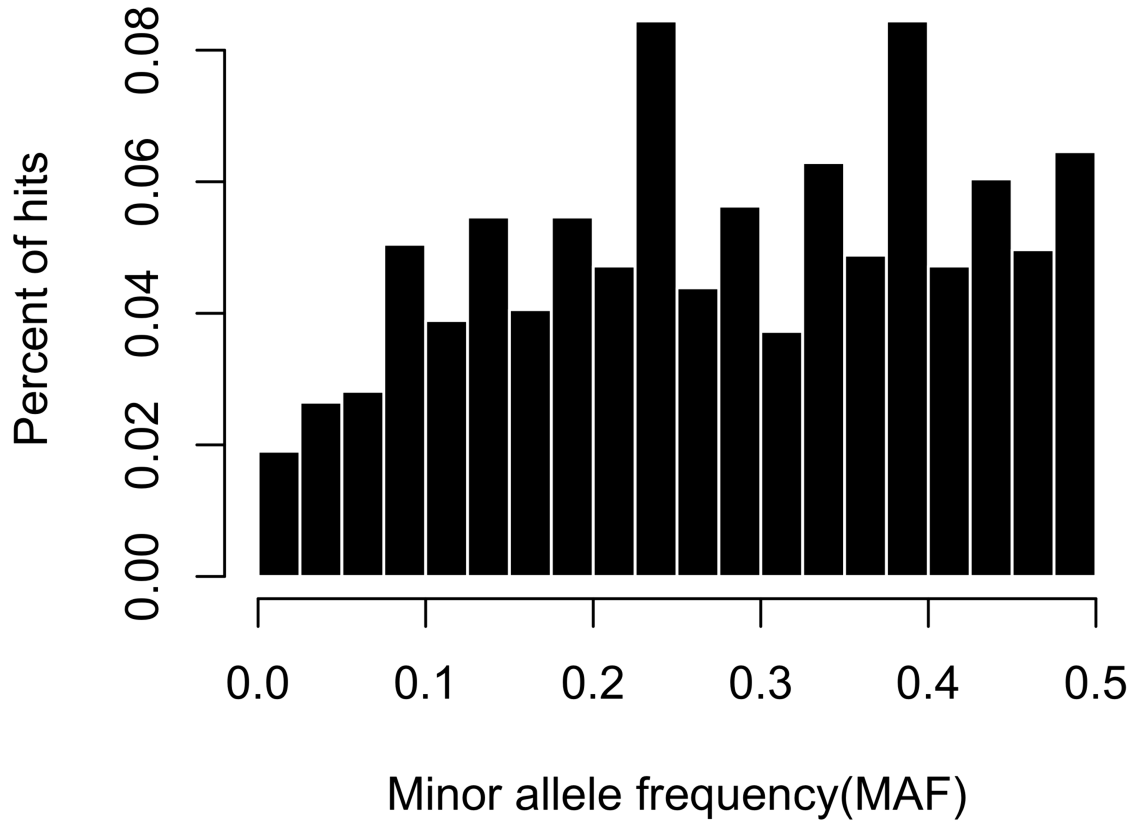

Figure A.16: Histogram GWAS hits (n=1208) obtained from the NHGRI-EBI GWAS database for disease discussed in [255]. Data are described in A.1.

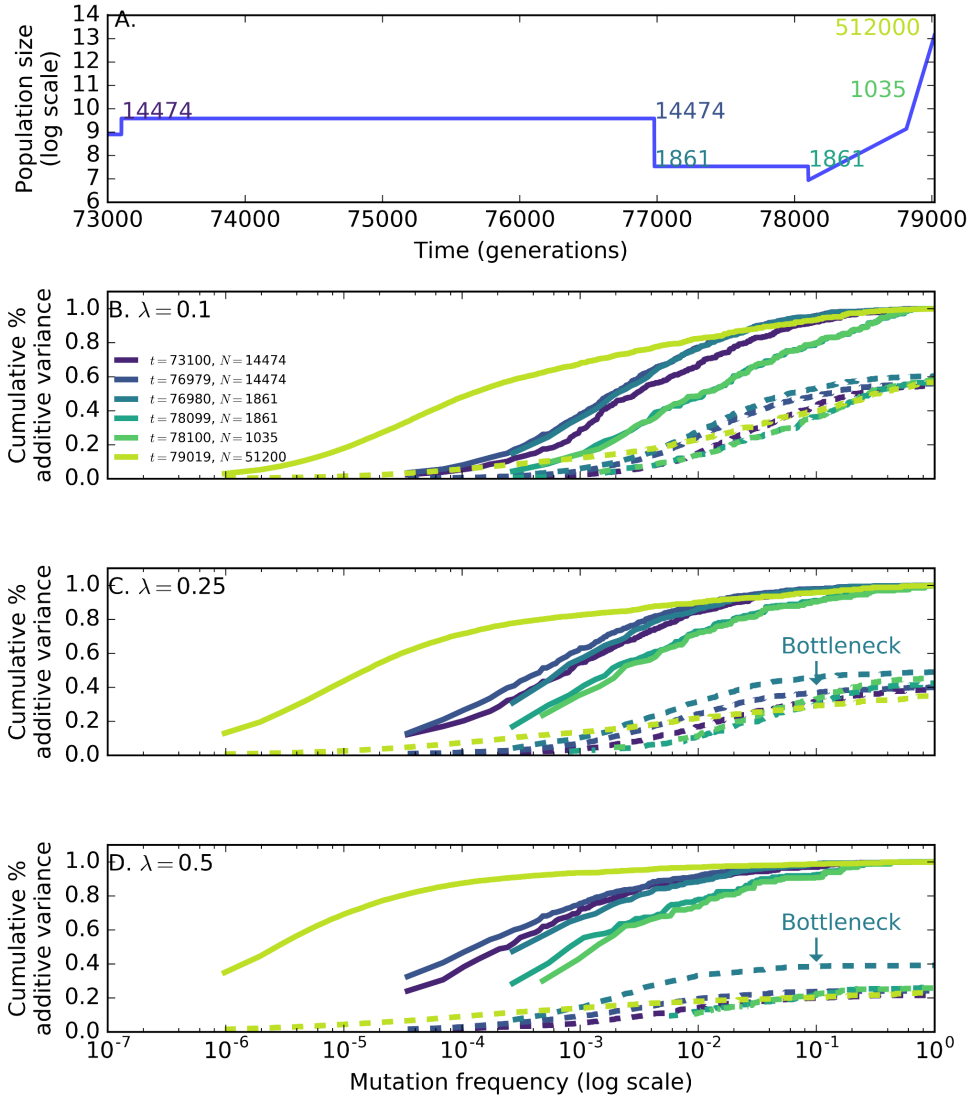

Figure A.17: (A) Population size change over time. Colored numbers represent population sizes at different times where we estimated the cumulative additive genetic variance (V_A) as a function of allele frequency using regression (see Materials and Methods). These time points represent key changes in population size in this model. (B-D) Estimated cumulative V_A as a function of frequency for three different mean effect sizes ($\lambda \in 0.1, 0.25, 0.5$). Solid lines are the standard additive model. Dashed lines are the GBR model of [232]. For all time points, the same total percent of variance is explained, with the exception of the line labelled "bottleneck". For larger effect sizes under the GBR model, the bottleneck increases the total V_A explained by all mutations. This effect is, however, short lived, and disappears by the end of the epoch defined by $N = 1,861$. This result is consistent with transient increases in variation under recessive models reported by [217].

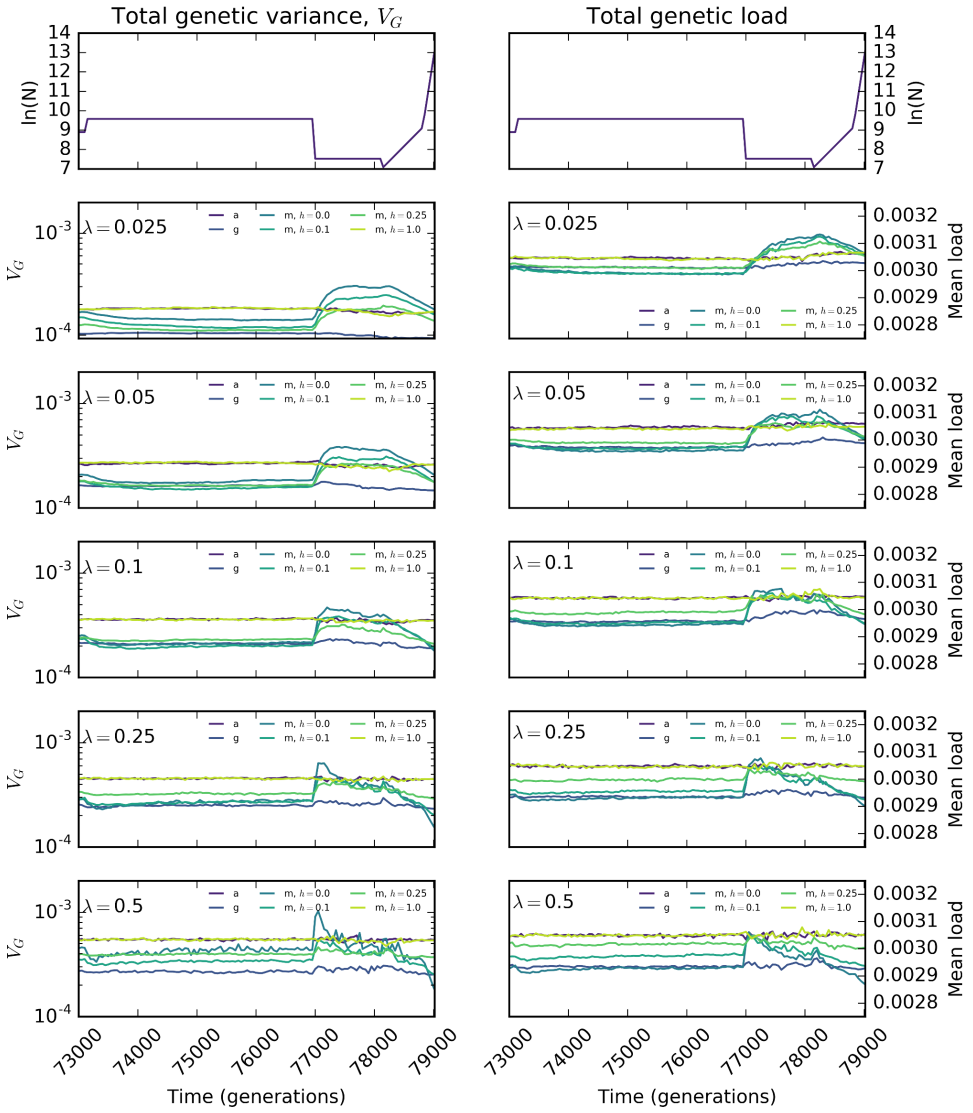

Figure A.18: The left column of panels shows how V_G changes over time under this model. The right column shows how the mean number of deleterious mutations per individual changes. The models shown are: a = additive, g = GBR, and m = multiplicative with varying degrees of dominance (h). The main difference is between additive models (a or m with $h = 1.0$) and recessive models (g or m with small h). The former models are largely insensitive to changes in N, while the recessive models show transient increases in V_G and "load" immediately following a bottleneck (consistent with [217]). However, at the final time point representing the "modern European population", all mean V_G is $\approx 4\mu$ for additive models and $\approx 2\mu$ for recessive models [237, 219], and recessive models show larger loads as expected [217].

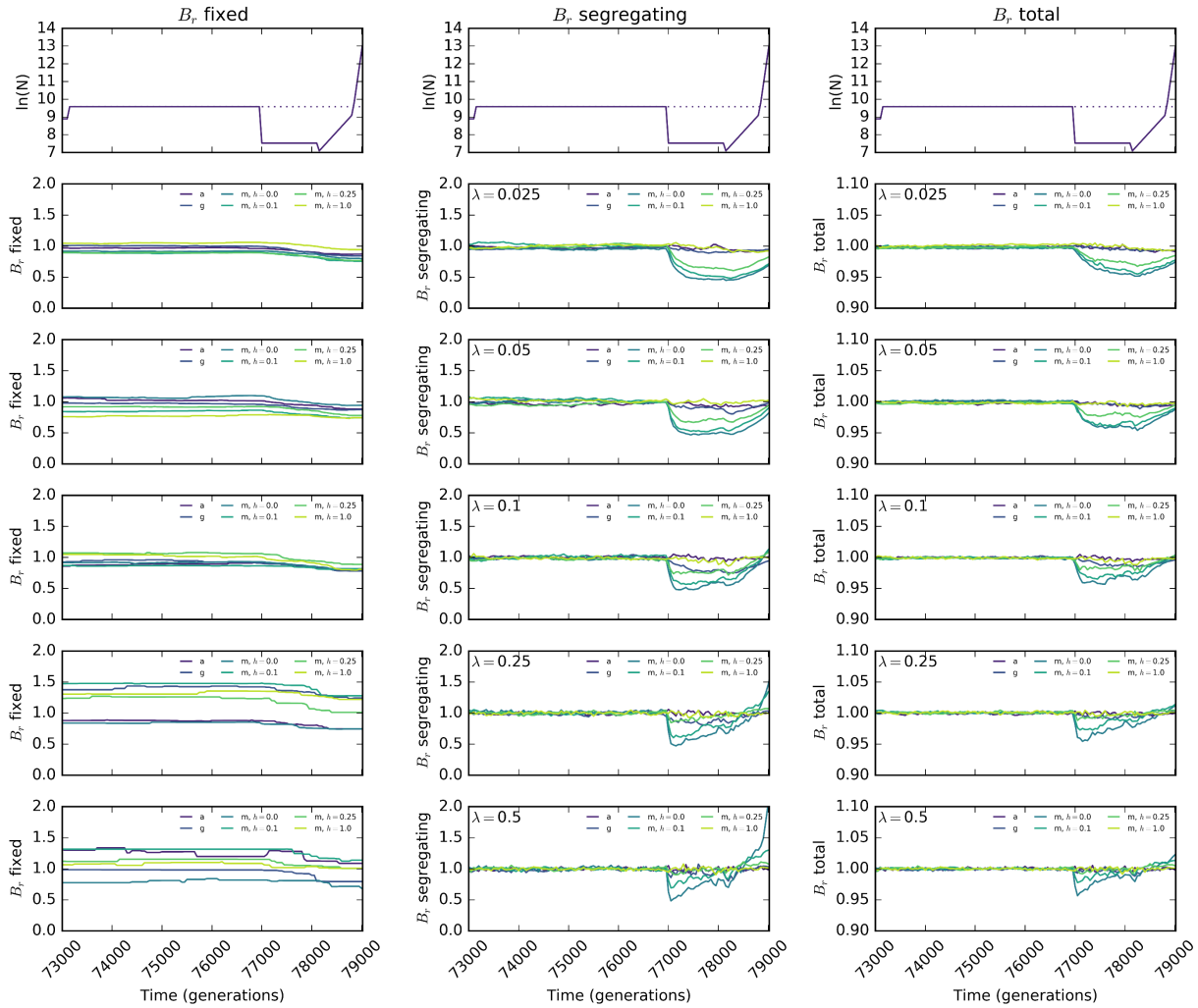

Figure A.19: The burden ratio [8] is calculated as the ratio of genetic load between simulations with only ancient growth and those with an additional recent bottleneck and growth. Here load is calculated as the average deviation from optimum fitness due to (left) fixed mutations, (middle) segregating mutations and (right) all mutations. Because of the use of the Gaussian fitness function, the total load is not the sum of the fixed and segregating load. The models shown are: a = additive, g = GBR, and m = multiplicative with varying degrees of dominance (h). For large effect size models, under which there are relatively more mutations that experience strong selection, we see the characteristic drop in the burden ratio following the bottleneck and rebound following re-expansion [8].

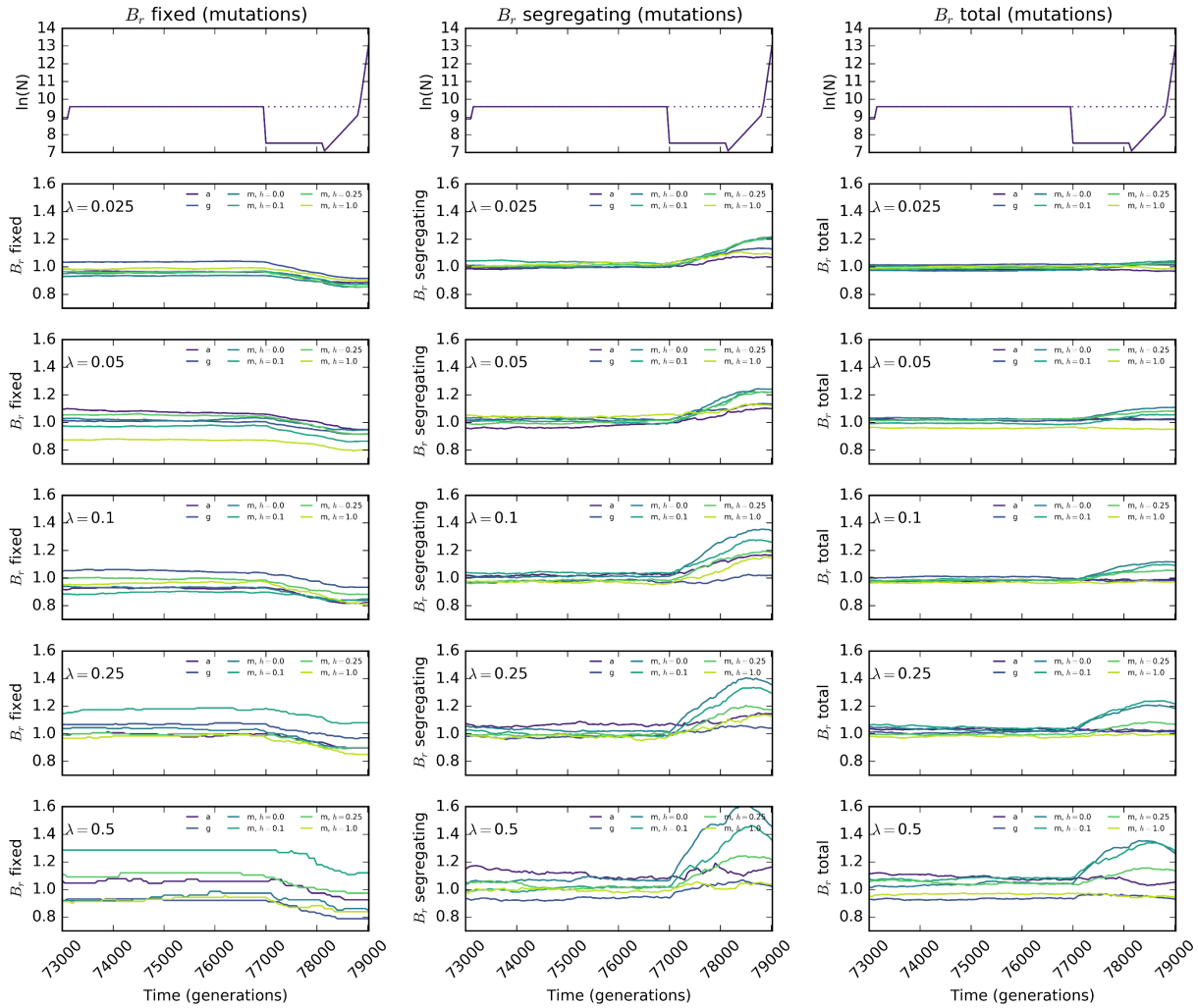

Figure A.20: The burden ratio [8] is calculated as the ratio of genetic load between simulations with only ancient growth and those with an additional recent bottleneck and growth. Here load is calculated as the average number of (left) fixed mutations, (middle) segregating mutations and (right) all mutations. The models shown are: a = additive, g = GBR, and m = multiplicative with varying degrees of dominance (h).

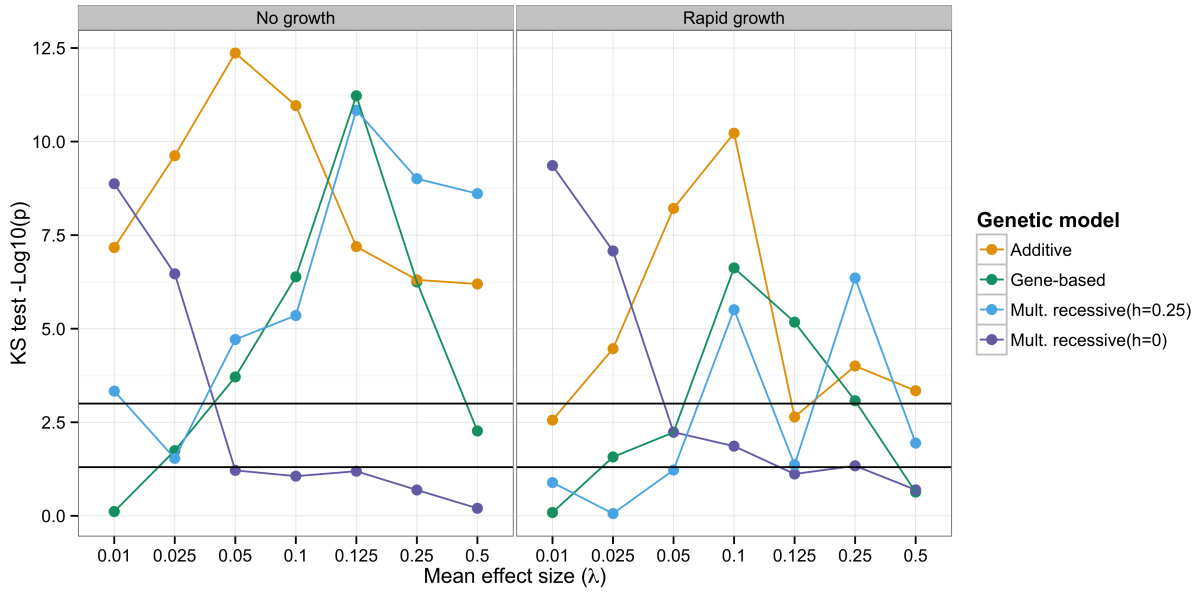

Figure A.21: A non-parametric comparison between distribution of allele frequencies between simulated and empirical GWAS hits. Shown are the -log10(p) values from the two-sample Kolmogorov-Smirnov test between the simulated and empirical allele frequencies. The lower and upper horizontal lines show where p=0.05 and p=0.001 respectively. Empirical data were downloaded from the NHGRI-EBI GWAS database (http://www.ebi.ac.uk/gwas/) on 02/03/2015, diseases and inclusion criteria are as in [255]. In cases where more than one marker was tied for the lowest p-value, one was chosen at random. Simulated data were subjected to ascertainment sampling such that the MAF distribution of all markers on the simulated genotyping chip was uniform. Specific information regarding the empirical data can be obtained in A.1.

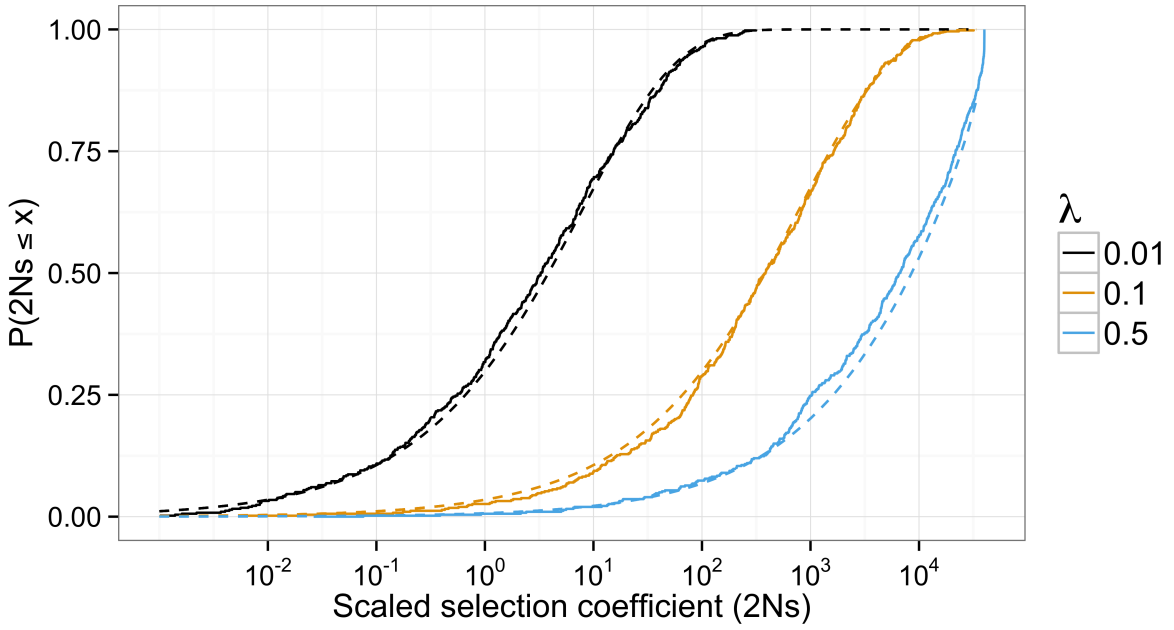

Figure A.22: The probability of a new mutation with $2Ns \leq x$ on a log scale for various values of λ. The dashed lines show the analytical result and the solid curves are empirical cumulative distribution functions based on a sample of 500 mutation effects from an exponential distribution. The analytical result is an approximation obtained by assuming there is only a single deleterious mutation.

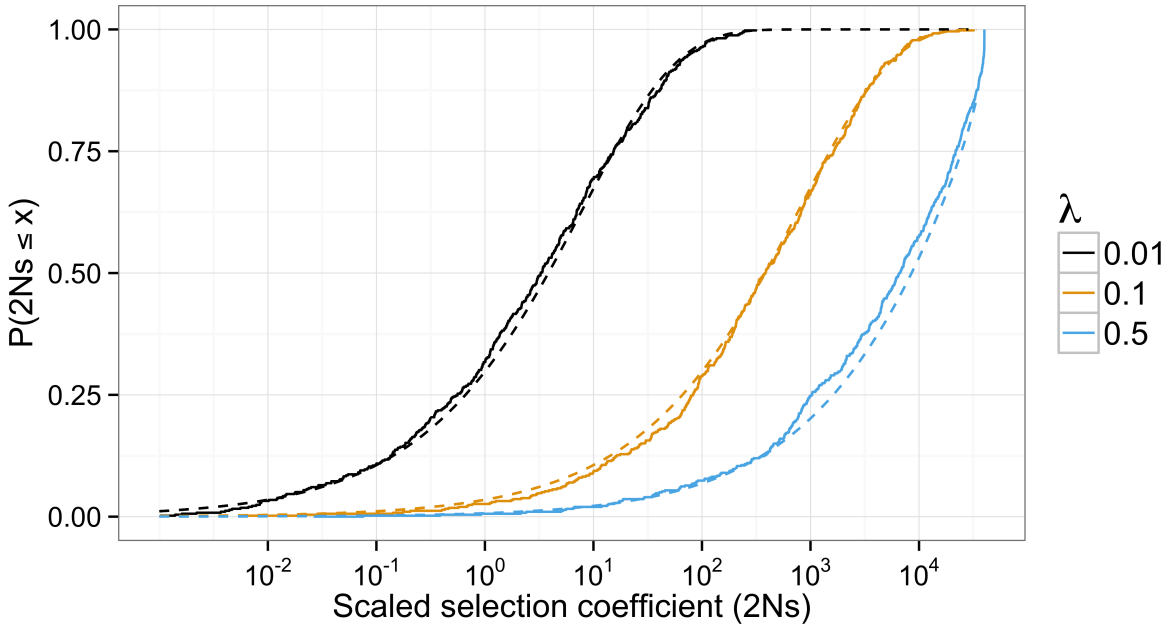

Figure A.23: The probability of a new mutation with $s = x$ on a log scale for various values of λ. The analytical result is an approximation obtained by assuming there is only a single deleterious mutation. For λ of 0.25 and 0.5 there is a large mass of lethals near $s = 1$.

A.3 Chapter 2 supplementary tables

Table A.1: This table contains information specifying the empirical studies used in the manuscript. The data were obtained from the NHGRI-EBI catalog of published genome-wide association studies.

PMID	First Author	Date	Journal	Disease/Trait
20385826	Neale BM	4/12/10	Proc Natl Acad Sci U S A	Age-related macular degeneration
21665990	Yu Y	6/10/11	Hum Mol Genet	Age-related macular degeneration
21909106	Arakawa S	9/11/11	Nat Genet	Age-related macular degeneration
22694956	Cipriani V	6/13/12	Hum Mol Genet	Age-related macular degeneration
23326517	Holliday EG	1/11/13	PLoS One	Age-related macular degeneration
23536807	Scheetz TE	3/11/13	PLoS One	Age-related macular degeneration
23455636	Fritsche LG	3/3/13	Nat Genet	Age-related macular degeneration
23577725	Naj AC	5/1/13	Ann Hum Genet	Age-related macular degeneration
20385819	Chen W	4/12/10	Proc Natl Acad Sci U S A	Age-related macular degeneration
20062062	Reveille JD	1/10/10	Nat Genet	Ankylosing spondylitis
21743469	Evans DM	7/10/11	Nat Genet	Ankylosing spondylitis
22138694	Lin Z	12/4/11	Nat Genet	Ankylosing spondylitis

17611496	Moffatt MF	7/26/07	Nature	Asthma
19426955	Himes BE	5/7/09	Am J Hum Genet	Asthma
20032318	Sleiman PM	12/23/09	N Engl J Med	Asthma
20159242	Li X	2/1/10	J Allergy Clin Immunol	Asthma
20860503	Moffatt MF	9/23/10	N Engl J Med	Asthma
21150878	Ferreira MA	12/8/10	Eur J Hum Genet	Asthma
21804548	Hirota T	7/31/11	Nat Genet	Asthma
21814517	Noguchi E	7/21/11	PLoS Genet	Asthma
21907864	Ferreira MA	9/10/11	Lancet	Asthma
23028483	Ramasamy A	9/28/12	PLoS One	Asthma
18711365	Ferreira MA	8/17/08	Nat Genet	Bipolar disorder
17554300	WTCCC	6/7/07	Nature	Bipolar disorder
17486107	Baum AE	5/8/07	Mol Psychiatry	Bipolar disorder
19416921	Scott LJ	5/5/09	Proc Natl Acad Sci U S A	Bipolar disorder
19488044	Smith EN	6/2/09	Mol Psychiatry	Bipolar disorder
21353194	Cichon S	2/23/11	Am J Hum Genet	Bipolar disorder
21771265	Yosifova A	7/19/11	Genes Brain Behav	Bipolar disorder
22925353	Lee HJ	8/25/12	J Affect Disord	Bipolar disorder
20386566	Lee MT	4/13/10	Mol Psychiatry	Bipolar I disorder
18463975	Kibriya MG	5/8/08	Breast Cancer Res Treat	Breast cancer
18326623	Gold B	3/11/08	Proc Natl Acad Sci U S A	Breast cancer
17529967	Easton DF	5/27/07	Nature	Breast cancer
17529973	Hunter DJ	5/27/07	Nat Genet	Breast cancer
17529974	Stacey SN	5/27/07	Nat Genet	Breast cancer

19219042	Zheng W	2/15/09	Nat Genet	Breast cancer
19330030	Thomas G	3/29/09	Nat Genet	Breast cancer
20453838	Turnbull C	5/9/10	Nat Genet	Breast cancer
20585626	Long J	6/24/10	PLoS Genet	Breast cancer
20852631	Antoniou AC	9/19/10	Nat Genet	Breast cancer
20872241	Li J	9/26/10	Breast Cancer Res Treat	Breast cancer
21263130	Fletcher O	1/24/11	J Natl Cancer Inst	Breast cancer
21424380	Sehrawat B	3/19/11	Hum Genet	Breast cancer
21908515	Cai Q	9/9/11	Hum Mol Genet	Breast cancer
22037553	Haiman CA	10/30/11	Nat Genet	Breast cancer
22383897	Long J	2/23/12	PLoS Genet	Breast cancer
22452962	Kim HC	3/27/12	Breast Cancer Res	Breast cancer
22923054	Chen F	8/25/12	Hum Genet	Breast cancer
22951594	Elgazzar S	9/6/12	J Hum Genet	Breast cancer
22976474	Siddiq A	9/13/12	Hum Mol Genet	Breast cancer
23354978	Rinella ES	1/25/13	Hum Genet	Breast cancer
23468962	Song C	2/28/13	PLoS One	Breast cancer
23535733	Garcia-Closas M	4/1/13	Nat Genet	Breast cancer
23535729	Michailidou K	4/1/13	Nat Genet	Breast cancer
24143190	Low SK	10/15/13	PLoS One	Breast cancer
18311140	Hunt KA	3/2/08	Nat Genet	Celiac disease
17558408	van Heel DA	6/10/07	Nat Genet	Celiac disease
20190752	Dubois PC	2/28/10	Nat Genet	Celiac disease
20383146	Kottgen A	4/11/10	Nat Genet	Chronic kidney disease

24351856	Nanayakkara S	12/18/13	J Occup Health	Chronic kidney disease
17634449	Samani NJ	7/18/07	N Engl J Med	Coronary heart disease
17554300	WTCCC	6/7/07	Nature	Coronary heart disease
19198612	Erdmann J	2/8/09	Nat Genet	Coronary heart disease
19198611	Tregouet DA	2/8/09	Nat Genet	Coronary heart disease
21088011	Erdmann J	11/18/10	Eur Heart J	Coronary heart disease
21239051	Reilly MP	1/14/11	Lancet	Coronary heart disease
21347282	Lettre G	2/10/11	PLoS Genet	Coronary heart disease
21378990	Schunkert H	3/6/11	Nat Genet	Coronary heart disease
21378988	The Coronary Artery Disease (C4D) Genetics Consortium	3/6/11	Nat Genet	Coronary heart disease
21606135	Wild PS	5/23/11	Circ Cardiovasc Genet	Coronary heart disease
21971053	Takeuchi F	10/5/11	Eur J Hum Genet	Coronary heart disease
22745674	Hager J	6/20/12	PLoS One	Coronary heart disease
22751097	Lu X	7/1/12	Nat Genet	Coronary heart disease
23364394	Lee JY	1/31/13	J Hum Genet	Coronary heart disease
18587394	Barrett JC	6/29/08	Nat Genet	Crohn's disease
17804789	Raelson JV	9/5/07	Proc Natl Acad Sci U S A	Crohn's disease
17684544	Franke A	8/8/07	PLoS One	Crohn's disease
17554300	WTCCC	6/7/07	Nature	Crohn's disease
17554261	Parkes M	6/6/07	Nat Genet	Crohn's disease
17435756	Rioux JD	4/15/07	Nat Genet	Crohn's disease

17447842	Libioulle C	3/5/07	PLoS Genet	Crohn's disease
20570966	McGovern DP	6/22/10	Hum Mol Genet	Crohn's disease
22412388	Kenny EE	3/8/12	PLoS Genet	Crohn's disease
21102463	Franke A	11/21/10	Nat Genet	Crohn's disease
22936669	Julia A	8/30/12	Gut	Crohn's disease
23128233	Jostins L	11/1/12	Nature	Crohn's disease
23266558	Yamazaki K	12/21/12	Gastroenterology	Crohn's disease
23850713	Yang SK	7/14/13	Gut	Crohn's disease
17660530	Hafler DA	7/29/07	N Engl J Med	Multiple sclerosis
18997785	Aulchenko YS	11/9/08	Nat Genet	Multiple sclerosis
19010793	Baranzini SE	11/14/08	Hum Mol Genet	Multiple sclerosis
19525953	De Jager PL	6/14/09	Nat Genet	Multiple sclerosis
19525955	Bahlo	6/14/09	Nat Genet	Multiple sclerosis
20159113	Jakkula E	2/12/10	Am J Hum Genet	Multiple sclerosis
20453840	Sanna S	5/9/10	Nat Genet	Multiple sclerosis
21654844	Briggs FB	6/9/11	Genes Immun	Multiple sclerosis
22190364	Patsopoulos NA	12/1/11	Ann Neurol	Multiple sclerosis
22457343	Martinelli-Boneschi F	3/28/12	Mult Scler	Multiple sclerosis
19648918	Amundadottir L	8/2/09	Nat Genet	Pancreatic cancer
20101243	Petersen GM	1/24/10	Nat Genet	Pancreatic cancer
20686608	Low SK	7/29/10	PLoS One	Pancreatic cancer

22158540	Wu C	12/11/11	Nat Genet	Pancreatic cancer
23180869	Wu C	11/24/12	Gut	Pancreatic cancer
19915575	Simon-Sanchez J	11/15/09	Nat Genet	Parkinson's disease
20711177	Hamza TH	8/15/10	Nat Genet	Parkinson's disease
21044948	Spencer CC	11/2/10	Hum Mol Genet	Parkinson's disease
21084426	Saad M	11/17/10	Hum Mol Genet	Parkinson's disease
21292315	Nalls MA	2/1/11	Lancet	Parkinson's disease
21738487	Do CB	6/23/11	PLoS Genet	Parkinson's disease
21812969	Liu X	8/3/11	BMC Med Genet	Parkinson's disease
22438815	Lill CM	3/15/12	PLoS Genet	Parkinson's disease
23793441	Davis MF	6/21/13	Hum Genet	Parkinson's disease
24511991	Hill-Burns EM	2/10/14	BMC Genomics	Parkinson's disease
24842889	Vacic V	5/19/14	Hum Mol Genet	Parkinson's disease
18264097	Eeles RA	2/10/08	Nat Genet	Prostate cancer
18264098	Gudmundsson J	2/10/08	Nat Genet	Prostate cancer
18264096	Thomas G	2/10/08	Nat Genet	Prostate cancer
17603485	Gudmundsson J	7/1/07	Nat Genet	Prostate cancer
17401366	Gudmundsson J	4/1/07	Nat Genet	Prostate cancer
17401363	Yeager M	4/1/07	Nat Genet	Prostate cancer
19117981	Sun J	1/1/09	Cancer Res	Prostate cancer
19767754	Gudmundsson J	9/20/09	Nat Genet	Prostate cancer

19767753	Eeles RA	9/20/09	Nat Genet	Prostate cancer
20676098	Takata R	8/1/10	Nat Genet	Prostate cancer
21602798	Haiman CA	5/22/11	Nat Genet	Prostate cancer
21743057	Schumacher FR	7/8/11	Hum Mol Genet	Prostate cancer
21743467	Kote-Jarai Z	7/10/11	Nat Genet	Prostate cancer
22923026	Cheng I	8/24/12	Cancer Epidemiol Biomarkers Prev	Prostate cancer
23023329	Xu J	9/30/12	Nat Genet	Prostate cancer
23535732	Eeles RA	4/1/13	Nat Genet	Prostate cancer
18677311	O'Donovan MC	7/30/08	Nat Genet	Schizophrenia
18347602	Sullivan PF	3/18/08	Mol Psychiatry	Schizophrenia
18332876	Kirov G	3/11/08	Mol Psychiatry	Schizophrenia
18282107	Shifman S	2/15/08	PLoS Genet	Schizophrenia
19571809	Shi J	7/1/09	Nature	Schizophrenia
19571808	Stefansson H	7/1/09	Nature	Schizophrenia
19571811	Purcell SM	7/1/09	Nature	Schizophrenia
21682944	Alkelai A	6/20/11	Int J Neuropsychopharmacol	Schizophrenia
21679298	Ma X	6/16/11	Genes Brain Behav	Schizophrenia
21926974	Ripke S	9/18/11	Nat Genet	Schizophrenia
22037555	Shi Y	10/30/11	Nat Genet	Schizophrenia

22883433	Irish Schizophrenia Genomics Consortium & the Wellcome Trust Case Control Consortium 2	8/7/12	Biol Psychiatry	Schizophrenia
23358160	Borglum AD	1/29/13	Mol Psychiatry	Schizophrenia
23894747	Aberg KA	2/1/13	JAMA Psychiatry	Schizophrenia
23974872	Ripke S	8/25/13	Nat Genet	Schizophrenia
19165918	Graham RR	8/1/08	Nat Genet	Systemic lupus erythematosus
18204446	Harley JB	1/20/08	Nat Genet	Systemic lupus erythematosus
18204098	Hom G	1/20/08	N Engl J Med	Systemic lupus erythematosus
18204447	Kozyrev SV	1/20/08	Nat Genet	Systemic lupus erythematosus
19838193	Han JW	10/18/09	Nat Genet	Systemic lupus erythematosus
20169177	Yang W	2/12/10	PLoS Genet	Systemic lupus erythematosus
21044949	Yang J	11/2/10	Hum Mol Genet	Systemic lupus erythematosus
21408207	Chung SA	3/3/11	PLoS Genet	Systemic lupus erythematosus

22291604	Okada Y	1/26/12	PLoS Genet	Systemic lupus erythematosus
23273568	Yang W	12/27/12	Am J Hum Genet	Systemic lupus erythematosus
24871463	Armstrong DL	5/29/14	Genes Immun	Systemic lupus erythematosus

A.4 Chapter 3 supplementary figures

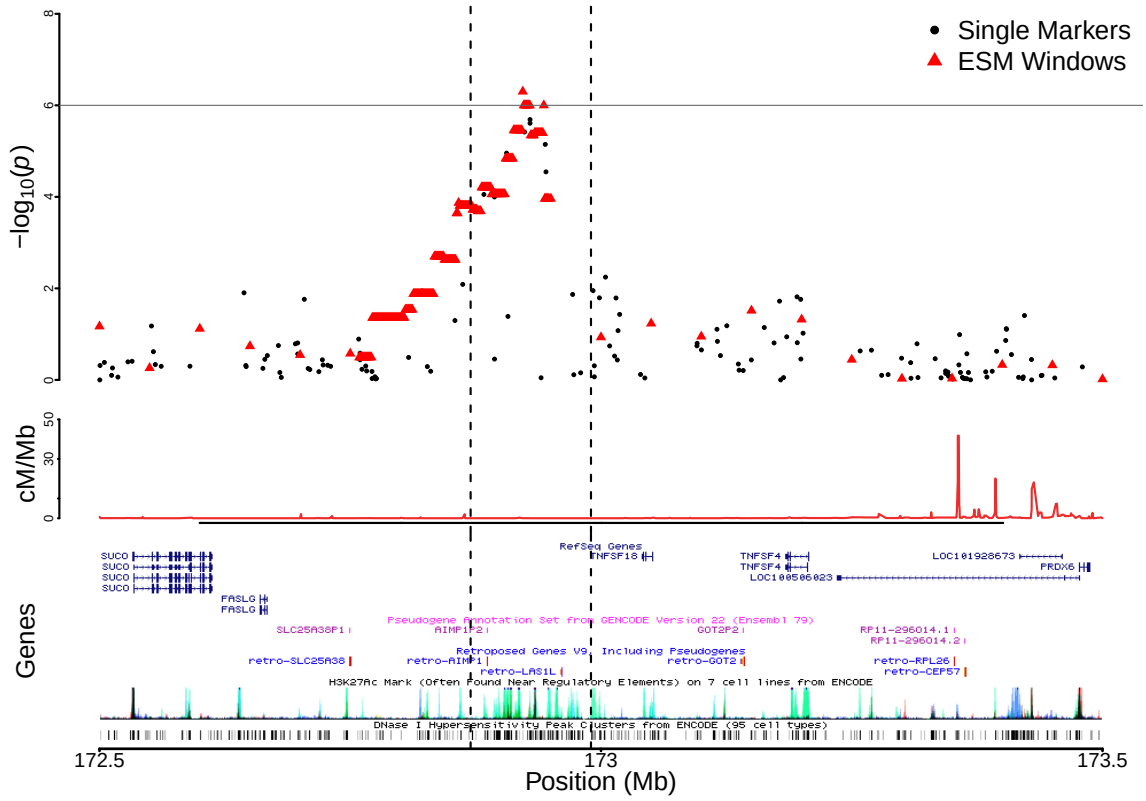

Figure A.24: The top panel contains single marker (black points) and ESM test (red triangles) $-log_{10}(p)$-values for inflammatory bowel disease versus chromosomal position in the region chr1:172.5-173.5 (Mb). Each ESM test point is plotted at the midpoint of a genomic window to which that $-log_{10}(p)$-values corresponds. The overlapping set of 100Kb ESM significant (ESM $p \leq$1e-6) regions which together span chr1:172.872-172.983 (Mb) are demarcated by vertical dashed lines, and the horizontal lines are placed at $-log_{10}(p) = 6$ to indicate the ESM test significance threshold. The middle panel contains the recombination rate in cM/Mb obtained from HapMap througout the same region. The lower panel shows UCSC genome browser tracks for the region; tracks shown include refseq genes, GENCODE psuedogene and retroposed gene annotations and ENCODE regulation.

A.5 Chapter 3 supplementary tables

Disease	Chromosome	Region(Mb)	SNP
CAD	9p21	21.93-22.13	rs1333049
IBD	1p31	67.06-67.24	rs11805303
IBD	2q37	233.35-233.43	rs10210302
IBD	3p21	49.29-49.86	rs9858542
IBD	5p13	40.28-40.62	rs17234657
IBD	5q33	150.79-150.95	rs1000113
IBD	10q21	62.63-62.881	rs10761659
IBD	10q24	99.51-99.57	rs10883365
IBD	16q12	50.43-50.814	rs17221417
RA	1p13	113.2-113.82	rs6679677
RA	6	MHC	rs6457617
T1D	1p13	113.2-113.82	rs6679677
T1D	6	MHC	rs9272346
T1D	12q13	55.96-56.41	rs11171739
T1D	12q24	110.9-112.57	rs17696736
T1D	16p13	10.93-11.37	rs12708716
T2D	6p22	20.52-20.73	rs9465871
T2D	10q25	112.96-113.06	rs4506565

Table A.2: Associations established in [252] which we find significant with the ESM test. Replicated associations are regions reported as showing strong associations under the standard analyis of [252] which are within regions with ESM test $p \leq$1e-6. Out of a total 21 SNP associations under the standard analysis of [252] we find that 18 replicate under the ESM test. Note that the exact coordinates do not match Table 3 of [252] because of the liftover to GRCh/hg38 coordinates.

Disease	Chromosome	Region(Mb)	SNP	Notes
BD	16p12	23.38-23.7	rs420259	Association not replicated in [236]
IBD	18p11	12.77-12.92	rs2542151	ESM p =9e-06,Association replicated in [233]
T2D	16q12	53.77-53.82	rs9939609	ESM p =5.2e-05, Replicated in the context of FTO gene effect on T2D [69]

Table A.3: Associations established in [252] which we do not find significant with the ESM test. Non-replicated associations are significant SNPs under the standard analyis of [252] which are within regions without genome-wide signficant ESM p-values. Out of a total 21 SNP associations under the standard analysis of [252] we find that 3 do not replicate under the ESM test. However, do note that two SNP, rs12708716 and rs2542151, are very close to being significant with ESM p =1.5e-06 and p =9e-06 respectively. Note that the exact coordinates do not match Table 3 of [252] because of the liftover to GRCh/hg38 coordinates.

Disease	Chr	Position(Mb)	Gene Region	Notes
CAD	3	193.6-193.8	OPA1	Plays role in heart disease, but not found through GWAS [36]. Not found via ESM test.
CAD	5	44.78-44.93	MRPS30	Reported for breast cancer and chronic kidney disease. May interact with SLC25A3 [89] which is indicated in diabetic cardiomyopathy [11]. ESM P = 2e-06.
IBD	19	45.73-45.85	SYMPK	Did not replicate in [67]

Table A.4: Three regions with SKAT test $p \leq$1e-6 with no corresponding hit from [252] are reported below. The SKAT test was implemented on the same windows as the ESM test, and using default parameters.

Disease	Chromosome	Region(Mb)	SNP
IBD	3p21	49.29-49.86	rs9858542
IBD	18p11	12.77-12.92	rs2542151
RA	1p13	113.2-113.82	rs6679677
RA	6	MHC	rs6457617
T1D	1p13	113.2-113.82	rs6679677
T1D	6	MHC	rs9272346
T2D	10q25	112.96-113.06	rs4506565

Table A.5: Associations established in [252] which we find significant with the SKAT test. Replicated associations are regions reported as showing strong associations under the standard analyis of [252] which are within regions with SKAT test $p \leq$1e-6. Out of a total 21 SNP associations under the standard analysis of [252] we find that 7 replicate under the SKAT test. The SKAT test was implemented on the same windows as the ESM test, and using default parameters. Note that the exact coordinates do not match Table 3 of [252] because of the liftover to GRCh/hg38 coordinates.

Disease	Chromosome	Region(Mb)	SNP	Notes
BD	16p12	23.38-23.7	rs420259	Association not replicated in [236],SKAT P¿0.2
CAD	9p21	21.93-22.13	rs1333049	Multiple studies show validation in GWAS catalog, SKAT P¿0.2
IBD	1p31	67.06-67.24	rs11805303	Validated, reviewed by [168], SKAT P¿0.2
IBD	2q37	233.35-233.43	rs10210302	Validated, reviewed by [168], SKAT P=1.6e-4
IBD	5p13	40.28-40.62	rs17234657	Validated in [250], SKAT P = 3.6e-5
IBD	5q33	150.79-150.95	rs1000113	Validated in [181], SKAT P = 1.5e-6
IBD	10q21	62.63-62.881	rs10761659	Validated in [66], SKAT P = 5.17 e-6
IBD	10q24	99.51-99.57	rs10883365	Validated in [250], SKAT P = 0.034
IBD	16q12	50.43-50.814	rs17221417	Validated, reviewed by [168], SKAT P¿0.2
T1D	12q13	55.96-56.41	rs11171739	Nearby SNP rs2292239 showed more robust signal in [233], SKAT P¿0.2
T1D	12q24	110.9-112.57	rs17696736	Validated in [233], SKAT P = 6e-5
T1D	16p13	10.93-11.37	rs12708716	Validated in [233], SKAT P = 0.005
T2D	10q25	112.96-113.06	rs4506565	Validated in [53], SKAT P = 0.00912
T2D	16q12	53.77-53.82	rs9939609	SKAT P ¿0.2,ESM p =5.2e-05, Replicated in the context of FTO gene effect on T2D [69]

Table A.6: Associations established in [252] which we do not find significant with the SKAT test. Non-replicated associations are significant SNPs under the standard analyis of [252] which are only within regions without genome-wide signficant SKAT p-values. Out of a total 21 SNP associations under the standard analysis of [252] we find that 14 do not replicate under the SKAT test. However, many do show marginally significant association signal. Note that the exact coordinates do not match Table 3 of [252] because of the liftover to GRCh/hg38 coordinates.

A.6 Chapter 4 supplementary texts

Supporting Materials and Methods

Study samples

The UK Biobank is a large prospective study of over 500,000 individuals in the United Kingdom(UK)[228]. Participants were 40-69 years of age during the recruitment phase (2006-2011). To avoid issues related to population structure, we studied only the 376,366 individuals of self-reported white-British ancestry. Unless otherwise notes we restricted our analysis to males over 50y old at assessment and females over 45y old, to ensure that number of children born to date is a good proxy for lifetime reproductive success. These filters resulted in 217,728 Female and 158,638 Male samples with phenotypic data. Of these individuals, there were 157,807 Female and 115,902 Male samples with genetic data available that were genetically unrelated (relatedness < 0.05).

Phenotypic data

The UKB contains data on the number of live births for females and the children fathered for males. These two variables were treated as life time reproductive success(LRS). To calculate relative lifetime reproductive success (rLRS) we followed the approach of [13]. Briefly, the samples were split into birth cohorts and LRS values were divided by the cohort specific mean value. We calculated rLRS within 4 non-overlapping birth cohorts, based on birth year. Specifically, the birth cohorts are: Cohort 1 (1934-1942), Cohort 2 (1943-1948), Cohort 3 (1949-1955) and Cohort 4 (1956-1965). In all subsequent regression analyses age, birth cohort and data collection assessment center were treated as covariates. All phenotypes, except LRS, measured in the set of 376,366 post-reproductive white-British ancestry samples

were split by sex and then scaled to mean zero and variance one. If a sample was measured on multiple visits to the assessment center then we used the mean value across measurements except in the case of educational attainment for which the maximum value was used. Unless otherwise noted, individuals more than 6 standard deviations from the mean were removed as outliers.

Genetic data

The UKB genetic data were collected using two similar genotyping arrays. Nearly 450,000 participants were genotyped on the custom Affymetrix UK Biobank Axiom (UKBA) array, while an additional 50,000 participants were genotyped using the UK BiLEVE (UKBL) array. The two arrays have over 95% common marker content, with the UKBA array having a small number of additional markers for genome-wide coverage. The genetic data was imputed using two different reference panels, by the UK Biobank team. The Haplotype Reference Consortium (HRC) panel was used as first choice option, but for SNPs not in that reference panel the UK10K + 1000 Genomes panel was used. A problem arose in the second set of imputed data from the UK10K + 1000 Genomes panel. The genotypes at these SNPs are imputed correctly, but have not been recorded as having the correct genome position in the files. We have established that the imputed data from the HRC panel is not affected and has the correct positions. This is about 40M sites and will include the majority of the common SNPs i.e. sites most likely to show genetic associations. These sites are readily identified since the HRC site list is public. The sample of White British ancestry individuals was derived using principal component analysis and the self-reported ancestry information. For our further genetic analyses, we selected 1,162,900 HapMap3 SNPs with info score ¿0.3, minor allele frequency ¿=0.01 and Hardy-Weinberg Equilibrium test p-value ¿=1e-6. We further constructed genetic relatedness matrices in GCTA [266] and removed one of each pair of individuals with estimated SNP marker relatedness greater than 0.05 or

if a genetically inferred gender of the sample did not match the self-reported gender. For some analyses only the UKB interim release was used which consisted of 108,402 unrelated White British individuals.

Phenotypic regression analyses

To estimate linear and quadratic selection gradients, we performed simple linear regressions of each phenotype and its square onto rLRS independently and through a multiple linear regression. In both cases, the phenotypes and their squared values were included and statistical significance was assessed by the Wald test.

The resulting regression coefficients are used to estimate the linear (β) and quadratic (γ) selection gradients [126, 224]. The value of β is simply equal to the regression coefficient on the phenotype itself. However, the value of γ is twice the regression coefficient on the square of the phenotype [224].

The particular subset of phenotypes used in the multiple linear regression was chosen to reduce the variance of the regression estimates. We observed the phenotypic correlations between traits (Fig. A.29 and A.30) and noticed some sets of highly correlated traits. Within each set we prioritized inclusion in the final model by (1) significance of genetic correlation with rLRS, (2) significance of phenotypic regression on rLRS and (3) sample size. The UKB has very large sample sizes, but the missing data is non-overlapping for each trait. As such, the data matrix became singular upon inclusion of all trait interaction terms. Therefore, we only include the traits interaction with itself (the quadratic term). To further address multi-collinearity in the data we calculated the variance inflation factor (VIF) for each trait. Individual traits were removed from the model, starting with the trait with the highest VIF, and the VIFs were recalculated. This process was repeated until all VIF values were below 2 for all included traits.

Genetic correlation analyses

Summary statistic based LD-score regression[22, 23] was performed on the full UKB dataset to calculate genetic correlations between various traits and rLRS. GWAS summary statistics were generated using a simple linear association test in plink[188]. Then, LD-score regression was performed using pre-computed LD-scores which are provided with the LD-score regression software.

In addition, a bivariate genetic variance component analysis was performed in the interim data release to establish genetic relationships between various traits and rLRS. The bivariate variance component model allows us to jointly estimate the genetic variance of each trait and their genetic covariance. Because of the large sample sizes of the UKB, BOLT-REML[139] was chosen for computational efficiency. Briefly, BOLT-REML estimates the genetic variance-covariance matrix via a Monte-Carlo Average Information REML approach. The genetic variance parameters are initially estimated using the related BOLT-LMM[139], which is a Bayesian linear mixed model methods. BOLT-LMM assumes a mixture-of-normals prior on the SNP effects such that most SNPs have small effects and others may have large effects. Given the BOLT-LMM initial estimates, BOLT-REML then applies a rejection sampling technique to obtain final estimates of the genetic variance-covariance matrices. We assessed the statistical significance of the BOLT-REML genetic correlations via the Wald test.

SI Text

BOLT-REML analysis of the interim UKB data release

We obtained genetic correlation estimates from a linear mixed modeling approach in addition to LD-score regression. Specifically, we used BOLT-REML, which gives very similar

estimates to the standard gREML procedure in GCTA [267] but scales more efficiently with large sample sizes. Following [234] a REML estimate of the genetic correlation between traits and rLRS, $r_{g,rLRS}$, was directly estimated using common (MAF$>$ 0.01) SNP markers in a bivariate linear mixed model (LMM) approach [230, 133] using BOLT-LMM [139]. The estimate of SNP heritability for rLRS varied across analyses and by sex. Due to the action of natural selection against deleterious mutations, the heritability of fitness components, such has reproductive success, is expected to be low and largely dominated by low frequency variants. Thus, our estimates of the common-SNP heritability of rLRS are most likely biased downward, which reduces the power of our genetic correlation analyses. Here we provide an overview of BOLT-REML results and a brief comparison to the LD-score regressions.

The BOLT-REML estimates of genetic correlation are summarized in Fig. A.37. Many traits in females show a $\hat{r}_{g,rLRS}$ in the same direction as the phenotypic regression estimate $\hat{\beta}$. Overall, there was a strong positive correlation between the $\hat{\beta}$ and $\hat{r}_{g,rLRS}$ in females only (Fig. A.40). Further, the total phenotypic correlation estimated from the mixed model is consistent with results from the regression analysis (Fig. A.39 and table A.9);see the following section for a more detailed discussion on the consistency between the phenotypic and genetic results.

In females, the median BOLT-REML estimate of $h^2_{SNP,rLRS}$ was 0.076, which on a relative scale is considerably larger than the value 0.0564 estimated from LD-score regression. While in males, the estimates of $h^2_{SNP,rLRS}$ from the two methods were quite close, with the BOLT-REML estimate being equal to 0.035 and the LD-score regression estimate being equal to 0.033. It is also known that, all else equal, LD-score regression estimates of genetic variance components will have larger standard errors than estimates obtained from mixed modeling approach. This means that there are multiple competing factors affecting power to detect non-zero genetic correlations including the sample size, the heritability explained by the model, and the precision of the estimate.

Four body-size related traits have a significant $\hat{r}_{g,rLRS}$ in females: WHR, WC, BFP and BMI. An additional two body-size traits, WT and HC were marginally significant in females. No traits show a significant $\hat{r}_{g,rLRS}$ at the FWER$\leq$ 0.05 level in males, but $\hat{r}_{g,rLRS}$ for male BMI is marginally significant and in the same direction as the phenotypic result ($\hat{\beta}$). Again consistent with the phenotypic results, $\hat{r}_{g,rLRS}$ values for EA and AFB in females are significant and negative.

The BOLT-REML and LD-score regression estimates of $\hat{r}_{g,rLRS}$ were highly correlated. However, the specific traits which passed the study-wise significance threshold varied considerably. More male traits were significant in the LD-score regression analysis while the opposite was true for females and the BOLT-REML analysis. However, it is important to emphasize that significance thresholds are somewhat arbitrary and we draw attention to the overwhelming consistency of the estimates obtained from the two approaches as demonstrated by Fig. A.38.

Consistency of phenotypic and genetic correlations

In the main text of this manuscript we present results from a phenotypic analyses in a large section of the UKB data and above we presented a genetic analysis from a reduced subset of that data. Specifically, we perform a linear regression for phenotypic analyses and bivariate linear mixed modeling for the genetic analyses. Here in this section we would like to provide a joint interpretation and discuss the issue of consistency between results of these two analyses. Below we provide calculations for various correlation coefficients obtained from our analyses; the empirical estimates of these coefficients are presented in table A.9.

The $\hat{\beta}$ estimates from a linear regression can be expressed in terms of phenotypic covariances and correlations. In the model

$$rLRS = \beta P + \epsilon$$

we have

$$\begin{aligned}\hat{\beta} &= \frac{cov(P, rLRS)}{V(rLRS)} \\ &= r_p \frac{\sigma_{rLRS}}{\sigma_P}\end{aligned}$$

Where r_p is the phenotypic correlation coefficient in the sample. Therefore, we obtain by simple algebra the first expression for the phenotypic correlation coefficient directly from our phenotypic analyses, which we call $r_{p,1}$.

$$r_{p,1} = \hat{\beta} \frac{\sigma_P}{\sigma_{rLRS}}$$

Given some assumptions we can obtain a similar expression for phenotypic correlation from the genetic results. We assume an additive polygenic model for both traits (P and $rLRS$) analyzed in the bivariate model such that the traits are expressed as additive genetic and environmental components.

$$\begin{aligned}rLRS &= A_{rLRS} + E_{rLRS} \\ P &= A_P + E_P\end{aligned}$$

We can then further decompose the additive genetic component into a portion explained by genotyped SNPs and a remainder.

$$A_{rLRS} = A_{s,rLRS} + A_{r,rLRS}$$
$$A_P = A_{s,P} + A_{r,P}$$

The covariance between $rLRS$ and P is

$$\begin{aligned} cov(rLRS, P) =& cov(A_{rLRS}, A_P) + cov(E_{rLRS}, E_P) \\ cov(rLRS, P) =& cov(A_{s,rLRS}, A_{s,P}) + \\ & cov(A_{r,rLRS}, A_{r,P}) + cov(E_{rLRS}, E_P) \end{aligned}$$

From the bivariate linear mixed model we obtain estimates of the correlation between the additive genetic components of both traits explained by SNPs and the covariance between the residual components.

$$r_{s,g} = \frac{cov(A_{s,rLRS}, A_{s,P})}{\sqrt{V(A_{s,rLRS})V(A_{s,P})}}$$
$$r_{s,e} = \frac{cov(A_{r,rLRS} + E_{rLRS}, A_{r,P} + E_P)}{\sqrt{V(A_{r,rLRS} + E_{rLRS})V(A_{r,P} + E_P)}}$$

We cannot assume that the environmental components of the two phenotypes is zero ($cov(E_{rLRS}, E_P) = 0$) because this is a strong untested assumption and it is one that would not be true under a causal relationship between P and $rLRS$. Therefore, it is not possible to extrapolate from the mixed model results to the true full genetic correlation (r_g). However, we can provide a second calculation of the full phenotypic correlation from the genetic results which we call $r_{p,2}$.

$$
\begin{aligned}
r_{p,2} &= \frac{cov(rLRS, P)}{\sqrt{V(rLRS)V(P)}} \\
&= \frac{cov(A_{s,rLRS}, A_{s,P}) + cov(A_{r,rLRS} + E_{rLRS}, A_{r,P} + E_P)}{\sqrt{V(rLRS)V(P)}} \\
&= \frac{r_{s,g}\sqrt{V(A_{s,rLRS})V(A_{s,P})}}{\sqrt{V(rLRS)V(P)}} \\
&+ \frac{r_{s,e}\sqrt{V(A_{r,rLRS} + E_{rLRS})V(A_{r,P} + E_P)}}{\sqrt{V(rLRS)V(P)}}
\end{aligned}
$$

The two calculations of phenotypic correlation should be closely related as one is obtained using a subset of the data used from the other. Indeed in a regression of $r_{p,1}$ on $r_{p,2}$ the $R^2 = 0.94$ (Fig. A.39). The residual variance-covariance estimates from the bivariate model contain both untagged genetic and non-genetic effects. Therefore we can not definitively demonstrate consistency between the pure phenotypic and genetic results. However, we can ask how well the phenotypic correlations predict the genetic correlations. To do so we regressed the mixed model genetic correlation estimates $r_{s,g}$ onto the phenotypic correlation estimates from the phenotypic regressions $r_{p,1}$. The regression coefficient in that model was 2.96 with an adjusted $R^2 = 0.27$. In other words the phenotypic correlation values explain 27 percent of the variance in genetic correlation estimates.

Correlation between rLRS and a polygenic predictor for height

As an alternative approach to finding genetic evidence for a relationship between rLRS and height we constructed a polygenic predictor for height based on a meta-analysis of published height GWAS and the interim UKB data. This meta-analysis has an effective sample size of 390,000. From the meta-analysis there were 1,371 SNPs that passed a clumped p-value threshold hold of 10^{-6}. We predicted height in the UKB samples using the sample genotypes and estimated effect sizes at these 1,371 SNPs. Our predictor explains 25 percent of phenotypic for height(Fig. A.47).

The rLRS values were regressed onto the predicted height and squared height values for males and females separately. In males neither the linear nor quadratic predictor were significantly associated with rLRS. However, in females both the linear and the quadratic predictor showed marginal significance. In females, the estimated effect size of the height predictor on rLRS was $-0.0081 \pm 0.004 (p = 0.0624)$ and effect size of the squared height predictor on rLRS was $-0.007919 \pm 0.003 (p = 0.0097)$. This result is qualitatively consistent with our phenotypic observations of a weak directional and quadratic relationship between rLRS and height.

We performed a simulation to better understand the behavior of the polygenic predictors for height and their relationship to rLRS. We simulated fitness values under a model of multivariate stabilizing selection[28] where three phenotypes contribute to fitness. One of the three underlying phenotypes was treated as being under directional selection by setting the phenotypic optimum to be different from 0 for that phenotype only.

We simulated genotypes at 20,000 unlinked biallelic variants. 10,000 of these variants were causal for the phenotype under directional selection. Variant effect sizes were estimated in a panel of 300,000 individuals using a simple linear association test. Using the estimated variant effects we created polygenic predictors for the phenotype and squared phenotype in

an independent validation panel of 50,000 individuals. The predictors were regressed against the simulated fitness values in the independent validation panel.

The simulations recapitulate an important qualitative signature of our empirical polygenic predictors. As the number of SNPs in the predictor is increased the p-value of a predictor goes down and then eventually goes back up(Fig. A.48. When the number of SNPs is too low, there is no statistical power to predict in a new panel. However, if too many SNPs are included then we are adding noise to the predictor and power is reduced. This effectively reflects a transition from a model that under-fits to one that has over-fit the data; both under and over fitting reduce prediction accuracy in an independent dataset.

According to our simulations, the transition from under-fitting to over-fitting as a function of number of variants happens much faster for the quadratic predictor than for the linear predictor. This is likely due to the propagation of measurement error through a quadratic function. By predicting the variant effects on the phenotype and then predicting its square we have propagated the error of the variant effect size estimates.

Mendelian randomization using summary statistics for Educational Attainment

We use Mendelian randomization based on summary statistics (GSMR) [276] to assess the evidence for a possible causal relationship between educational attainment and reproductive phenotypes. In the main text of this manuscript we show that there is a strong phenotypic and genetic correlation between educational attainment and relative lifetime reproductive success. Briefly, a Mendelian randomization (MR) analysis estimates and tests a causal relationship of trait X on trait Y by using known SNP associations for trait X as instruments. The rationale is that if trait X causes Y then any perturbation that affects X will have the same proportional effect on Y.

Using summary statistics from GWAS for educational attainment [175] and the UKB data on

rLRS used in the main text, we tested the hypothesis that educational attainment, or a trait genetically highly correlated with it, is causal for rLRS. Using 50 instruments (genome wide significant SNPs for educational attainment) we estimate that $\hat{\beta_{EA,rLRS}} = -0.2(p < 10^{-5.8})$.

This instrument variable analysis implies that an increase of one standard deviation in educational attainment leads to a 0.2 decrease in rLRS. While this limited analysis is insufficient to fully demonstrate causality, the results are clearly consistent with the hypothesis that educational attainment, or a trait such as cognitive ability (which is genetically correlated with EA and might itself be causal for EA), has a negative causal relationship with lifetime reproductive success.

We also performed a similar GSMR analysis between educational attainment and age at first birth. Using the same educational attainment summary statistics, we had 51 instruments and estimated that $\hat{\beta_{EA,AFB}} = 0.653(p < 10^{-21})$. The results are consistent with the hypothesis that educational attainment (or a highly correlated trait) causally increases age at first birth.

Linear regression sensitivity analysis

The phenotypic results presented in the main text followed the default data filter and QC pipeline. In the defaults pipeline we used age cutoff of 50 and 45 for males and females respectively, did not perform inverse noram transformation on the data, used 6 standard deviations to define outliers for removal and did not remove known related individuals. We were concerned that, while rare, it is possible for males and females to have children above the ages of 50 and 45 respectively. When the age inclusion thresholds were increased to 55 and 50 for males and females respectively, we did not see many major changes to the results although the specific magnitudes of the selection gradient estimates did change. Similarly, we increased stringency of the outlier inclusion criteria by removing individuals outside of 4 standard deviations from the mean. The increased outlier stringency had little qualitative

effect on the our results.

We found that our results were also robust to normalization via inverse-normal transformation and the removal of known related individuals (Fig. A.45 and A.46). Additionally, a logistic regression analysis was performed using a binary encoding of LRS in which zero indicates no children and one indicates one or more children. Many of the phenotypes appear to be associated with this binary phenotype. This indicates that some of the our phenotypic regression results can be explained by whether people end up having children or not. These results are contained in Dataset A.10 along with all other regression results.

The broad sense heritability of a squared phenotype

In the main text, we argue that the narrow sense heritability for a squared trait will necessarily be much lower than the heritability for the trait itself. This stems from a few fundamental features of the squared phenotype including gene-by-environment interactions, over-dominance and epistasis. We will not derive the general case here, but instead will illustrate a couple of informative special cases. First, we will demonstrate the reduction in broad-sense heritability for a general trait with independent genetic and environmental components. Second, we will derive the dominance and epistatic variance components under purely additive single-locus and two-locus trait models. We finally appeal to our empirical results to support our claim (Fig. A.44).

We begin with the simple phenotypic model with independent genetic and environmental

components.

$$P = G + E$$
$$H^2 = \frac{V(G)}{V(P)} = V(G)$$

Where $P \sim \mathcal{N}(0,1)$, $G \sim \mathcal{N}(0, V(G))$, and $E \sim \mathcal{N}(0, V(E))$. Assume that G and E are independent. Therefore when we take the squared phenotype we get:

$$P^2 = G^2 + 2GE + E^2$$

I want to find expressions for $V(P^2)$, $V(G^2)$, $V(E^2)$, and $V(2GE)$. Because G and E both have mean zero and are independent:

$$V(2GE) = 4V(G)V(E)$$

From the definition of variance:

$$V(X^2) = E(X^4) - E(X^2)^2$$

This is the difference between the fourth central moment and the square of the second central moment. For a normal random variable the fourth central moment is 3 times the squared variance

$$\mu_4 = 3\sigma_X^4$$
$$V(X^2) = 3\sigma_X^4 - (\sigma_X^2)^2$$
$$V(X^2) = 2(\sigma_X^2)^2$$

The remaining expressions clearly follow.

$$\sigma_P^2 = 1$$
$$\sigma_G^2 = V(G) = H^2$$
$$\sigma_E^2 = V(E) = 1 - H^2$$

Therefore:

$$V(P^2) = 2$$
$$V(G^2) = 2(H^2)^2$$
$$V(E^2) = 2(1 - H^2)^2$$
$$V(GE) = 4 * H^2(1 - H^2)$$

If we decompose the squared phenotype, we can get expressions for the proportions of vari-

ance due to each component:

$$\frac{V(G^2)}{V(P^2)} = (H^2)^2$$
$$\frac{V(E^2)}{V(P^2)} = (1 - H^2)^2$$
$$\frac{V(2GE)}{V(P^2)} = 2 * H^2(1 - H^2)$$

Thus the broad sense heritability of squared phenotype will be the square of broad sense heritability of the phenotype.

Finite locus models for a squared phenotype

Next, we will show how the genetic variance for the squared trait decomposes under single and two locus models of a trait. Consider a single biallelic locus contributing to purely additive trait P with alleles and at frequencies p and q respectively.

	A_1A_1	A_1A_2	A_2A_2
G	a	0	$-a$
G^2	a^2	0	a^2
G^{2*}	0	$-a^2$	0

The additive variance in P is given by the classic formula.

$$V_{G^2}(A) = 2pqa^2$$

In the case of the P^2 there is no additive effect, but there is a dominance deviation ($d = -a^2$) such that:

$$V_{G^2}(A) = 2pq(-a^2(q-p))^2$$
$$V_{G^2}(D) = (-2pqa^2)^2$$

$$\begin{aligned}\frac{V_{G^2}(A)}{V_{G^2}(A) + V_{G^2}(D)} &= \frac{2pq(-a^2(q-p))^2}{2pq(-a^2(q-p))^2 + (-2pqa^2)^2} \\ &= \frac{a^2(q-p)^2}{a^2(q-p)^2 + 2pqa^2} \\ &= \frac{1}{1 + \frac{2pq}{(q-p)^2}}\end{aligned}$$

The expression for the percent of total genetic variance which is due to additive effects only depends on the allele frequencies .The absolute magnitude of the additive variance for the squared genetic component is maximized for $p = 0.5 \pm \frac{1}{2\sqrt{2}}$. However, the relative magnitude of the of the additive component is maximized as p goes to zero. This composition of the genetic variance across the range of allele frequencies is shown in Fig. A.41. Thus, for a given trait architecture the additive component of squared trait will be more highly influenced by rare variants. This fact further reduced the power of present study, in which we use common SNPs to estimate genetic relatedness.

To illustrate the inclusion of epistasis, we derive the variance components under a two-locus

model with equal additive effects. The loci have equal effects, a, and allele frequencies p_1 and p_2.

The genetic values for the trait and the squared trait are

$$G = \begin{pmatrix} 2a & a & 0 \\ a & 0 & -a \\ 0 & -a & -2a \end{pmatrix}$$

$$G^2 = \begin{pmatrix} 4a^2 & a^2 & 0 \\ a^2 & 0 & a^2 \\ 0 & a^2 & 4a^2 \end{pmatrix}$$

For this model, we follow the approach of Kojima [121] for the derivation of genetic variance components based on partial derivatives of the population mean with respect to allele frequencies. Given the population mean genetic value μ, the (L-additive X Q-dominance) variance due to a particular locus of set of loci can be defined as:

$$a_{LQ} = \frac{1}{2^{L+Q}} \frac{\delta^{L+2Q}\mu}{\prod_i^L \delta p_i \prod_j^Q p_j^2}$$
$$\sigma^2_{LQ} = 2^L \prod_i^L p_i(1-p_i) \prod_j^Q (p_j(1-p_j))^2 a^2_{LQ}$$

For example, the additive variance for a trait with M loci would be found by setting L=1 and Q = 0, such that:

$$a_{10,i} = \frac{1}{2}\frac{\delta\mu}{\delta p_i}$$
$$\sigma^2_{10} = \sum_i^M 2p_i(1-p_i)a^2_{10,i}$$

In the case of our two locus squared trait model, we will have additive, dominance and additive by additive epistatic terms. It can be shown that dominance interactions go to zero, i.e DxA and DxD epistasis.

$$\begin{aligned}\mu_{G^2} =& 4a^2\left(1-p_2\right)^2\left(1-p_1\right)^2 + 2a^2\left(1-p_2\right)p_2\left(1-p_1\right)^2 + \\ & 2a^2p_1\left(1-p_2\right)^2\left(1-p_1\right) + 2a^2p_1p_2^2\left(1-p_1\right) + \\ & 4a^2p_1^2p_2^2 + 2a^2p_1^2 + \left(1-p_2\right)p_2\end{aligned}$$

$$\begin{aligned}V_{G^2}(A) =& \sum_{i=1}^{2} 2p_i(1-p_i)\frac{\delta\mu_{G^2}}{\delta p_i} \\ V_{G^2}(D) =& \sum_{i=1}^{2}(p_i(1-p_i))^2\frac{\delta^2\mu_{G^2}}{\delta p_i^2} \\ V_{G^2}(AA) =& 4p_1(1-p_1)p_2(1-p_2)\frac{\delta\mu_{G^2}}{\delta p_1\delta p_2}\end{aligned}$$

In Fig. A.42, we illustrate how much of the total genetic variance is attributable to the additive component across the allele frequencies at each locus. The conclusions are similar to the single locus model: at intermediate allele frequencies there is a substantial reduction in the relative contribution of additive variance to total genetic variance for a squared phe-

notype. This implies that, even if the additive variance is preserved at some loci, our study using common variants will be severely under powered. However, as the number of loci in the model increase, the importance of the additive variance component should also increase [149]. Regardless, our empirical results show that narrow-sense heritability of a squared trait is severely reduced compared to that of the trait itself.

Additive heritability of a squared trait under a polygenic model

In the previous section, we derived estimates of genetic variance under one and two locus model. However, the traits studied in this work are highly polygenic and we sought an alternative approximation based on the infinitesimal model. Under the infinitesimal model, we can approximate the additive heritability of a trait as the limit of the correlation between relatives as the correlation goes to zero(unrelated). This approximation should remain valid for functions a traits.

Let there be a phenotype Y which is purely additive and another phenotype $Z = Y^2$. We then consider both phenotypes Y and Z in a set of relatives with relatedness r for

$$
\begin{aligned}
Y &= A + E \\
E[Y] &= 0 \\
var[Y] &= 1 \\
Z_1 &= Y_1^2 = (A_1 + E_2)^2 \\
Z_2 &= Y_2^2 = (rA_1 + E_2)^2 \\
var[Z] &= 2
\end{aligned}
$$

First, we need to derive a statement for the covariance between Z_1 and Z_2. Given that we $E[Y] = 0$, this derivation is relatively straightforward. The results follow below.

$$
\begin{aligned}
cov(Z_1, Z_2) &= 2r^2 V(A)^2 \\
\frac{cov(Z_1, Z_2)}{V(Z)} &= corr(Z_1, Z_2) = r^2 (h_Y^2)^2 \\
\lim_{r \to 0} \frac{corr(Z_1, Z_2)}{r} &= h_Z^2 \\
h_Z^2 = \lim_{r \to 0} \frac{corr(Z_1, Z_2)}{r} &= \lim_{r \to 0} r(h_Y^2)^2 = 0
\end{aligned}
$$

Therefore, under a highly polygenic model we expect there to be no additive genetic variance for a squared phenotype. To check this result we performed simple simulations in R. We sampled genotypes at M markers for N individuals (N>>M), assuming equal allele frequencies as allelic effects at each marker. Then we estimated the additive genetic variance using the R^2 from simple linear regression. We performed the regression on either the raw phenotypes or the phenotypes scaled to mean zero and variance one. Fig. A.43 shows that the two locus analytical expression for additive variance is accurate when M=2 in a regression on the raw phenotypes. Fig. A.43 also shows that when the phenotypes is scaled, the additive variance decreases rapidly as M (the number of markers) increases. From this we infer that, in accordance with our derivation under a polygenic model, as M gets very large there is no additive genetic variance for a squared phenotype.

Genetic control of variability

We first considered the behavior of the broad sense heritability of the square of a trait and found that it is the square of the broad-sense heritability of the trait. Next we derived expressions for genetic variance components under finite locus models following the classic approach of Kojima [121] (revisited in [149]). Our single locus derivations here are equivalent to those done by Yang, et al. [267]. The finite locus models suggested that for a purely additive trait, the additive genetic variance of the square of trait will be less than genetic variance by an amount which depends on allele frequencies and number of loci. Under an infinitesimal model, based on the correlation between relatives, we expect there to be zero additive genetic variance for the square of a trait. Using simple simulations we validate our finite locus expressions and show that as the number of loci increases the additive genetic variance approaches zero, in agreement with the derivations under the infinitesimal model (Fig. A.43).

Further, without making assumptions about the genetic architecture of the trait, any observed additive genetic variance for a squared phenotype can not be easily disentangled from the effects of loci that explicitly control phenotypic variability. In other words, when genotypic classes differ in phenotypic variability there will additive genetic variance for the square of the phenotype–a fact that has been previously appreciated in the literature[220, 24, 274, 272, 94, 247].Yang et al [267] showed that it is possible to determine whether the additive effect of a SNP on the square of the trait is too large to be induced by that SNPs direct effect on the trait. However, this approach does not easily allow us to interpret variance components of the square of the trait, because of the confounding effect of the number of loci–we cannot say exactly what the heritability of the square of a purely additive trait should be without knowing both the heritability of the trait and the number of loci involved.

Converting from scaled phenotype to real phenotype estimates

For us to interpret out regression estimates in term of theoretical parameters, we scaled all phenotypes to mean zero and unit variance. However, for visualization of the predicted relationship between real phenotypes and rLRS, we must convert our regression estimates back to the real scale. While this is relatively straight forward algebra, the presence of the squared term add some minor additional complexity, which we illustrate here.

Given a sample of paired rLRS and trait values y and x, we convert x to z-scores:

$$z = \frac{x - x_0}{\sigma_x}$$

We have the multiple linear regression model:

$$y = \beta_0 + \beta_1 z + \beta_2 z^2 + \epsilon$$

$$y = \beta_0 + \beta_1(\frac{x - x_0}{\sigma_x}) + \beta_2(\frac{x - x_0}{\sigma_x})^2 + \epsilon$$

$$y = (\beta_0 - \frac{x_0\beta_1}{\sigma_x} + \frac{\beta_2 x_0^2}{\sigma_x^2}) + (\frac{\beta_1}{\sigma_x} - \frac{2\beta_2 x_0}{\sigma_x^2})x + \frac{\beta_2}{\sigma_x^2}x^2 + \epsilon$$

$$y = \beta_0^* + \beta_1^* x + \beta_2^* x^2 + \epsilon$$

A.7 Chapter 4 supplementary figures

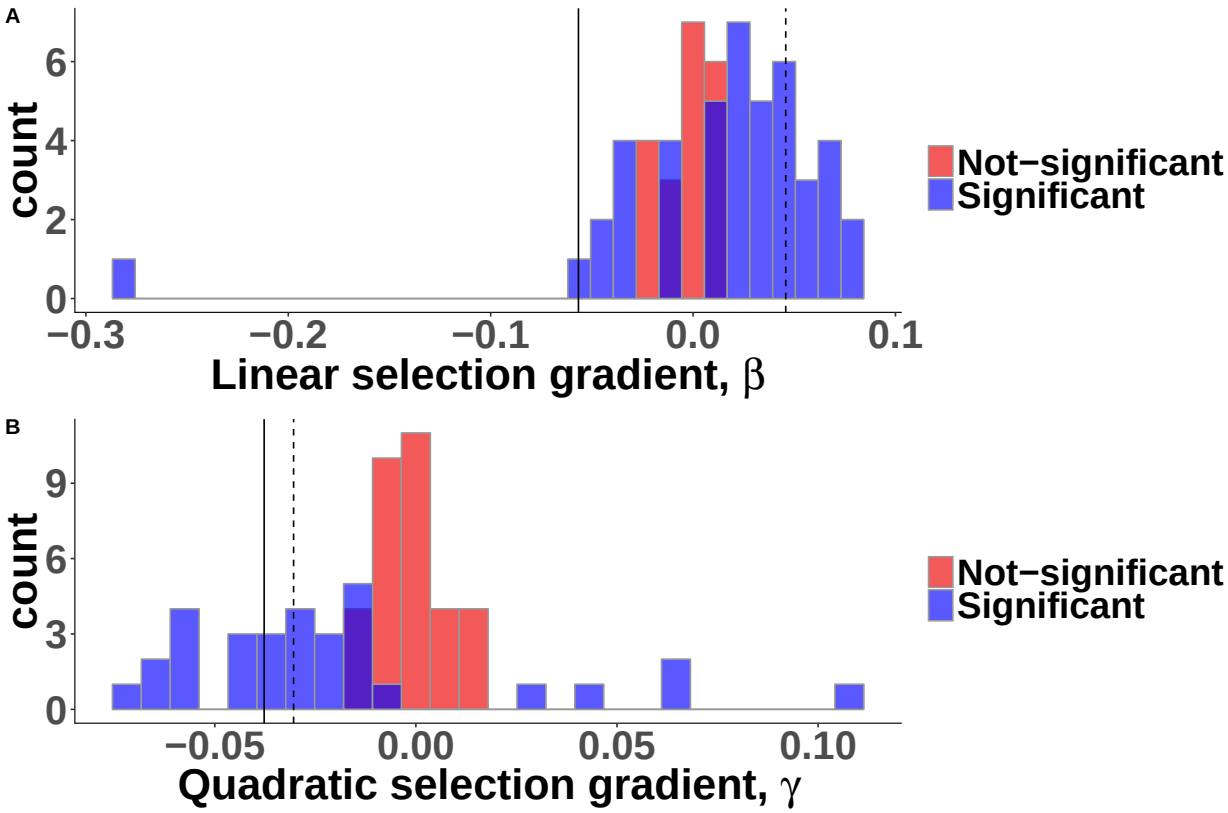

Figure A.26: Histograms showing the distributions of (A) linear and (B) quadratic selection gradients estimated from single trait regressions. Results are not split by sex, i.e. each data point is a result for a specific sex-trait pair. Linear selection gradients are equal to the regression coefficients estimates. Quadratic selection gradients are equal to twice the value of the regression coefficient estimates. A significance cut-off of FWER$\leq$ 0.05 (Bonferroni correction) was chosen for visualization. Vertical lines show the values for Female (solid) and Male (dashed) height.

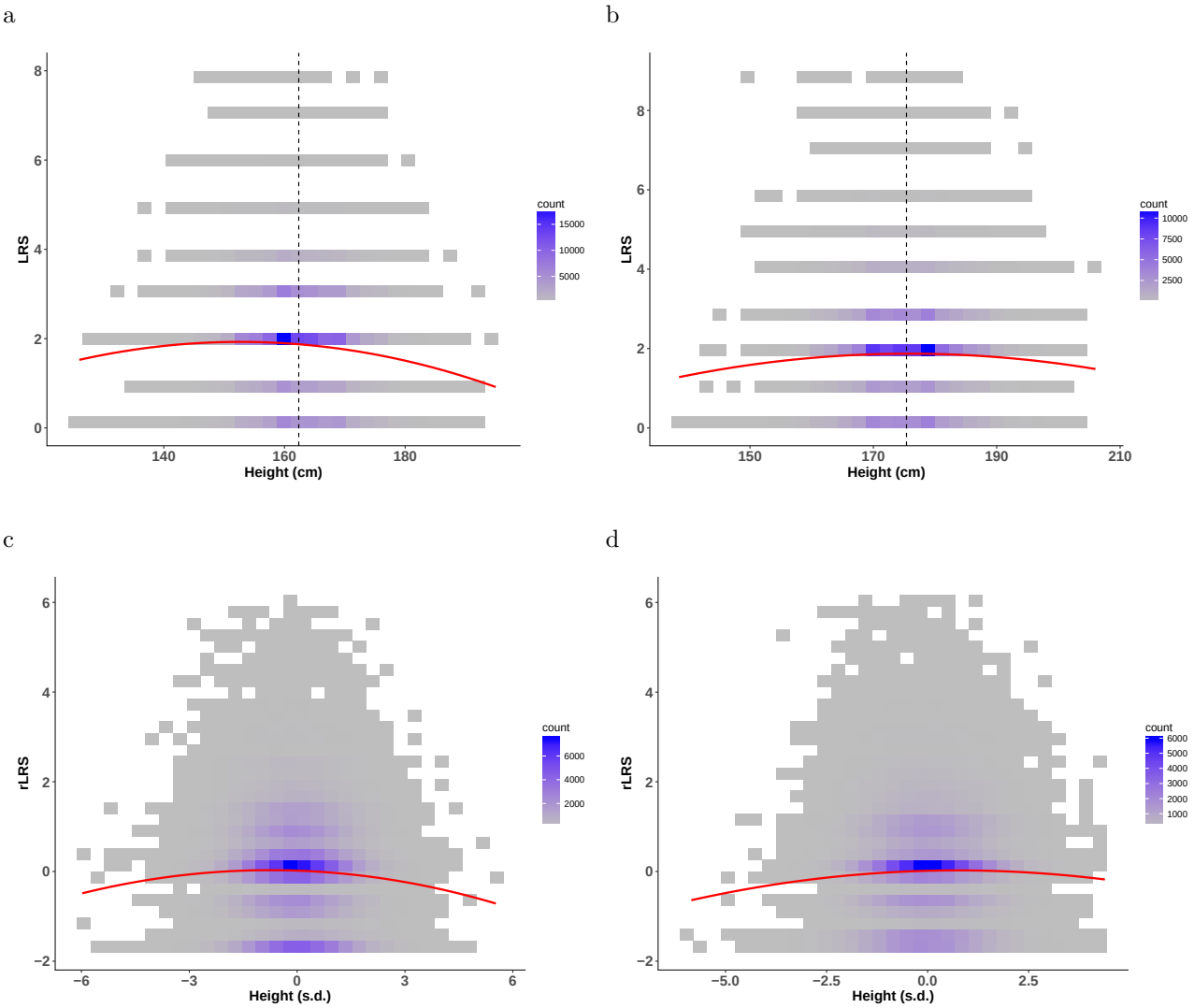

Figure A.27: Empirical relationships between LRS and Height. (Top row) Raw LRS values plotted against raw height values with a quadratic regression line fit to the data for (A) Females and (B) Males with a dashed horizontal line at the sex-specific population mean. (Bottom row) rLRS adjusted for age, birth cohort and assessment center values plotted against centered and scaled height values with a quadratic regression line fit to the data for (C) Females and (D) Males.

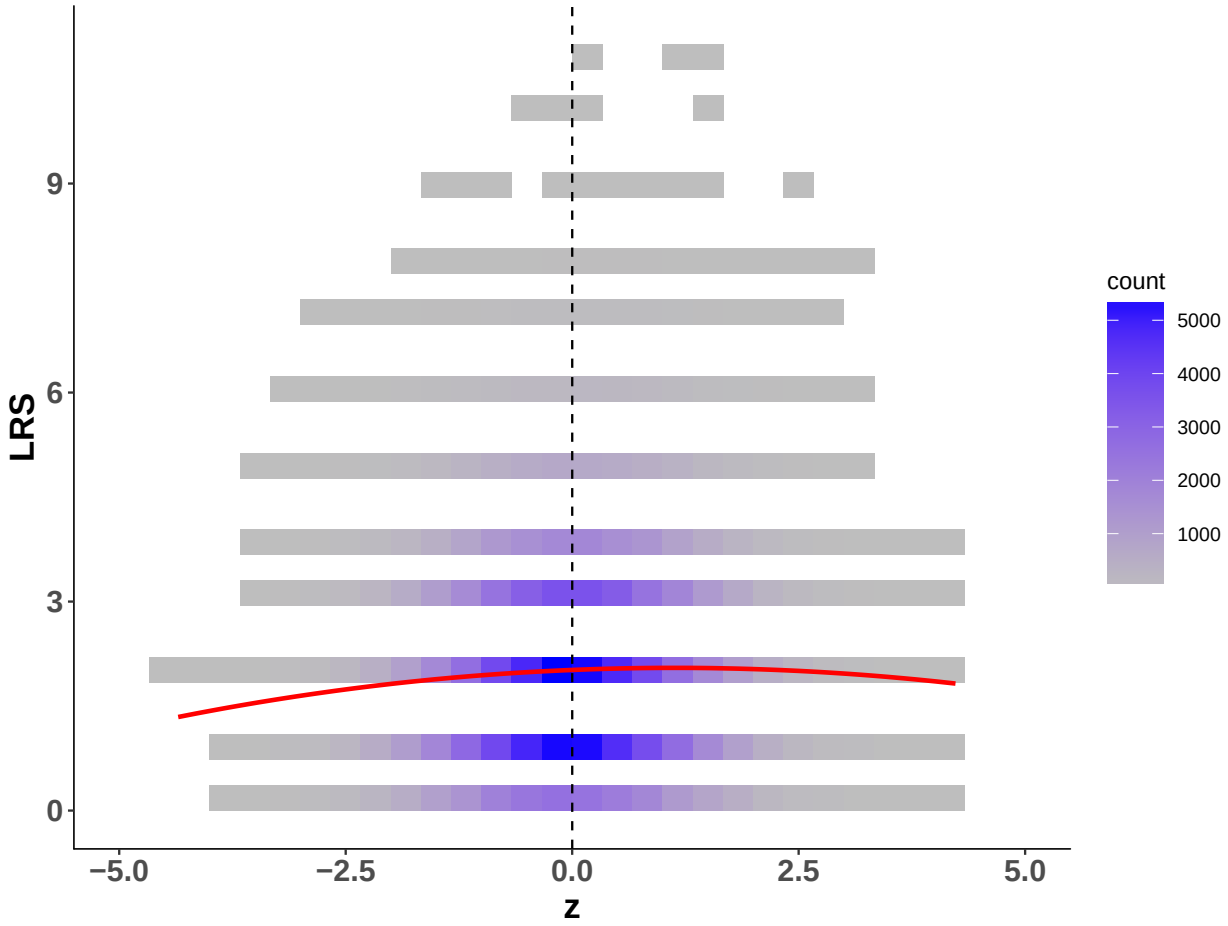

Figure A.28: Simulated relationship between LRS and a trait under stabilizing selection. Trait values for 150,000 individuals were drawn from a unit normal distribution. Fitness values were calculated with a Gaussian stabilizing selection fitness model with an optimum at $z_{opt} = 1$ and $V_s = 40$. Then LRS values were drawn from a poisson distribution with a mean equal to twice the relative fitness of each individual. We use twice the relative fitness so that the mean number of offspring per individual is 2, reflecting a constant population size in a sexual system. This poisson model for LRS will closely approximate the results of a Wright-Fisher model.

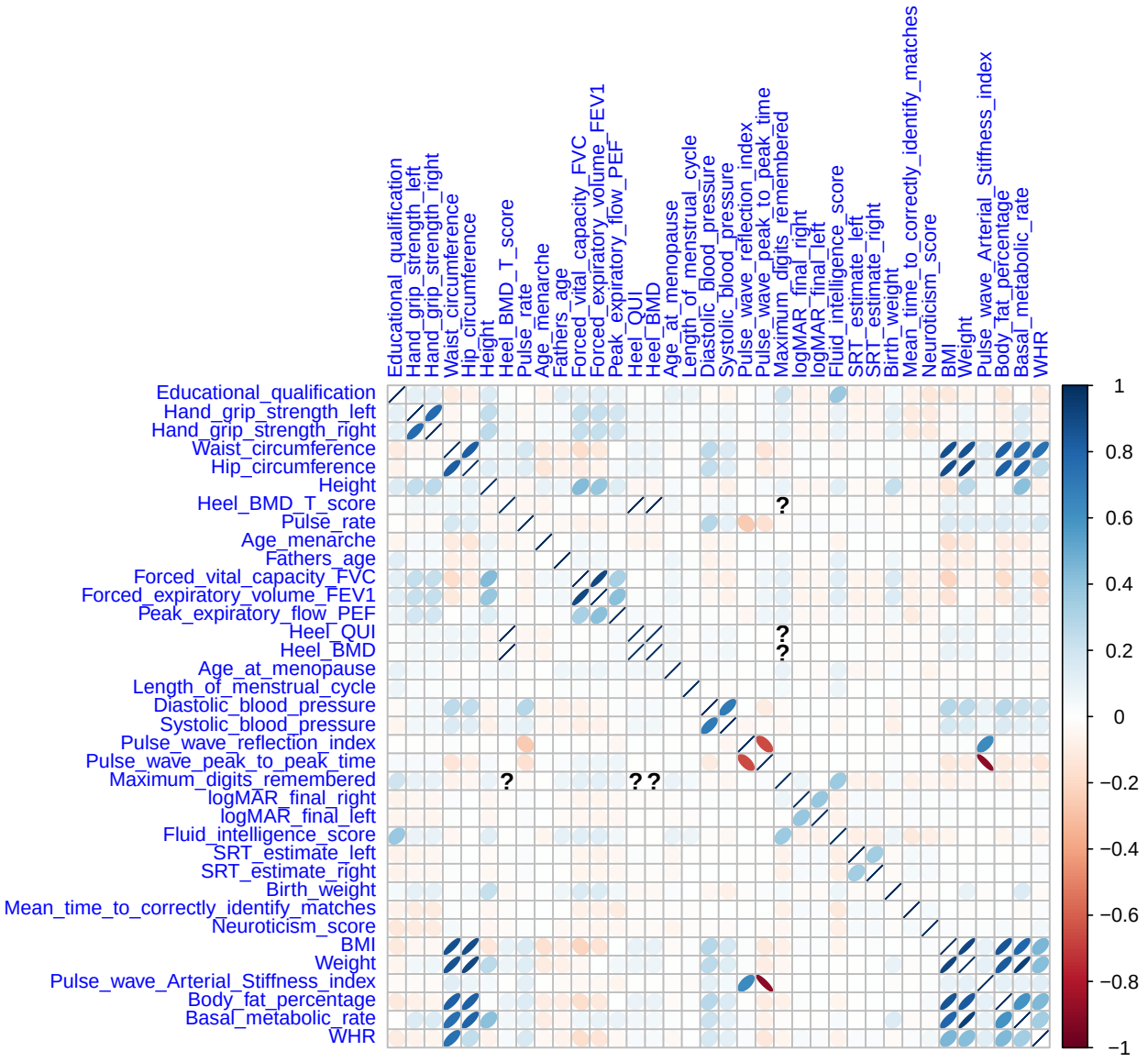

Figure A.29: Phenotypic correlation matrix for Females. Shows the correlation coefficient for the measured phenotypes in females. The color legend is shown on the right hand side, with dark blue and dark red representing strong positive and negative correlations, respectively.

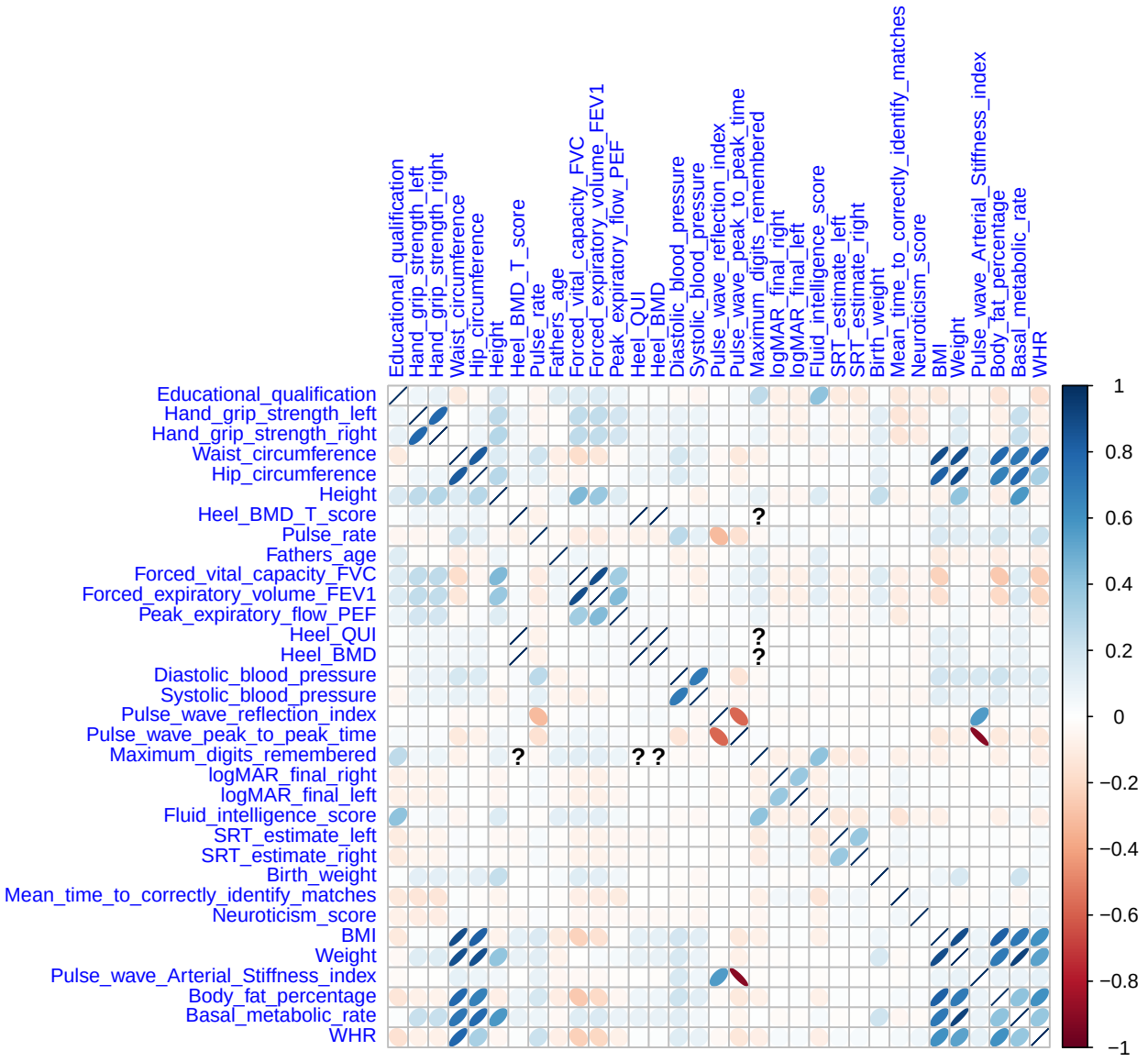

Figure A.30: Phenotypic correlation matrix for Males. Shows the correlation coefficient for the measured phenotypes in males. The color legend is shown on the right hand side, with dark blue and dark red representing strong positive and negative correlations, respectively.

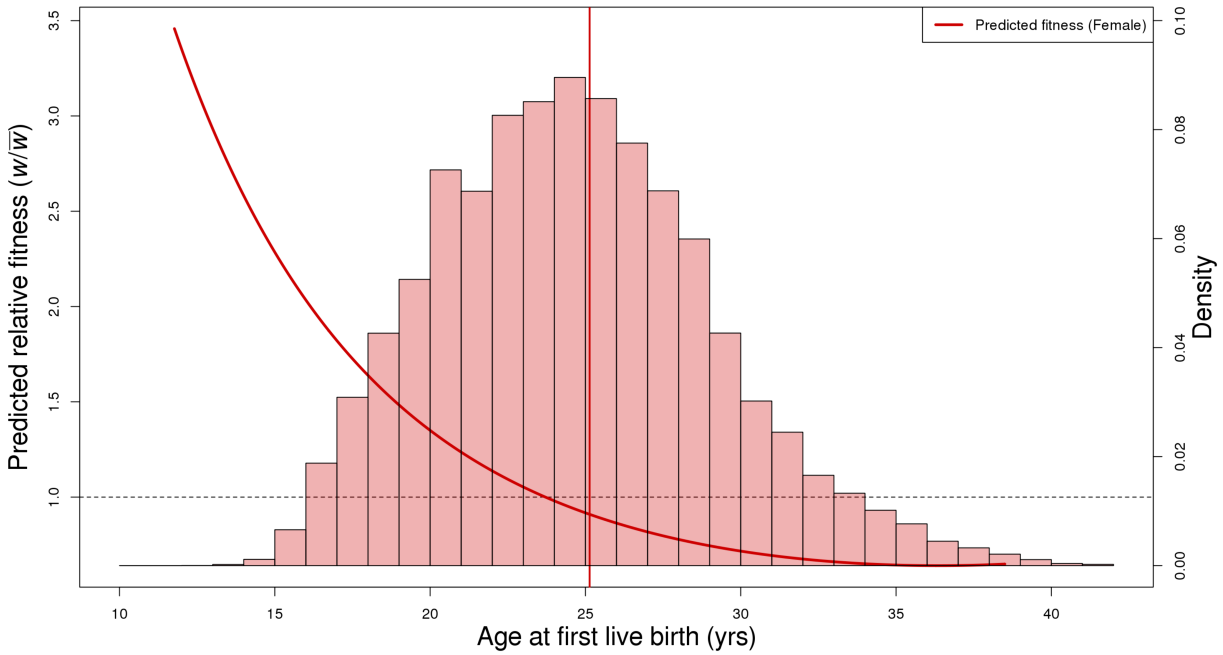

Figure A.31: Predicted relative fitness as a function of age at first live birth. Linear and quadratic selection gradients were converted into parameters of a Gaussian fitness function. Using the parameterized Gaussian fitness function, relative fitness values across the observed phenotypic range were predicted and shown by solid red (female) line. The population means are indicated by vertical solid red (female). Histograms of female (red) phenotypes are overlaid with an axis on the right hand side. The horizontal dashed line indicates a relative predicte fitness of 1.

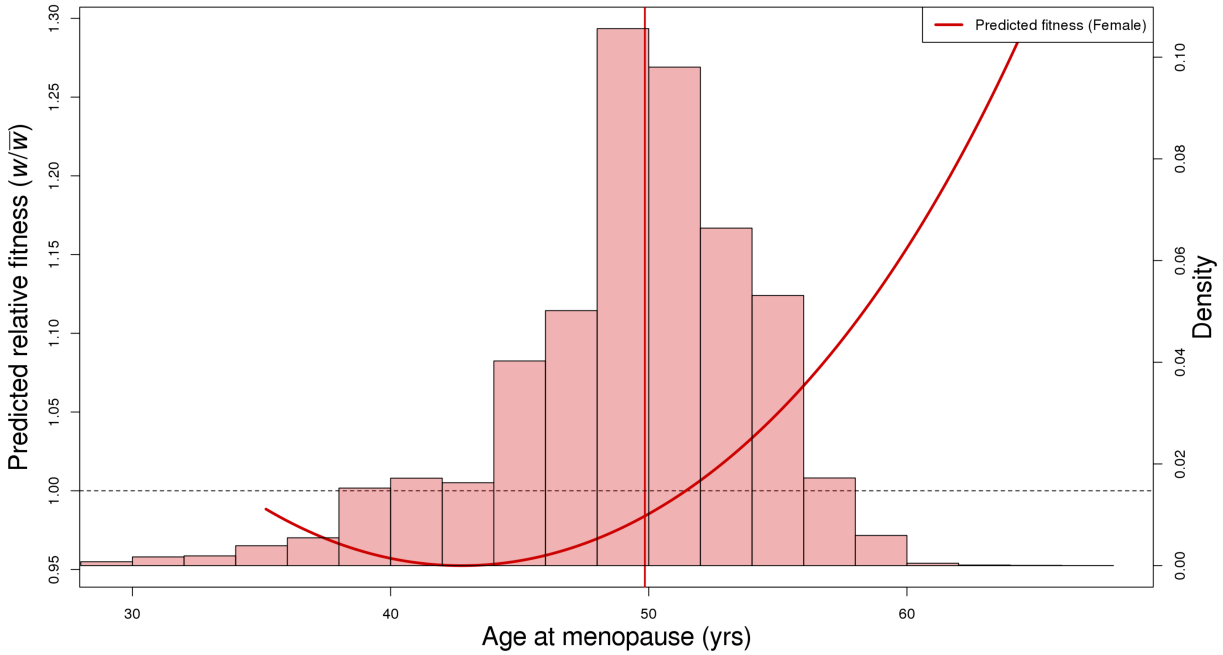

Figure A.32: Predicted relative fitness as a function of age at menopause. Linear and quadratic selection gradients were converted into parameters of a Gaussian fitness function. Using the parameterized Gaussian fitness function, relative fitness values across the observed phenotypic range were predicted and shown by solid red (female) line. The population means are indicated by vertical solid red (female). Histograms of female (red) phenotypes are overlaid with an axis on the right hand side. The horizontal dashed line indicates a relative predicted fitness of 1.

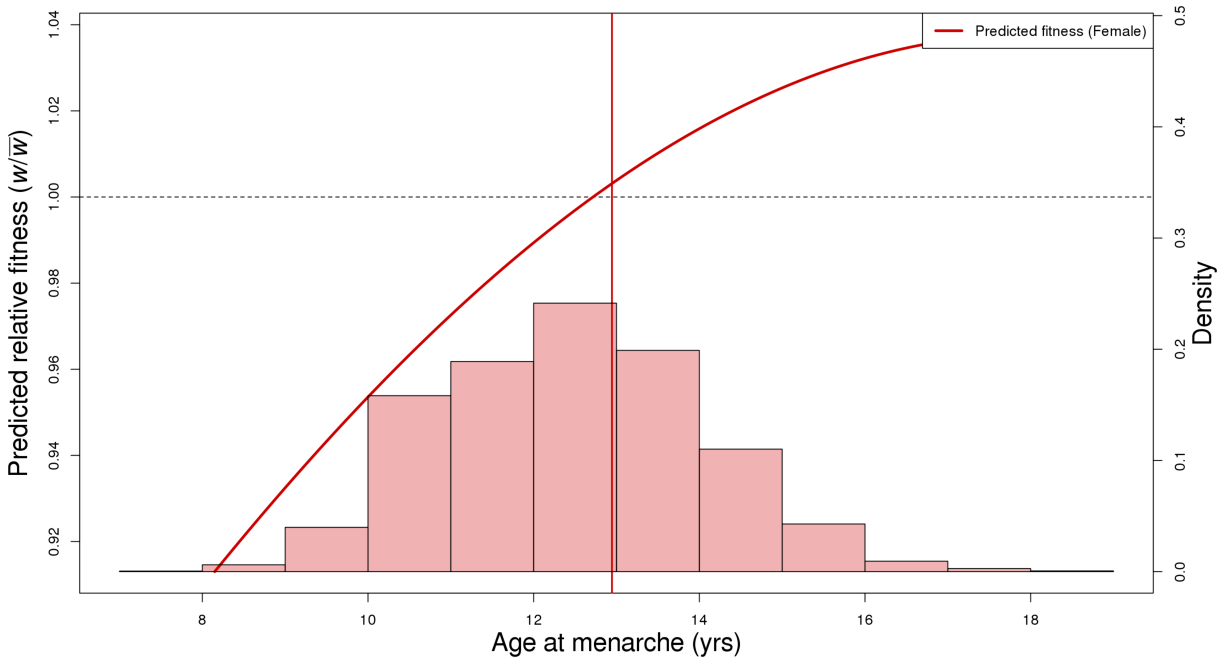

Figure A.33: Predicted relative fitness as a function of age at menarche. Linear and quadratic selection gradients were converted into parameters of a Gaussian fitness function. Using the parameterized Gaussian fitness function, relative fitness values across the observed phenotypic range were predicted and shown by solid red (female) line. The population means are indicated by vertical solid red (female). Histograms of female (red) phenotypes are overlaid with an axis on the right hand side. The horizontal dashed line indicates a relative predicted fitness of 1.

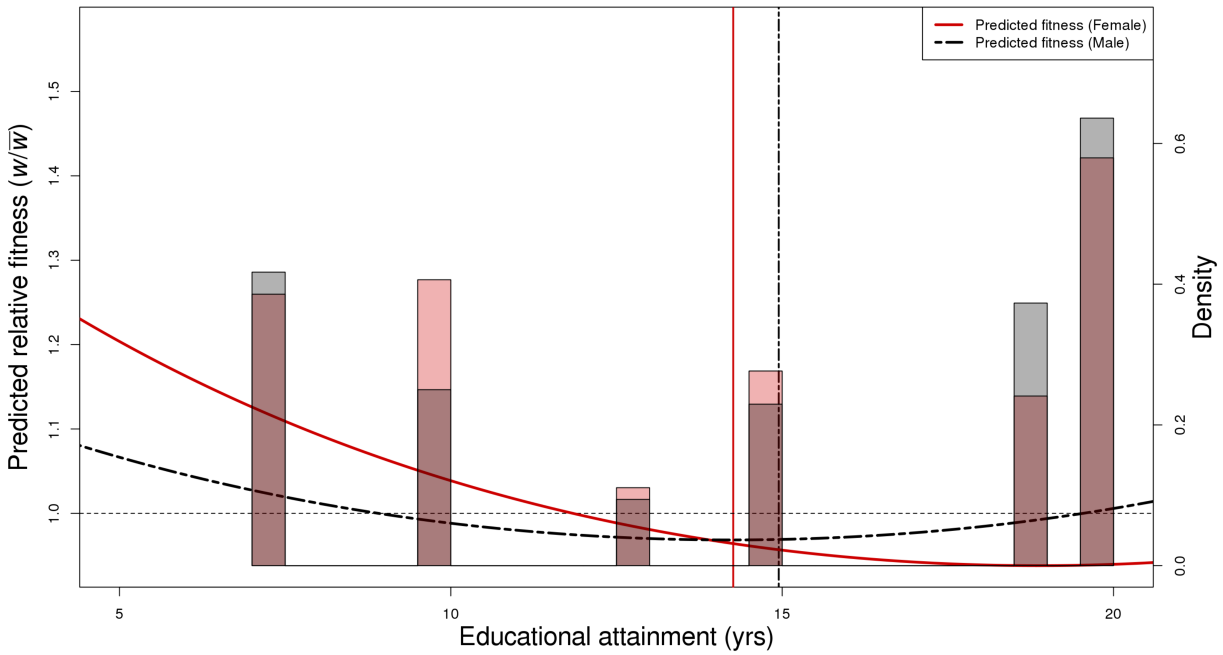

Figure A.34: Predicted relative fitness as a function of educational attainment. Linear and quadratic selection gradients were converted into parameters of a Gaussian fitness function. Using the parameterized Gaussian fitness function, relative fitness values across the observed phenotypic range were predicted and shown by solid red (female) and dashed black (male) lines. The population means are indicated by vertical solid red (female) and dashed black (male) lines. Histograms of female (red) and male (gray) phenotypes are overlaid with an axis on the right hand side. The horizontal dashed line indicates a relative predicted fitness of 1.

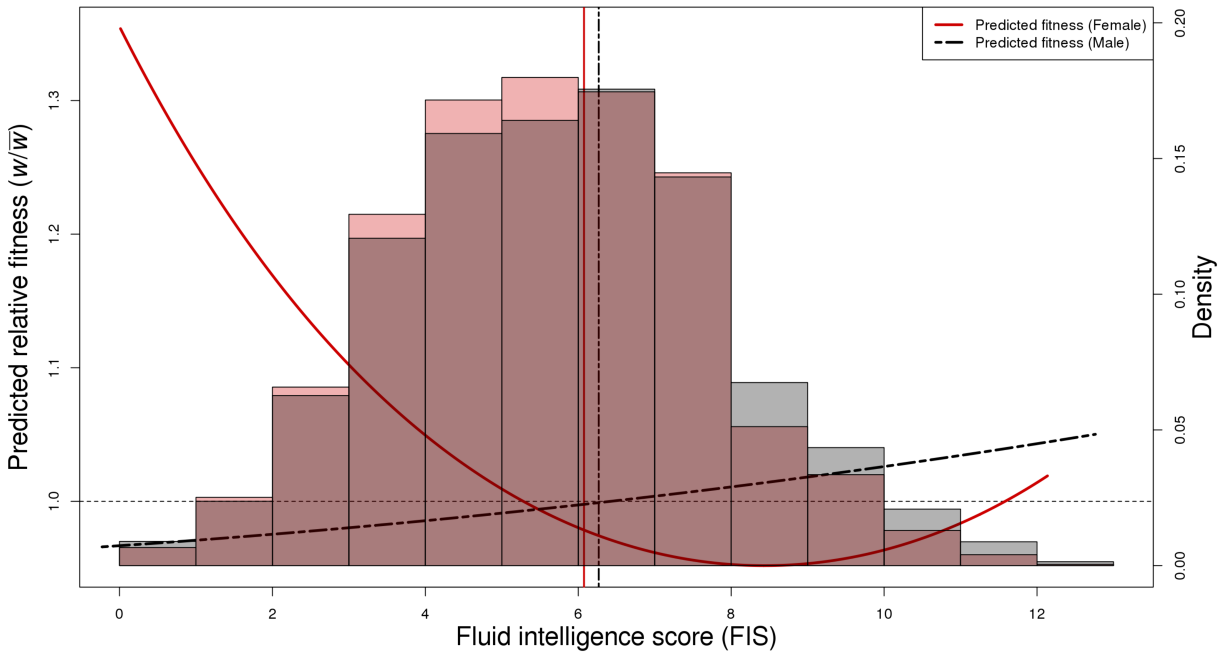

Figure A.35: Predicted relative fitness as a function of Fluid intelligence. Linear and quadratic selection gradients were converted into parameters of a Gaussian fitness function. Using the parameterized Gaussian fitness function, relative fitness values across the observed phenotypic range were predicted and shown by solid red (female) and dashed black (male) lines. The population means are indicated by vertical solid red (female) and dashed black (male) lines. Histograms of female (red) and male (gray) phenotypes are overlaid with an axis on the right hand side. The horizontal dashed line indicates a relative predicted fitness of 1.

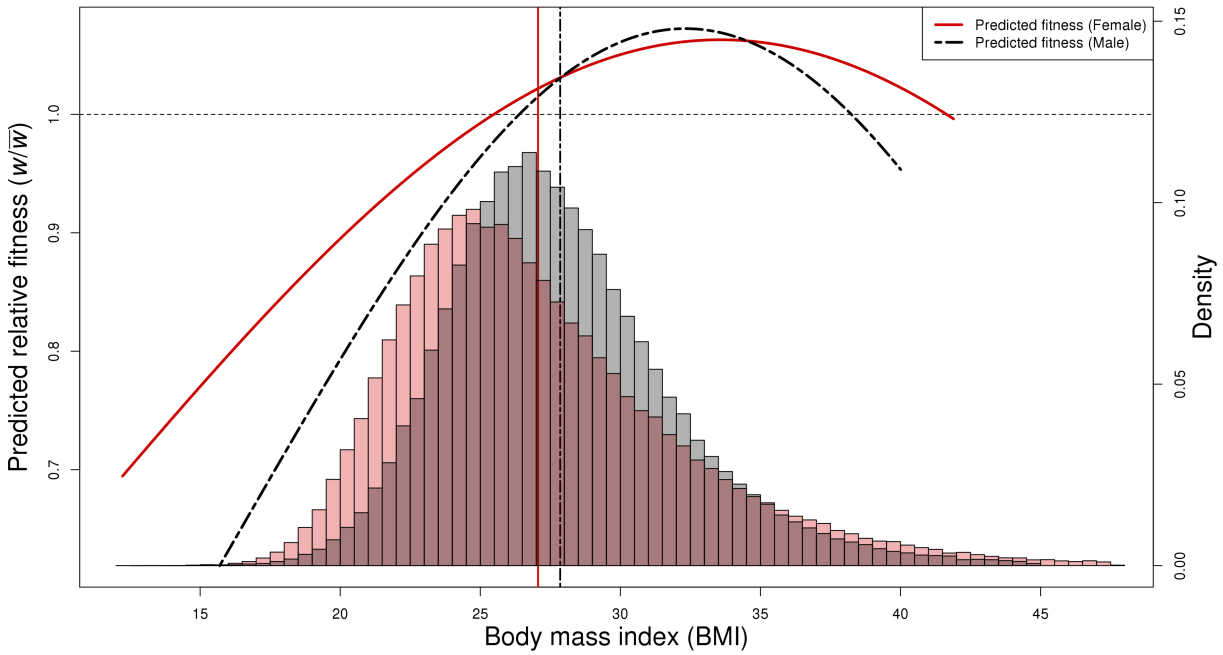

Figure A.36: Predicted relative fitness as a function of BMI. Linear and quadratic selection gradients were converted into parameters of a Gaussian fitness function. Using the parameterized Gaussian fitness function, relative fitness values across the observed phenotypic range were predicted and shown by solid red (female) and dashed black (male) lines. The population means are indicated by vertical solid red (female) and dashed black (male) lines. Histograms of female (red) and male (gray) phenotypes are overlaid with an axis on the right hand side. The horizontal dashed line indicates a relative predicted fitness of 1.

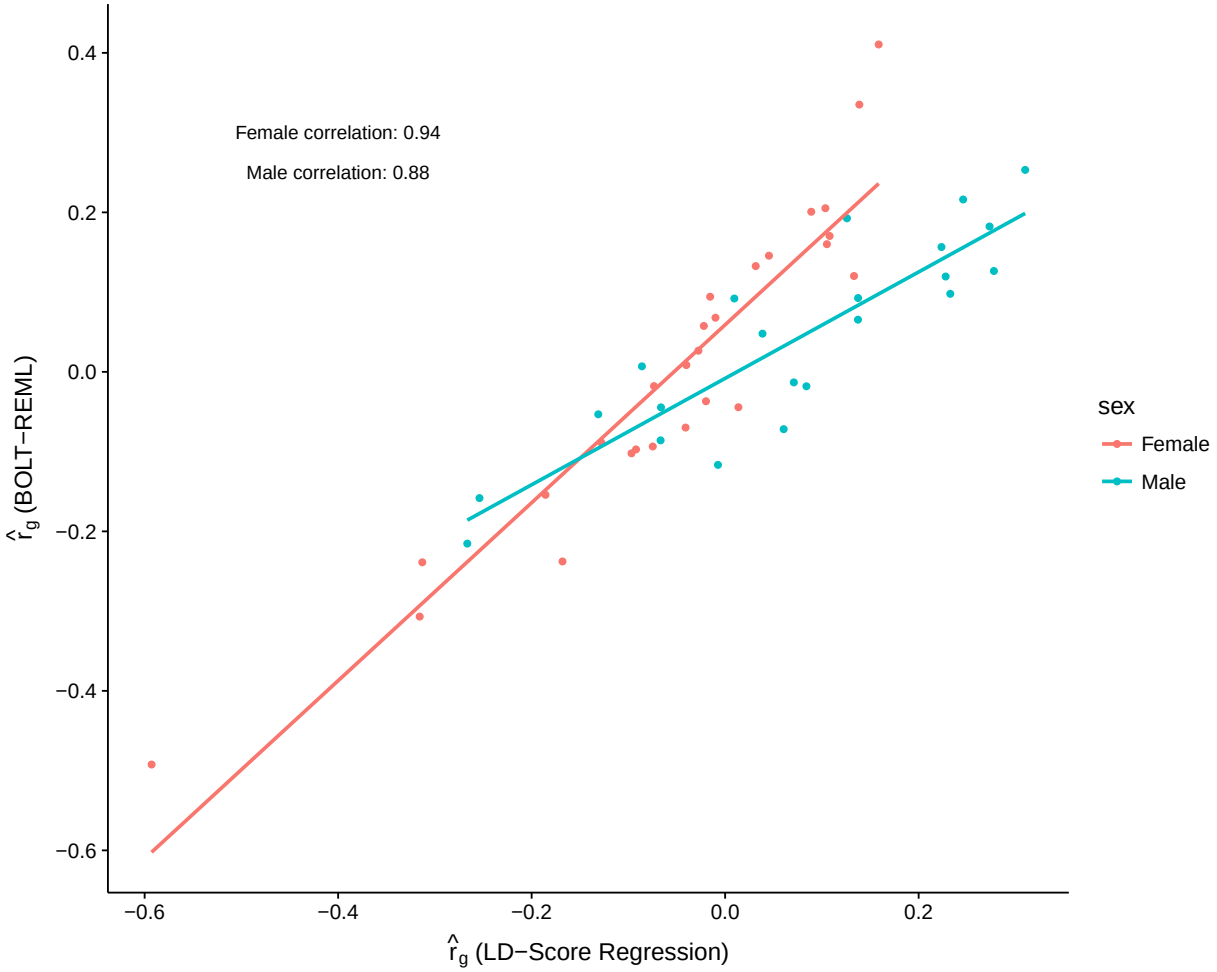

Figure A.38: Relationship between genetic correlation estimates. A scatter plot of the genetic correlation estimates from the UKB interim data release using a full REML approach versus genetic correlation estimates from the full UKB data release obtained from LD-score regression.

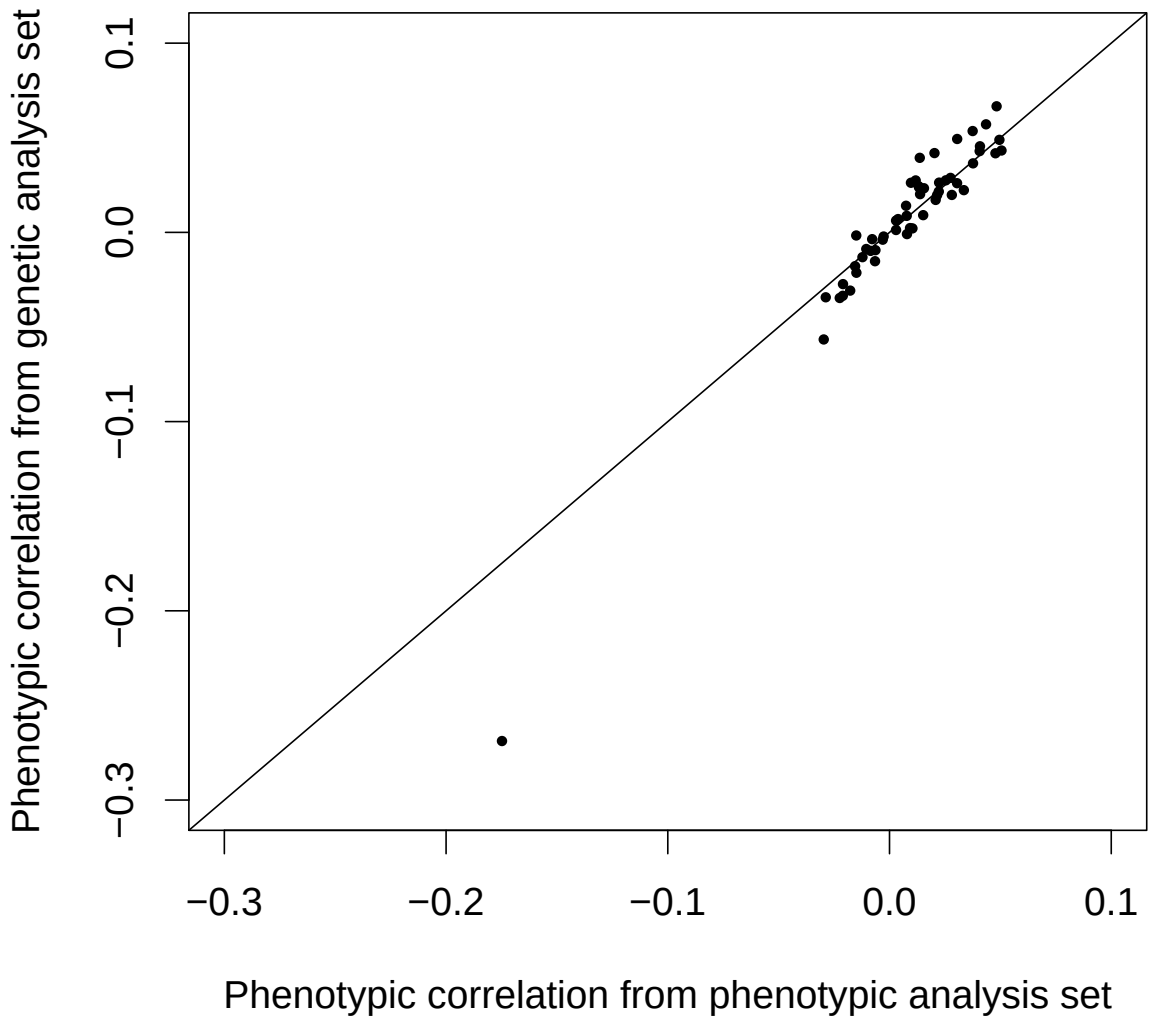

Figure A.39: Relationship between phenotypic correlation estimates. Phenotypic correlations are estimated from the phenotypic and genetic analyes($R^2 = 0.945$).

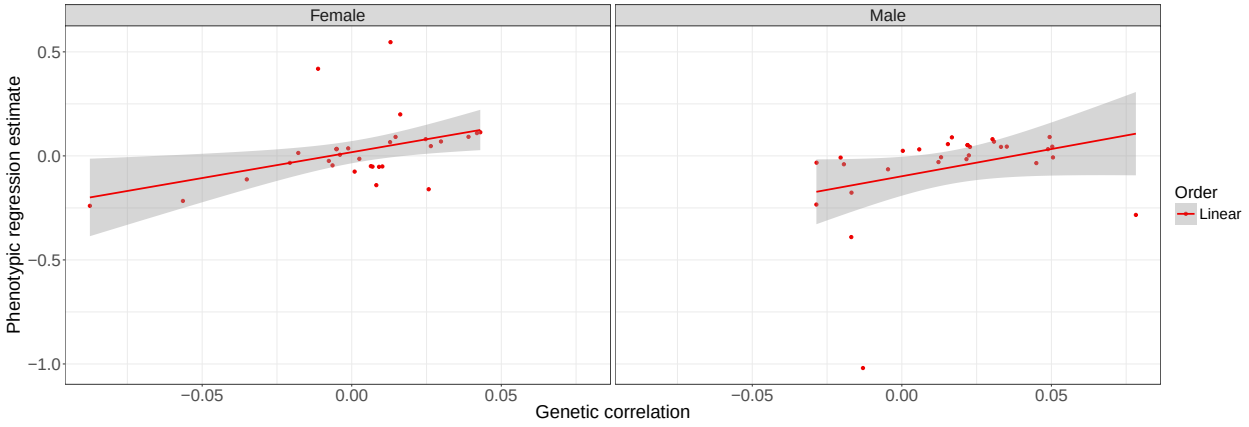

Figure A.40: Correlation between phenotypic regression estimates and genetic correlations. Phenotypic regression estimates and genetic correlations were grouped by sex and order. For females, there is significant correlation between phenotypic regressions and genetic correlations (Linear: $R^2 = 0.427$, $p < 10^{-4}$; Quadratic: $R^2 = 0.376$, $p = 0.0011$). For males,only the linear terms are marginally significantly correlated (Linear: $R^2 = 0.174$, $p = 0.014$; Quadratic: $R^2 = 0.092$, $p = 0.105$).

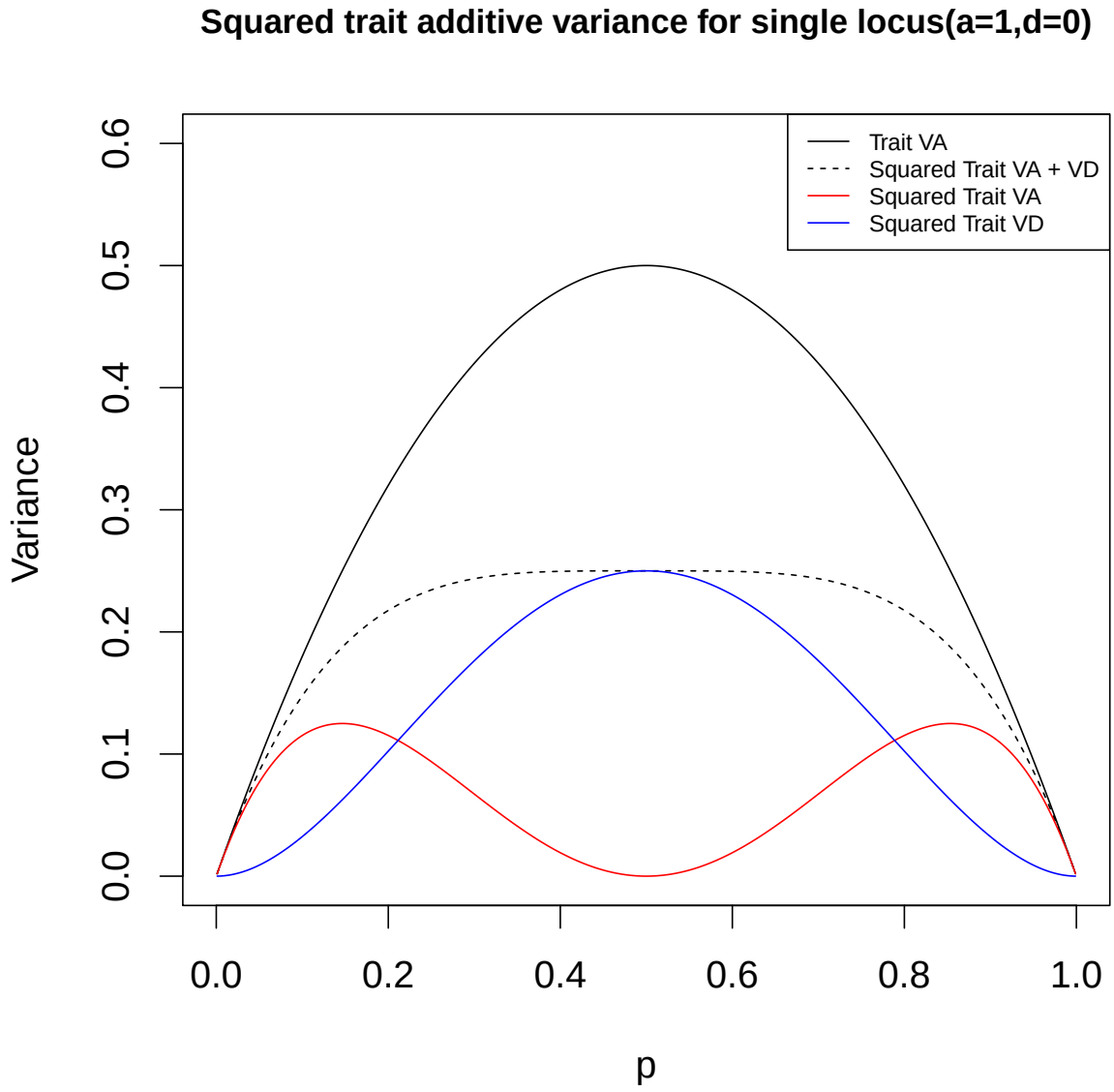

Figure A.41: Genetic variance for a squared trait determined by a single biallelic locus. In this case, the effect of the alleles is set to be $a = 1$.

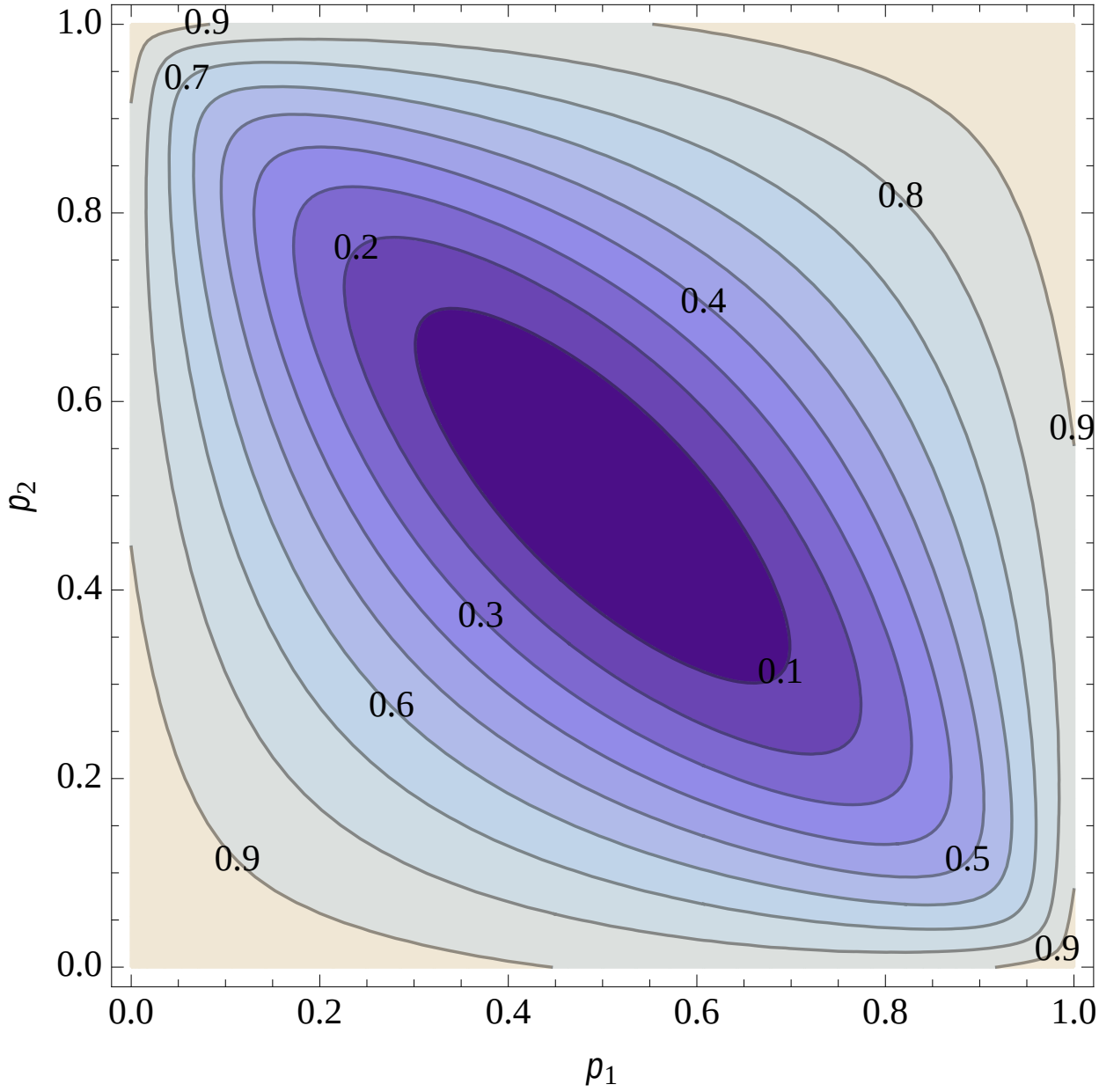

Figure A.42: The percentage of genetic variance for a squared trait which is purely additive. Under a purely additive two locus model with equal effects ($a = 1$), the squared trait contains additive, dominance and additive by additive epistatic variance components. The relative magnitudes of each component, depends on the allele frequencies. The axes represent allele frequencies at each locus and the color shows the relative magnitude of the additive variance component. Lighter colors show a large additive component.

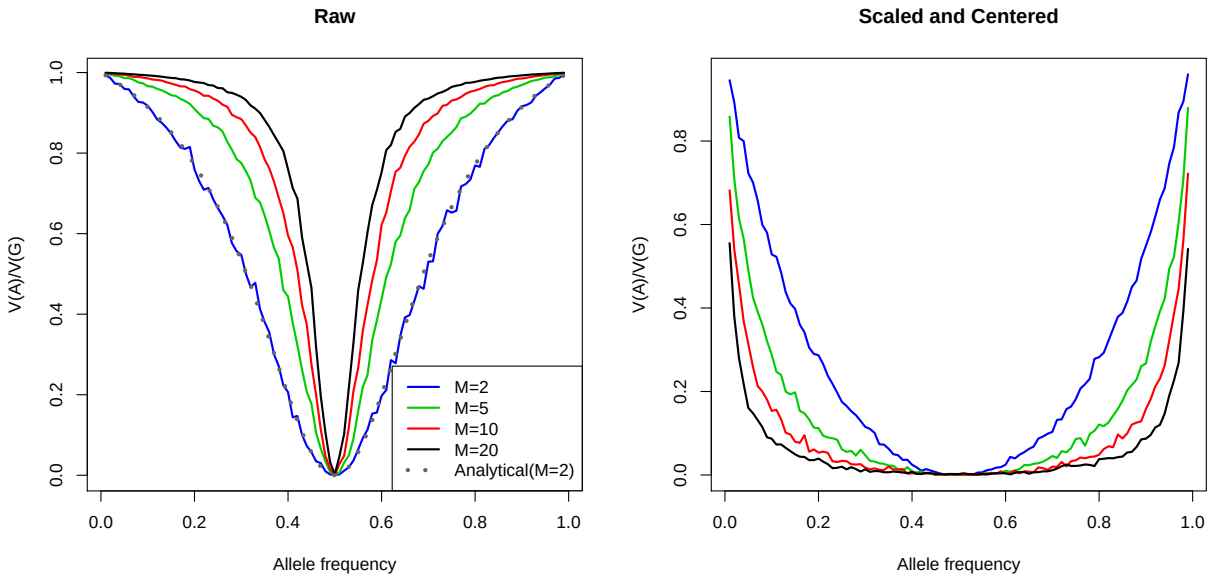

Figure A.43: Regression based estimates of the percentage of genetic variance for a squared trait which is purely additive, under a purely additive multilocus model with equal effects ($a = 1$). The regression was performed on the raw unscaled trait values (left) and the scaled(right) trait values. The number of markers was varied from M=2 to M=20. The analytic expression for M=2, which does not assume scaling, is displayed as a dotted line.

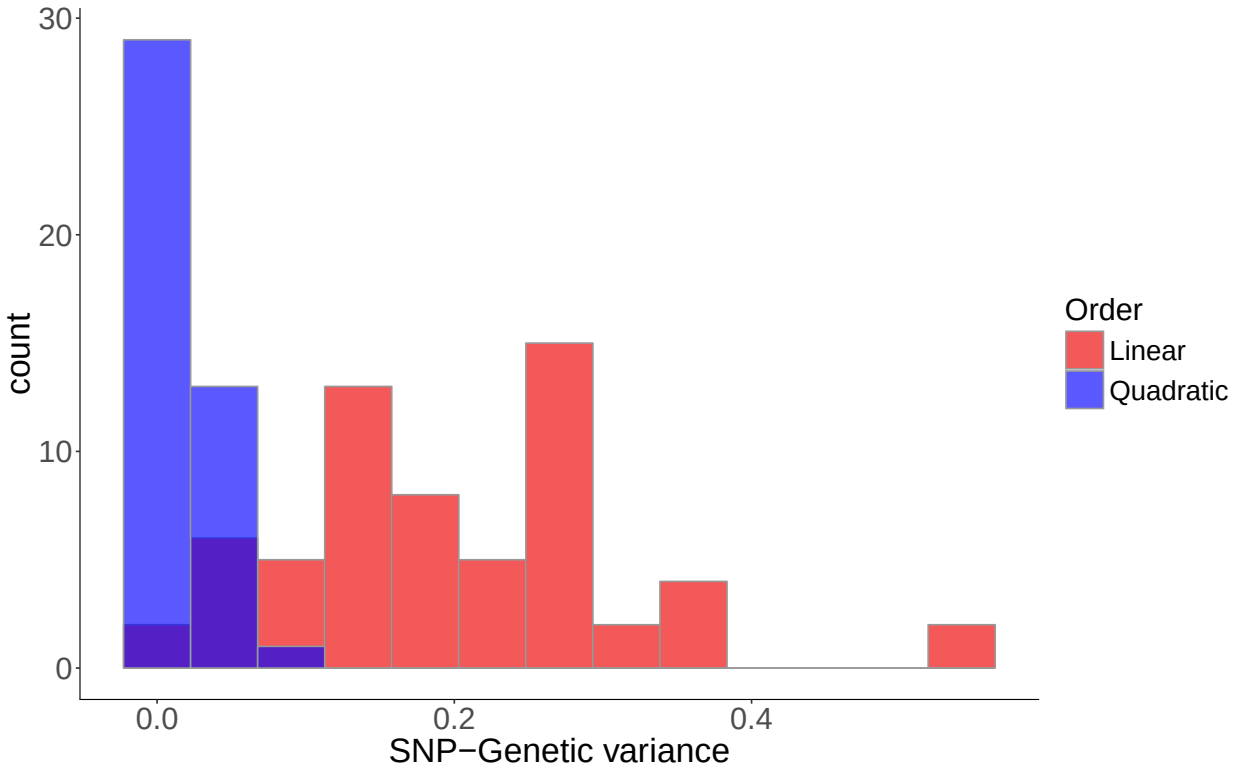

Figure A.44: The distribution of genetic variance explained by common SNPs. Estimates of genetic variance were for traits (red) and squared traits(blue) were obtained from the bivariate REML analysis in BOLT-LMM. The same samples were used in each case. Data for males and females were pooled here.

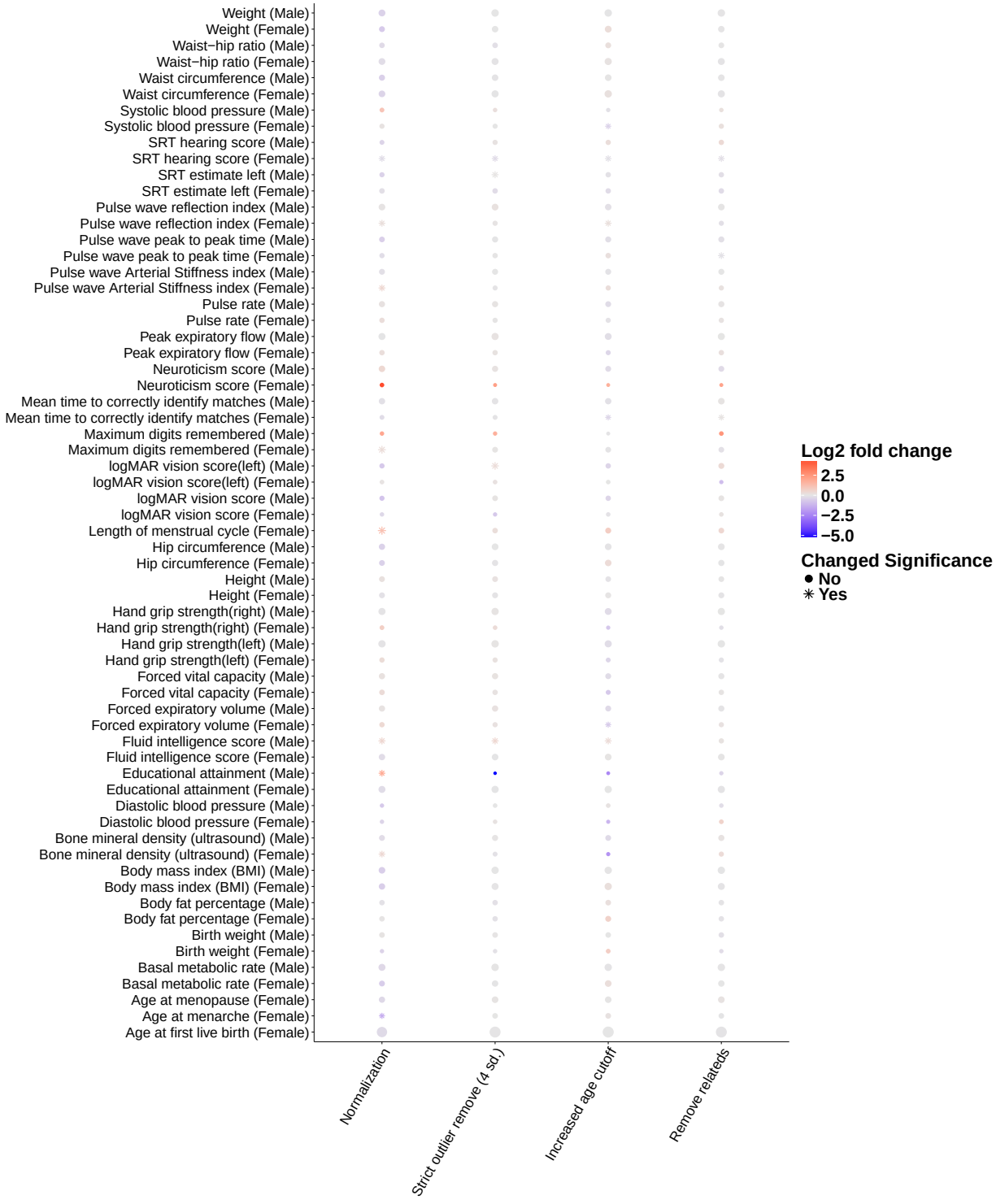

Figure A.45: Sensitivity analysis of linear selection gradients. Each set of β estimates is compared to a baseline analysis described in the the text. For each data QC pipeline we display the log fold-change in the value of β compared to and whether the estimate changed significance status. Red coloration implies that β increased in absolute magnitude and thus became more significant and vice versa for blue coloration.

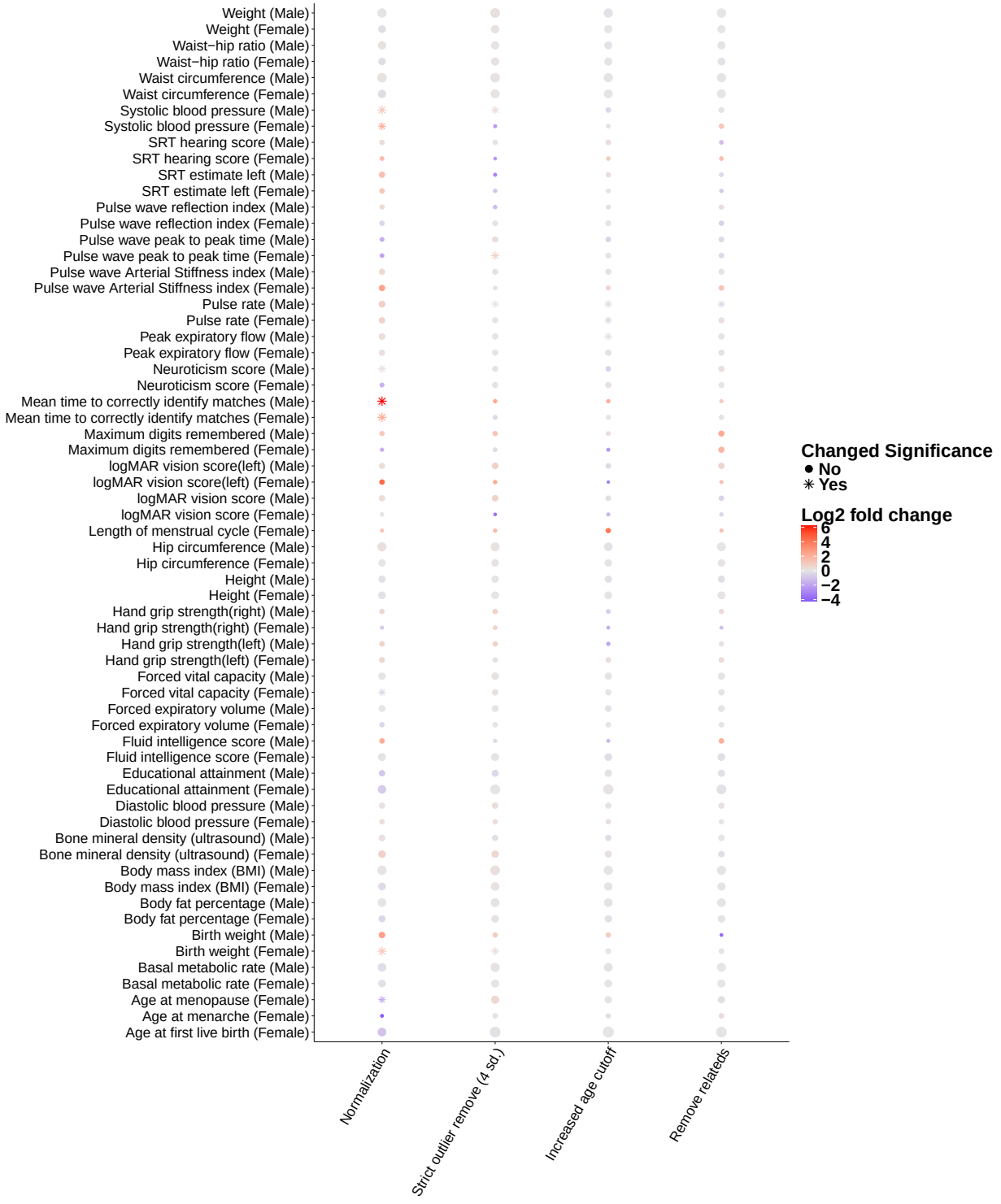

Figure A.46: Sensitivity analysis of quadratic selection gradients. Each set of γ estimates is compared to a baseline analysis described in the the text. For each data QC pipeline we display the log fold-change in the value of γ compared to and whether the estimate changed significance status. Red coloration implies that γ increased in absolute magnitude and thus became more significant and vice versa for blue coloration.

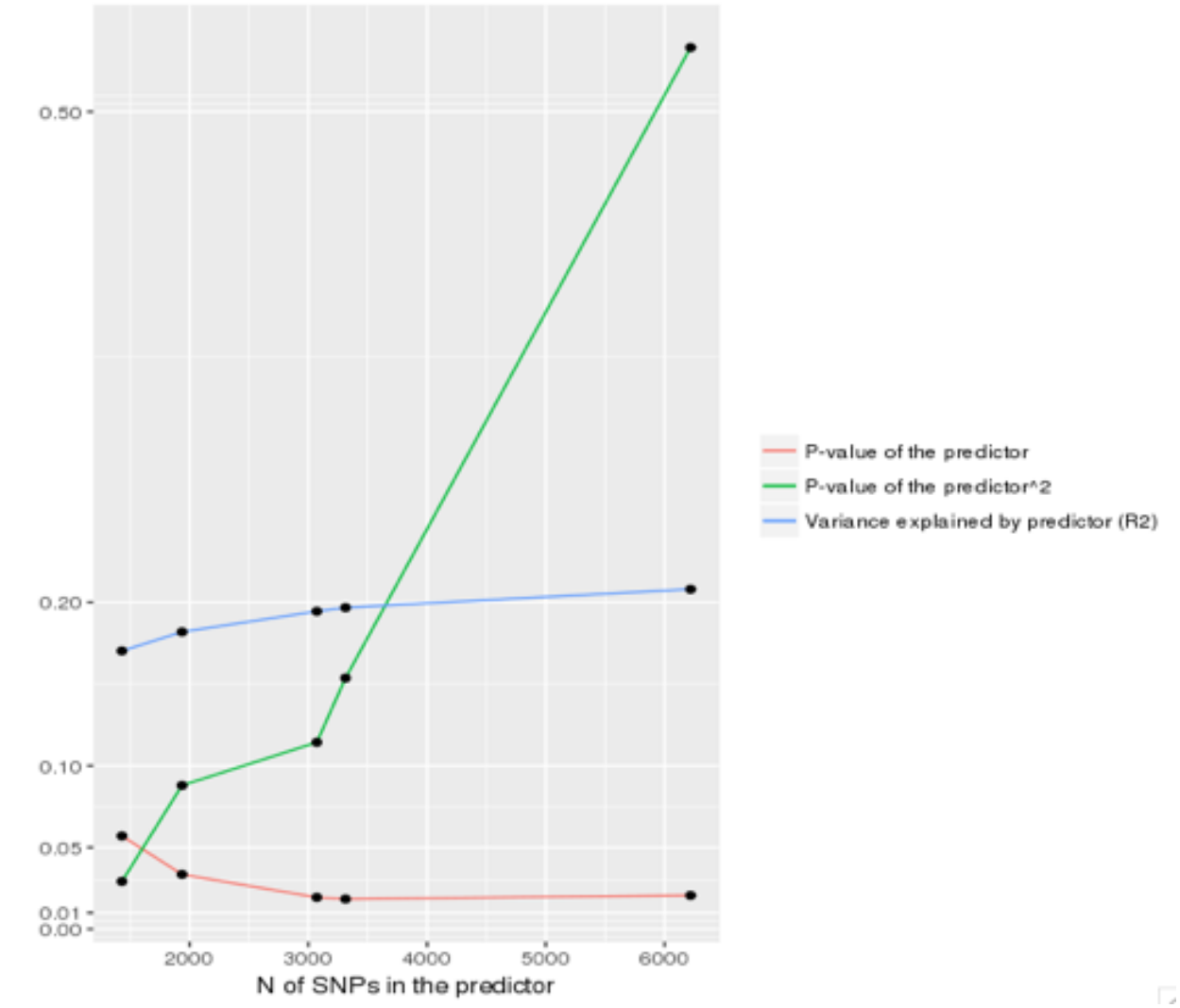

Figure A.47: Performance of polygenic predictor for height. Polygenic predictors for height and squared height were constructed based on genetic association statistics obtained from a meta-analysis of the UKB and GIANT data. The R^2 score and p-values for each predictor are plotted against the number of SNPs included in the model.

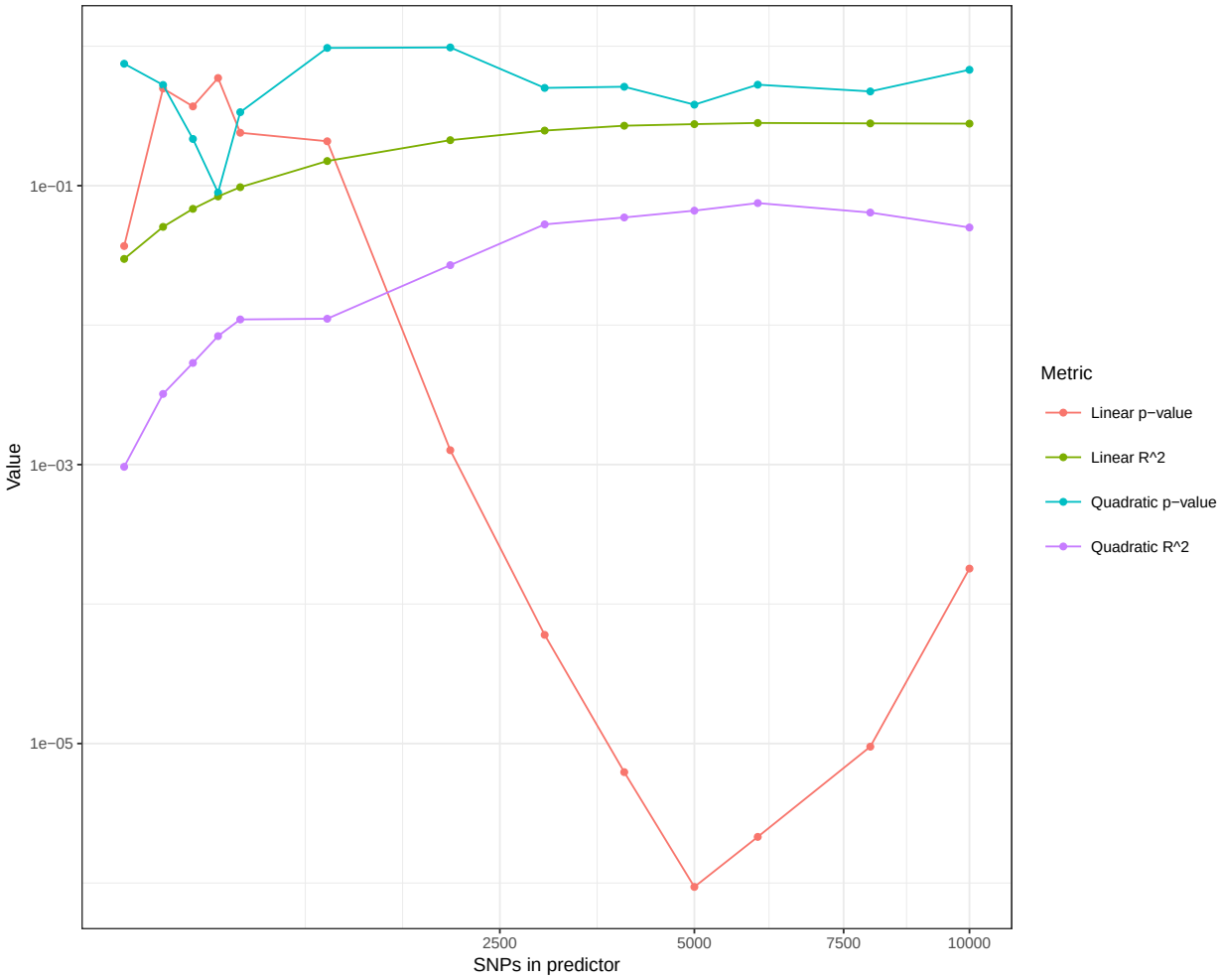

Figure A.48: Performance of simulated polygenic predictors. A phenotype was simulated under a polygenic model with 20,000 causal markers. SNP effects were estimate in a discovery panel of 300,000 individuals. Polygenic predictors for the phenotype and squared phenotype were constructed in a test panel of 50,000 individuals. The R^2 score and p-values for each predictor are plotted against the number of SNPs included in the model.

A.8 Chapter 4 supplementary tables

Trait	Sex	Order	Estimate	$-log_{10}(p)$
Educational attainment	Female	1	0.04	23
Age at menarche	Female	1	0.02	4
Age at first live birth	Female	1	-0.18	>220
Bone mineral density (ultrasound)	Female	1	-0.02	5
Systolic blood pressure	Female	1	-0.03	8
Waist-hip ratio	Female	1	0.04	14
Hand grip strength(right)	Male	1	0.05	15
Pulse rate	Male	1	-0.04	7
Systolic blood pressure	Male	1	-0.03	4
Body mass index (BMI)	Male	1	0.08	17
Age at first live birth	Female	2	0.04	43
Body mass index (BMI)	Female	2	0.01	5
Educational attainment	Male	2	0.05	6
Body mass index (BMI)	Male	2	-0.03	13

Table A.7: Summary of multiple regression of traits onto rLRS. Traits which were marginally significant in a multiple regression model are displayed below. The multiple regression was carried out separately for each sex. The full results of multiple regression are contained in the Dataset A.10.

Predictor	Estimate	$-log10(p)$
Educational Attainment (EA)	0.032	75
Age at First Birth (AFB)	0.167	> 220
Interaction (EA:AFB)	0.038	112

Table A.8: Summary of multiple regression of educational attainment and age at first birth onto rLRS. The linear model $rLRA\ EA + AFB + EA * AFB + \epsilon$ was fit to the female samples. Both predictors as well as their interaction term were found to be statistically significant.

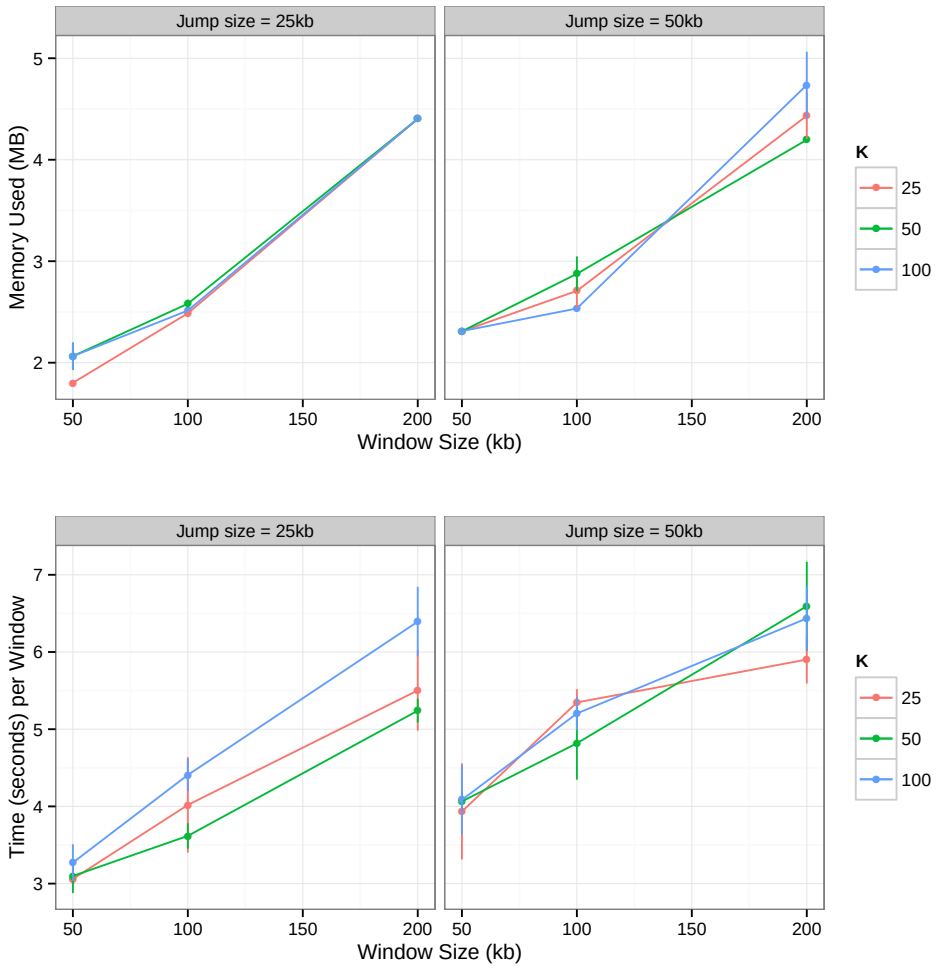

Figure A.25: This shows the peak memory use (top) and time per window (bottom) as a function of window size for a panel of ESM test runs with 2,000,000 permutations of chromosome 15 for bipolar disorder. The permutations are done in PLINK 1.90a and takes approximately 12 CPU-hours per million permutations to perform and write to disk; typically we performed analysis on an HPC cluster. We varied the size of the genomic window, the jump size (related to window overlap), and K (the number of markers used per window). Each point is the mean over only three replicates, but each replicate involves thousands of windows. In general, the window size is the primary contributor to time per window. Benchmark runs were performed were performed as 4-core jobs as was done in the full scale analysis. Peak memory use is quite low and the multithreading was primarily used to save user time. A larger jump size between windows is less efficient per window because of the HDF5 chunking strategy. When windows have significant overlap, data from an HDF5 chunk may be read once and used in analyzing multiple windows, thus driving down average time per window. Additionally, when using multiple cores, a larger jump size means more data is needed per set of windows and peak memory use is also higher.

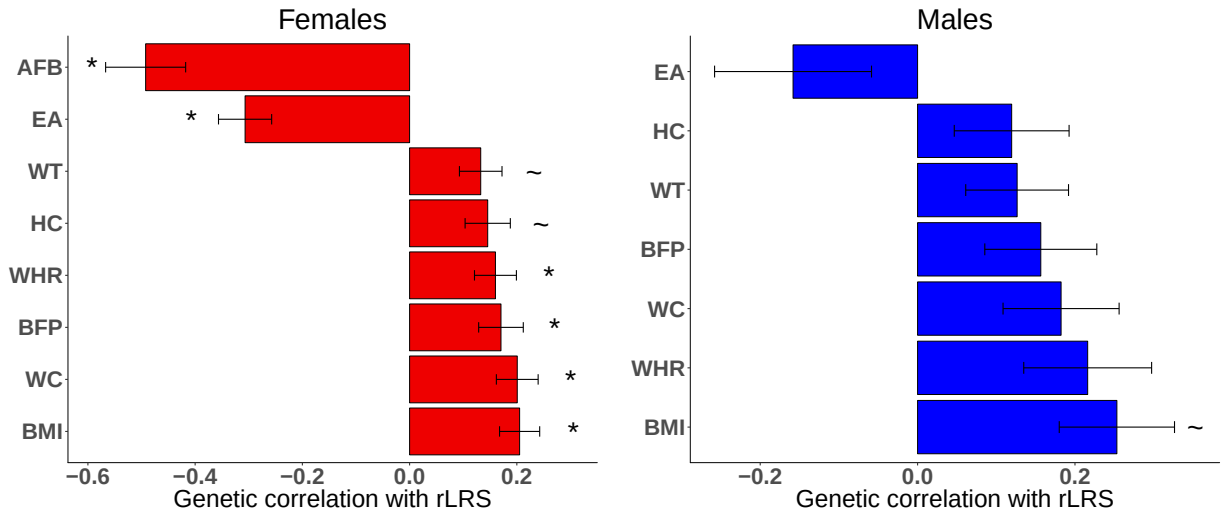

Figure A.37: Bar plots showing BOLT-REML estimates of genetic correlations between a selection of traits and rLRS for Females (red) and Males (blue) . Traits were selected on the basis of being marginally significant ($p \leq 0.001$) in at least one sex, and were sorted in ascending order of the estimate for each sex. Data are displayed as the correlation estimate plus or minus the standard error.($\sim$ $p \leq 0.001$, *FWER$\leq$ 0.05) Bars are labeled with abbreviated trait descriptions described in the text.

Predictor	Sex	$r_{g,LDSC}$	$r_{p,full}$	$r_{p,BOLT}$	$r_{g,BOLT}$	$r_{e,BOLT}$
Age at first live birth	Female	-0.593	-0.175	-0.269	-0.492	-0.254
Age at menopause	Female	-0.168	0.028	0.020	-0.238	0.052
Age at menarche	Female	0.133	0.013	0.024	0.120	0.008
Basal metabolic rate	Female	-0.015	0.025	0.028	0.094	0.016
Birth weight	Female	-0.073	0.003	0.001	-0.018	0.003
Body mass index (BMI)	Female	0.104	0.038	0.036	0.205	0.009
Body fat percentage	Female	0.108	0.012	0.027	0.170	0.004
Diastolic blood pressure	Female	-0.022	-0.003	-0.004	0.058	-0.013
Fluid intelligence score	Female	-0.313	-0.030	-0.057	-0.239	-0.032
Forced expiratory volume	Female	-0.075	0.008	-0.001	-0.094	0.015
Forced vital capacity	Female	-0.092	0.010	0.002	-0.097	0.020
Hand grip strength(right)	Female	-0.097	0.003	0.006	-0.102	0.019
Bone mineral density (ultrasound)	Female	-0.010	-0.007	-0.015	0.068	-0.035
Height	Female	-0.128	-0.018	-0.031	-0.089	-0.019
Hip circumference	Female	0.045	0.022	0.026	0.146	0.008
Maximum digits remembered	Female	-0.028	-0.016	-0.018	0.027	-0.022
Mean time to correctly identify matches	Female	0.014	-0.006	-0.009	-0.044	-0.007
Peak expiratory flow	Female	-0.041	0.009	0.002	-0.070	0.011
Pulse rate	Female	-0.020	-0.009	-0.010	-0.037	-0.006
Pulse wave Arterial Stiffness index	Female	-0.186	0.007	0.014	-0.154	0.026
Pulse wave peak to peak time	Female	0.159	-0.010	-0.009	0.411	-0.039
SRT hearing score	Female	0.139	0.010	0.026	0.335	0.015
Systolic blood pressure	Female	-0.040	-0.008	-0.004	0.009	-0.005
Waist circumference	Female	0.089	0.041	0.043	0.201	0.019
Weight	Female	0.032	0.028	0.029	0.133	0.012
Waist-hip ratio	Female	0.105	0.037	0.054	0.160	0.038
Basal metabolic rate	Male	0.233	0.050	0.049	0.098	0.048
Birth weight	Male	0.071	0.014	0.020	-0.013	0.023
Body mass index (BMI)	Male	0.310	0.048	0.042	0.253	0.020
Body fat percentage	Male	0.224	0.013	0.024	0.157	0.010
Diastolic blood pressure	Male	0.137	0.004	0.007	0.065	0.003
Forced expiratory volume	Male	-0.086	0.022	0.022	0.007	0.025
Forced vital capacity	Male	-0.066	0.021	0.020	-0.045	0.028
Hand grip strength(right)	Male	0.060	0.044	0.057	-0.072	0.069
Bone mineral density (ultrasound)	Male	0.126	0.021	0.017	0.193	0.003
Height	Male	-0.007	0.015	0.009	-0.117	0.038
Hip circumference	Male	0.228	0.034	0.022	0.120	0.013
Mean time to correctly identify matches	Male	-0.131	-0.029	-0.034	-0.053	-0.033
Neuroticism score	Male	-0.067	-0.021	-0.033	-0.086	-0.031
Peak expiratory flow	Male	0.009	0.041	0.046	0.092	0.043
Pulse rate	Male	0.084	-0.021	-0.027	-0.018	-0.029
Pulse wave Arterial Stiffness index	Male	0.137	0.020	0.042	0.093	0.038
Pulse wave peak to peak time	Male	-0.267	-0.022	-0.035	-0.215	-0.024
Systolic blood pressure	Male	0.039	-0.003	-0.002	0.048	-0.006
Waist circumference	Male	0.273	0.030	0.026	0.182	0.010
Weight	Male	0.278	0.050	0.043	0.126	0.036
Waist-hip ratio	Male	0.246	0.016	0.023	0.216	0.005

Table A.9: Summary of correlation coefficients. This table contains the calculated phenotypic, genetic, and residual correlations from the analyses presented in the main text. Columns correspond to the genetic correlation from LD score regression $r_{g,LDSC}$, phenotypic correlation from regression analyses $r_{p,full}$, the phenotypic correlation from mixed model analyses $r_{p,BOLT}$, the genetic correlation from mixed model analyses $r_{g,BOLT}$, and the residual correlation from the mixed model $r_{e,BOLT}$. Note that these are correlation coefficients not covariances and are thus normalized by total variance components. This means that the residual correlation can be smaller than the genetic correlation but still have a greater contribution to the phenotypic correlation if the residual variances are larger than the genetic variance explained by genotyped SNPs.

Table A.10: Supplemental excel file containing simple and multiple regression results.

Table A.11: Supplemental text file containing genetic correlation results.