

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Examining the Minor Allele Frequency Spectrum of the Heritability of Complex Human Traits

Permalink

<https://escholarship.org/uc/item/1071n97q>

Author

Hartman, Kevin Alexander

Publication Date

2019

Peer reviewed|Thesis/dissertation

Examining the Minor Allele Frequency Spectrum of the Heritability of Complex Human Traits

by
Kevin Hartman

THESIS

Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY


in

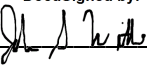
Biological and Medical Informatics


in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

B71F851AEAF541C... Ryan D Hernandez
Chair

DocuSigned by:

John S Witte

DocuSigned by:

A91A262CA37C486... Noah Zaitlen

Committee Members

Copyright 2019

by

Kevin Hartman

Acknowledgements

To Ryan Hernandez, my thesis advisor, thank you for the scientific and emotional guidance I needed to complete my PhD. You taught me the skills I needed from statistical genomics, to high performance computation, to critically reading scientific literature. The faith that you had in me gave me the confidence I needed to persevere when I lost faith in myself.

To the Hernandez Lab, thank you for being my primary scientific community. I loved coming into lab to see the garden that we cultivated together. To Dominic Tong, thank you for discussions regarding career and life strategy. To Melissa Spear, thank you for making the lab a more vibrant place with your positive energy.

To my parents, Dora Hartman and Roger Hartman, thank you for setting me on the course I need to even undertake this journey. From an early age you fostered my scientific interest and skill through the activities we did together. You gave me the ambition to imagine more for myself and the strength to achieve it.

To my sister, Erika Hartman, thank you for paving the way for me. We share both genetics and environment and, I like to think, have become similar people because of it. But you made my journey easier because I had your example to follow and you broke the barriers before me.

To my Caltech community, thank you for the decade of friendship. The challenges we overcame together have shaped the people we will be for the rest of our lives. To Ioana

Aanei, thank you for the yoga, the dancing, and the meaningful conversations. To Justine Chia, thank you for welcoming me into the Gardeners of Missed Connections; you are all my family.

To Frances Arnold, I am honored by the ways you have shaped my scientific journey. My interest in the chemical engineering and my work in the biofuels industry was inspired by your incredible achievements in renewable chemistry. I am continually astounded by your completely-deserved success.

To my cohort of fellow students in the iPQB, thank you for being the first community I was a part of at UCSF. To Ilan Chemmamma, thank you for 6 years and three apartments of living together; I appreciated coming home to someone who completely understood how my day was been. To Zairan Liu, thank you for showing me the path to consulting and the various adventures you joined me on. To Clint Cario, Nima Emami, and Garret Gaskins, thank you for all the late nights of both commiseration and celebration; here's to having more to celebrate.

To the Witte Lab, thank you for being my second home during the most challenging portion of my PhD. The body of knowledge and the enthusiastic aid that you collectively offered solved problem that would have taken me hours in minutes. I felt truly welcomed by the friendship and caring that you all showed me.

To all the other communities I was a part of during my time at UCSF—Gilliam Fellows, S4D, GQA, Vocal Chords—thank you for helping me lead a balanced life. *Integer vitae.*

Abstract

Examining the Minor Allele Frequency Spectrum of the Heritability of Complex Human Traits

By Kevin A Hartman

The genetic architecture of complex human traits remains largely unknown. The distribution of heritability across the minor allele frequency (MAF) spectrum for a trait will be a function of the MAF of its causal variants and their effect sizes. Assumptions about these relationships underpin the tools used to estimate heritability. We examine the performance of two widely used tools, Haseman-Elston (HE) Regression and genomic-relatedness-based restricted maximum-likelihood (GREML). Our simulations show that HE is less biased than GREML under a wide-variety of models, and that the estimated standard error for HE tends to be substantially overestimated. We then applied HE Regression to infer the heritability of 72 quantitative biomedical traits from up to 50,000 individuals with genotype and imputation data from the UK Biobank. We found that adding each individuals' geolocation as covariates corrected for population stratification that could not be accounted for by principal components alone (particularly

for rare variants). The biomedical traits we analyzed had an average heritability of 0.27, with low frequency variants ($MAF \leq 0.05$) explaining an average of 47.7% of the total heritability (and lower frequency variants with $MAF \leq 0.02$ explaining a majority of our increased heritability over previous estimates). Variants in regions of low LD accounted for 3.3-fold more heritability than the variants in regions of high LD, an effect primarily driven by low frequency variants. These findings suggest a moderate action of negative selection on the causal variants of these traits.

Table of Contents

Chapter 1 Introduction	1
Chapter 2 Methods and Materials for the Simulations of Complex Traits and Inference of Complex Trait Heritability	6
Genomic and Phenotypic Data	6
Computing MAF and LD Score Partitions	9
Simulation Framework	10
Simulation Parameters	12
Heritability Inference	13
Summarizing Inference Performance	14
Software	14
Chapter 3 Evaluating the Performance of Heritability Inference	18
The Role of Genetic Architecture on the Distribution of Heritability	18
Comparing HE Regression and GREML	19
Heritability Inference Quality as a Function of MAF Partitioning	21
Impact of Sample Size on Heritability Inference Quality	23
Chapter 4 Heritability of Complex Human Diseases in the Wellcome Trust Case Control Consortium	30
Chapter 5 Heritability of Complex Human Traits in the UK Biobank	34
Chapter 6 Discussion	37
References	40

Supplementary Materials	49
Inclusion of Covariates	49

List of Figures

Figure 3.1: Distribution of Simulated Heritability Varying Fraction of Rare Causal Alleles	24
Figure 3.2: Simulations comparing GREML and HE	25
Figure 3.3: Impact of MAF Partitioning on Heritability Inference	26
Figure 3.4: Impact of Excluding Low Frequency Variants on Heritability Inference	27
Figure 3.5: Impact of Sample Size on Bias and Mean Squared Error of Estimates	29
Figure 4.1: Total Heritability of WTCCC Disease with Different MAF Threshold	31
Figure 4.2: Reverse Cumulative Heritability of WTCCC Disease for Different MAF Thresholds	32
Figure 4.3: Heritability per Bin of WTCCC Disease for Different Imputation Quality Thresholds	33
Figure 5.1: Heritability of Human Traits in UK Biobank	36
Figure S.1: Total inferred heritability of WTCCC disease for different imputation quality and MAF thresholds	50
Figure S.2: Distribution of Simulated Heritability Varying Fraction of Rare Causal Alleles across Different Total Heritabilities	51
Figure S.3: Simulations comparing GREML and HE	52
Figure S.4: Impact of MAF Partitioning on Heritability Inference for High MAF	54
Figure S.5: Impact of MAF Partitioning on Heritability Inference for Intermediate MAF	55
Figure S.5: Inferred Total Heritability of Different Quantitative Measurements in UK Biobank	56
Figure S.6: Total Heritability of Different Quantitative Measurements in UK Biobank with Differing Covariates Used	57

List of Tables

Table 1.1: Quality Control of UK Biobank Genomic Data	15
Table 2.2: Simulation Parameters	16
Table S.1: Prevalence of Diseases in WTCCC	59
Table S.2: MAF Breakpoints for WTCCC Heritability Inference	60

Chapter 1 Introduction

Complex traits are caused by a combination of environmental factors and genetic variants scattered throughout the genome of an organism. The mechanisms by which the alleles at those sites induce differences in traits among individuals in a population is often unknown, and can be intertwined with many loci influencing many traits (Boyle, Li, & Pritchard, 2017). The collective fraction of the variance of a trait between individuals in a population that can be explained by the genetic variance between people is known as heritability, specifically the trait's broad-sense heritability, H^2 . Family studies have measured the heritability of many complex human traits to be as high as 90% for height (Silventoinen et al., 2003), 72% for type 2 diabetes (Willemsen et al., 2015), and 83% for autism (Sandin et al., 2017).

In the search for causal loci, genome-wide association studies (GWAS) are performed (typically assuming an additive model), and many sites have been statistically associated with a bevy of traits. However, while the collective fraction of a trait's variance explained by the additive effects of all causal variants (the narrow-sense heritability, h^2) can be approximated by the statistically associated variants (h^2_{GWAS}), this estimate often remains much lower than the estimates of broad sense heritability [e.g. only 16% for height (Wood et al., 2014), and 10% for type 2 diabetes (Ali, 2013)]. Even the collective fraction of variance in height explained by additive effects across all genotyped and imputed sites in these GWAS is only 60% in cohorts with $n > 250,000$ (Wood et al., 2014). One of the potential explanations for this so-called "missing heritability" problem is the contribution of rare variants. Indeed, recent studies have

implicated rare variants as a major source of missing heritability for height and BMI (Wainschtein et al., 2019) as well as gene expression (Hernandez et al., 2019), but a broader understanding of the extent to which rare variants contribute to the heritability of complex traits is needed.

The minor allele frequency (MAF) of a variant represents the frequency of the less-common allele in a sample of individuals from a population. Populations that have recently experienced rapid population growth will exhibit a larger fraction of rare alleles than populations that have not been rapidly growing. However, population genetic theory suggests that population growth alone is insufficient to drive rare variants to constitute a substantial fraction of heritability (Uricchio, Zaitlen, Ye, Witte, & Hernandez, 2016; Uricchio, 2019; Sanjak, Long, & Thornton, 2017). Natural selection is the evolutionary force that puts pressure on deleterious alleles to stay at low frequency (or be eliminated from the population) and increases the chance that advantageous alleles will increase in frequency (toward fixation in the population). If alleles that have major causal effects on a phenotype are evolutionarily deleterious, then natural selection will preferentially keep large effect alleles at low frequency, and this process can indeed drive rare variants to constitute a substantial fraction of heritability (Pritchard, 2001; Eyre-Walker, 2010; Simons, Turchin, Pritchard, & Sella, 2014; Uricchio et al., 2016). When strong effect alleles are deleterious in a population that has recently expanded (like many European and Asian populations), these evolutionary forces can act in concert to cause the genetic architecture of a trait to be dominated by rare variants (Uricchio et al., 2016; Hernandez et al., 2019; Lohmueller, 2014).

Note that a particular trait itself does not need to be under selective pressure directly to drive an effect of rare variants. If pleiotropy is common, then causal variants for a trait will have widespread phenotypic effects through interconnected networks [e.g. an omnigenic model, (Boyle et al., 2017)], and if any one of the affected traits negatively impacts reproductive fitness, then the causal alleles could be evolutionarily deleterious. Indeed, much evidence supports the omnigenic model: 1) conserved regions of the genome tend to account for a disproportionate fraction of heritability of several complex traits (Finucane et al., 2015), 2) several attempts to infer the contribution of rare variants to heritability have found substantial evidence for it (Mancuso et al., 2016; Hernandez et al., 2019; Wainschtein et al., 2019), and 3) efforts to model the genetic architecture of complex traits as a function of purifying selection have argued that purifying selection is a prevalent force acting on causal variants (Gazal et al., 2018; Gazal et al., 2017; Zeng et al., 2018).

The primary tools for inferring heritability from genotypes of unrelated individuals are variance component models: Haseman-Elston (HE) regression (Haseman & Elston, 1972; Elston, Buxbaum, Jacobs, & Olson, 2000; Sham & Purcell, 2001; Bulik-Sullivan, 2015; Golan, Lander, & Rosset, 2014), Genome-based Restricted Estimation Maximum Likelihood (GREML) (Yang, Lee, Goddard, & Visscher, 2011; Yang et al., 2010), and Linkage Disequilibrium Adjusted Kinships (LDAK) (Speed, Hemani, Johnson, & Balding, 2012). A separate category of tools, LD Score Regression, make use summary statistics from genome-wide association studies to estimate the same (Bulik-Sullivan, 2015). Each approach makes assumptions regarding the genetic architecture of

complex traits (such as the number of causal sites, the distribution of effect sizes, or the relationship between effect size and MAF or linkage-disequilibrium), and the estimates from these techniques can be biased when models are misspecified (Evans et al., 2018; Speed, Cai, Johnson, Nejentsev, & Balding, 2017; Speed & Balding, 2019).

Unfortunately, since the true underlying genetic architecture is not known in advance for a given trait, correcting for biases introduced by model misspecification may be challenging. A particularly common form of bias for variance component models is introduced when sites with different statistical properties are pooled together (e.g. heteroscedasticity). While the true causes of heteroscedasticity are often unknown, a first step to alleviate such biases is to partition sites by MAF and degree of linkage disequilibrium (LD) (Yang et al., 2015; Evans et al., 2018). Additionally, we have noted that partitioning sites based on the MAF inferred from a larger external cohort can further reduce bias for rare variants (Hernandez et al., 2019).

The ability to study the effect of rare alleles is fundamentally limited by the difficulty and expense of accurately identifying and collecting rare variants in sufficiently large cohorts. Investigators have leveraged information from large whole genome sequencing databases such as the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016) to impute millions of rare variants in cohorts of hundreds of thousands of samples [e.g. the UK Biobank (Howie, Donnelly, & Marchini, 2009; Bycroft et al., 2018)]. The UK Biobank in particular has measured a wide variety of phenotypes that we can use to ask about heritability and the genetic architecture of complex traits. However, before estimating the contribution of rare and common variants to the heritability of complex

traits, we must first understand the accuracy of various inference procedures. We conducted thousands of simulations of phenotypes from genetic data and assessed how well two methods for heritability inference perform. We then explored the impact of sample size on the bias and standard errors of the estimated heritability. Lastly, we explored the impact of excluding rare MAF partitions on the inference of heritability for common variant partitions. We then applied our framework for studying variants across the MAF spectrum to infer the heritability of 72 quantitative human traits from the UK Biobank.

Chapter 2 Methods and Materials for the Simulations of Complex Traits and Inference of Complex Trait Heritability

Genomic and Phenotypic Data

The primary genomic data for both simulations and the heritability inference came from the UK Biobank (Bycroft et al., 2018). The UK Biobank consists of a cohort of roughly 500,000 individuals recruited from the United Kingdom (UK) National Health Services. Individuals were recruited on the basis of age between 40 and 69 at the time of assessment. The total dataset collected included blood samples, urine samples, body measurements, self-reported ancestry, medical history, and lifestyle exposures.

The blood samples allowed the extraction and genotyping of DNA on one of two genotyping arrays designed for the UK Biobank. These genotype data were quality controlled then imputed to the HRC (McCarthy et al., 2016) with additional sites imputed to a whole genome sequence reference panel consisting of UK10K haplotype reference panel and the 1000 Genomes Phase 3 reference panel (Chou et al., 2016). These imputed allelic dosages were retrieved as BGEN filetype (Band & Marchini, 2018). We used PLINK 2.0 (Chang et al., 2015) for further quality control and to export variants to PLINK 1 format for downstream analysis. Post-imputation quality control consisted of restricting to sites with imputation info score greater than 0.3 (Howie et al., 2009), with greater than 95% genotype hard-calls from dosage, and no deviation from Hardy-Weinberg Equilibrium (p -values $> 1 \times 10^{-5}$) (Winkler et al., 2014).

From all the individuals of the UK Biobank we applied the filtration steps described in Table 2.1. These filters retained a total of 366,647 high-quality, unrelated individuals. For computational reasons, we selected a subset of 50,000 of these individuals at random for both our inference of heritability and for our simulation studies. To evaluate the role of sample size, we also selected random subsets of 500 and 5,000 individuals to be used for some of the simulations.

We examined all 72 quantitative phenotypes that had at least 25,000 reported values within our 50,000 person cohort. This included 42 blood measurements, 22 anthropometric traits, 5 respiratory traits, and 3 urinary traits.

A preliminary set of experiments were conducted using the Wellcome Trust Case Control Consortium (WTCCC) data (Burton et al., 2007). Specifically, this includes a shared set of 3,000 controls (1,500 from the 1958 British Birth Cohort and 1,500 from the UK Blood Service Collection) and 2,000 samples from each of seven diseases: bipolar disorder, coronary heart disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. Genotypes of these individuals were retrieved from the European Genotype Archive (Lappalainen et al., 2015). Pre-imputation quality control consisted of restricting to sites with <5% missingness, no deviation from Hardy-Weinberg Equilibrium (p -values $> 5.7 \times 10^{-5}$) (Winkler et al., 2014), no evidence of trend between the two control groups (p -values $> 5.7 \times 10^{-5}$). The remaining genotypes were imputed to the HRC (McCarthy et al., 2016) using the Michigan Imputation Server (Das et al., 2016). Post-imputation quality control consisted of restricting to sites with imputation info score > 0.9 (Howie et al., 2009) though a

subset of experiments were conducted with a more permissive information threshold. Individuals were filtered to include only individuals without 3rd degree or closer relationships.

Computing MAF and LD Score Partitions

Variable sites in the 50,000 individual UK Biobank cohort were partitioned in two ways. First by MAF computed using PLINK 2.0 across the full set of >360,000 unrelated, quality-controlled individuals into 17 MAF bins according to the following upper (closed) breakpoints: 2×10^{-6} , 5×10^{-6} , 1×10^{-5} , 2×10^{-5} , 5×10^{-5} , 1×10^{-4} , 2×10^{-4} , 5×10^{-4} , 1×10^{-3} , 2×10^{-3} , 5×10^{-3} , 0.01, 0.02, 0.05, 0.1, 0.2, 0.5. We next used GCTA (Yang et al., 2011) to compute LD scores across the full set of quality-controlled individuals within each MAF bin, in sliding windows of 10 megabases along each chromosome. We then partitioned each MAF bin into high and low LD score bins using the median value LD score for that partition. This procedure resulted in a total of 34 bins of sites.

Variable sites from the imputed WTCCC data were filtered to one of three different minimum MAF thresholds based on HRC MAF: minor allele count ≥ 5 , MAF ≥ 0.001 , and MAF ≥ 0.01 . For each of these MAF thresholds, the remaining MAF spectrum was divided in to 5 bins with roughly logarithmic spacing (with the breakpoint closes to MAF 0.01 explicitly set at 0.01). The resulting MAF bounds are listed in Table S.2.

Simulation Framework

Simulations were performed to compare two inference methods, Haseman-Elston (HE) Regression and genomic-relatedness-based restricted maximum-likelihood (GREML), as well as to identify the most suitable conditions for inference. We used PLINK 1.9 (Purcell et al., 2007) to recode the genotypes for the selected individuals into a genotype matrix, \mathbf{X} , where the genotype of individual i at variant j , x_{ij} is 0, 1, or 2 copies of the non-reference allele. In each simulation we selected a specified number of causal variants. For each causal variant, we drew effect sizes, β_j from a standard normal distribution, $\beta_j \sim N(0, 1)$, with the effect sizes of the non-causal variants implicitly 0. The unscaled genetic component of the phenotype for individual i , g_i was then the sum of the

product of the effect sizes with their corresponding genotypes, $g_i = \sum_j \beta_j x_{ij}$ or $\vec{g} = \mathbf{X}\vec{\beta}$.

This unscaled genetic component was rescaled to give the appropriate variance,

$\vec{g}' = \sqrt{\frac{h^2}{Var(\vec{g})}} \vec{g}$, where h^2 is the simulated heritability. The phenotype of individual i , p_i

was the sum of the scaled genetic component and a residual of appropriate variance

$p_i = g'_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1 - h^2)$.

In simulations where total heritability, h^2 , was partitioned across m collections of variants

(or bins) as $h^2 = \sum_{k=1}^m h_k^2$, we represented each collection of variants as genotype

matrices: $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m$. Letting, $\vec{\beta}_k$ be the vector of effect sizes in collection k with the

specified number of causal sites drawn from that partition as $\beta_k \sim N(0, 1)$ and the

remaining non-causal sites with effect size 0, the unscaled genetic component of the phenotype from collection k was $\vec{g}_k = \mathbf{X}_k \vec{\beta}_k$, which we rescale by the appropriate

heritability, $\vec{g}_k = \sqrt{\frac{h_k^2}{\text{Var}(\vec{g}_k)}} \vec{g}_k$. The phenotypes were then the sum of the genetic

components and a residual term: $\vec{p} = \sum_{k=1}^m \vec{g}_k + \vec{\epsilon}$, where $\epsilon_i \sim N(0, 1 - h^2)$ as before.

Simulation Parameters

We conducted a series of 5 sets of simulations, the parameters of which are summarized in Table 2. For each parameter combination or distribution of heritability we conducted 500 simulations.

For Sets 1-3, causal variants were drawn from the entire genome of 500 unrelated individuals. The variants were partitioned by MAF computed within the 500 individual cohort itself. In Set 1 we varied the total heritability as well as the fraction of causal variants drawn from $MAF < 0.01$ with uniform effect size across the MAF partitions. In sets 2 and 3 we simulated 500 individuals with heritability distributed across 7 and 8 MAF partitions respectively.

For Set 4, we simulated phenotypes for 50,000 individuals using genotypes from chromosomes 18-22. We partitioned these variants by their MAF in the >360,000 unrelated individuals into 17 MAF partitions. We further subdivided each of these by LD, and simulated heritability on each of the 34 MAF-LD bins.

Heritability Inference

We used GCTA to compute the genetic relatedness matrices of individuals from the variants of each partition described. We inferred GREML heritability using GCTA's unconstrained restricted maximum likelihood method (“--reml-no-constrain” flag) using multiple GRMs. For HE heritability, we used either HE Regression as implemented in GCTA or in our own implementation in R, which we verified gave the same results to single floating-point precision. Heritability inferred on the observed scale for the WTCCC data were converted to liability scale using the prevalence values listed in Table S.1. When inferring the heritability of the UK Biobank quantitative traits, we progressively included the first 15 principal components (PCs) of genetic variation and three geographic parameters of the subjects location (North-South coordinate, East-West coordinate, and distance to coast) as covariates. As the HE method of GCTA did not allow the inclusion of covariates directly, these were included as pseudo-GRMs [as per (Hernandez et al., 2019); See Supplemental Methods].

Summarizing Inference Performance

Two metrics were used to summarize the quality of the inference for each set of simulations: bias and empirical standard errors. The bias reported represents the mean of the difference between the estimated and true value. Empirical standard errors were calculated as the average of the standard deviation of the inferred heritability for each set of simulations weighted by the number of simulations in each set.

Software

We used PLINK1.9 v1.90b6.9 and PLINK 2.0 v2.00a2LM to manipulate the genomic data including computing MAF, filtering sites, and exporting to formats. We used GCTA version 1.92.0 to compute GRMs and to perform REML and HE regression. We used R version 3.5.1 with packages ggplot2_3.0.0, dplyr_0.8.0.1 to analyze results and generate figures. We used Python version 2.7.5 to compute covariate GRMs.

Table 1.1: Quality Control of UK Biobank Genomic Data

Quality control step	Remaining Individuals
Initial	473,850
Restrict to samples where self-reported and genetic sex match	473,482
Restrict to self-reported ethnicity	445,826
Restrict to samples with principal components 1 and 2 within 5 standard deviations from the mean	440,222
Remove samples with inappropriate sex-specific cancers	440,148
Restrict to samples in imputation sample file	439,317
Restrict sample those with Dish Quality Control scort (DQC) ≥ 0.82	439,317
Restrict samples to those with hard call rates $\geq 97\%$	438,287
Restrict to samples with heterozygosity within 5 standard deviations of the mean	437,331
Exclude at least one of any pair of individuals with 3rd degree or closer relationship (kinship $\geq (1/2)^{9/3}$), prioritizing exclusion of those with more relationships	366,647

Table 2.2: Simulation Parameters

Set	Number of individuals	Total h^2	MAF Partitions	Distribution of causal variants	Distribution of heritability	Total Number of Simulations
1	500	{0.15, 0.5, 0.8}	(0,0.01], (0.01, 0.05], (0.05, 0.2], (0.2, 0.5]	1000 Total Fraction of causal variants with MAF < 0.01 each of {0.1, 0.5, 0.9}	Uniform effect size	4,500
2	500	0.8	(0, 0.002] (0.002, 0.005], (0.005, 0.01], (0.01, 0.05], (0.05, 0.1], (0.1, 0.2], (0.2, 0.5]	1000 Total 143 from each of the 6 lowest MAF partitions and 142 from the last	All 42 permutations of the set: {0.4, 0.2, 0.04, 0.04, 0.04, 0.04}	21,000
3	500	0.8	(0, 0.002] (0.002, 0.005], (0.005, 0.01], (0.01, 0.02], (0.02, 0.05], (0.05, 0.1], (0.1, 0.2], (0.2, 0.5]	625 Total 125 from each partition with non-zero h^2	1000 permutations of the set: {0.4, 0.2, 0.1, 0.05, 0.05, 0, 0, 0}	500,000

Set	Number of individuals	Total h^2	MAF Partitions	Distribution of causal variants	Distribution of heritability	Total Number of Simulations
4	50,000	0.68	(0, 0.000002], (0.000002, 0.000005], (0.000005, 0.00001], (0.00001, 0.00002], (0.00002, 0.00005], (0.00005, 0.0001], (0.0001, 0.0002], (0.0002, 0.0005], (0.0005, 0.001], (0.001, 0.002], (0.002, 0.005], (0.005, 0.01], (0.01, 0.02], (0.02, 0.05], (0.05, 0.1], (0.1, 0.2], (0.2, 0.5] With each sub-divided by LD	50014 Total 1471 from each MAF-LD partition	0.02 from each MAF-LD partition	500

Chapter 3 Evaluating the Performance of Heritability Inference

The Role of Genetic Architecture on the Distribution of Heritability

The first set of simulations we conducted evaluated the impact of varying the fraction of causal variants that were rare ($MAF < 0.01$), when all variants had the same distribution of effect sizes. The distributions of these simulated heritabilities are shown in Figure 3.1. In this size cohort, rare variants accounted for roughly 10% of variants. Under a “neutral model” where causal variants are randomly selected from the set of all variants, ~10% of causal variants are rare, yet they accounted for less than 1% of the simulated heritability. When we push the simulation to have an extreme excess of rare causal variants (e.g. when 90% of causal variants were rare but effect sizes maintain the same distribution across frequencies), rare variants still account for only 13% of the total heritability. These trends held regardless of total heritability (Figure S.1). In all cases, the majority of heritability came from the (0.2, 0.5] MAF partition, ranging from 67% of heritability when 10% causal variants were rare to 58% when 90% causal variants were rare.

Rare variants can account for a greater fraction of heritability if the distribution of effect sizes is allowed to be a function of MAF. However, the actual model relating number of causal alleles, effect size, and MAF for actual complex traits is unknown. Instead of specifying such a model and in order to test the tools of heritability inference on the full range of possible heritability distributions, we simulated phenotypes where we directly specified the heritability coming from each MAF bin.

Comparing HE Regression and GREML

We compared the accuracy of two common methods for heritability inference: HE and GREML (both implemented in GCTA, see Methods). Specifically, we examined how well the two methods inferred heritability across partitions of MAF when the true underlying heritability was known. We simulated a wide range of genetic architectures with heritability distributed across 8 MAF partitions using a sample size of 500 individuals and a total heritability of 0.8 (Simulation Set 3). We found that when $n=500$ individuals are simulated and analyzed using 8 MAF bins, GREML fails to converge ~65% of the time, regardless of the fraction of heritability deriving from rare variants (Figure 3.2 panel A). When GREML does converge, the resulting heritability estimates can be biased (Figure 3.2 panel B). In contrast, the regression framework of HE always provides a heritability estimate, and the inferred values tend to be unbiased under a broad range of conditions (Figure 3.2 panel C). Figure S.2 shows a direct comparison of heritability estimates across simulated parameters for the two algorithms, and shows that the standard deviation of the heritability estimates across simulations tend to be comparable between HE and GREML.

Both HE and GREML report theoretical standard errors (SE) of the estimated heritability, but we found that neither algorithm report estimates of the SE that reliably reflected the empirical standard errors. While the SE reported by both algorithms are comparable for the higher MAF bins analyzed ($MAF > 0.02$), the reported SEs for the lowest MAF bin analyzed ($0.001 \leq MAF < 0.002$) exhibit conflicting patterns (Figure S.2). When compared to the empirical standard error across simulations in a set, HE

tends to grossly overestimate the SE of the estimate for the lowest MAF bin, while GREML tends to underestimate the SE of the estimate. As a result, approximate 95% confidence intervals ($\hat{h}^2 \pm 2SE$) of the estimates for the lowest MAF bin are highly conservative for HE (100% of confidence intervals overlap the true h^2), but become anti-conservative for GREML as the simulated h^2 increases (only 83.8% of confidence intervals overlap the true h^2 when the true $h^2 = 0.4$; Figure S.3). Given that HE tends to be more unbiased than GREML and not suffer from convergence issues, we exclusively used HE for the remainder of our analyses.

Heritability Inference Quality as a Function of MAF Partitioning

Prior research has suggested that bias can be introduced when sites of differing MAF are pooled into the same GRM (Lee et al., 2013; Yang et al., 2015). We assessed this form of bias in a cohort of 500 individuals using heritability simulated across 8 MAF partitions (Simulation Set 3). We inferred the heritability of these simulated phenotypes either with the same 8 MAF partitions upon which they were simulated or pooled MAF bins (diagrammed in Figure 3.3 panel A). The results of these inferences show that when variants are finely partitioned by MAF, the estimates are unbiased. As more of the MAF spectrum is included with the rarest partition, the estimate is upwardly biased, by as much as 0.24 (30% of the total simulated heritability) when sites $0.001 \leq \text{MAF} < 0.1$ were pooled together. These biases in the total h^2 estimates were driven by the estimates from the pooled variants, with the estimates from the remaining bins being relatively unbiased.

Using the same set of simulations, we assessed the impact of pooling high and intermediate MAF partitions on the performance of HE regression (Figure S.4 and Figure S.5, respectively). We found that inference of heritability showed moderate downward bias when the highest MAF partitions are pooled, with the worst bias occurring when pooling MAF range $(0.005, 0.5]$ with a bias of -0.08 (-10% of the total simulated heritability). Pooling variants of intermediate MAF resulted in less bias than the pooling of high MAF variants.

In any given study, issues of genotyping error, imputation, and MAF-dependent standard errors limit the lowest MAF that can be examined, and such sites are often excluded. We examined whether excluding the lowest MAF bins would bias the heritability estimates from the remaining MAF bins. We simulated phenotypes on 500 individuals using heritability distributed across 7 partitions (Simulation Set 2). We inferred heritability across the full 7 original partitions and successively excluding rare variants. The biases and mean square errors of inferred heritability are shown in Figure 3.4. We found that exclusion of rare variants did not induce a substantial bias in the estimates of heritability of the included bins (less than 0.02), rather than the total estimated heritability would be an unbiased estimate of the variants that are included. As a result, any heritability attributed to the excluded MAF bins would simply remain as missing heritability.

Impact of Sample Size on Heritability Inference Quality

The forces of natural selection will drive causal variants to different frequencies in the population. We sought to investigate how finely we can explore the population level MAF-heritability spectrum for different sample sizes. To this end, we simulated heritability partitioned across 34 LD-MAF partitions of quality-controlled, unrelated UK Biobank European population (17 MAF partitions each split by median LD score) on 50,000 individuals. We then inferred the heritability of these 34 partitions using the full cohort of 50,000 individuals, as well as subsets of 5,000 and 500 individuals. The magnitude of bias (Figure 3.5 panel A) was generally larger for the lower MAF bins, and the scale of the bias was much higher for smaller sample sizes. Standard error (Figure 3.5 panel B) generally increased for more rare partitions, and decreased dramatically with larger sample sizes (dropping by more than a factor of 10 for each factor of 10 increase in sample size in many of the partitions).

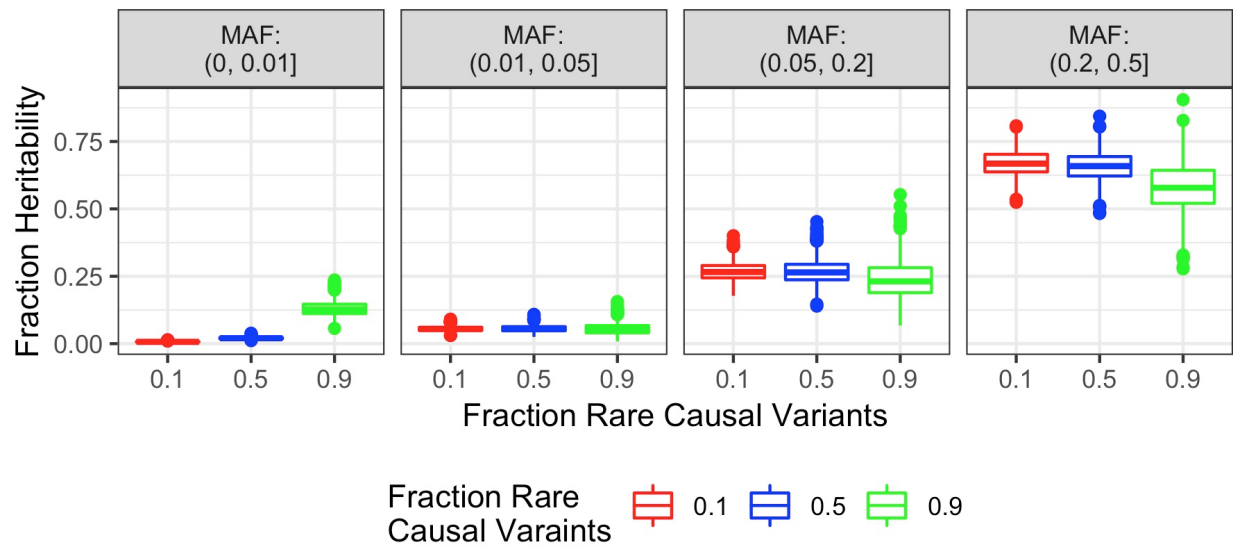


Figure 3.1: *Distribution of Simulated Heritability Varying Fraction of Rare Causal Alleles*

The fraction of the simulated heritability coming from different MAF partitions (horizontal panels) when varying the fraction of causal rare (MAF < 0.01) shows that under “neutral” models where variants have uniform effect sizes across the MAF spectrum, the rare variants account for very little heritability. Even when 90% of causal variants are rare, more common variants account for the majority of heritability.

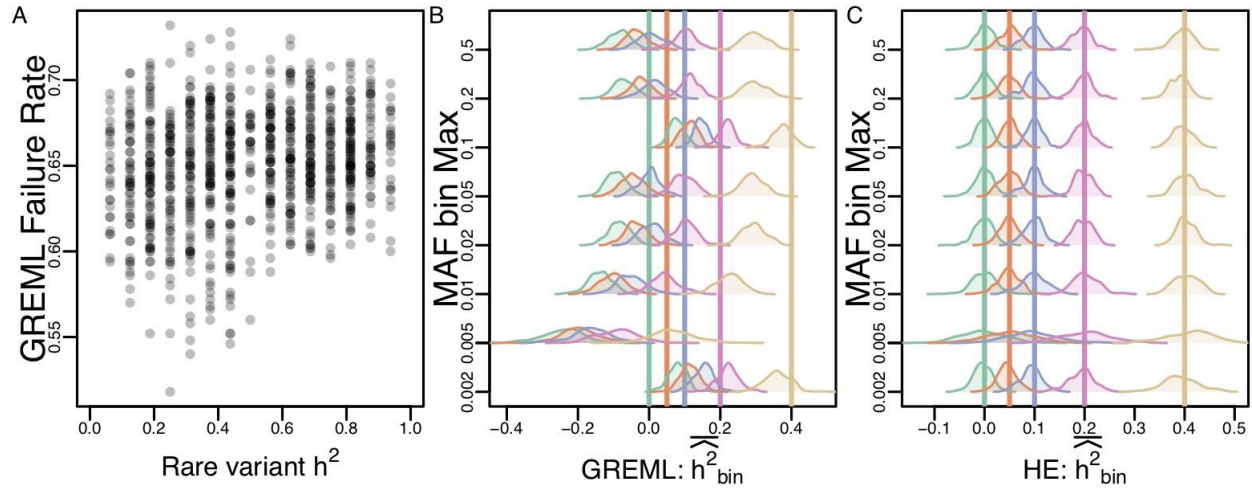


Figure 3.2: Simulations comparing GREML and HE

(A) The fraction of simulations that failed to converge as a function of the fraction of h^2 that derives from rare variants ($MAF < 0.02$). Each point represents 500 simulations of a different genetic architecture (see methods). For the GREML iterations that did converge, the distribution of mean h^2 inferred across genetic architectures is shown for each MAF bin analyzed. True h^2 shown as vertical bars. Similarly, (C) shows the distribution of mean h^2 inferred for HE. Direct comparisons of point estimates and standard errors are shown in Figure S.3.

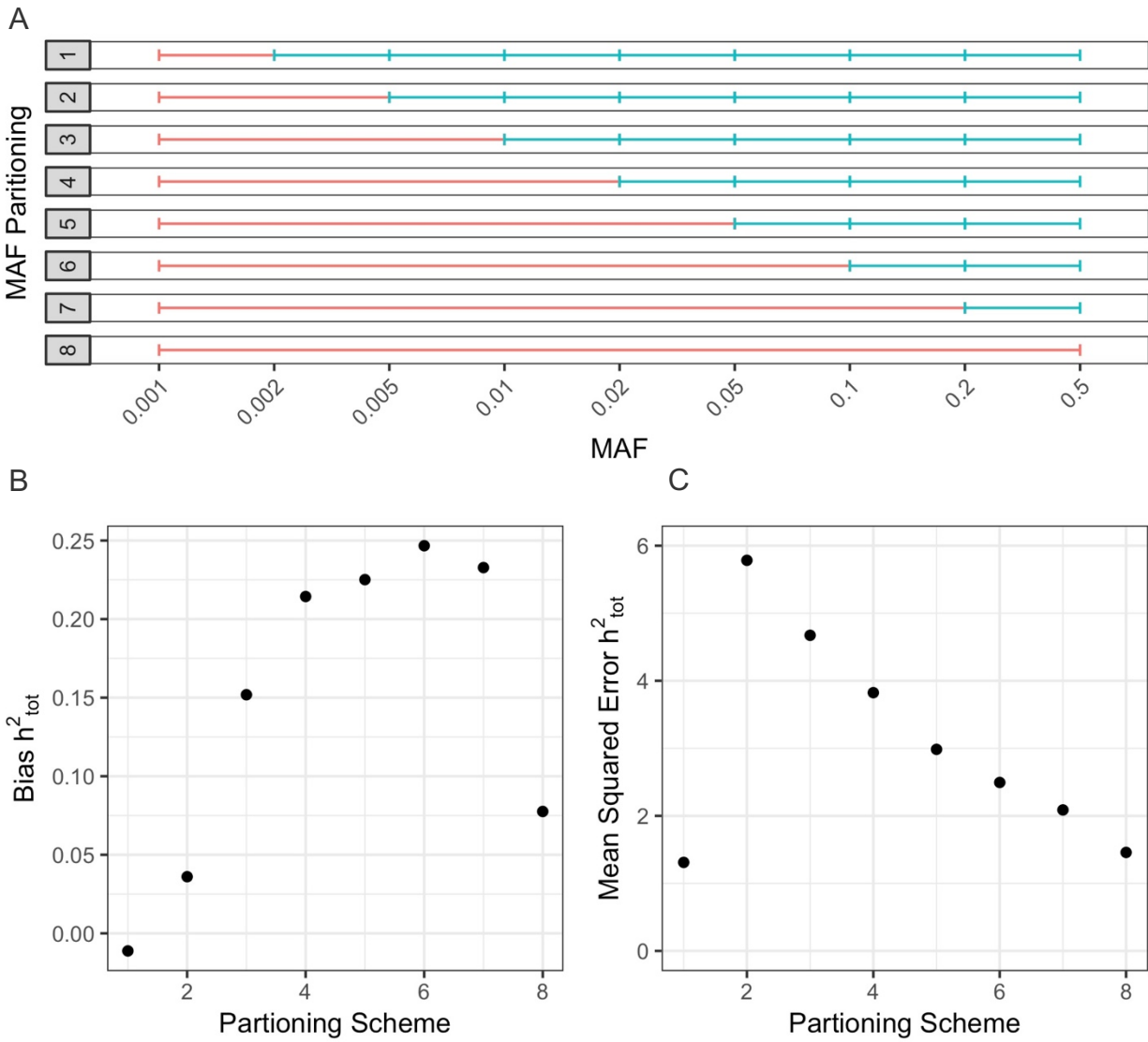


Figure 3.3: Impact of MAF Partitioning on Heritability Inference

(A) The partitioning scheme of the MAF spectrum. (B) Bias of the total inferred heritability for different partitioning schemes. (C) Mean squared error of different partitioning schemes.

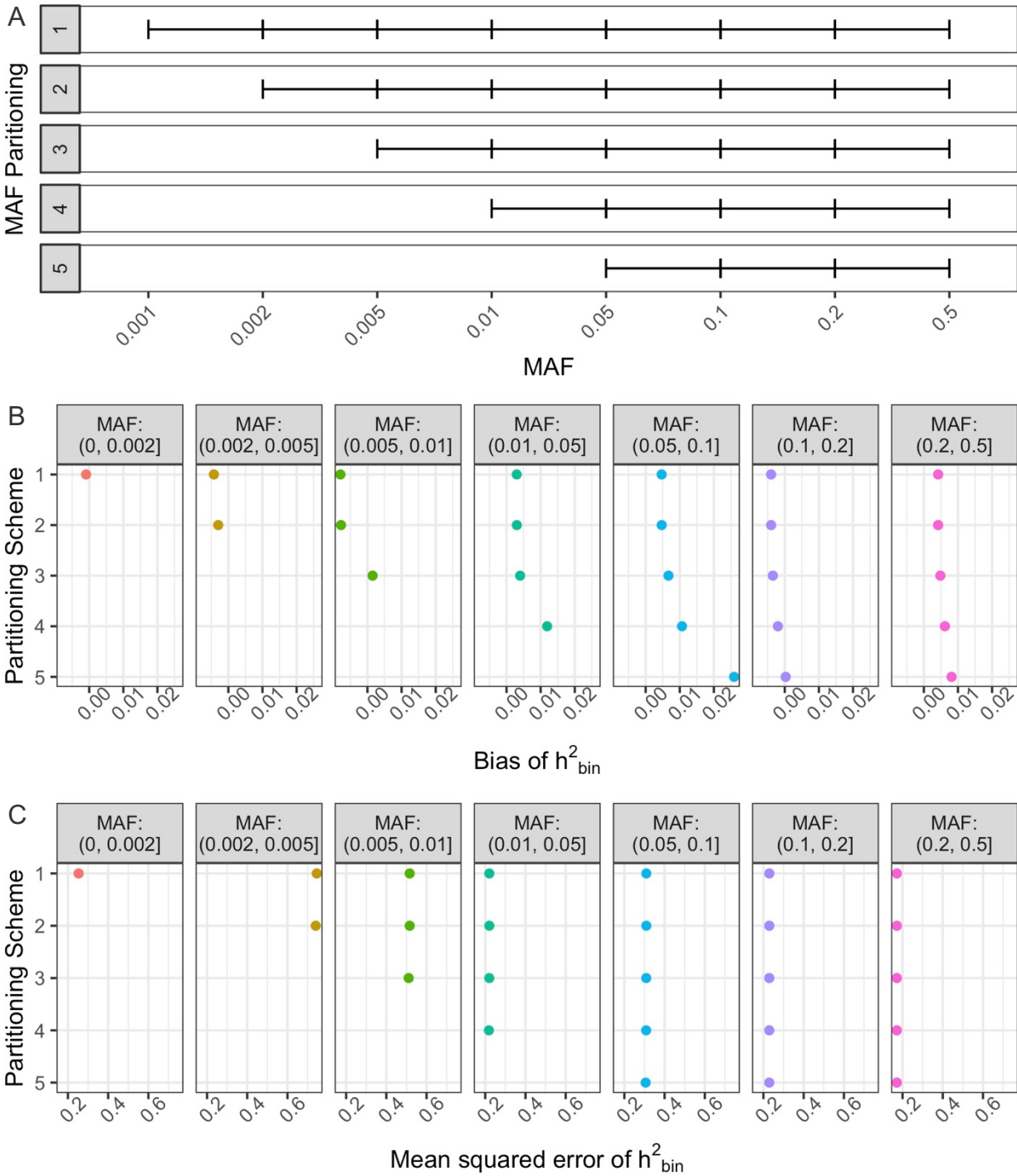


Figure 3.4: Impact of Excluding Low Frequency Variants on Heritability Inference

(A) The partitioning scheme of the MAF spectrum showing the exclusion of increasing

range of the MAF spectrum. (B) The average bias of the inferred heritability of each

partition included in the inference. (C) The mean squared error of the inferred heritability of each partition included in the inference.

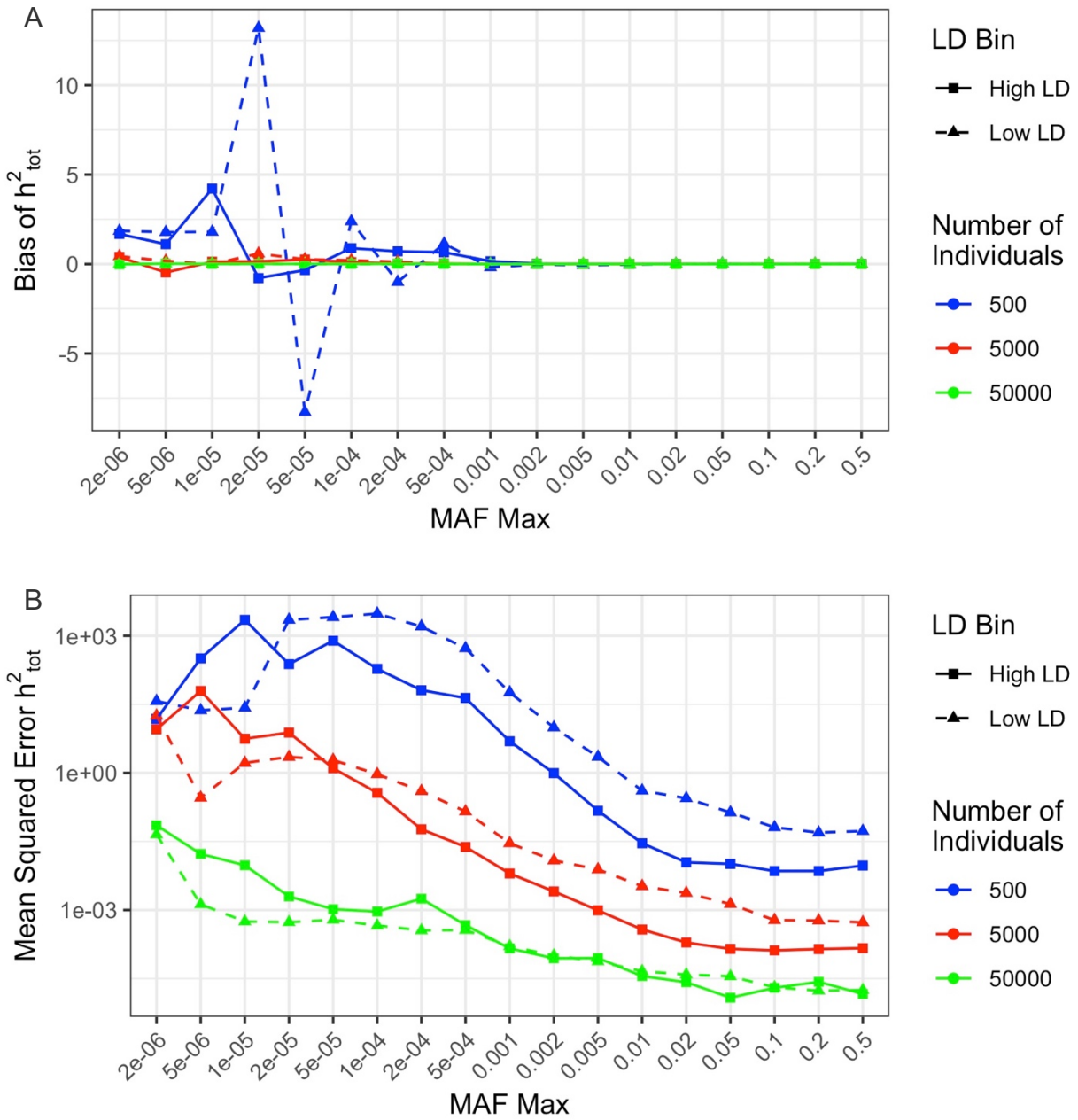


Figure 3.5: Impact of Sample Size on Bias and Mean Squared Error of Estimates

Chapter 4 Heritability of Complex Human Diseases in the Wellcome Trust Case Control Consortium

We examined the heritability of 7 diseases in the WTCCC: bipolar disorder, coronary heart disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. We inferred heritabilities using approximately 2,000 individuals with each disease and 3,000 shared controls for a total of 5,000 individual in each study. We inferred the heritability from 5 partitions of the HRC MAF spectrum with different minimum MAF thresholds. These diseases had an average total heritability of 0.133 when thresholding at $MAF \geq 0.01$. However, the estimates of total heritability were sensitive MAF threshold with lower estimates for lower MAF threshold (Figure 4.1). Examination of the reverse cumulative distribution of heritability (Figure 4.2) shows that the discrepancy in total estimate comes from negative estimates in the lower MAF bins and that the contribution of the higher MAF bins is relatively consistent.

We investigated the impact of imputation quality on the estimates from different MAF partitions using a MAF threshold of 0.001. The per bin heritability estimates (Figure 4.3) increased moderately for the highest three MAF bins (those with $MAF \geq 0.01$) with more permissive imputation filtering, and decreased substantially for the lowest MAF bin. The total effect was that the estimates of total heritability were negative for 6 of the diseases with the most permissive imputation filter (Figure S.1). Ultimately these results indicated that we would need large cohort sizes to investigate the MAF ranges we were interest in.

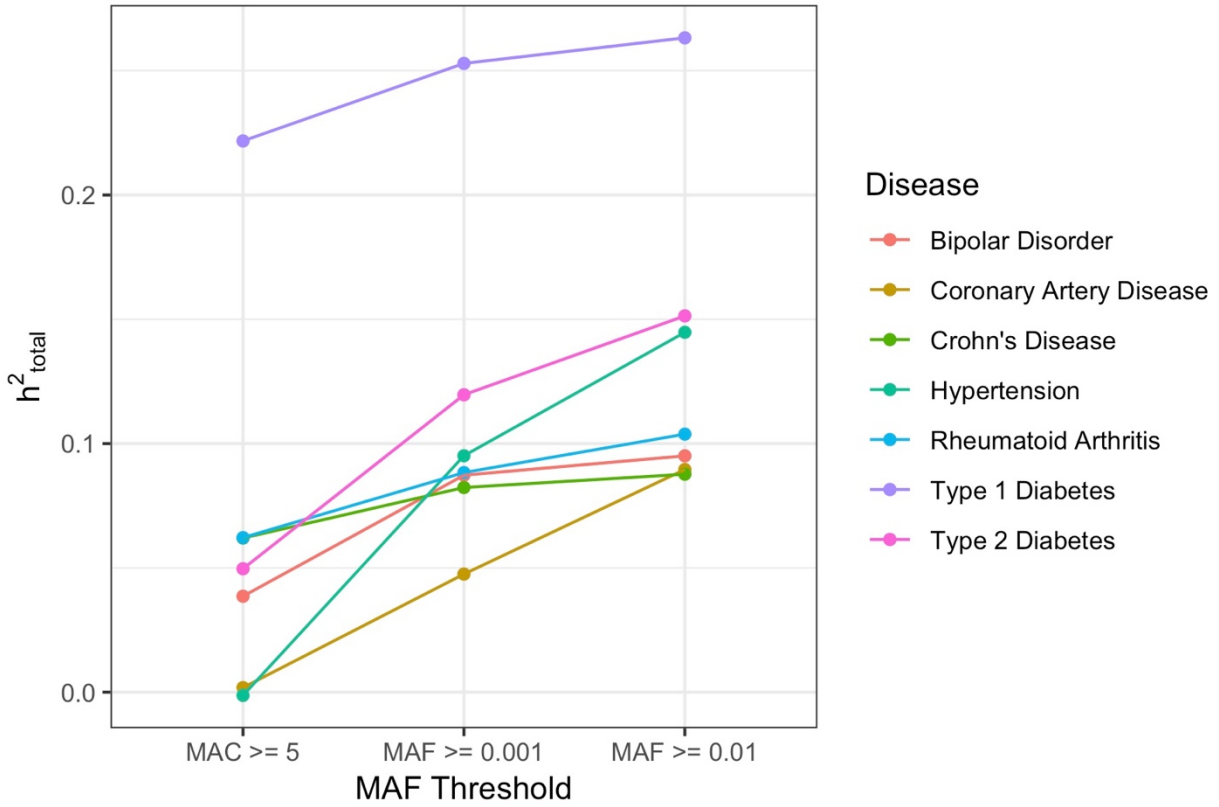


Figure 4.1: Total Heritability of WTCCC Disease with Different MAF Threshold

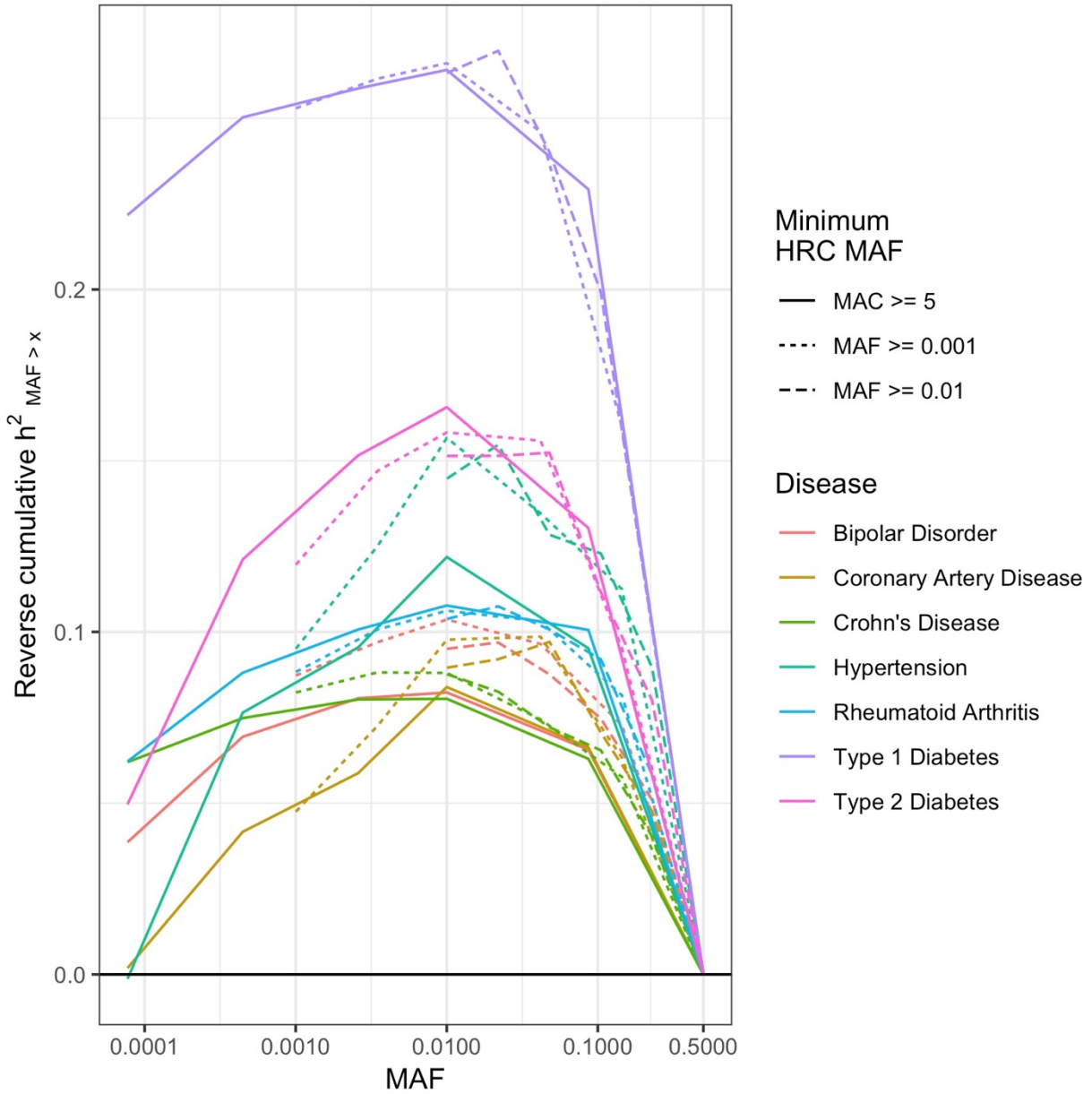


Figure 4.2: Reverse Cumulative Heritability of WTCCC Disease for Different MAF Thresholds

The cumulative heritability above a given MAF value, inferred for each disease.

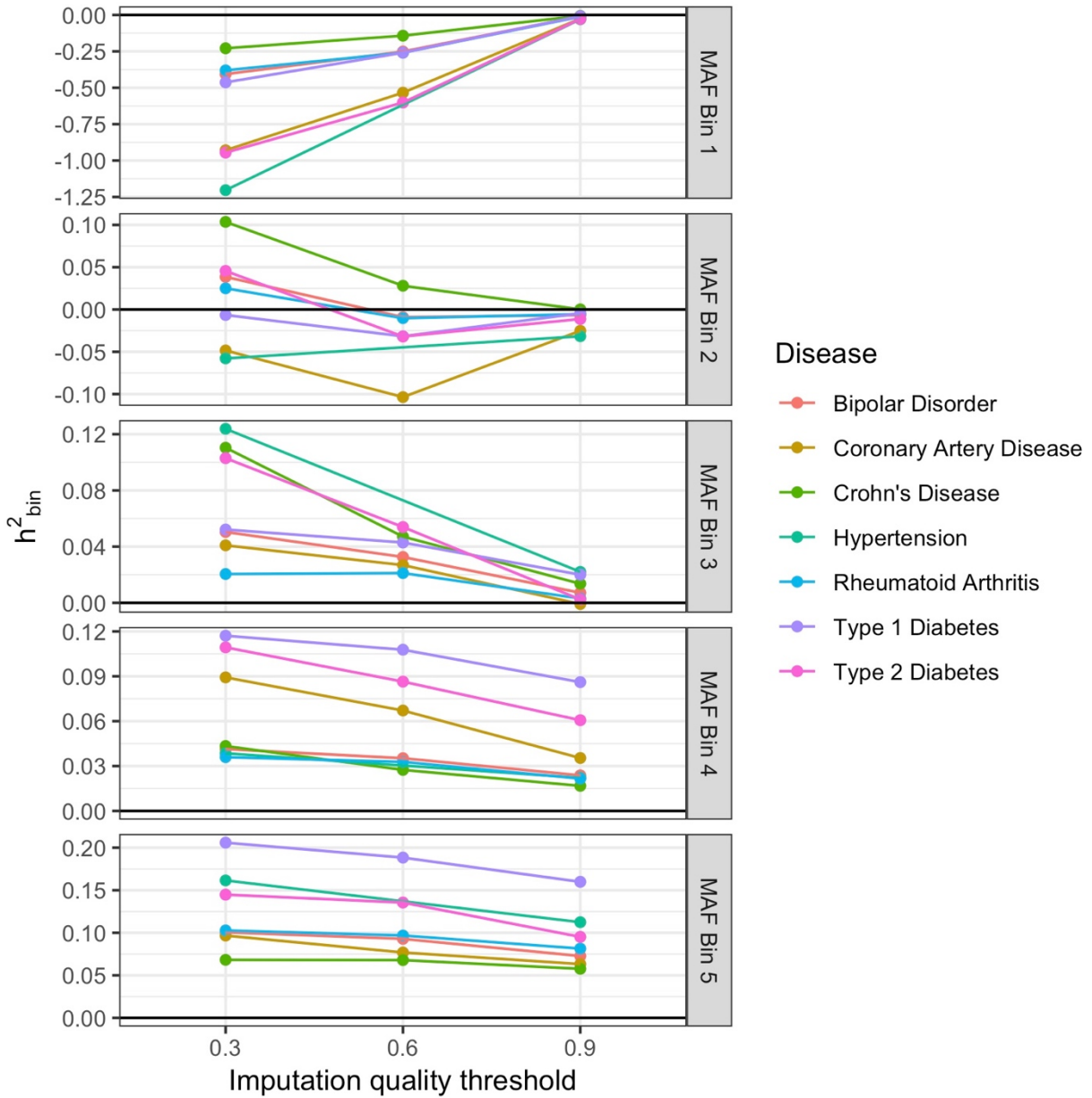


Figure 4.3: Heritability per Bin of WTCCC Disease for Different Imputation Quality Thresholds

Heritability across 5 MAF bins with a lower MAF threshold of 0.001.

Chapter 5 Heritability of Complex Human Traits in the UK Biobank

We randomly selected 50,000 unrelated individuals to infer the genetic architecture of quantitative human traits. We restricted the analysis to the 72 quantitative traits among the biomedical categories blood, body, breath, and urine that were measured in at least 25,000 individuals. We used HE regression to infer the heritability of each trait using variants with $MAF \geq 1 \times 10^{-4}$, partitioned across 11 MAF bins, each split into 2 LD bins (see methods). To correct for population structure, we progressively added principal components (PCs) as covariates up to 15 PCs. We then added three geolocation covariates that describe where each individual lives (north/south, east/west, and distance from the coast; Figure S5). We found that there is only a subtle effect of adding additional PCs beyond the fifth PC. However, geolocation covariates corrected for an additional source of rare variant stratification (particularly for variants with low LD). For further analysis, we focus on the inclusion of 15 PCs and the three geolocation covariates.

The average total heritability of these traits was 0.269 (full list of \widehat{h}_{tot}^2 in Figure S.4). Figure 5.1 panel A shows the heritability estimates across MAF/LD bins. The plurality of heritability derives from the most common MAF bin ($h_{0.2 < MAF \leq 0.5}^2 = 0.092$, representing 34.3% of the average total heritability; Figure 5.1 panel A). However, there is considerable variation in the contribution of different MAF bins to heritability of different traits (Figure 5.1 panel B, which shows the cumulative, left, and reverse-cumulative, right, heritability across MAF bins for each of the 72 traits). Averaging across traits (Figure 5.1 panel C), we find that little heritability derives from ultrarare variants.

Superimposing the cumulative and reverse-cumulative heritability plots allows us to easily identify the MAF at which half the heritability has been described (the intersection of the cumulative and reverse-cumulative heritabilities). Overall, approximately half the heritability is explained by variants with $MAF \leq 0.05$. Partitioning alleles by low versus high LD, we find that low LD variants constitute 3.3-fold more heritability than high LD variants, which is largely driven by low frequency variants (approximately half the heritability of low LD variants is explained by variants with $MAF \leq 0.02$), while heritability of high LD variants is primarily driven by common variants (approximately half the heritability of high LD variants is explained by the highest MAF bin alone).

Previous estimates of heritability from these data have been calculated using LD Score (LDS) regression (Walters et al., n.d.). Our estimates of total heritability using HE regression have a reasonable concordance with the LDS estimates (Figure 5.1 panel D), with a correlation of $r^2 = 0.75$. The discrepancies between our HE estimates and the LDS estimates are mostly driven by the contribution of low frequency variants ($MAF \leq 0.02$) to our HE-based estimates (Figure 5.1 Panel E).

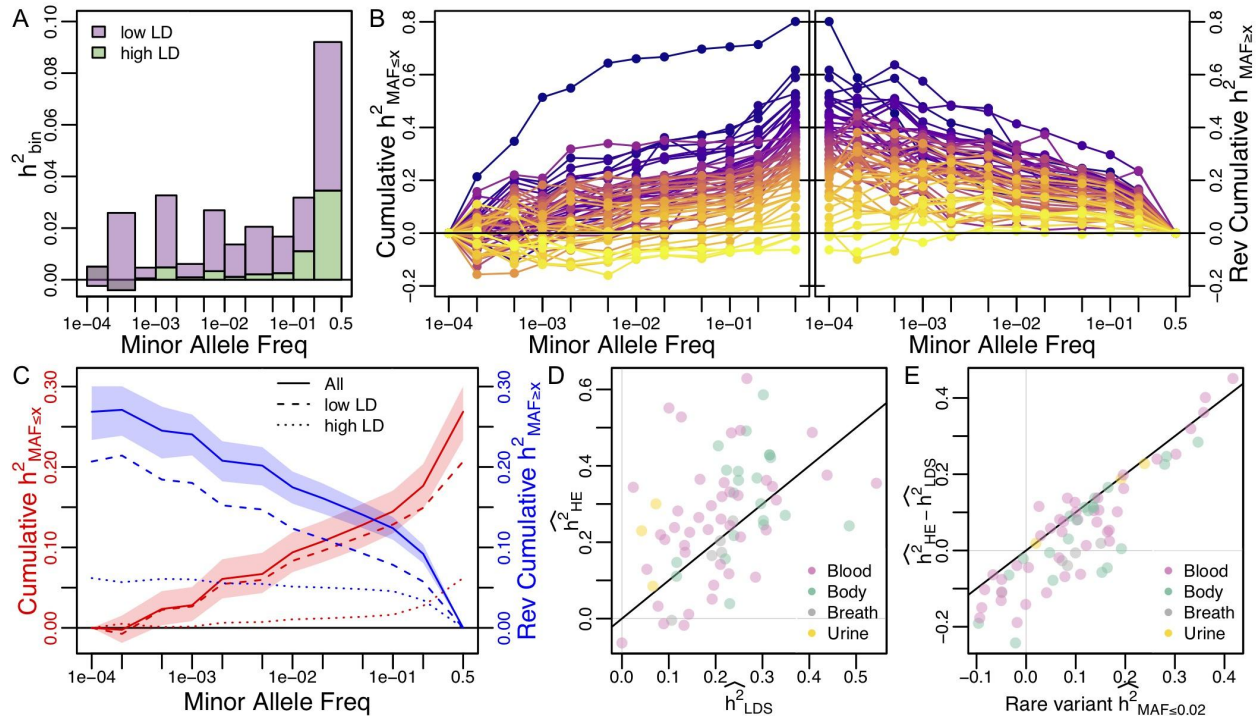


Figure 5.1: Heritability of Human Traits in UK Biobank

(A) Stacked bar plot of average heritability in each MAF-LD partition across 69 biomedical traits. (B) Cumulative and reverse cumulative heritability of all biomedical traits (with traits colored according to their total heritability, see Figure S4). (C) Average cumulative and reverse cumulative heritability across traits (solid line) with envelope showing the 95% quantile range from 1000 bootstrap samples. Dashed and dotted lines represent low and high LD partitions, respectively. (D) Comparison of the inferred total heritability across traits using HE regression (y-axis) versus LD Score (LDS) regression (x-axis). (E) Difference between HE and LDS heritability estimates versus our inferred rare variant (MAF \leq 0.02) heritability estimate. In D-E, points are colored according to the four biomedical categories of traits, with diagonal line show for reference.

Chapter 6 Discussion

Our simulations show that drawing causal alleles and the effect sizes for those alleles independently of MAF will result in the majority of heritability arising from common alleles. While certain models could propose a relationship between probability of being drawn as a causal allele, effect size distribution, and minor allele frequency the actual relationship underlying actual traits remains unknown. If heritability inference procedures are tested and calibrated on a small subset of possible models, the performance on traits that do not fit that model may not be accurate. Indeed we found that REML exhibited substantial bias in many of our simulations. HE Regression, in contrast, was much more robust to a variety simulated heritabilities.

Our investigation into the performance of HE Regression underscored the importance of partitioning variant by MAF. The simulations we conducted also highlighted the importance of sample size in assessing the contribution of rare variants. A ten-fold increase in sample size reduced standard errors by more than a factor of ten for rare variants. The computational efficiency of HE Regression based methods should allow for examination of greater sample sizes, and therefore the examination of the contribution of rarer variants, as compared to REML.

Sample size ultimately limited our ability to describe the contribution of rare alleles complex human diseases in the WTCCC. Based on our simulations of 5,000 individuals, we cannot reliably infer the contribution of variants with $MAF < 1\%$. These samples also were potentially more susceptible to population structure as they gathered for a case-

control cohort and therefore a less random subset than the UK Biobank data. As these are dichotomous traits, their heritabilities are reported on a liability scale which utilizes a correction base on prevalence of the disease. If the prevalence used does not match the population described, the liability scale of these traits will be inaccurate.

Using a cohort of 50,000 individuals from the UK Biobank, we were able to examine the heritability of 72 biomedical traits down to a MAF of 0.01%. We found that these traits had average heritability was 0.269. Of this, 34.3% of the total heritability was found in the highest MAF partition and 34.9% of the total heritability was explained by variants with $MAF \leq 1\%$. These data are inconsistent with simulations that have independent and identically distributed effect sizes across MAF bins (where we inferred 67% of heritability to be due to the highest MAF bin; Figure 1). This suggests that causal variants are disproportionately at low frequency or that these low frequency causal variants have larger effect sizes than common causal variants. The variants in regions of low LD accounted for 3.3-fold more heritability than those in regions of high LD, consistent with past findings (Zeng et al., 2018; Wainschtein et al., 2019) and is considered evidence of negative selection. That the variants with $MAF \leq 0.02$ explain roughly half of the heritability of the low LD variants may be further suggestive of negative selection acting upon the genetic architecture of these traits.

One important caveat to our analysis is that we have only considered variants identified through genotyping and imputing samples to an external reference panel. This means that a majority of ultrarare variants that are carried by the 50,000 individuals we studied were not included in our analysis. Indeed a recent study showed that there were more

than ten times as many variants with MAF < 0.01% revealed through whole exome sequencing in a cohort of 50,000 UK Biobank individuals than in a genotyped and imputed comparable cohort (Hout et al., 2019). Therefore, our ability to infer the contribution of these ultrarare variants to heritability of complex traits is nonexistent. While we did not conduct simulations directly to assess the impact of genotyping and imputation error, these effects would mostly be observed in the most rare MAF bins, where we only observed modest amounts of heritability. As technologies for collection of genetic material improve and computational feasibility of ever-larger cohorts is achieved, we will be better able to examine the contribution of ultrarare variants to heritability of human traits.

The findings here relate to the specific population studied, a non-random sample of the UK population. While findings may have some sensitivity to the inclusion of additional covariates, covariates must be examined on a case-by-case basis to avoid altering the interpretation of particular phenotypes. Future work can examine how these findings generalize to other populations.

References

- Ali, O. (2013). Genetics of type 2 diabetes. *World Journal of Diabetes*, 4(4), 114–123.
<https://doi.org/10.4239/wjd.v4.i4.114>
- Band, G., & Marchini, J. (2018). BGEN: a binary file format for imputed genotype and haplotype data. *BioRxiv*, 308296. <https://doi.org/10.1101/308296>
- Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7), 1177–1186.
<https://doi.org/10.1016/j.cell.2017.05.038>
- Bulik-Sullivan, B. (2015). Relationship between LD Score and Haseman-Elston Regression. *BioRxiv*, 018283. <https://doi.org/10.1101/018283>
- Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., ... Primary Investigators. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. <https://doi.org/10.1038/nature05911>
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203–209. <https://doi.org/10.1038/s41586-018-0579-z>

- Chou, W.-C., Zheng, H.-F., Cheng, C.-H., Yan, H., Wang, L., Han, F., ... Hsu, Y.-H. (2016). A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Scientific Reports*, 6(1), 1–9. <https://doi.org/10.1038/srep39313>
- Elston, R. C., Buxbaum, S., Jacobs, K. B., & Olson, J. M. (2000). Haseman and Elston revisited. *Genetic Epidemiology*, 19(1), 1–17. [https://doi.org/10.1002/1098-2272\(200007\)19:1<1::AID-GEPI1>3.0.CO;2-E](https://doi.org/10.1002/1098-2272(200007)19:1<1::AID-GEPI1>3.0.CO;2-E)
- Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G. R., Das, S., Gazal, S., ... Keller, M. C. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, 50(5), 737–745. <https://doi.org/10.1038/s41588-018-0108-x>
- Eyre-Walker, A. (2010). Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(suppl 1), 1752–1756. <https://doi.org/10.1073/pnas.0906182107>
- Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., ... Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11), 1228–1235. <https://doi.org/10.1038/ng.3404>

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., ... Price, A. L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, 49(10), 1421–1427. <https://doi.org/10.1038/ng.3954>

Gazal, S., Loh, P.-R., Finucane, H. K., Ganna, A., Schoech, A., Sunyaev, S., & Price, A. L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nature Genetics*, 50(11), 1600–1607. <https://doi.org/10.1038/s41588-018-0231-8>

Golan, D., Lander, E. S., & Rosset, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences of the United States of America*, 111(49), E5272–E5281. <https://doi.org/10.1073/pnas.1419064111>

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1), 3–19. <https://doi.org/10.1007/bf01066731>

Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., & Zaitlen, N. (2019). Ultrarare variants drive substantial cis heritability of human gene expression. *Nature Genetics*, 51(9), 1349–1355. <https://doi.org/10.1038/s41588-019-0487-7>

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, 5(6), e1000529.
<https://doi.org/10.1371/journal.pgen.1000529>

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., ur-Rehman, S., ... Flicek, P. (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, 47(7), 692–695.
<https://doi.org/10.1038/ng.3312>

Lee, S. H., Yang, J., Chen, G.-B., Ripke, S., Stahl, E. A., Hultman, C. M., ... Wray, N. R. (2013). Estimation of SNP Heritability from Dense Genotype Data. *American Journal of Human Genetics*, 93(6), 1151–1155.
<https://doi.org/10.1016/j.ajhg.2013.10.015>

Lohmueller, K. E. (2014). The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLOS Genetics*, 10(5), e1004379.
<https://doi.org/10.1371/journal.pgen.1004379>

Mancuso, N., Rohland, N., Rand, K. A., Tandon, A., Allen, A., Quinque, D., ... Reich, D. (2016). The contribution of rare variation to prostate cancer heritability. *Nature Genetics*, 48(1), 30–35. <https://doi.org/10.1038/ng.3446>

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... the Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10), 1279–1283. <https://doi.org/10.1038/ng.3643>

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69(1), 124–137. <https://doi.org/10.1086/321272>

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., & Reichenberg, A. (2017). The Heritability of Autism Spectrum Disorder. *JAMA*, 318(12), 1182–1184. <https://doi.org/10.1001/jama.2017.12141>

Sanjak, J. S., Long, A. D., & Thornton, K. R. (2017). A Model of Compound Heterozygous, Loss-of-Function Alleles Is Broadly Consistent with Observations from Complex-Disease GWAS Datasets. *PLOS Genetics*, 13(1), e1006573. <https://doi.org/10.1371/journal.pgen.1006573>

Sham, P. C., & Purcell, S. (2001). Equivalence between Haseman-Elston and Variance-Components Linkage Analyses for Sib Pairs. *American Journal of Human Genetics*, 68(6), 1527–1532.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., ... Kaprio, J. (2003). Heritability of adult body height: A comparative study of twin cohorts in eight countries. *Twin Research: The Official Journal of the International Society for Twin Studies*, 6(5), 399–408.
<https://doi.org/10.1375/136905203770326402>

Simons, Y. B., Turchin, M. C., Pritchard, J. K., & Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature Genetics*, 46(3), 220–224. <https://doi.org/10.1038/ng.2896>

Speed, D., & Balding, D. J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics*, 51(2), 277–284.
<https://doi.org/10.1038/s41588-018-0279-5>

Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., & Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, 49(7), 986–992.
<https://doi.org/10.1038/ng.3865>

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved Heritability Estimation from Genome-wide SNPs. *American Journal of Human Genetics*, 91(6), 1011–1021. <https://doi.org/10.1016/j.ajhg.2012.10.010>

Uricchio, L. H. (2019). Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. *Human Genetics*. <https://doi.org/10.1007/s00439-019-02040-6>

Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S., & Hernandez, R. D. (2016). Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Research*, gr.202440.115. <https://doi.org/10.1101/gr.202440.115>

Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, T.-O. for P. M. C., Cupples, L. A., ... Visscher, P. M. (2019). Recovery of trait heritability from whole genome sequence data. *BioRxiv*, 588020. <https://doi.org/10.1101/588020>

Walters, R., Baya, N., Tashman, K., Chen, D., Abbott, L., Carey, C., ... Neale, B. (n.d.). Heritability of >4,000 traits & disorders in UK Biobank. Retrieved November 29, 2019, from https://nealelab.github.io/UKBB_Idsc/index.html

Willemsen, G., Ward, K. J., Bell, C. G., Christensen, K., Bowden, J., Dalgård, C., ... Spector, T. (2015). The Concordance and Heritability of Type 2 Diabetes in

34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*, 18(6), 762–771.

<https://doi.org/10.1017/thg.2015.83>

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Mägi, R., ... Loos, R. J. F. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, 9(5), 1192–1212.

<https://doi.org/10.1038/nprot.2014.071>

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., ... Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11), 1173–1186.

<https://doi.org/10.1038/ng.3097>

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., ... Visscher, P. M. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*,

47(10), 1114–1120. <https://doi.org/10.1038/ng.3390>

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7), 565–569.

<https://doi.org/10.1038/ng.608>

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1), 76–82. <https://doi.org/10.1016/j.ajhg.2010.11.011>

Zeng, J., de Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., ... Yang, J. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*, 50(5), 746–753. <https://doi.org/10.1038/s41588-018-0101-4>

Supplementary Materials

Inclusion of Covariates

As GCTA has not implemented the inclusion of covariates in their HE Regression method, these were included as “pseudo GRMs.” Letting c_i be the value of the i^{th} individual for the covariate \mathbf{C} , the mean-centered, unit-variance-adjusted covariate, \tilde{c}_i is:

$$\tilde{c}_i = \frac{c_i - \text{Mean}(\mathbf{C})}{\text{Var}(\mathbf{C})}.$$

The entry of the covariate matrix for the pair of individuals i and j , Γ , would be

$\Gamma_{i,j} = \tilde{c}_i \tilde{c}_j$. These covariate matrices were computed in Python and exported in a format matching that of GCTA's GRMs. Individuals missing values for covariates were replaced with median of the remaining values.

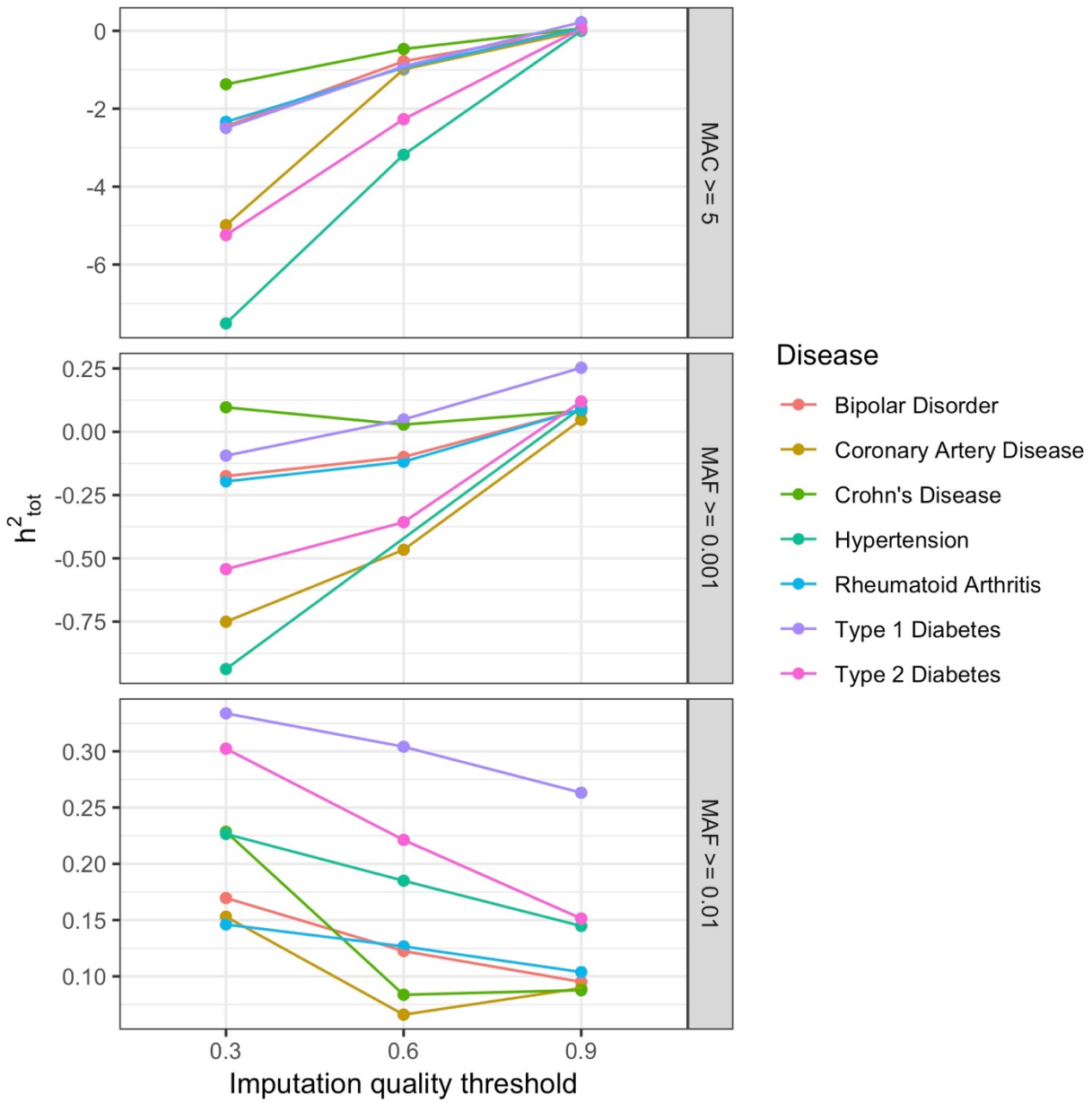


Figure S.1: Total inferred heritability of WTCCC disease for different imputation quality and MAF thresholds

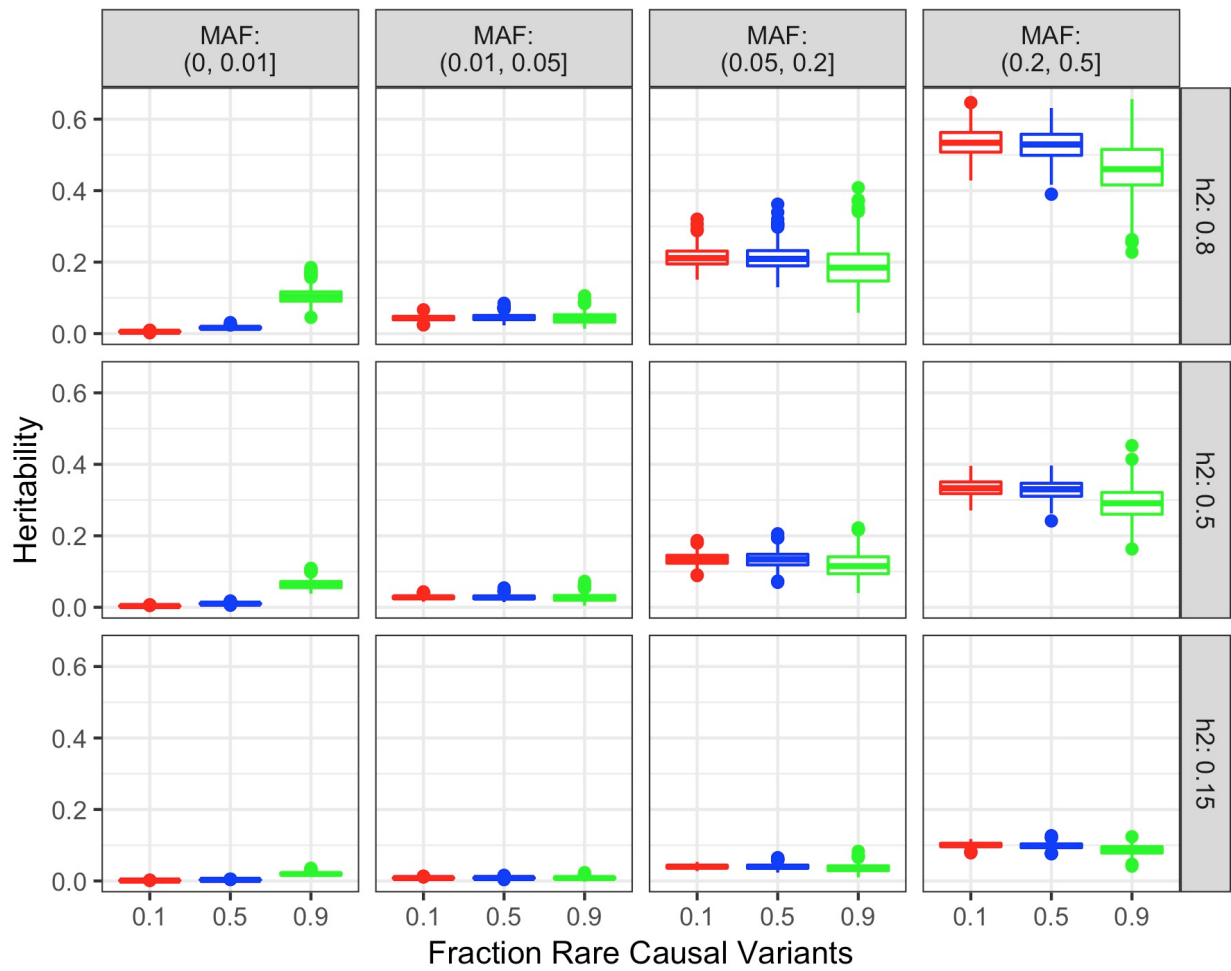


Figure S.2: Distribution of Simulated Heritability Varying Fraction of Rare Causal Alleles across Different Total Heritabilities

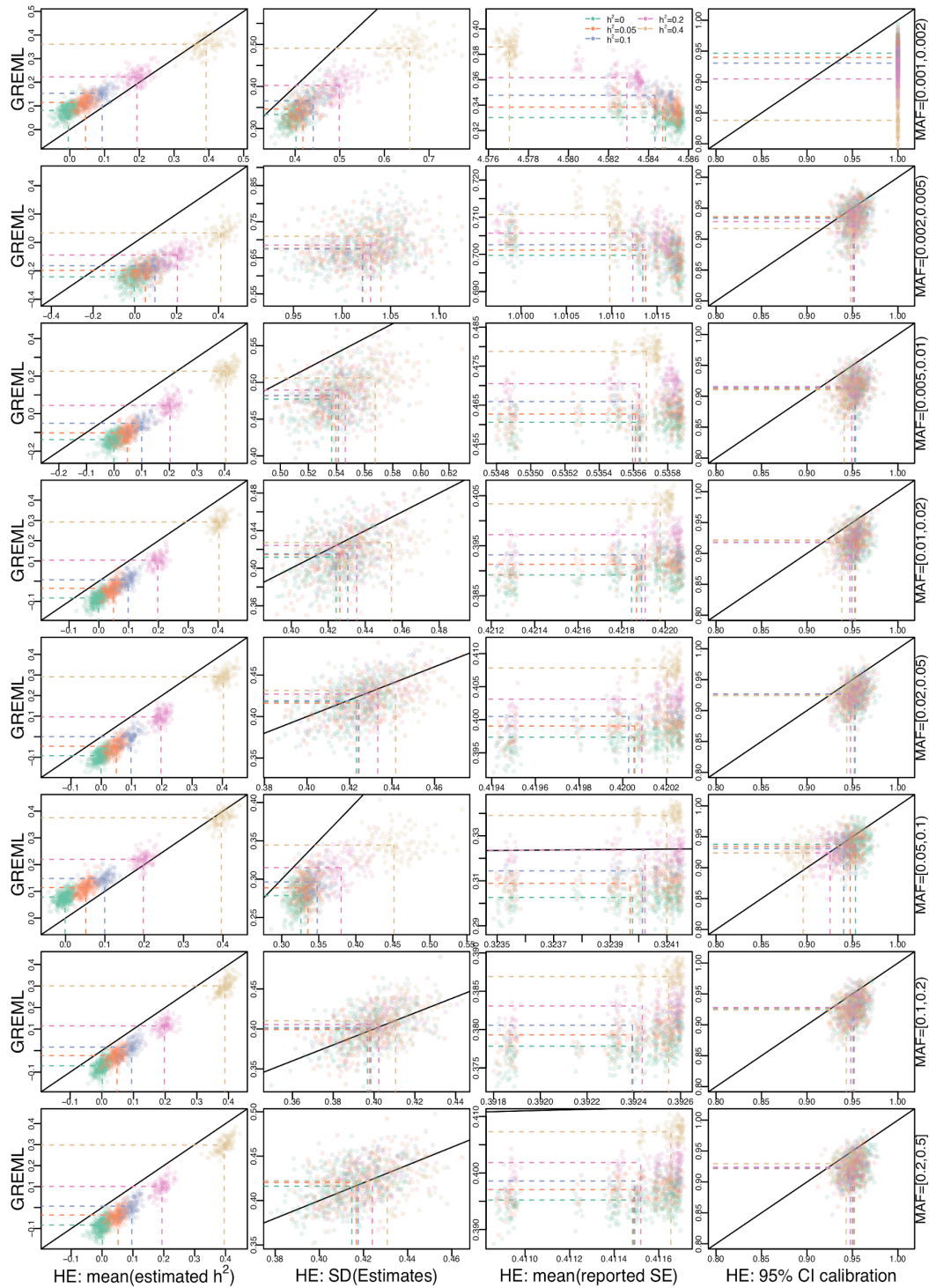


Figure S.3: Simulations comparing GREML and HE

In all plots, each point represents 500 simulations of a single genetic architecture when the true total $h^2=0.8$. Each row of figures represents a different MAF bin (rare variants at

the top, common variants on the bottom), where each point is colored by the true h^2 that derives from that MAF bin and is one of: green ($h^2=0$), orange ($h^2=0.05$), blue ($h^2=0.1$), pink ($h^2=0.2$), or brown ($h^2=0.4$). Plots in the first column (left) compare the mean estimated h^2 (across 500 simulations, or the number that converged, see main text Figure 4.2 panel A) for GREML (y-axis) versus HE (x-axis). Note that the density functions in main text Figure 4.2 panel B-C represent the marginal distributions of these points. The 2nd column of plots compare the standard deviation of the estimates for each genetic architecture. The third column of plots compare the reported standard errors from GREML vs HE. The fourth (right) column of plots compare the fraction of approximate 95% confidence intervals (CI) that overlap the true h^2 for a given bin. In all plots, the dashed lines connect the average across all sets of simulations with the same true h^2 in a bin to their axis, and the black line represents the $y=x$ line.

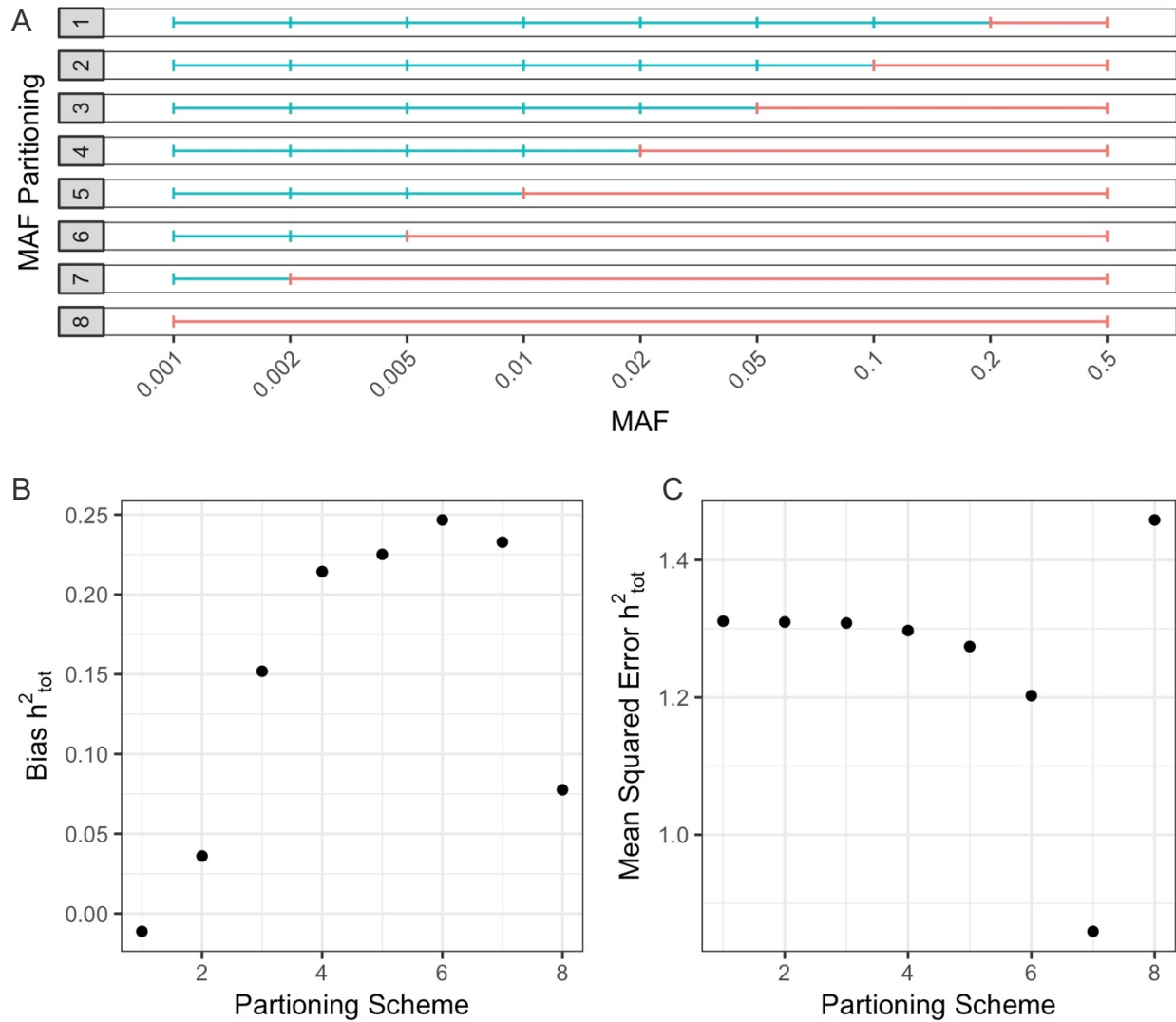


Figure S.4: Impact of MAF Partitioning on Heritability Inference for High MAF

(A) The partitioning scheme of the MAF spectrum used for inference to investigate the impact of pooling variants of high MAF. (B) Bias of the total inferred heritability for different partitioning schemes. (C) Mean squared error of different partitioning schemes.

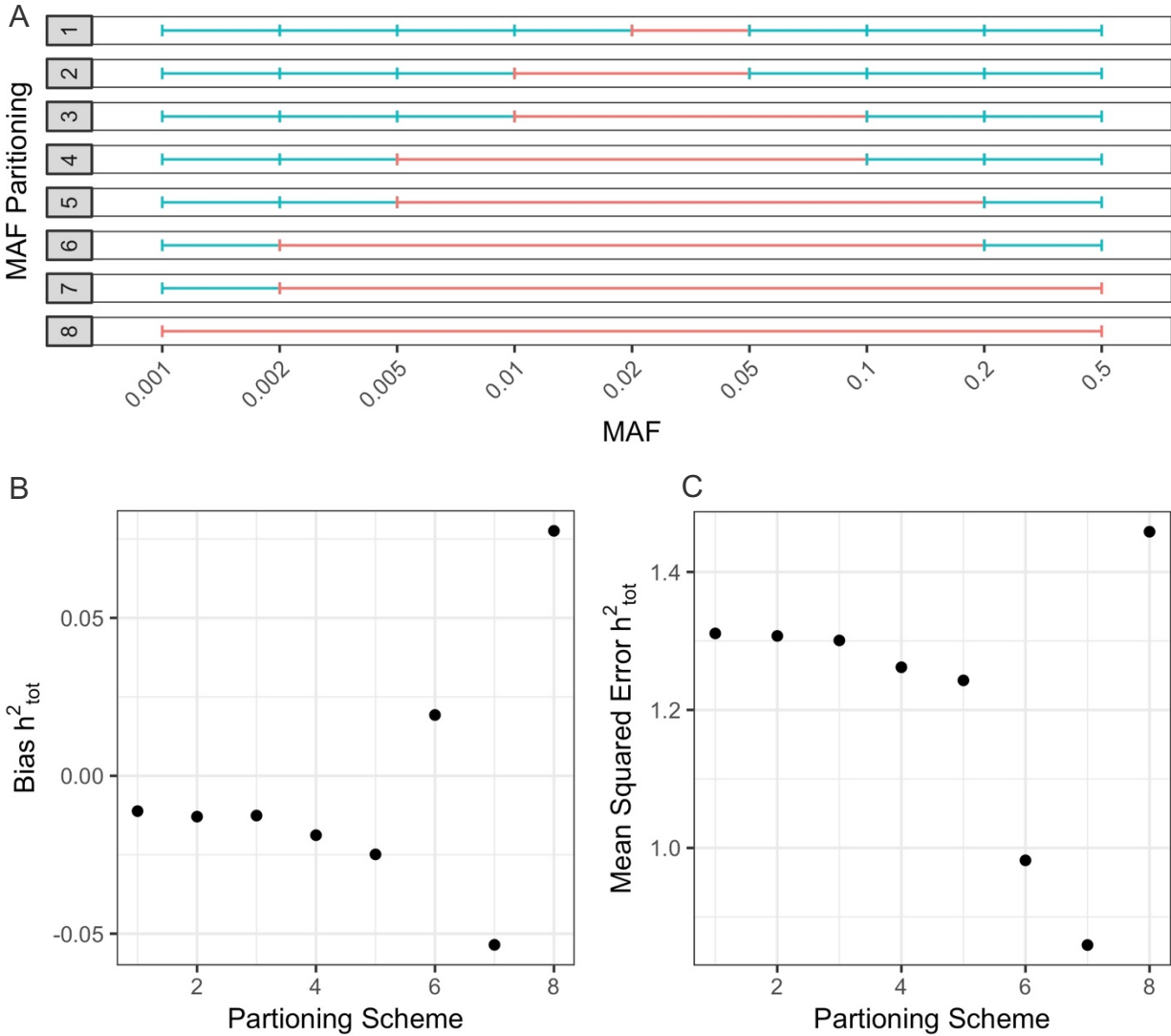


Figure S.5: Impact of MAF Partitioning on Heritability Inference for Intermediate MAF

(A) The partitioning scheme of the MAF spectrum used for inference to investigate the impact of pooling variants of intermediate MAF. (B) Bias of the total inferred heritability for different partitioning schemes. (C) Mean squared error of different partitioning schemes.

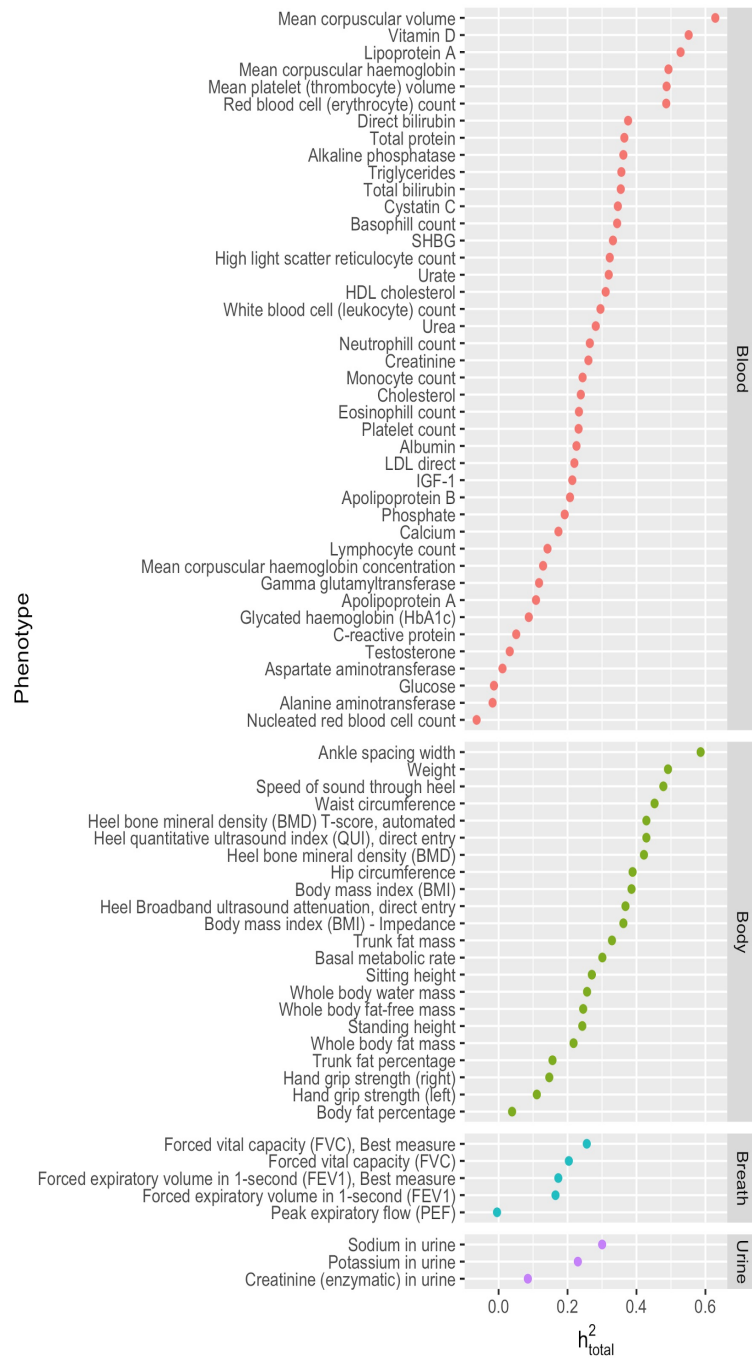


Figure S.6: Inferred Total Heritability of Different Quantitative Measurements in UK Biobank

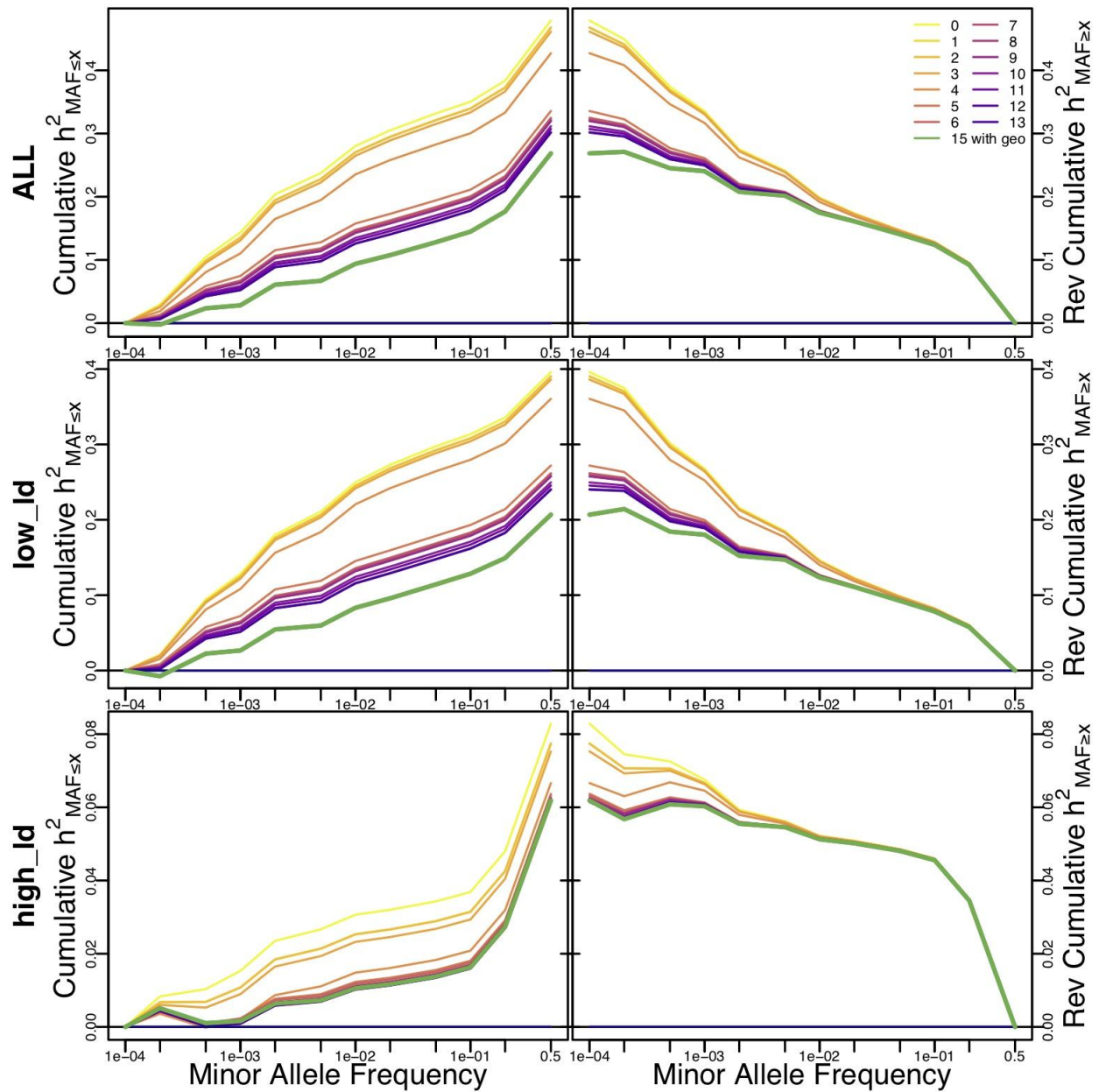


Figure S.7: Total Heritability of Different Quantitative Measurements in UK Biobank with Differing Covariates Used

The left panels show the cumulative heritability below a given MAF, and the right panels show the reverse cumulative heritability above a given MAF. Top panels show the average total heritability, while the middle and bottom panels examine the low LD and high LD bins (respectively).

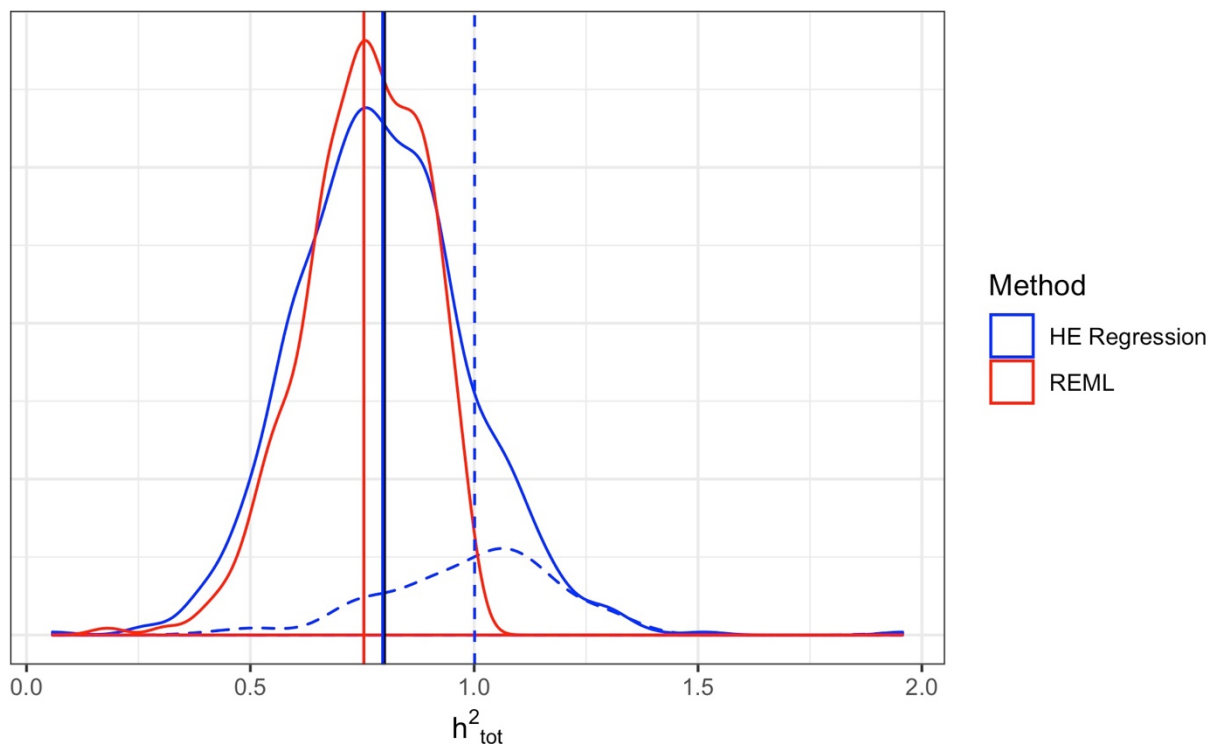


Figure S.8: Distributions of Total Heritability Inferred for Simulations of 5,000 Individuals using HE Regression and REML

We simulated 800 phenotypes for 5,000 individuals (Simulation Set S1) with 0.8 heritability distributed across 11 MAF Bins. We inferred heritability distributed across the 11 bins using either HE Regression (blue) or REML (red). 125 of 800 simulations (15.6%) did not converge for REML. The distribution of heritability inferred by HE Regression for these 125 phenotypes is shown in blue with a dashed line. The vertical lines correspond to the means of the associated distributions. We observed that HE Regression is relatively unbiased compared to REML. The phenotypes that did not converge for REML had higher inferred heritabilities by HE regression than those that did converge.

Table S.1: Prevalence of Diseases in WTCCC

Phenotype	Prevalence
Bipolar Disorder	0.005
Coronary Artery Disease	0.06
Crohn's Disease	0.001
Hypertension	0.26
Rheumatoid Arthritis	0.005
Type 1 Diabetes	0.005
Type 2 Diabetes	0.08
Bipolar Disorder	0.005
Coronary Artery Disease	0.06

Table S.2: MAF Breakpoints for WTCCC Heritability Inference

Minimum MAF/MAC	MAF Bin	Lower MAF	Upper MAF
MAC \geq 5	1	7.69941E-05	0.000447
	2	0.000447	0.00258
	3	0.00258	0.00100
	4	0.0100	0.0865
	5	0.0865	0.5
MAF \geq 0.001	1	0.00100	0.00346
	2	0.00348	0.00100
	3	0.0100	0.0417
	4	0.0417	0.144
	5	0.144	0.5
MAF \geq 0.01	1	0.0100	0.0218
	2	0.0218	0.0478
	3	0.0478	0.104
	4	0.104	0.228
	5	0.228	0.5

Table S.3: Supplementary Simulation Parameters

Set	Number of individuals	Total h^2	MAF Partitions	Distribution of causal variants	Distribution of heritability	Total Number of Simulations
S1	5000	0.8	(1e-04, 2e-04], (2e-04, 5e-04], (5e-04, 0.001], (0.001, 0.002], (0.002, 0.005], (0.005, 0.01], (0.01, 0.02], (0.02, 0.05], (0.05, 0.1], (0.1, 0.2], (0.2, 0.5]	1100 total (100 per MAF Bin)	See Table S.4	800 (50 for each heritability level in Table S.4)
S2	50,000	0.68	(0, 0.000002], (000002, 0.000005], (0.000005, 0.00001], (0.00001, 0.00002], (0.00002, 0.00005], (0.00005, 0.0001], (0.0001, 0.0002], (0.0002, 0.0005], (0.0005, 0.001], (0.001, 0.002], (0.002, 0.005], (0.005, 0.01], (0.01, 0.02], (0.02, 0.05], (0.05, 0.1], (0.1, 0.2], (0.2, 0.5] With each sub- divided by LD	50014 Total (100 from each MAF-LD partition)	0.02 from each MAF-LD partition	500

Table S.4: Heritability Levels for Simulation Set S1

Heritability Distribution	MAF Bin 1	MAF Bin 2	MAF Bin 3	MAF Bin 4	MAF Bin 5	MAF Bin 6
1	0	0.1	0.1	0.1	0.1	0.1
2	0.125	0.0875	0.0875	0.0875	0.0875	0.0875
3	0.25	0.075	0.075	0.075	0.075	0.075
4	0.5	0.05	0.05	0.05	0.05	0.05
5	0.1	0.1	0.1	0	0.1	0.1
6	0.0875	0.0875	0.0875	0.125	0.0875	0.0875
7	0.075	0.075	0.075	0.25	0.075	0.075
8	0.05	0.05	0.05	0.5	0.05	0.05
9	0.1	0.1	0.1	0.1	0.1	0.1
10	0.0875	0.0875	0.0875	0.0875	0.0875	0.0875
11	0.075	0.075	0.075	0.075	0.075	0.075
12	0.05	0.05	0.05	0.05	0.05	0.05
13	0.1	0.1	0.1	0.1	0.1	0.1
14	0.0875	0.0875	0.0875	0.0875	0.0875	0.0875
15	0.075	0.075	0.075	0.075	0.075	0.075
16	0.05	0.05	0.05	0.05	0.05	0.05

Heritability Distribution	MAF Bin 7	MAF Bin 8	MAF Bin 9	MAF Bin 10	MAF Bin 11
1	0.1	0.1	0.1	0.1	0.1
2	0.0875	0.0875	0.0875	0.0875	0.0875
3	0.075	0.075	0.075	0.075	0.075
4	0.05	0.05	0.05	0.05	0.05
5	0.1	0.1	0.1	0.1	0.1
6	0.0875	0.0875	0.0875	0.0875	0.0875
7	0.075	0.075	0.075	0.075	0.075
8	0.05	0.05	0.05	0.05	0.05
9	0	0.1	0.1	0.1	0.1
10	0.125	0.0875	0.0875	0.0875	0.0875
11	0.25	0.075	0.075	0.075	0.075
12	0.5	0.05	0.05	0.05	0.05
13	0.1	0.1	0.1	0.1	0
14	0.0875	0.0875	0.0875	0.0875	0.125
15	0.075	0.075	0.075	0.075	0.25
16	0.05	0.05	0.05	0.05	0.5

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

DocuSigned by:

D1F7DA76FCF64AA... Author Signature

12/18/2019
Date