Imputation and simulation of large genetic datasets



Sinan Shi

Corpus Christi College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Acknowledgements

I am deeply grateful to my supervisors, Professor Pier Palamara and Professor Simon Myers, for their invaluable guidance and unwavering support throughout my DPhil. Their extraordinary dedication to science has been a great source of inspiration to me as an emerging researcher. I feel extremely fortunate to have worked under their mentorship.

I also extend my thanks to Professor Jonathan Marchini for his significant contributions to the Genomics England project. His expertise was crucial to the project's success. I am thankful for the support of the researchers and staff at Genomics England, especially Dr. Loukas Moutsiana and Alex Stuckey, for their technical assistance. I would like to acknowledge Dr. Simone Rubinacci, Dr. Sile Hu, and Professor Olivier Delaneau for their contributions at various stages of the project. My gratitude goes to Faseeha Ayaz, Christopher Gallacher, Tara Lee, and Kira Schützenhofer for their meticulous proofreading of my thesis. Their keen attention to detail has greatly enhanced the quality of my work. I also appreciate the collaboration and camaraderie of the lab members from the Myers and Palamara groups, particularly Romain Fournier, Yiorgos Kalantzis, Miriam Stricker, with whom I have shared many productive hours in various Oxford colleges and libraries.

I am thankful to Corpus Christi College for not only providing a stimulating and vibrant environment throughout my DPhil, but also bestowed me with nearly all my closest friends—Faseeha Ayaz, Tyson Jones, David Murakami, Tara Lee, and Kira Schützenhofer — for making the DPhil journey an enjoyable and memorable adventure.

Lastly, I must express my deepest gratitude to Faseeha and my parents for their love and support. Their encouragement has been my anchor and source of strength.

Abstract

During the past two decades, human genomics has increasingly become a data-driven field, resting on the availability of large-scale datasets and computational tools for their analysis. This thesis built infrastructure and data resources to facilitate the analysis of large human genomic datasets, focusing on genotype phasing and imputation, as well as data simulation. Genotype phasing and imputation are crucial steps in genome-wide association studies (GWAS), which allow for the augmentation of an existing dataset with additional genomic variation, thereby increasing the statistical power to detect associations.

A significant fraction of this thesis is devoted to the processing and analysis of data from the Genomics England (GEL) 100,000 Genomes project, one of the largest high coverage sequencing datasets in the UK. We built a reference panel consisting of 78,195 GEL samples and 342 million autosomal variants. This panel has been made available to the wider research community, constituting a novel resource for genotype imputation and phasing of existing datasets. We used the GEL reference panel to impute autosomal variants in 488,315 samples from the UK Biobank (UKB) data set, increasing the number of available variants from 631,081 to 342 million. We obtained state-of-the-art accuracy for both phasing and imputation, with a mean phasing switch error rate of 0.18% in European samples and an imputation r^2 of 0.75 at MAF$= 2 \times 10^{-4}$ in White British samples.

To demonstrate the potential of downstream analyses performed using this GEL-imputed UKB data set, we conducted a GWAS of four exemplar traits: standing height, BMI, systolic and diastolic blood pressure. Across these four traits, the GEL-imputed UKB data allowed the discovery of 988 more rare associations (MAF ≤ 0.05) compared to a previous version of imputed using HRCUK10K [1]. In addition, imputation using the GEL resource allowed the discovery of 76% of the exome-wide rare associations that were previously identified using direct exome sequencing, including several novel associations of missense and TF-binding variants. We also show that the GEL imputation benefits downstream analyses such as Bayesian fine-mapping.

Finally, this thesis develops a novel software package for the simulation of large-scale genomic data sets, called ARGON2. It allows fast backward discrete-time Wright Fisher (DTWF) simulation, with a performance comparable or superior

to state-of-the-art DTFW simulators across different scenarios. We show that, under complex demographic models involving multiple populations, ARGON2 is significantly faster than currently available coalescent simulators. In addition, ARGON2 enables performing data augmentation of existing data sets, by simulating new samples over inferred genealogical data. This strategy may be used to generate large genomic data sets to support study design and methodological development, particularly for populations that are under-represented in large sequencing studies.

Contents

List of Figures **xi**

List of Abbreviations **xiii**

1 Introduction **1**
1.1 Genetics and genetic datasets 1
1.1.1 Heredity and variation 1
1.1.2 Genome-wide association studies 3
1.1.3 Rare variants associations 5
1.1.4 DNA genotyping and sequencing technologies 6
1.1.5 Large DNA datasets and reference panels 7
1.2 Haplotype phasing and genotype imputation 9
1.2.1 The phasing problem 9
1.2.2 Genotype imputation and its application in GWAS 10
1.2.3 The Li-Stephens model for phasing and imputation 12
1.2.4 Phasing related samples and scaffolding 17
1.3 Genetic variation data simulation 19
1.3.1 Population genetics and simulation 19
1.3.2 The discrete-time Wright-Fisher and the coalescent processes 19
1.3.3 An overview of genome simulators 22
1.4 Thesis outline 24

2 Building the Genomics England (GEL) Reference Panel **27**
2.1 Overview 27
2.2 Genomics England reference panel 28
2.2.1 The aggregated variant (aggVCF) calls 28
2.2.2 Population structure 29
2.2.3 Relatedness 30
2.2.4 GEL reference panel with 342 million autosomal variants . . 33
2.3 GEL reference panel quality control 34
2.3.1 Lenient filters for common variants 37
2.4 Haplotype phasing 40

2.4.1 Phasing pipeline for autosomes 40
2.5 GEL reference panel phasing accuracy 41
2.6 GEL reference panel imputation accuracy 43
2.7 Discussion 45

3 Imputing UK Biobank samples using the GEL reference panel 47
3.1 Overview 47
3.2 UK Biobank imputation pipeline 48
3.2.1 UK Biobank SNP array data quality control and phasing . . 48
3.2.2 UK Biobank imputation using GEL reference panel 48
3.3 GEL imputed UK Biobank data 49
3.3.1 Imputation quality 49
3.3.2 GEL Imputed UKB dataset 49
3.4 Discussion 52

4 UK Biobank imputed data genome-wide association studies and analysis 55
4.1 Overview 55
4.2 Genome-wide association studies 56
4.2.1 The genome-wide association studies for four quantitative traits 56
4.2.2 Rare variant associations 59
4.2.3 Common variant associations 63
4.3 Fine mapping rare variant associations 68
4.3.1 Conditional joint analysis 68
4.3.2 Results 69
4.4 Fine-mapping common variant associations 72
4.4.1 Bayesian fine-mapping 72
4.4.2 Results 74
4.5 Discussion 75

5 Efficient Discrete time Wright Fisher Simulation 77
5.1 Overview 78
5.2 The ARGON2 DTWF simulator 78
5.2.1 The ancestral recombination graph 78
5.2.2 Simulation algorithm 79
5.2.3 The merging algorithm for coalescence events 81
5.2.4 Recombination and gene conversion 82
5.2.5 Mutation 84
5.3 ARGON2 validation and performance 85
5.3.1 Generating the ARG 85

5.3.2 Performance on a single constant sized population 92
5.3.3 Performance on multiple populations with migration 94
5.4 ARG augmentation: simulating new samples on an existing ARG . 94
5.4.1 New data structure: Genealogy and fixed blocks 95
5.4.2 Simulation heuristics 95
5.4.3 Simulation results on simulated ARGs 97
5.4.4 Simulation results on inferred ARG 99
5.5 Discussion 101

6 Conclusion **103**

Appendices

A Additional Figures **109**

Bibliography **117**

List of Figures

1.1 The schematic of pre-phased genotype imputation 18

2.1 Mean coverage of the GEL sequencing samples 29
2.2 Ancestry components and residential postcode of GEL samples . . . 31
2.3 Relatedness in the GEL reference panel 32
2.4 Venn diagram comparing variants from GEL, HRC and TOPMed reference panel 33
2.5 The effect of the quality control filters applied to chromosome 20 . . 37
2.6 Variant distribution along chromosome 20 38
2.7 Compare phasing switch error rate with the GEL and HRC phased 1000G trio data 43
2.8 Imputation r^2 results 45

3.1 Imputation INFO histogram 50
3.2 Comparing IMPUTE INFO with GEL imputed and HRCUK10K imputed UKB data 50
3.3 Imputation INFO score comparison in terms of variant types and allele frequency. 51

4.1 Z-score comparison between G-UKB and H-UKB GWAS 58
4.2 Standing height GWAS associations 61
4.3 The histograms of squared correlation r^2 between the exome sequencing variants and GEL imputed variants in 200,000 UK Biobank samples. 62
4.4 Rare variant associations with moderate or high functional impact . 63
4.5 Common variant associations with moderate or high functional impact 64
4.6 Allele frequency concordance to the TOPMed allele frequency: genome-wide significant sites 65
4.7 Allele frequency concordance to the TOPMed allele frequency: standard filter sites 66
4.8 Allele frequency concordance to the TOPMed allele frequency: standard filter sites 67
4.9 Step-wise regression significant associations 70

4.10 The Manhattan plots showing the novel synonymous association of rs757561770 with SBP discovered in G-UKB. 72
4.11 Compare 95% credible sets size ranging from 1 to 20 variants for standing height. This plot used the prior of $W = 0.02^2$. The size of the dots represents the number of fine-mapping regions. 75

5.1 The schematic of coalescent and recombination implementation in ARGON2 81
5.2 Block fragmentation in ARGON2 83
5.3 The schematic of crossover recombination and non-crossover gene conversion 84
5.4 The performance of ARG building, mutation generation and total runtime. 93
5.5 The total runtime of Island model 94
5.6 Allele frequency spectrum comparison between ARGON2 ARG augmentation and HAPGEN2 100
5.7 Allele frequency spectrum of the mutation simulation of ARG inferred by ARG-Needle 101

A.1 Common and rare variants effect size 110
A.2 The Manhattan plots showing the novel association of rs151036307 with standing height discovered in G-UKB. 111
A.3 The Manhattan plots showing the novel association of rs150391548 with BMI discovered in G-UKB. 111
A.4 The Manhattan plots shows the novel TF-binding site association of rs577620983 with SBP discovered in G-UKB. 112
A.5 The Manhattan plots showing the novel missense association of rs9332745 with SBP discovered in G-UKB. 112
A.6 The Manhattan plots showing the novel synonymous association of rs151036307 with SBP discovered in G-UKB. 113
A.7 The Manhattan plots showing the association of rs1755838777 with SBP discovered in G-UKB. 113
A.8 The Manhattan plots showing the novel association of rs575127487 with SBP discovered in G-UKB. 114
A.9 The Manhattan plots showing the novel association of rs151036307 with DBP discovered in G-UKB. 114
A.10 The Manhattan plots showing the novel association of rs151036307 with SBP discovered in G-UKB. 115

List of Abbreviations

1000G . . . . . The 1000 Genomes Project

AF . . . . . Allele frequency

GEL . . . . . Genomics England

GnomAD . . . The Genome Aggregation Database

G-UKB . . . . The Genomics England imputed UK Bionbank data

HRC . . . . . Haplotype reference panel consortium

H-UKB . . . . The HRCUK10K imputed UK Biobank data

LD . . . . . Linkage disequilibrium

NGS . . . . . Next generation sequencing technology

INDEL . . . . Insertion and deletion

MAF . . . . . Minor allele frequency

SNP . . . . . Single nucleotide polymorphisms

TOPMed . . . NHLBI Trans-Omics for Precision Medicine

UKB . . . . . UK Biobank

WGS . . . . . Whole genome sequencing

WES . . . . . Whole exome sequencing

I suggest ... the term Genetics, which sufficiently indicates that our labours are devoted to the elucidation of the phenomena of heredity and variation.

William Bateson, Address to the third Conference on Hybridisation, 1906

1 Introduction

Contents

1.1 Genetics and genetic datasets 1
1.1.1 Heredity and variation 1
1.1.2 Genome-wide association studies 3
1.1.3 Rare variants associations 5
1.1.4 DNA genotyping and sequencing technologies 6
1.1.5 Large DNA datasets and reference panels 7
1.2 Haplotype phasing and genotype imputation 9
1.2.1 The phasing problem 9
1.2.2 Genotype imputation and its application in GWAS . . . 10
1.2.3 The Li-Stephens model for phasing and imputation . . . 12
1.2.4 Phasing related samples and scaffolding 17
1.3 Genetic variation data simulation 19
1.3.1 Population genetics and simulation 19
1.3.2 The discrete-time Wright-Fisher and the coalescent processes 19
1.3.3 An overview of genome simulators 22
1.4 Thesis outline 24

1.1 Genetics and genetic datasets

1.1.1 Heredity and variation

Genetics was a well established scientific field long before the discovery of the helical structure and nucleotide compositions of DNA molecules in 1953 [2]. The

relationship between the hereditary process and the variation of biological traits is always at the centre of genetic research. As early as the late 19$^{\text{th}}$century, the Mendelian school derived a set of hereditary principles based on the observations of discrete and monogenic traits such as the colour of peas. However, it was soon challenged by the biometric school, focusing on continuous traits. The biometricians pointed out that the observation of the continuous traits such as height cannot be fully explained by Mendelian inheritance with monogenic large effect genes [3].

One of the most notable examples is Galton's 1886 observation of "regression towards mediocrity" in height measurements. It was until the 1918 landmark paper by Fisher [4] reconciled the two views, showing that in his famous infinitesimal model, where numerous genes affects a trait under the Mendelian laws, each with a small effect, can yield theoretical results that fit the observation of continuous traits. The model proposed by Fisher shaped our understanding of genetics for the rest of the century. The infinitesimal model is also known for the modern day geneticists as the additive effect polygenic model [3].

Mirroring the debate from a century ago about heredity, the ongoing division between monogenic traits and complex traits continues to drive advancements in human genetics research well into the late 20$^{\text{th}}$century. In the 1980s and 1990s, Sanger sequencing technique allowed mapping genes to rare and monogenic diseases by linkage analysis and fine-mapping within pedigrees [5]. During this period, more than 1000 monogenic diseases were mapped with a single causal gene, including most notably Huntington's disease and cystic fibrosis [5]. However, little success were made for complex traits and common diseases, such diabetes and cardiovascular disease, until large genetic datasets became available in the 2000s [6], making it possible to perform genome-wide association studies, testing millions of genetic markers from thousands of individuals.

Besides the intellectual pursuit of the knowledge of "heredity and variation", the research on how genetic variation affects the variation of the phenotypes opened the doors for many applications. The associated sites, found by GWAS, provided many potential drug and therapeutical targets for genomics medicine [7]. It is

also one of the focal points of the personalized medicine and healthcare with the ever improving accuracy of polygenic risk score, fuelled by better algorithms and bigger datasets [8, 9].

The future progress in human genetics research relies on the large genetic datasets. We are witnessing the data being collected at an unprecedented speed using various sequencing technologies. The academia is no longer the only source of the genetic data, private sector and public healthcare system are increasingly involved in this endeavour. To contribute to the grand narrative in the human genetics research, this thesis is focus on the imputation reference panel building on large sequencing dataset. Genotype imputation refers to a procedure in which a sparsely genotyped data, typically using microarray chips, are imputed using a densely sequenced reference panel, which is widely used in GWAS to boost the number of testing sites and statistical power [10].

Academia is no longer the main source of novel genetic data, and the private sector and public healthcare systems are increasingly involved in the endeavour of gathering such data. To contribute to the grand narrative in human genetics research, this thesis to focus on building the infrastructure and data resources for Genome-wide association studies.

1.1.2 Genome-wide association studies

The linear mixed model for GWAS

Quantitative traits can be tested using linear regression, while binary traits use a logistic regression [11]. Covariates such as age, sex and ancestry are often introduced to account for stratification and population confounding effect. The state of the art GWAS methods use a linear mixed model to increase statistical power by introducing an additional random effect term u to account for the genetic relatedness. Such methods include BOLT-LMM [12], fastGWA [13], and REGENIE [14] Suppose we try to test a single SNP in a GWAS with N samples, the linear mixed model can be stated as follows,

$$\mathbf{Y} \sim \mathbf{X}\beta + u + \epsilon \tag{1.1}$$

where $\mathbf{Y}$ is a phenotype vector with length of N, $\mathbf{X}$ is $N \times (1 + c)$ genotype and covariate matrix. $u \sim \mathcal{N}(0, \sigma_g^2 \mathbf{A})$, where $\mathbf{A}$ is a $N \times N$ general relationship matrix, for instance for parent-child relationship, the value is 0.5. $\epsilon \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$ is residual variance due to the environmental effects.

The main goal of GWAS is to obtain the P-value in association tests and estimate the effect size β for each SNP to identify robust association and elucidate clues about diseases. Allele frequencies and effect sizes could affect the power to find associations. Rare variant associations are particularly difficult to find.

Linkage Disequilibrium and fine-mapping

Linkage disequilibrium (LD) refers to non-random correlation between alleles in different positions along the genome [15]. It can be affected by the population history, breeding system, recombination, natural selection, mutation, etc [16]. GWAS significant SNPs are not necessarily causal for traits. However, they might be highly correlated with the neighbouring causal SNPs, in which case they can be served as "tag SNPs" [17–19].

Fine-mapping methods including, heuristic fine-mapping which is also known as LD-clumping [11, 17], conditional forward stepwise regression or conditional joint analysis [20, 21], penalized regression fine-mapping [22] and Bayesian fine-mapping [23–26], are widely used in order to give a better indication of the causal variants [17]. Functional annotations are also informative for fine-mapping analyses [27–29]. We used conditional forward step-wise regression analysis and a rudimentary Bayesian fine-mapping method for fine-mapping rare and common variants (see Chapter 4).

Missing heritability

Despite the early success, the GWAS approach encounters an issue of "missing heritability", which was difficult to explain at the time. Heritability broadly refers to the proportion of phenotype variation explained by the genetic variation [30–32]. The narrow-sense heritability takes only additive genetic effect into account, whereas the broad-sense heritability takes both additive and non-additive effects, such as gene-gene interactions into account [30–32].

The early GWAS tagged markers can only explain a very small fraction of heritability estimated by twin studies [33]. For instance, the heritability estimated by GWAS is 0.03, in contrast to the twin study estimation (0.64) for human standing height. Complex traits are not mere aggregations of Mendelian conditions, but thousands of common variants with small to modest effects [7, 34, 35]. The potential places where we can find the missing heritability are common variants with small effects, often under-powered by the GWAS and rare variants often cannot be observed or grossly under-powered with smaller datasets [7, 33, 34, 36]. Increasing the number of variants and sample sizes is the key to gain more insights on the genetic architecture of complex traits. A recent successful study showed, by testing on 5.4 million samples, a significant improvement of heritability estimation has been made for height, with the heritability estimate to be 40% - 50% [35].

1.1.3 Rare variants associations

Rare variants are of great importance to understand the genetic architecture and human diseases. It is a less explored area than common variants, largely due to the unavailability of large genetic datasets earlier. Rare variants with additive small effects might contribute to the missing heritability, while those with large effect sizes could be biologically important. Causal variants for Mendelian diseases tend to be rare [34]. In addition, it was predicted by the evolutionary theory that disease mutations should be rare due to the negative selection [34]. Rare variations are much more numerous than the common variations across the human species, as shown in many large genetic datasets [37, 38], including the GEL reference panel we built in the next chapter, where 84% of the variants were with allele frequency lower than 10^{-4}. When common variants are shared across all populations, rare variants are often private [39], and more geographically localized [40] to individual populations.

For a causal variant with moderate an effect size, the rarer it goes in terms of the allele frequency, the bigger the sample size it requires to be detected by association tests. Not to mention that it is also hard to estimate the haplotype and impute the rare variants, which are essential to modern GWAS.

The two main topics of this thesis revolved around the accuracy of estimating rare variants. First, we built a reference panel to impute the rare variants at a higher precision, measured by the squared correlation between the imputed variants and the ground truth directly observed by sequencing, in order to find more rare associations, so that benefits the downstream analyses. Second, we built a more precise and efficient simulation specialized in simulating geographically localized rare variants in a large scale.

1.1.4 DNA genotyping and sequencing technologies

The first method for sequencing DNA is developed in 1977, known as Sanger-sequencing. , the first human genome was sequenced at a cost of $2.7 billion using the labour-intensive Sanger-sequencing method in 2001 [41, 42]. Reported by NHGRI the cost of sequencing one human genome has reduced from $100 million per human genome to $450 in 2021 [43], largely thanks to the development of high throughput "next-generation sequencing (NGS)", allowing massive parallel sequencing of a single molecule DNA in the early 2000. NGS is currently the most widely-used DNA sequencing technology, including whole genome sequencing (WGS), whole exome sequencing (WES), epignome sequencing, and targeted sequencing, depending on the sequencing template used [42]. NGS is also called as the short-reads sequencing, which is capable of producing reads with the length of a few hundreds of base pairs or lower. Inaccurate readings of long structural variants may happen when ligating and assembling millions of short reads. The error of NGS is typically between 1% to 0.1% [44]. In addition, short-reads creates huddle for de-novo assembly [42] and computational phasing (see Section 1.2).

Microarrays DNA testing emerged in the early 2000s [45], is popular for customized research projects and large-scale biobank datasets because of its cost efficiency. Unlike NGS, it collects mRNA and then compares the cDNA of the sample to the reference samples on a selected set of loci using florescent DNA probes. This process is called hybridization [46]. Microarray is more cost-effective than WGS and WES, and it has been the first choice for Biobank scale datasets over a

decade [1], despite the fact the cost gap between microarray and WGS is shrinking due to the rapid decrease of the price of WGS and WES.

The third-generation sequencing, focusing on long-reads over thousands of base pairs, is maturing over the last decade, represented by Pacific Biosciences (PacBio) and Oxford Nanopore technologies. It was reported 5% of PacBio can be greater than 100 Kb and Oxford Nanopore can reach 1 Mb in length. The third-generation sequencing technologies are usually more time efficient than NGS. Apart from the advantages, currently, the main problem of these new sequencing methods is that they tend to have much higher sequencing error rate than NGS.

1.1.5 Large DNA datasets and reference panels

Since the first human genome being sequenced in 2001, there are millions of DNA sequencing or genotyping data available for genome-wide association studies [5]. Launched in 2002, HapMap project sequenced 270 individuals from diverse populations, allowing to study the genetic architecture [47] and selection [48] at a population level. Started in 2007, the Wellcome trust Case Control Consortium genotyped over 19,000 individuals using microarray technology. It conducted the largest GWAS and turned out to be a huge success, advanced our understanding of the genetic architecture of common diseases [49]. Following by this, 1000 Genomes Project, gnomAD, UK Bionbak, HRC, TOPMed and GEL are the major sequencing datasets, among which the 1000 Genomes Project, HRC, TOPMed are popular reference panel for phasing and imputation [50]. These datasets are used to cross-compare various measures in Chapter 2 and 3. We will describe these datasets in more detail and leave the GEL description in Section 2.2.

The 1000 Genomes Project (1000G)

The landmark sequencing dataset of the 1000 Genomes Project was accomplished in 2015, combining low coverage sequencing, deep exome sequencing, and microarray genotyping on 2,504 individuals from 26 diverse populations [51]. The 1000 Genomes phase 3 increased the sample size to 5,008 samples and 88 million variants. Recently,

the 1000 Genomes samples are resequenced at a 30x coverage by the New York Genome Center [38]. In the later chapter, we will mainly use the high coverage data for benchmarks of GEL phasing and imputation accuracy.

Genome Aggregation Database (gnomAD)

GnomAD is a high quality aggregated dataset, containing 125,748 exomes and 15,708 genomes from WGS in v2.1.1. The GnomAD v3.1.2 is called with GRCh38 reference genome, consisting of 76,156 low coverage WGS samples, among which 34,029 are categorized as Non-Finnish Europeans [52]. The non-Finnish European site list and allele frequency was used in the quality control of the GEL reference panel (see Section 2.3).

UK Biobank (UKB)

The UK Biobank (UKB) is one of the most comprehensive and deeply phenotyped dataset used by thousands of researchers since it was published in 2017 [1]. The UK Biobank has around 500,000 samples, mainly from the UK, genotyped by microarrays with approximately 800,000 markers. The imputation using HRCUK10K reference panel [1] increased the testable markers by over 100 folds, reaching 96 million markers. The UKB sequencing data has been produced in the last years. The WES for 450,000 UKB samples are made available in 2021 [53] and the high coverage WGS for 150,000 individuals are made available in 2022 [54]. The rest of the UK Biobank samples are planned to be sequenced by late 2023. Similar to the UKB, there are biobanks around the world, typically using microarray technology such as FinnGen [55], Chinese Kadoori Biobank [56] and Biobank Japan [57].

Haplotype reference consortium (HRC)

Haplotype reference panel consortium (HRC) is an aggregation of low coverage datasets from 30 populations, consisting of 32,000 samples and 39 million variants. HRC and 1000 Genomes reference phase 3 panel are the most popular low coverage sequencing reference panel.

Panel	Year	Haplotypes	Populations	Variants (M)
HapMap3	2009	2,368	11	1.4
1000 Genomes phase 1	2012	2,186	14	38.2
1000 Genomes Phase 3	2014	5,008	26	88
UK10K	2014	7,562	1	26
HRC	2016	64,976	30	39.2
TOPMed r2	2021	194,512	Multi-ethnic	308.1*
GEL	2021	156,390	Multi-ethnic	342.6*

Table 1.1: Summary of imputation reference panels, showing the release year, the number of population, the number of haplotypes, and the number of variants, including SNPs and INDELs, in millions. TOPMed and GEL are multi-ethnic dataset, without a specified number of populations. (*) indicates the autosome variants only.

NHLBI Trans-Omics for Precision Medicine (TOPMed)

TOPMed consists of over 180,000 deep whole genome sequencing samples from 85 different studies, focusing on heart, lung, blood, and sleep disorders. TOPMed samples are collected from the US with a diverse ancestry and ethnicity, approximately 40% European, 29% African and 19% Hispanic ancestries. TOPMed is widely used as an imputation reference panel. The high coverage sequencing data allowed TOPMed to impute rare and ultra-rare variants, which were not possible to be observed in lower coverage sequencing panels such as HRC [37]. The latest publicly available TOPMed reference panel is TOPMed r2, consisting 97,256 samples with 308 million autosomal variants [37].

1.2 Haplotype phasing and genotype imputation

1.2.1 The phasing problem

Human genomes consists of maternal and paternal genetic material or alleles, as a result of the recombination process during meiosis [58]. A haplotype is a set of alleles inherited from a single parent, and it can be regarded as a fundamental unit of inheritance [59]. Haplotypes has its application to a wide range of statistical and population genetics, to answer questions such as, relationship between genetic variation and diseases [60], the detection of selection [48], recombination rate estimate [61], and ancestry inference [62–64].

Sequencing and SNP array doesn't include explicit haplotype information. The current laboratory methods to obtain haplotypes are either not sufficiently accurate or not cost-effective to operate on a large scale [65]. It is a common practice to use computational methods to estimate the haplotypes, as we call it phasing. Phasing is also a necessary step for many modern imputation algorithms [10].

The coalescent based hidden Markov models (HMM) consider new haplotypes as a product of mutation and recombination processes of the known haplotypes. The modern computational phasing methods are proven to be cheap and highly accurate, especially for the common variants [59, 65]. The first algorithm that uses this method is PHASE [66], released in 2001. The latest methods include Eagle2 [67], beagle5[68], SHAPEIT4 [69]. We choose to use SHAPEIT4 to phase the Genomics England samples (Section 2.4), as it is the fastest phasing algorithm to phase a large population by the time we carried out the experiment, 3.6x faster than the second best phasing algorithm Beagle 5 [69]. Multi-allelic SNPs are phased with one alternative allele at a time, and short Indels and structural variants are treated the same as SNPs. Diploid samples can be phased with or without a reference panel. Details of coalescent based methods will be discussed in Section 1.2.3.

1.2.2 Genotype imputation and its application in GWAS

The amount of genetic variation captured by genotyping array is limited, typically up to several million markers [70], leaving hundreds of millions of less common variants in human populations remained unobserved [71]. Genotype imputation is to statistically infer these unobserved genotypes. It is a common practice used in association studies. For example, the UK Biobank data increased the number of testable markers from 800,000 to 96 million by imputing from the SNP array data [1]. Imputation can also boost the power of GWAS and fine-mapping and enable the integration of different datasets in meta analysis [59].

A combination of SNP array and imputation is still the most affordable way to create large genetic data, since the SNP array data is considerably cheaper than WGS or WES and computation cost for imputation is as low as one US penny per genome

[72]. In Chapter 4, we will show the sample size is still a major limitation of GWAS, despite a relatively large sample size of the UK Biobank (0.5 million). Sample size is very much in tandem with the cost. The cost benefit is one of the most attractive feature of SNP array + imputation approach, compared to WGS and WES.

The state-of-the-art imputation methods are based on Li-Stephens model, requiring a phased reference panel and phased target SNP array data [10, 70, 72, 73]. The Li-Stephens HMM estimates a phased genotype by matching an imperfect mosaic of copying segments among the reference panel samples (Figure 1.1). IMPUTE5 [74], minimac4 [50] and Beagle5 [72] implemented such model, which are suitable for imputing with large reference panels. Online imputation servers, such as the Michigan imputation server [50] and TOPMed imputation server [37], allowed the data to be imputed by a wide range of reference panels. In Chapter 3, we used IMPUTE5 to impute the UK Biobank data by the GEL reference panel, as it is up to 30x faster than minimac4 and 3x faster than Beagle5 [74].

Measuring imputation quality

MaCH $\hat{r}^2$ and IMPUTE information (INFO) are the most commonly used imputation quality metrics. Both are ranged from 0 to 1, with 0 standing for full uncertainty and 1 no uncertainty. MaCH $\hat{r}^2$ measures a ratio of the variance of the imputed alleles to the expected binomial variance under Hardy-Weinberg equilibrium. The IMPUTE information (INFO) metric measures the information of imputed allele frequency [10]. Suppose we try to impute N individuals,

$$\text{INFO} = \begin{cases} 1 - \frac{\sum_{i=1}^{N}(p_{1,i}+4p_{2,i}-\text{dosage}_i^2)}{2N\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1 \end{cases} \tag{1.2}$$

where $\hat{\theta}$ is the imputed allele frequency of the imputed marker. The dosage_i is calculated as in equation 1.3 for individual i. $p_{1,i}$ and $p_{2,i}$ are the imputed posterior probability of imputed alleles being 1 and 2 for individual i, respectively. MaCH r^2 and INFO are highly correlated and can be used interchangeably [10]. INFO can be used as a post-imputation quality control measure. If a GWAS has N=500,000

samples, in which an SNP with INFO=0.3 can be roughly interpreted as having the equivalent statistical power of 150,000 (INFO $\times$ N) samples.

Testing imputed data for association

Imputation methods usually produce a posterior distribution $p_g, g \in \{0, 1, 2\}$ for each marker per individual to determine the imputed genotype. It is not recommended to use genotype estimates by using a simple threshold in genome-wide association studies, as it may cause the loss of power and spurious associations [59], instead using expected allele count or allele dosage is often used to test imputed data for associations [59], which can be computed directly from the posterior distribution,

$$\text{dosage} = \sum_{g=0}^{2} p_g g \tag{1.3}$$

Factors affecting the accuracy of imputation

Essentially, imputation models copy segments of genetic data from the reference panel. The higher the number of haplotypes in the reference panel, the higher the chance of finding a longer matching segments. For similar reason, the imputation accuracy is better, when the reference panel ancestry matches the target sample ancestry [59], which effectively increases the chance of finding distant relatives. Therefore, a good general purpose reference panel is expected to have high genetic diversity and large sample sizes [59]. Other factors include the SNP array density and allele frequency. Imputation algorithm tends to perform better in common variants and worse in rare variants [75]. Squared correlation r^2 is considered to be a good measure for imputation accuracy, since it directly linked to the power of association tests. The power of detecting a causal imputed variant is r^2 times the power, had it been directly observed [59].

1.2.3 The Li-Stephens model for phasing and imputation

The Li-Stephens model [76] is a widely adopted approach in the latest development of phasing and imputation algorithms [10, 77]. It uses a hidden Markov model (HMM)

framework, modelling a diploid sample as a mosaic of haploid segments copied from a haploid reference panel. The transition probability models the recombination and the emission probability to model the mutation process. Before talking about the details of the method, we will first review the hidden Markov model.

Hidden Markov model

The hidden Markov model (HMM) is the backbone of phasing and imputation algorithms [10, 65, 68, 78, 79]. We will use this section to briefly introduce the concepts and notations of HMM.

A Markov model consists of the following components,

- a sequence of L observations $X = [x_1, x_2, ..., x_L]$, in the phasing and imputation, observation is the observed genotypes.
- a set of K states $Z = [z_1, z_2, ..., z_K]$
- transition probability matrix A, where a_{ij} represents the probability of moving from i state to j state.
- emission probability b, where $b_i(o_l)$ is the probability of observing at position l, o_l, from state i.
- initial probability $\pi = [\pi_1, \pi_2, ...\pi_K]$, signify the probability that the Markov chain will start in state i.

Rabiner pointed out in his 1989 tutorial [80] that there are three basic problems that HMM are trying to solve. First the **(1) likelihood problem**, that we want to know the likelihood $P(X|\theta)$ of the observation sequence, given the parameters $\theta = (A, b)$; the second problem is the **(2) decoding problem**, that we want to find the best hidden state Z to explain the observation $\mathbf{X}$, given the parameter θ; the third question is the **(3) learning problem**, that we want to learn the parameter θ, given the observation X and states Z [80]. The solution to the first problem is a forward algorithm, to the second problem is Viterbi algorithm and

forward-backward algorithm can solve the third problem [80]. We will go through the forward algorithm and forward-backward algorithm in more detail.

To solve the first problem, we can calculate the likelihood using the forward algorithm. The forward algorithm taking advantage of the Markovian assumption that each state only depends on its immediate previous state, $P(z_i|z_1, ..., z_{i-1}) = P(z_i|z_{i-1})$, the likelihood can be expressed as the sum of forward probability $\alpha_l(i)$ at the state i, at the last position n.

$$\alpha_1(j) = \pi_j b_j(o_1)) \tag{1.4}$$

$$\alpha_l(j) = \sum_{i=1}^{N} \alpha_{l-1}(i) a_{ij} b_j(o_t) \tag{1.5}$$

$$P(\mathbf{X}|\theta) = \sum_{i=1}^{K} \alpha_n(i) \tag{1.6}$$

Similarly, the backward probability is derived from the backward recursion,

$$\beta_l(j) = \sum_{i=1}^{K} a_{ij} b_j(X_l) \beta_{l+1}(i) \tag{1.7}$$

forward-backward algorithm

$$\begin{aligned} P(Z_l|X) &= \frac{P(X|Z_l)P(Z_l)}{P(X)} \\ &= \frac{P(X_{1:l}, Z_l)P(X_{l+1:L}|Z_l)}{P(X)} \\ &= \frac{\alpha(Z_l)\beta(Z_l)}{P(X)} \propto \alpha_l(Z_l)\beta_l(Z_l) \end{aligned} \tag{1.8}$$

The Li-Stephens model

When a reference panel has a sufficiently large number of haplotypes, sampled from a particular population, then new haplotypes from this population can be modelled as a mosaic of copying segments from a haploid reference panel. This intuition is formalized by the Li-Stephens model. Suppose there are K haplotypes with length L in the reference panel and each are labelled from 0 to $K-1$. The label from which the current haplotype is copied at position l is Z_l, the hidden state in HMM. H is

denoted as the haplotype reference panel, where the haplotype of u^{th} the sample at l^{th} position is expressed as $H_{i,l} \in \{0, 1\}$. The observed haplotype is denoted as O.

A switching between copying state can happen, if there is a recombination happened in the genealogy. Derived from Kingman's coalescent [81], the transition probability of the Li-Stephens model is specified as follows [76],

$$\begin{aligned} P(Z_1 = u) &= \frac{1}{K} \\ P(Z_l = u | Z_{l-1} = v) &= \begin{cases} e^{-\rho_l/K} + \frac{1-e^{-\rho_l/K}}{K} & \text{if } u = v \\ \frac{1-e^{-\rho_l/K}}{K} & \text{if } u \neq v \end{cases} \end{aligned} \tag{1.9}$$

where $\rho_l = 4N_e r_l$ and r_l is the per generation average rate of crossover recombination per unit physical distance between sites l and $l-1$. N_e is the effective population size, which is insensitive to changes for phasing and imputation purposes [82]. By default, the N_e is set to be 15,000 in SHAPEIT2 [82]. It has also been shown that N_e is not sensitive to the phasing results [82].

The emission model follows the mutation process,

$$P(O_l = a | Z_l = u) = \begin{cases} \frac{\theta}{2(\theta+K)} & H_{u,l} \neq a \\ 1 - \frac{\theta}{2(\theta+K)} & H_{u,l} = a \end{cases} \tag{1.10}$$

where θ is the Watterson's coefficient $(\sum_{n=1}^{N-1} \frac{1}{N})^{-1}$.

Haplotype phasing without a reference panel

Phasing diploid samples without a reference panel is more complex than the Li-Stephens model we described earlier. It requires the modelling of two hidden states and the MCMC iterations to improve the haplotype estimate. To demonstrate how Li-Stephens HMM works with phasing without a reference panel, we will describe the phasing model, largely based on SHAPEIT2 [83]. Suppose we try to estimate the haplotype of N diploid samples with genome length of L, the individual's genotypes are denoted as $G = [G_1, G_2, ..., G_L]$, where $G_l \in \{0, 1, 2\}$ and each individual is assigned with a label from 0 to $N-1$. We also define X as the labels of consistent haplotypes to the genotypes based on the current haplotype estimate H, where H is a matrix containing K current haplotype estimates. The copying state is denoted

as $Z = Z_1, Z_2, ..., Z_l$. This phasing model is an HMM with two hidden states X and Z. We are mainly interested in the conditional distribution of $P(X|H)$.

The joint distribution of both hidden states X and Z can be expressed by a Markov chain,

$$P(X, Z|H) = P(X_1|Z_1, H)P(Z_1)\prod_{l=2}^{L} P(X_l, Z_l|X_{l-1}, Z_{l-1}, H) \tag{1.11}$$

where $P(X_l, Z_l|X_{l-1}, Z_{l-1}, H)$ is the product of the transition probability (e.q. 1.9) and emission probability (e.q. 1.10) in the Li-Stephens model. Since the direct computation of such distribution is intractable, we will use the forward-backward algorithm. The forward probabilities can be calculated through the following iterations.

$$\begin{aligned}\alpha_{l+1}(i,u) &= P(X_l = i, Z_l = u, G_{1:l}|H) \\ &= P(X_{l+1} = i|Z_{l+1} = u, H)\sum_{v=1}^{K}\alpha_l(i,v)P(Z_{l+1} = u|Z_l = v)\end{aligned} \tag{1.12}$$

In the right-hand side of the second equation, $P(X_{l+1} = i|Z_{l+1} = u, H)$ is the emission probability (e.q. 1.10) and $P(Z_{l+1} = u|Z_l = v)$ (e.q. 1.9) is the transition probability of the Li-Stephens model. The forward probabilities is initialized as follows,

$$\alpha_1(i,u) = P(X_1 = i|Z_1 = u, H)P(Z_1 = u) \tag{1.13}$$

The backward probabilities can be calculated as

$$\begin{aligned}\beta_l(i,u) &= P(G_{l+1:L}|X_m = i, Z_m = u, H) \\ &= \sum_{v=1}^{K}\beta_{l+1}(i,v)P(X_{l+1} = i|Z_{l+1} = v)P(Z_{l+1} = v|Z_l = u)\end{aligned} \tag{1.14}$$

and it is initialized as

$$\beta_L(i,u) = 1 \tag{1.15}$$

The marginal probability of the unobserved states X and Z at position l is

$$\begin{aligned}P(X_l = i, Z_l = u|H) &\propto P(X_l = i, Z_l = u, G_{1:l}|H)P(G_{l+1:L}|X_l = i, Z_l = u, H) \\ &= \alpha_l(i,u)\beta_l(i,u)\end{aligned} \tag{1.16}$$

Hence the marginal probability of X_l can be calculated as,

$$P(X_l = i|H) \propto \sum_{u=1}^{K} \alpha_l(i,u)\beta_l(i,u) \tag{1.17}$$

Pre-phased genotype imputation with a reference panel

Genotype imputation requires a densely sequenced haplotype reference panel and a phased set of target haplotypes [10, 59]. The phasing algorithm uses a haploid Li-Stephens' model, modelling one haploid copying state along the genome. A haplotype target (T) is modelled as an imperfect mosaic of copying states (Z). Following equation 1.8, the marginal posterior distribution of copying state can be written as a product of forward and backward probability

$$P(Z_l|T,H) \propto P(T_{1:l}, Z_l|H)P(T_{l+1:L}|Z_l,H) \tag{1.18}$$

The marginal posterior distribution of the unobserved site (T_l) can be imputed as [59],

$$P(T_l = d|T,H) = \sum_{T_l=1}^{M} P(X_l = d|T_l)P(T_l|X,H) \tag{1.19}$$

The advantage of haploid imputation algorithm is that it doesn't require updating the haplotype estimate of the reference panel, as it does in phasing. Therefore, imputation is an order of magnitude faster than phasing. Phasing and imputation algorithm assumes the reference panel and target samples are called by the same reference genome [69, 74]. In case of misalignment of the reference genome, for instance, UK Biobank SNP array data is called on hg19 and the GEL reference panel is hg38 (see Chapter 3). Tools converting genome coordinates between assemblies are called LiftOver tools, such as Picard LiftOver [84]. During this process, some variants might be lost due to conversion failures [84].

1.2.4 Phasing related samples and scaffolding

There are many methods to phase related samples [77], yet the easiest way is to infer by transmission. In table 1.2, in a trio family nearly all genotype cases can be inferred by transmission, apart from triple-het case. The phased related

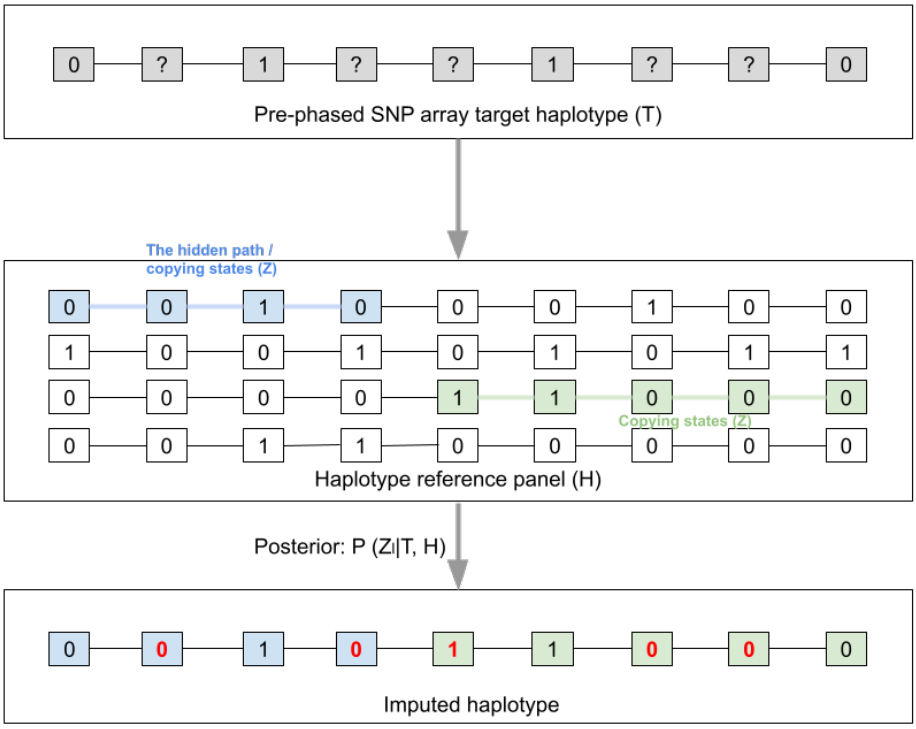


Figure 1.1: The schematic of pre-phased genotype imputation.

Unphased Genotypes			Phased Haplotypes		
Father	Mother	Child	Father	Mother	Child
0	1	0	0|0	0|1	0|0
0	1	1	0|0	0|1	0|1
0	2	1	0|0	1|1	0|1
1	0	0	1|0	0|0	0|0
1	0	1	0|1	0|0	1|0
1	1	0	1|0	1|0	0|0
1	1	1	?	?	?
1	1	2	0|1	0|1	1|1
1	2	1	1|0	1|1	0|1
1	2	2	0|1	1|1	1|1
2	0	1	1|1	0|0	1|0
2	1	1	1|1	1|0	1|0
2	1	2	1|1	0|1	1|1

Table 1.2: In this table we listed all the consistent genotype cases in a trio family and their phasing results.

samples then can be used as scaffold, which was a concept introduced as part of SHAPEIT2 [85]. It sets the transition probability of haplotype pairs that are inconsistent to the scaffold pre-phased results to zero. This means not only the final phasing dataset would not be produce inconsistent result to the scaffold, but also brings a huge advantage when ligate phasing segments, as the order of haplotypes of a sample will be fixed. SHAPEIT2 [83] and SHAPEIT4 [69] allows phasing with pre-phased scaffold. We used makeScaffold program to produce the partially phased trio/duo dataset. For parents, the transmitted alleles are on the right and for children, the alleles transmitted from the mother are on the right and from the father on the left, as shown in table 1.2. For phasing the GEL samples, all the undetermined sites, including the ones with Mendelian errors are treated as unphased, but not removed from the dataset, instead their phases are determined by computation phasing using the Li-Stephens model.

1.3 Genetic variation data simulation

1.3.1 Population genetics and simulation

Population genetics is a field that studies the genetic composition within a population, modelling quantities such as allele frequencies and the genealogical history of a population. There are four key drivers which give arise to the genetic variation in a population, namely mutation, recombination, genetic drift and selection. Simulations that model these factors enable creating synthetic genetic data sets, which may be used to assist theoretical modelling and to test experimental designs in statistical genetics [86–88]. Chapter 5 of this thesis is dedicated to the development of an efficient simulator of the discrete-time Wright-Fisher process, called ARGON2.

1.3.2 The discrete-time Wright-Fisher and the coalescent processes

Independently developed by Wright (1931) and Fisher (1930), the Wright-Fisher model is one of the simplest and most commonly used models in population genetics, describing the evolutionary and transmission dynamics in an idealized population.

In each discrete, non-overlapping generation, individual genes are copied at random from an individual in the previous generation. In its simplest form, the model assumes neutrality, namely that each individual has equal fitness, as well as no recombination, random mating, and that the effective population size is constant. The number of offspring of an individual, X, follows a binomial distribution, causing random drift in the frequency of genetic variatiants (genetic drift), which are created through random mutation events. A new generation is created by repeatedly copying randomly sampled individuals. For a population of effective size $2N$, each individual is chosen with a probability of $p = 1/2N$, having 1 child in expectation.

$$P(X = k) = \binom{2N}{k} p^k (1-p)^{2N-k} \tag{1.20}$$

$$\mathbb{E}(X) = 2Np = 1 \tag{1.21}$$

Kingman (1983) introduced a stochastic process, called the coalescent, that approaches the Wright-Fisher process as the effective population size goes to infinity. The coalescent starts at generation 0 (present time) and proceeds backward until all the genetic material reaches a most recent common ancestor (MRCA) [81, 89]. Under this model, the coalescence times T_i at which two lineages merge by finding a common ancestor are independent and exponentially distributed [81],

$$f_{T_i}(t_i) = \binom{i}{2} e^{-\binom{i}{2} t_i} \tag{1.22}$$

The coalescent process operates on a continuous timescale, in contrast to the Wright-Fisher model [90]. Though the coalescent can be seen as an approximation of the Wright-Fisher process, it is often more tractable, and can be used to obtain significant insights into the behaviour of quantities of interest, such as the expected height of a marginal coalescence tree (1.23) or the expected number of segregating sites (1.24). These easy-to-calculate from the coalescent model are useful in examine the topology and the variation pattern, during in the development of ARGON2 in Chapter 5.

Under the coalescent, the time for the most recent common ancestor T_{MRCA} [91, 92] may be computed as

$$\mathbb{E}[T_{\text{MRCA}}] = 2\left(1 - \frac{1}{n}\right). \tag{1.23}$$

If we know the population-scaled mutation rate θ, Watterson's estimator [93] can be used to estimate the number of segregating sites S:

$$S = \theta \sum_{j=1}^{n-1} \frac{1}{j}. \tag{1.24}$$

The coalescent process may also be used to model recombination [91], and the expected number of recombination events in a sample history may be computed as

$$\mathbb{E}(R_n) = \rho \sum_{i=1}^{n-1} \frac{1}{n}. \tag{1.25}$$

In 1990, Hudson proposed an algorithm to sample from the coalescent process [91]. Many modern simulators such as ms [94] and msprime [95, 96] are developed based on Hudson's algorithm. Since both coalescence and recombination events independently follow a Poisson process, sampling in the coalescent with recombination may be described as waiting for competing events with exponentially distributed arrival times. A brief sketch of Hudson's backward algorithm is as follows (also see [97]). Going backwards in time, first draw a random number as the waiting time for the next event to happen from an exponential distribution with parameter $k(k-1)/2 + k\rho/2$. Upon reaching the waiting time, draw a Bernoulli random sample to determine whether the event is a coalescent or recombination event, with the probability of coalescent event to be $(k-1)/(k-1+\rho)$. In case of a coalescent event, randomly choose two individuals and merge them into one ancestral genome, while, in the case of recombination, randomly choose an individual and break the genome at a position chosen uniformly at random. These operations are repeated until all present day genetic material reaches a most recent common ancestor.

Although the coalescent with recombination process may be used to effectively approximate the discrete-time Wright-Fisher (DTWF) process, this approximation is only accurate if the sample size is small compared to the effective population

size. The two processes may otherwise diverge significantly [98–102]. The coalescent may for instance underestimate the count of rare variants expected in a population, especially singletons (count=1) and doubletons (count=2), or distort patterns of long-range LD. Some of these distortions may be ascribed to the fact that the coalescent does not allow multiple lineages to be merged in a single generation to form a polytomy [100]. These cases affect the simulation of the recent evolutionary past more than the distant past, due to a gradual decrease of the number of lineages. This observation has been leveraged to improve accuracy by performing hybrid calculations involving both the DTWF and coalescent processes [100, 102].

1.3.3 An overview of genome simulators

Genomic simulators of the processes described above can be broadly categorized as forward simulators, simulating from the past to recent generations; backward simulators, simulating from the recent generations to the past; and spatial simulators, simulating along the genome. The modeling of natural selection is more difficult in backwards simulation, so complex models of selection are often simulated forward in time. This section will provide a brief overview of existing simulators that are most closely related the work described in Chapter 5, focusing on coalescent, spatial, and discrete-time simulators, as well as other paradigms based on the Li-Stephens model [76].

In the early 2000s, ms [94, 103] introduced recombination into the simulation process, allowing the simulation of short segments with LD for a limited number of samples. Several authors also developed models of the coalescent process where events are sampled along the genome, rather than backwards in time [104–106]. Such a spatial approach led to the development of several simulators, such as MaCH [107], Fastsimcoal [108, 109] and SCRM [110], that rely on the sequential Markovian coalescent (SMC) approximation of the coalescent, making the simulation and genealogical inference more tractable [105, 106].

A number of simulators have also been developed to extend and improve the efficiency of Hudson's backwards-in-time simulation algorithm. msHOT [111] added

the possibility to perform simulations including crossover and gene conversion hotspots. The COSI simulator [86] was used in conjunction with data from the HapMap Project [47] to obtain realistic simulation parameters for human populations. A subsequent implementation, COSI2 [112], made significant improvements to computational performance, and enabled the simulation of some modes of natural selection. The msprime simulator provided further improvements in computational efficiency using better data structure, enabling to simulate larger sample sizes and chromosome lengths [95, 96]. msprime is currently one of the most popular genomic simulators; it is extensively documented, actively maintained, and implements several features.

While Hudson's algorithm efficiently skips between subsequent coalescence or recombination events (see 1.3.2), discrete-time Wright-Fisher simulators require processing genomic and genealogical information at each discrete generation. Because the coalescent is an inaccurate approximation of the DTWF process when the sample size is increased [98–102], however, DTWF simulation may still be desirable. Furthermore, as further discussed in Chapter 5, sampling from the DTWF can be more efficient than sampling from the coalescent process under complex demographic or evolutionary models.

Several DTWF simulators have been developed. The GENOME simulator [113] allowed efficiently sampling from the DTWF by relying on an approximation where genomic regions are only allowed to recombine at user-specified intervals. The ARGON simulator, extended in Chapter 5 and hereafter referred to as ARGON1, enabled DTWF simulations involving large sample sizes and chromosome lengths, including support for non-crossover gene conversion. The mspime software package was also recently extended to support efficient DTWF and hybrid coalescent/DTWF simulation [96, 102].

Progress in coalescent and DTWF simulation has been driven by several algorithmic advances, including the use of graph-based representations of the simulated genealogies, which underpins some of the work described in Chapter 5. The msprime [95] coalescent simulator and the ARGON [101] DTWF simulator,

in particular, relied on graph based representations of the simulated genealogies, referring to these as "succinct tree sequences" and "ancestral recombination graphs" (ARG [114]), respectively. This kind of representation is significantly more compact than a representation based on storing a new marginal tree whenever a recombination event affects the genealogy [95] and enables to efficiently compute several quantities of interest, such as identical-by-descent (IBD) segments [101] and several other summary statistics [115].

The synthetic data generated using the above simulators has no relationship with any existing genomic data set. In some applications, it may be useful to generate data that is closely related and in some sense extends an existing data set. The Li-Stephens model [76] further approximates the genealogical processes described above. It underpins several algorithms used for genotype phasing an imputation, and has also been used to enable reference-based simulation in software packages such as HAPGEN2 [87]. This approach is fast and can capture the unique genetic characteristics of a set of reference genomes, although compared to coalescent and DTWF simulation the generated data may more strongly deviate from what would be expected under realistic evolution. The ARGON2 simulator, developed in Chapter 5, enables simulating new samples on top of an ARG that has been inferred from real genomic data (see 5.4).

1.4 Thesis outline

The outline of this thesis is as follows. The introduction chapter describes the evolution of DNA sequencing technologies, landmark genetic datasets, and provides background information on genotype phasing and imputation. In Chapter 2 we describe how the GEL reference panel was built using $78,195$ WGS samples from the $100,000$ Genome project, as well as its performance in genotype phasing and imputation. Chapter 3 describes the imputation of $488,315$ UK Biobank samples using the GEL reference panel. The GEL imputed UKB data are then used to perform genome-wide association studies on four continuous traits, including height, BMI, systolic and diastolic blood pressure; GWAS and fine-mapping analysis are

shown in Chapter 4. Finally, we describe the development of ARGON2, a new algorithm for DTWF simulation, in Chapter 5.

2

Building the Genomics England (GEL) Reference Panel

Contents

2.1 Overview 27
2.2 Genomics England reference panel 28
2.2.1 The aggregated variant (aggVCF) calls 28
2.2.2 Population structure 29
2.2.3 Relatedness 30
2.2.4 GEL reference panel with 342 million autosomal variants 33
2.3 GEL reference panel quality control 34
2.3.1 Lenient filters for common variants 37
2.4 Haplotype phasing 40
2.4.1 Phasing pipeline for autosomes 40
2.5 GEL reference panel phasing accuracy 41
2.6 GEL reference panel imputation accuracy 43
2.7 Discussion 45

2.1 Overview

Genomics England 100,000 Genomes project sequenced over 120,000 genomes from over 80,000 individuals across the UK. We built a GEL reference panel from 78,195 high coverage WGS samples. The main challenge comes from the quality control of the variants. We designed variant filters to ensure that predominantly high quality

variants were included in the reference panel. This resulted in a reference panel containing 342 million autosomal variants, a reduction from 722 million variants prior to the quality control. In Section 2.2, we will summarize the main features of the GEL reference panel, including the population structure, relatedness and allele frequency spectrum. The site list will be compared to other reference panels, such as HRC and TOPMed. The details of QC filters and the effect of these filters are discussed in Section 2.3. Leveraging the high relatedness of the GEL dataset, the phasing of 47,340 related and 30,855 unrelated samples are carried out separately, using SHAPEIT4 [69]. The details of the phasing pipeline will be described in Section 2.4. In Section 2.5, Finally, we will examine the phasing accuracy and imputation accuracy of the GEL reference panel in Section 2.5 and 2.6, respectively.

2.2 Genomics England reference panel

2.2.1 The aggregated variant (aggVCF) calls

The Genomics England 100,000 Genomes Project was launched in 2013, focusing on rare diseases and cancer patients. Over 120,000 genomes, including germline and cancer genomes, have been sequenced by the time of writing. It comprises genomes from 73,700 rare diseas patients (disorders affecting ≤ 1 in 2000 persons) and their close relatives, and 46,539 genomes from cancer patients[116]. The GEL reference panel is built on the aggregated dataset (aggV2), comprising 78,195 samples from both rare disease and cancer germline genomes. Samples are sequenced with 150bp paired-end reads on the IlluminaHiSeq X and processed with the Illumina North Star Version 4 Whole Genome Sequenced Workflow (iSAAC Aligner v03.16.02.19 and Starling small variant caller v2.4.7). The resulting gVCF files are aligned to the GRCh38 human reference genome. Then, the individual gVCF files are aggregated into multi-sample VCF files, using Illumina gVCF genotyper and normalized with vt v0.57721. Sample level quality control has been carried out by Genomics England, including sample contamination less than 0.03, ratio of SNP heterozygous to homozygous calls less than 3, total number of SNPs between 3.2M to 4.7M per sample, array concordance greater than 90%, median fragment size greater than

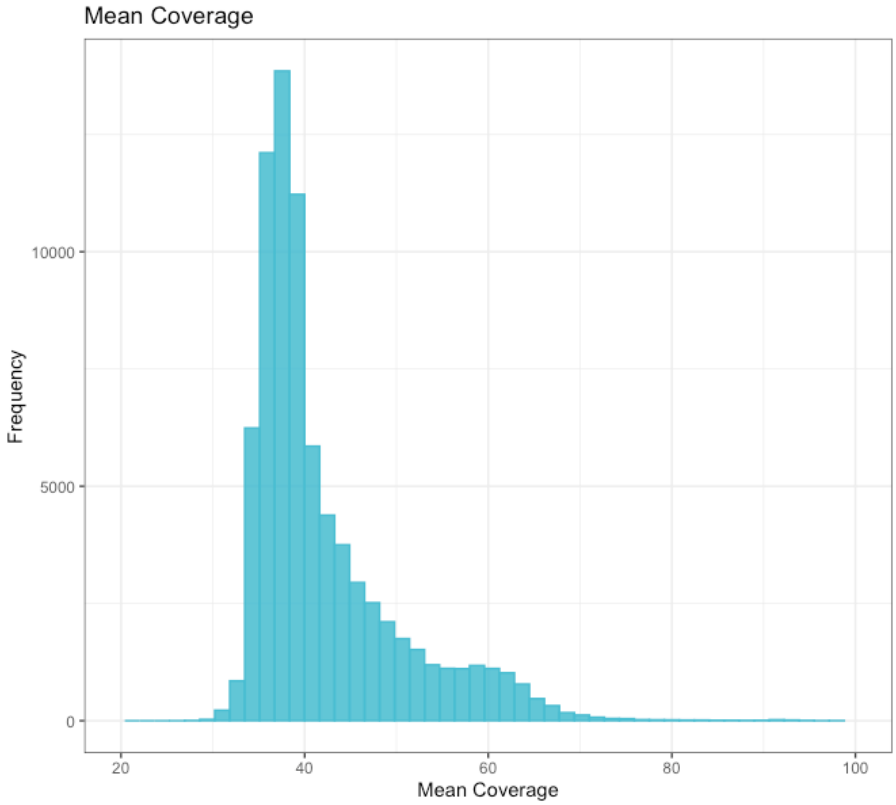


Figure 2.1: The mean genome-wide coverage for the 78,195 GEL reference panel sequencing samples. This figure is downloaded from the GEL website, https://re-docs.genomicsengland.co.uk/sample_qc/

250bp, excess of chimeric reads less than 5%, percentage of mapped reads greater than 60% and the percentage of AT dropout less than 10%. Finally, the resulting aggregated VCF dataset (aggVCF) was created, comprising over 722 million SNPs and short indels ($\leq$ 50bp). Multi-allelic variants were decomposed into bi-allelic variants. The average genome-wide coverage of the aggVCF samples is 42.2 and median mean coverage is 39.11 (Figure 2.1).

2.2.2 Population structure

The reference panel consists of 78,195 samples, recruited from hospitals across England (Figure 2.2). It has a wide representation of ethnic groups, including 49,641 samples (63.48%) self-identified as White British, 4,100 (5.24%) as other white, 2,885 (3.69%) as Pakistani and 1,751 (2.24%) as Indian. Table 2.1 provides a breakdown of self-reported ancestry. The large White British/Irish and relatively large South

	Self-reported ethnicity	**Number of samples (percentage)**
White or White British	White British	49,641(63.48%)
	White Irish	1,048 (1.34%)
	Other White	4,100 (5.24%)
Asian or Asian British	Pakistani	2,885 (3.69%)
	Indian	1,751 (2.24%)
	Bangladeshi	647 (0.82%)
	Chinese	209 (0.27%)
	Other Asian	1,180 (1.5%)
Black or Black British	African	991 (1.2%)
	Caribbean	652 (0.83%)
	Other Black	217 (0.27%)
Other	Other ethnic groups	1,151 (1.47%)
Mixed	Mixed	1,446 (1.85%)
Unknown	Unknown	12,277 (15.7%)

Table 2.1: Breakdown of self-reported ethnicity in reference panel.

Asian sample size will later prove helpful in boosting phasing and imputation accuracy for these populations [10, 77]. Notably, the population structure of GEL reference panel resembles that of the UK Biobank [1], making it an ideal reference panel for the UK Biobank imputation.

In addition, we also used the unpublished software provided by Hu et al., to obtain ancestry components of the GEL individuals. The admixture plots (Figure 2.2-b) show that the inferred ancestry components are consistent with the self-reported ethnicity. The ancestry component (defined by Hu et al.) results also indicate the vast majority of the self-reported unknown group have British and South Asian ancestry. In this thesis, we will use the self-reported ethnicity for QC and phasing (see Section 2.3).

2.2.3 Relatedness

The sample relatedness in the reference panel is high. According to the self-reported data, 27,346 samples (34.97%) are unrelated, whereas 11,584 (14.81%), 32,679 (41.79%), and 6,586 (8.43%) samples have 2, 3 and >3 first degree relatives in the dataset, respectively. Among the related samples, 17,871 (22.85%) are marked

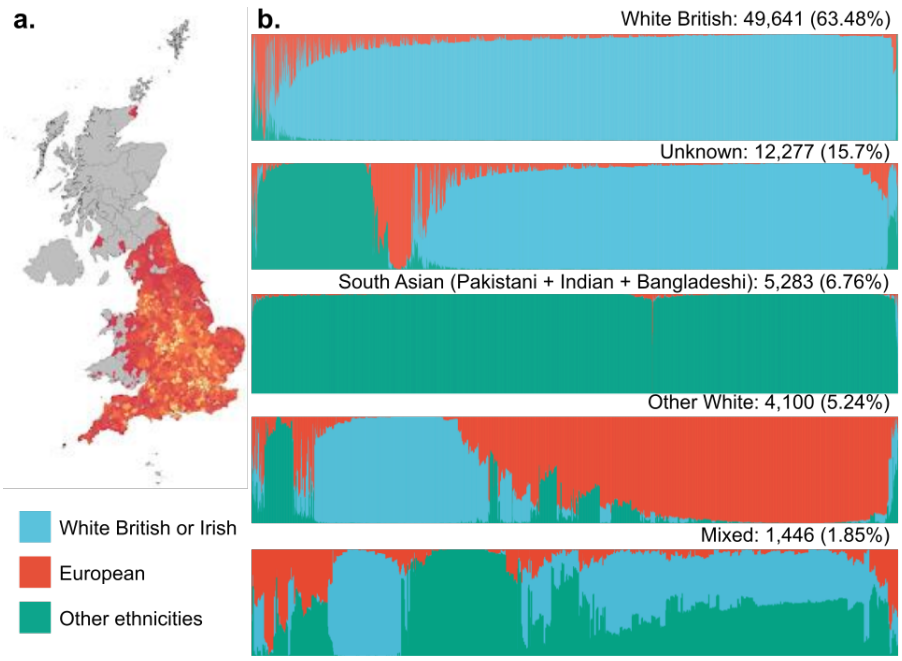


Figure 2.2: a) GEL sample residential postcode density map. The colour scale is from red to yellow, with yellow representing more densely sampled regions. b) Admixture plots grouped by self-reported ethnicity. The label consists of self-reported ethnicity, the number of sample and the proportion of the ethnicity group in GEL reference panel. The ancestry component assessment for each individual is produced by unpublished software (Hu et al., shared with permission) and then grouped using UPGMA hierarchical clustering (hcluster) in R.

as proband, 15,908 (20.34%) as mother to the proband, 12,409 (15.8%) as father to the proband, 3,149 (4.03%) as siblings to the proband, and 1,512 (1.93%) as other relatedness, such as grandparents or cousin to the proband. High relatedness improves the performance of Mendel error filter and phasing accuracy, for example allowing phasing of singletons by transmission: singletons cannot otherwise be phased by the phasing algorithms. We estimated that the reference panel contains approximately 63,000 unique genomes (a 20% reduction in sample size).

To identify parent-child relationship for phasing, we combined information from self-reported relatedness, IBD (identity by descent)[11] and Mendel errors. Firstly, 30,000 autosomal variants that meet the following criteria are randomly selected for the analysis: (1) passing the mean genotype quality and depth filter; (2) passing allele balance filter; (3) missingness $< 1\%$; (4) inbreeding coefficient $>$ -0.1; (5) LD-pruned r2 < 0.1 with window size of 500Kb; (6) the Hardy Weinberg equilibrium

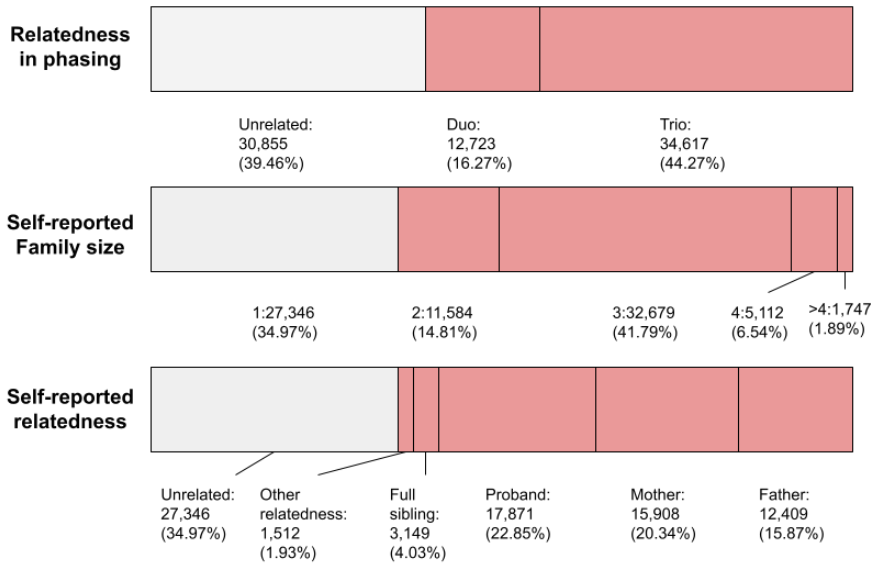


Figure 2.3: Relatedness in 78,195 GEL reference panel samples. The first row shows samples being allocated in duo/trio/unrelated families for phasing and imputation. The second and third row are derived from self-reported data, showing samples being allocated in family with different sizes and relationship within the family. As one might have noticed, some related samples by self-reporting are treated as unrelated in phasing and imputation, due to stringent screening process described in Section 2.2.3

test p-value > 0.01; (7) in the set of 1000 Genomes phase 3 data; (8) excluding high LD sites identified in Price et al. (2008) [117]. We then carried out the following procedure on the selected variants. We selected samples with pairwise IBD0 < 0.1 and IBD1 > 0.7 as potential parent/child pairs inferred by PLINK [11]. For all potential parent/child pairs matching the self-reported relationships, we calculated the Mendel errors, separating duo (parent-child) and trio (mother-father-child) families. The Mendel error cut-offs are Q3+1.5IQR, and Q3+4.5IQR in terms of their mean for trios and duos in order to identify mislabelled and uniparental disomy (UDP) cases. There are 54 potentially UDP cases, often occurring on one chromosome in an affected child-parent pair, yet we still removed all UDP affected pairs and treated them as unrelated.

Furthermore, we marked samples as unrelated when parenthood was inconsistent with the self-reported age, i.e. the parent should be at least 14 years older than

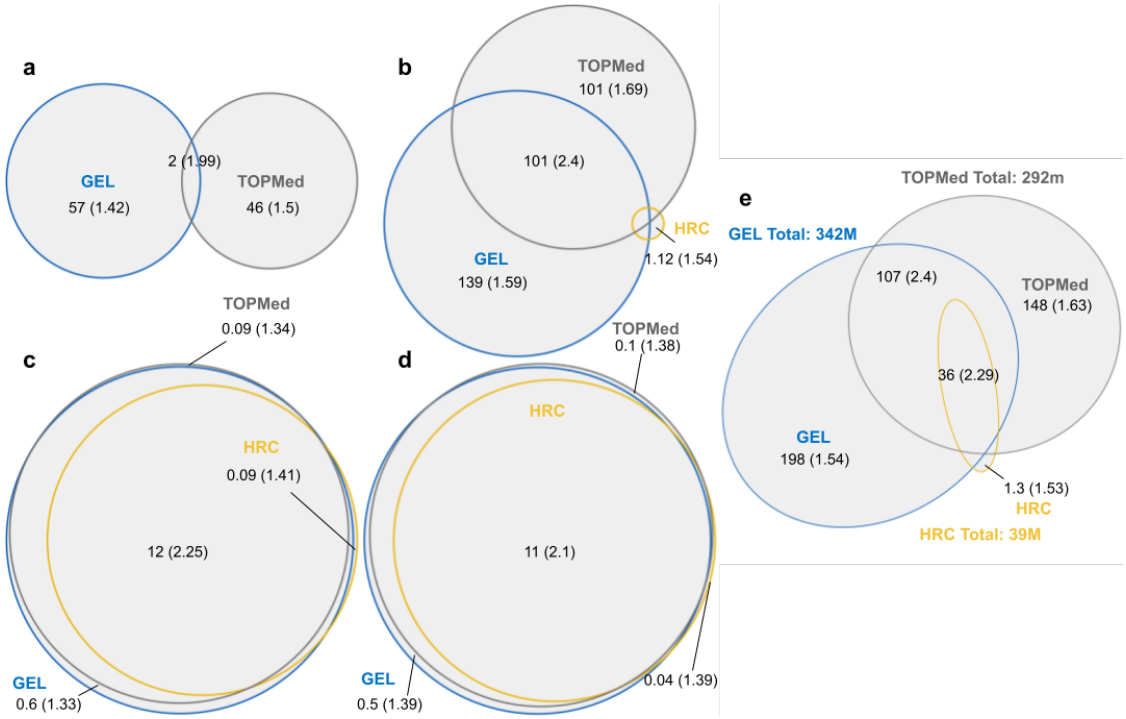


Figure 2.4: Venn diagram comparing variants from GEL, HRC and TOPMed reference panel. Allele frequency for variants existing in more than one reference panel is assigned with the highest allele frequency among all the panels. The Venn diagrams show variants with (a) $AF < 10^{-5}$, (b) $10^{-5} \leq AF < 10^{-4}$, (c) $10^{-4} \leq AF < 10^{-2}$, (d) $10^{-2} \leq AF < 1$, and (e) all variants. The numbers show the variants count of each region (in millions of variants), followed by Ts/Tv ratio of these variants.

the child. Through this procedure, we identified 12,816 (16.39%) samples as in a duo family and 35,106 (44.9%) in a trio family. As such, 30,273 (38.71%) samples were treated as unrelated for phasing. (Figure 2.3)

2.2.4 GEL reference panel with 342 million autosomal variants

The resulting GEL reference panel (see Section 2.3 for the QC detail) consisting of 341,922,205 autosomal variants, among which 31,502,703 (9.26%) are INDELs with an average length of 4.54 and maximum length of 50 bases. The rare variants constitute the bulk of the GEL reference panel, 287.2 million (84.1%) variants are rare with frequency lower than 10^{-4}, among which 66.7 million (19.5%) and 91.1 million (26.7%) are singletons and doubletons, mutations that have only been observed once and twice in the dataset, respectively.

We compared the GEL reference panel to the widely used TOPMed [37] and HRC reference panels [118]. The GEL reference panel has 8 times and 1.1 times more autosomal variants than the HRC v2 and TOPMed r2[1] panels, respectively. Figure 2.4 compares the three datasets overall and in different frequency bins. Limited by the low coverage sequencing technology, HRC has very few ultra-rare variants with allele frequency lower than 10^{-4}. The number of rare variants captured are more comparable between TOPMed and GEL, as both used high coverage sequencing data. Despite the different sequencing depths and sample sizes, all three panels captured a similar set of relatively common variants ($AF > 10^{-4}$), with less than 4% unique to each panel (Figure 2.4 c-d), whereas around half of the rarer variants ($AF \leq 10^{-4}$) from GEL and TOPMed cannot be found in each other (Figure 2.4 a-b).

2.3 GEL reference panel quality control

Since the variants are called sample-by-sample, there are many false calls on the original list of variants. Even though false calls are rare in one genome, the aggregated number of such calls can be large. The QC pipeline has been run in chunks on the aggregated VCF files (aggVCFs), creating a merged QC'd VCF per chromosome for phasing. The following filters are applied to SNPs and INDELs alike. The filters are applied sequentially, reducing the number of variants from 722 million to 341 million. bcftools v1.9 was used to filter and create new VCF files, and the Mendelian error was calculated using PLINK v1.90b6.16.

- **Genotype quality (GQ) + depth (DP)** : Individual genotypes with either GQ < 15 or DP <10 were marked as missing.
- **Missingness**: remove sites that have a missing rate higher than 5%, including the missing genotypes flagged by GQ + DP filter.
- **Allele balance (ABhet)**: allele depths (AD) for REF and ALT are expected not to have a huge discrepancy for each heterozygous individual genotype.

[1]TOPMed r2 site list is not provided to the public. We obtained the site list from the imputation results of the TOPMed imputation server [37]

We first obtained the allele balance for each genotype, i.e. $AD_{REF}/(AD_{REF} + AD_{ALT})$. We then counted the number of sites where $0.25 < ABhet < 0.75$ and marked them as pass. Sites of less than a 75% pass rate were removed.

- **Mendel Errors (MErr)**: No more than 3 Mendelian errors among all duo and trio families for sites with $AF < 0.001$ and 7 Mendelian errors for sites with $AF \geq 0.001$.

- **Hardy-Weinberg equilibrium (HWE)**: Sites where the Hardy-Weinberg Equilibrium (HWE) p-value $< 10^{-5}$ in self-reported White British samples were removed.

- **gnomAD allele frequency (gnomAD)**: We removed sites that showed discrepancy in allele frequency between White British GEL and non-Finnish European gnomAD r3.0 samples[52]. To do this, we used a Fisher's exact test of the allele frequency difference and a p-value threshold of 10^{-10}.

- **Unrelated singletons**: We removed singletons that did not occur in related families, due to no phasing information at such sites.

- **Additional filters** : We chose a set of more lenient filters for those relatively common sites (AF>0.001) overlapping with external datasets (TOPMed[37], HRC[118], 1000 Genomes[116] and GnomAD (non-Finnish European and South Asian samples only)). For these sites, we used the missingness threshold of 25%, Mendel error threshold of 250 per site and gnomAD allele frequency filter p-values of 10^{-20}. All other filters on GQ, DP, ABHet and HWE were kept the same as above.

Before we start delving into the efficacy of the filters, it's important to understand one widely used QC metric, transition vs transversion ratio (Ts/Tv) [84]. Adenine (A) and Guanine (G) are two-ring purine-based nucleotides, while Cytosine (C) and Thymine (T) are one-ring pyrimidine-derived nucleotides. Interchanges within the same type of nucleotide are called transitions, i.e. C/T and G/A mutations,

while interchanges between types are called transversion, which involves A/C, A/T, C/G and G/T mutations. Transversions are less likely to happen in the human genome than transition. In addition, Cytosine (C), especially when adjacent to a Guanine (G), can be easily methylated, turning C into uracil (U), and consequently causing C to T mutations. Different regions in the genome have different Ts/Tv ratios. For example intergenic regions usually have lower Ts/Tv ratios than the genome-wide average, while exomes have higher Ts/Tv than the average [119]. The genome-wide Ts/Tv ratio is thought to be around 2.1 [84]. We examined the TOPMed r2 and HRC reference panel, finding that for sites with allele frequency $> 10^{-5}$, the Ts/Tv ratios were all around 2.1 (Figure 2.4). However, rarer variants in TOPMed have a much lower ratio of 1.6, and a similar ratio can be observed in the GEL dataset. Lower sequencing quality of rare variants could be the reason causing large sequencing datasets to have low Ts/Tv ratios. It is also possible that the Ts/Tv ratio is pushed down due to the lack of CpG variants, i.e. C to T mutation (transition), in rare variants. In this study, we will use the 2.1 overall Ts/Tv ratio only as a guideline of the whole genome Ts/Tv ratio.

The **GQ+DP** filter and **missingness** filter combined were responsible for 32% of the site removal from the raw data (Table 2.2), with the removed variant Ts/Tv ratio to be 0.54, improving the overall Ts/Tv ratio from 1.28 to 1.84 (Figure 2.5). Though the rest of the filters did not change the overall Ts/Tv ratio, they are still crucial to the detection of low quality sites at a different allele frequency. The **MErr** filter with over 47 thousands duo/trio samples is extremely powerful at detecting problematic variants with AF $> 10^{-3}$. However, **MErr** cannot capture errors when the variants are too rare, due to lack of mutations observed within a family. In addition, the **MErr** filter is also not effective to exclude common variants with excessive heterozygosity or homozygosity. The following two filters are orthogonal to **MErr**, covering the area within the allele frequency spectrum that are not covered by **MErr**. The allele balance filter (**ABHet**) is responsible for the ultra-rare side of the spectrum (AF $< 10^{-3}$), removing 30 million rare variants, 4% of the raw data, with the removed sites Ts/Tv ratio to be around 1.0. (Figure 2.2).

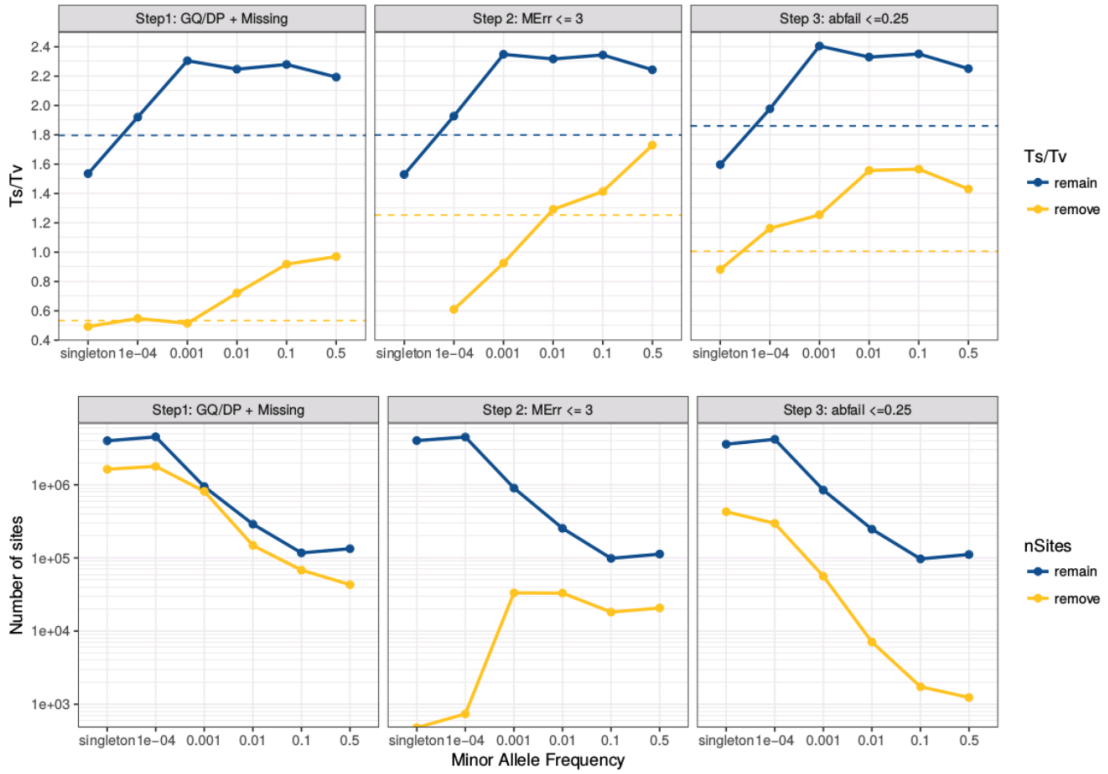


Figure 2.5: The effect of the quality control filters applied to chromosome 20 in terms of the number of removed and remaining sites (First row) and the Ts/Tv (Second row), stratified by minor allele frequencies. The filters GQ/DP + missingness, Mendel (MErr) and allele balance (ABhet) are sequentially applied to the raw data. The dotted lines are the overall Ts/Tv ratio of remaining (blue) and removing (yellow) variants of the filter.

The Hardy Weinberg filter (**HWE**) is mainly set to detect excessive homozygote or heterozygote issues, removing a further 189,000 variants.

The variants are expected to be roughly evenly distributed along the chromosomes. We often observe an increased number of variants in regions that are difficult to sequence, such as centromeres or telomeres. In Figure 2.6, we plotted the number of variants in chromosome 20 for raw aggregated VCF (aggVCF) and the data after QC (GEL). Excessive variants in aggVCF around centromeres and telomeres were removed, resulting in more evenly distributed variant count along the chromosomes.

2.3.1 Lenient filters for common variants

By cross-comparing the common variants to other datasets (Table 2.3), we noticed that 4.1%, 7.7%, 7.4%, and 17% of common variants (AF $>$ 0.1) in 1000 Genomes,

	Number of SNPs left after applying the filter (removed %)	**Number of Indels / SVs left after applying the filter (removed %)**	**Total number of variants after applying the filter (removed %)**
Raw	630,967,910	91,374,497	722,342,407
GQ/DP + missingness	428,701,462 (-32%)	55,702,335 (-39%)	484,403,797 (-32%)
ABhet	411,285,423 (-3%)	42,963,226 (-14%)	454,248,649 (-4%)
Mendelian errors	410,854,761 (-0.07%)	41,905,560 (-1%)	452,760,321 (-0.2%)
HWE	410,764,722 (-0.01%)	41,868,797 (-0.04%)	452,633,515 (-0.01%)
gnomaAD	410,628,878 (-0.02%)	41,815,306 (-0.05%)	452,444,184 (-0.02%)
Singleton	309,825,243 (-16%)	31,639,011 (-11%)	341,464,254 (-15%)
Lenient filters	310,844,262 (+0.16%)	31,716,292 (+0.08%)	342,560,554 (+0.15%)

Table 2.2: : Variant filtering. The table shows the effect of each filter applied sequentially from top to bottom. The number of variants (SNPs, Indels/SVs and Total variants) and the percentage removed is shown in each row.

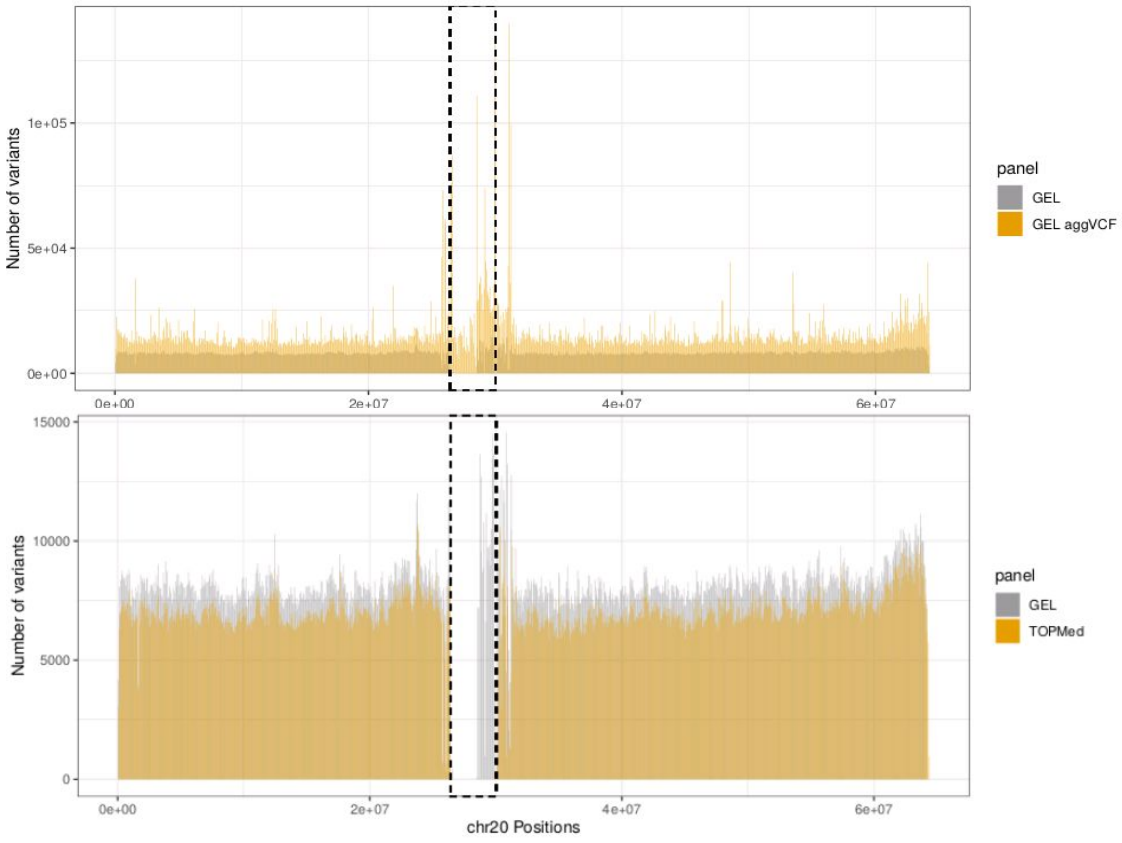


Figure 2.6: Number of site along chromosome 20. The top plot compares the site count before QC (GEL aggVCF) and after QC (GEL). The bottom plot compares GEL reference panel (after QC) with TOPMed. The black dashed box signifies the centromere.

	1000G (%)	HRC (%)	TOPMed (%)	UKB (%)
Not in aggVCF	0.073	0.039	0.049	1.98
GQ/DP	3.5	1.27	3.12	7.99
AB	0.028	0.018	0.042	0.079
MErr	2.3	2.07	2.82	4.54
HWE	0.5	0.13	0.44	0.1
GnomAD	0.22	0.16	0.31	0.41
Multiple filters	0.85	0.36	0.56	1.69
Total missing	7.47	4.04	7.34	16.79

Table 2.3: The effect of standard filters on the common variants (AF>0.01). This table shows the percentage of 1000G, HRC, TOPMed and UKB common variants in chromosome 20 that cannot be found in the standard filter GEL dataset, breaking down by the filters that are responsible for the removal of the variants. UKB is the UK Biobank data imputed by HRCUK10K reference panel [1].

	Not overlapping with				
	GEL	TOPMed	HRC	UK10K	UKB
GEL (9,284,802)	0	482,255	1,497,927	853,110	403,068
TOPMed (12,830,650)	197,955	0	1,628,650	2,663,743	4,538,480
HRC (8,021,801)	85,462	96,365	0	244,174	373,528
UK10K (9,769,637)	1,170,315	1,428,019	2,099,068	0	1,255,342
UKB (9,884,806)	919,686	1,121,825	2,079,515	1,203,865	0

Table 2.4: Common variants (AF>0.01) overlapping situation across reference panels, showing the number of common variants unique to row datasets that are not present in the column datasets. For instance, the second row and second column shows the number of TOPMed common variants that cannot be found in the entire GEL dataset is 197,955. The numbers in parentheses show the number of common variants in each dataset. The GEL variant count in this table include the sites added by the lenient filters. UKB is the imputed UK Biobank data by HRCUK10K.

HRC, TOPMed, and HRCUK10K, respectively, are not present in the GEL data with the standard filters. The removal of the common variants in the GEL standard filter set is largely explained by the GQ/DP and Mendel error filters. We relaxed the filters by increasing the missingness threshold from 5% to 25%, the **MErr** threshold from 7 per site to 250 per site, and in addition, decreased the GnomAD allele frequency concordance (**GnomAD**) p-value to 10^{-20}. These filters are applied on a set of GEL common variants (in raw aggVCF) that have AF > 0.001 in either HRC, TOPMed, 1000G, or the GnomAD dataset. The lenient filters introduced 1,096,300 additional sites which were phased and imputed separately (See Section

2.4). The lenient filter sites are marked in the GEL reference panel and UKB imputed data to remind the users to take caution on these sites.

Common variants with AF > 0.01 constitutes the bulk of GWAS findings (see Chapter 4). The final GEL reference panel contains 9.2 million common variants (AF>0.01), comparing to 12.8, 9.7, and 9.8 million in TOPMed, HRC, UK10K and UKB datasets (Table 2.4). Over 1 million UK10K common variants (more than 10% of total common variants call in UK10K) are not found in GEL and TOPMed. Given that the UK10K is a much smaller reference panel in terms of sample size, and lower coverage (7×) [120] compared to GEL and TOPMed, this indicates an inflation of the number of common variant calls in the UK10K reference panel, which subsequently affected the imputation of UKB. In Chapter 4, we will discuss the implication of these inflated common variants to the GWAS results in detail.

Table 2.4 also showed that GEL has over 3 million less common variants than TOPMed, yet almost all these excessive TOPMed common variants can be found in GEL. Similarly, TOPMed has 4.8 million more common variants than HRC, among which 3 million variants were present in HRC with a lower allele frequency. GEL and HRC consist of predominantly samples with European ancestry, while nearly half of the samples in TOPMed either have African or South American ancestry. The difference in the sample ethnicity can explain the shift of allele frequencies.

2.4 Haplotype phasing

2.4.1 Phasing pipeline for autosomes

Haplotype phasing was carried out using SHAPEIT4.2.2 [69]. We used a multi-stage strategy to leverage the relatedness within the GEL dataset.

In the first stage, makeScaffold[2] was used to determine the phase of samples in a duo and trio family by transmission (see Section 1.2.4). The vast majority of genotypes can be phased using this process, with a few genotypes whose phase is ambiguous due to heterozygosity or missingness patterns. These genotypes were phased using SHAPEIT4.2.2.

[2]https://github.com/odelaneau/makeScaffold

To phase the unrelated samples, we first created a phased scaffold of common variants, with the remaining variants phased onto this scaffold. To create the scaffold, common variants with the minor allele frequency > 0.01 are phased using the phased related samples as the reference panel.

The remaining rarer variants were then phased onto the scaffold in chunks, containing approximately 300,000 sites with 30,000 sites on each side as buffers. The phased duo/trio data was used as a reference panel in this step. The chunks and are merged and concatenated using bcftools [121]. Concatenation of the phased chunks is possible as each set of variants have been phased onto the scaffold. The phasing step was computationally intensive and took about 6500 CPU days to accomplish.

The sites identified using the additional lenient filters (Section 2.3) were phased separately. The additional variants were initially selected from a common variant set (AF > 0.001), variants appeared in either 1000G, HRC, TOPMed, or GnomAD datasets. The aggVCF variants passing the lenient filters and are in common with the common variant set were chosen as the additional common variants. To ensure phase consistency to the initial phased data, the additional common variants were phased independently, using the initial phased data at UKB Axiom array sites as scaffold. The phased reference panel and additional common variant panel are available on the GEL server.

2.5 GEL reference panel phasing accuracy

Phasing accuracy could affect the quality of imputation and other downstream applications. We phased the 1,178 parents from 589 mother-father-child trio families in the 1000 Genomes Project high coverage sequencing data [38], using HRC and GEL reference panel. We assessed the phase accuracy by comparing the phase estimate to the phase derived by transmissions. We measured the phasing accuracy using switch error rate, which is the ratio of the number of possible switches required to obtain the true haplotype phase and the number of heterozygotes minus one [65].

The HRC reference panel was lifted over from the GRCh37 to the GRCh38 reference genome using GATK Picard LiftoverVCF [84]. The original GRCh37

HRC reference panel has 39,131,578 autosomal variants. 13,813 variants were removed either due to the incompatibility between reference genomes or mismatching chromosome between the two reference genomes. The resulting autosomal GRCh38 HRC reference panel contains 39,115,765 variants and 27,165 samples. To make a fair comparison of phasing and imputation performance, the 1000 Genomes samples were removed from the HRC reference panel.

The site quality control was carried out on 2,405 high coverage sequencing 1000 Genomes samples. It is based on the missingness and Hardy Weinberg equilibrium test (HWE). The missingness threshold is 5% and the HWE P-value threshold is 10^{-10} on 26 1000 Genomes populations[3]. This QC'd 1000 Genomes dataset will also be used in the imputation experiment in the coming section.

The phasing experiment was carried out using SHAPEIT 4.2.2 [69]. Parents from the trio families are phased using the reference panels. We tested 1.04×10^9 (1.76 million per trio family) and 1.16×10^9 (1.9 million per trio family) heterozygotes in HRC and GEL phased samples, respectively.

Figure 2.7 shows the switch error of samples, phased by GEL vs HRC. The GEL phased haplotypes obtained a lower switch error rate than HRC phased haplotypes for all samples with CEU (Northern European from Utah), African, South Asian and East Asian ancestry. The mean GEL phased haplotype switch error rate is 0.18%, 0.33%, 0.31% and 0.73% for European, African, South Asian and East Asian samples, in comparison to 0.22%, 0.43%, 0.31%, 1.07% using the HRC reference panel. The population structure of the reference panel is a key factor in determining the phasing

[3]Populations in 1000 Genomes high coverage sequencing data: Han Chinese South (CHS), Finnish in Finland (FIN), Kinh in Ho Chi Minh City, Vietnam (KHV), African Caribbean in Barbados (ACB), Puerto Rican in Puerto Rico (PUR), Bengali in Bangladesh (BEB), African Ancestry in Southwest US (ASW), Gambian in Western Division, The Gambia - Mandinka (GWD), Mende in Sierra Leone (MSL), Han Chinese in Beijing, China (CHB), Esan in Nigeria (ESN), Japanese in Tokyo, Japan (JPT), Toscani in Italy (TSI), Iberian populations in Spain (IBS), Colombian in Medellin, Colombia (CLM), Peruvian in Lima, Peru (PEL), Mexican Ancestry in Los Angeles, California (MXL), Yoruba in Ibadan, Nigeria (YRI), Punjabi in Lahore, Pakistan (PJL), Luhya in Webuye, Kenya (LWK), Utah residents (CEPH) with Northern and Western European ancestry (CEU), Chinese Dai in Xishuangbanna, China (CDX), Gujarati Indians in Houston, TX (GIH), British in England and Scotland (GBR), Indian Telugu in the UK (ITU), and Sri Lankan Tamil in the UK (STU).

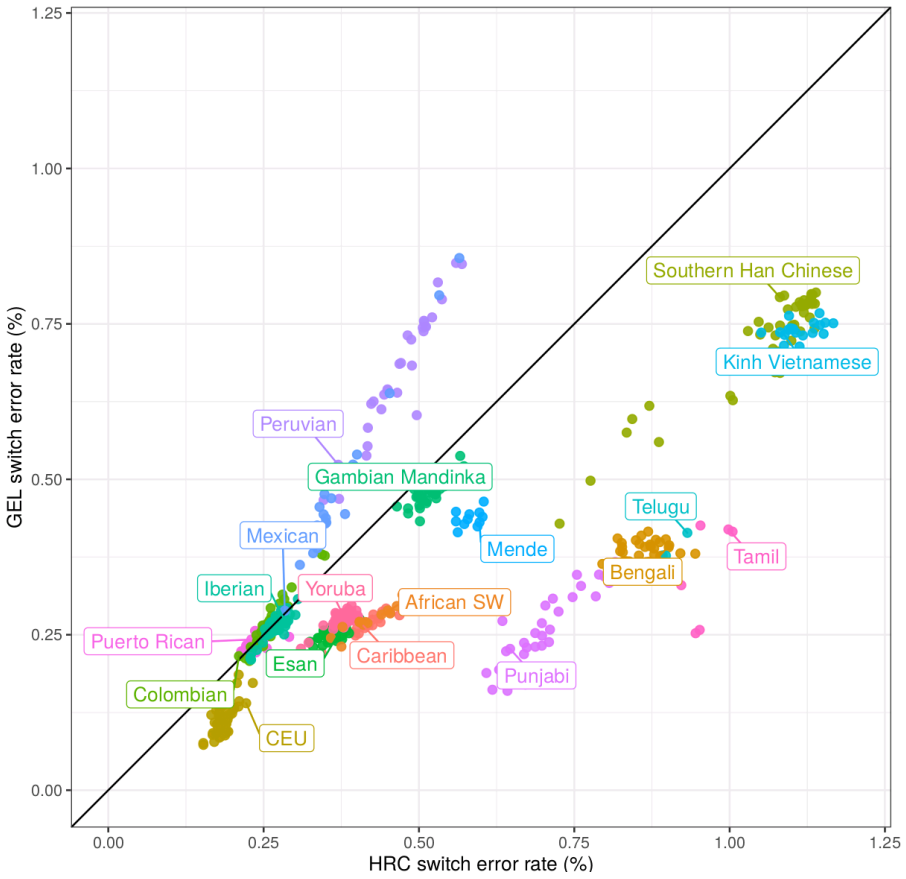


Figure 2.7: 1000 Genomes high coverage sequencing trio data phasing switch error rate comparing haplotypes inferred by trio transmission and computationally inferred using reference panels.

performance [65]. Due to the absence of the South American samples in GEL, HRC outperforms GEL when phasing Peruvians, Mexicans, and Puerto Ricans.

2.6 GEL reference panel imputation accuracy

We assessed the utility of the GEL reference panel for imputation, compared to the TOPMed and HRC reference panels. The aim is to assess the imputation accuracy of these three panels on 2,405 highly diverse 1000 Genomes samples [38]. The data went through the same quality control procedure as in the phasing experiment (see Section 2.5).

SNP positions in the UK Biobank Axiom array [1] were used to create a pseudo-SNP array dataset, masking genotypes in 1000 Genomes sequencing samples, except the sites existing in the Axiom array, resulting in 716,473 bi-allelic SNPs across

all autosomes. The pseudo-SNP array dataset was then phased one chromosome at a time using SHAPEIT4.1.2 without a reference panel [69]. The liftOver HRC reference panel used here is the same as that used in Section 2.5. TOPMed imputation was carried out using the TOPMed imputation server with TOPMed r2 reference panel[37] with the imputation software minimac 1.5.7 [122]. IMPUTE5 [74] was used to impute from the GEL and HRC panels. We converted the reference panels to the IMP5 format to facilitate efficient imputation [74].

The imputed genotypes from the GEL, TOPMed and HRC reference panels were then compared to the sequencing data (QC'd) as the ground truth, stratified by allele frequency. Squared correlation r^2 between the imputed allele dosages D and the 1000 Genomes sequencing data true genotype G were calculated, stratified by gnomAD (v3.3.1) minor allele frequency. For numerical stability, we used an online algorithm to calculate the variance and covariance [74].

$$r^2 = \frac{\mathrm{Cov}(D, G)}{\mathrm{Var}(D)\mathrm{Var}(G)} \tag{2.1}$$

As we focus on showing the overall performance of the reference panel across different allele frequencies, only variants that are in common with the gnomAD variants are taken into account. As a result, the number of variants measured may differ across reference panels. We also stratified the imputation results from the 2,405 1000G samples into 6 groups : 661 African (AFR), 347 American (AMR), 504 Eastern Asian (EAS), 489 South Asian (SAS), 313 non-Finnish European (NFE) samples and 91 British (GBR) samples.

Figure 2.8 shows the imputation results. The GEL imputation accuracy outperforms HRC panel in all allele frequency bins for all ethnicities. The GEL panel out-performs the TOPMed panel in GBR and SAS samples, especially for rarer variants. This is likely due to the GEL panel having a larger number of individuals with British and South Asian ancestry than the TOPMed panel. At MAF $< 10^{-5}$, the GEL imputation performance (r^2) for GBR samples is 0.6, compared to 0.3 and 0.29 using TOPMed and HRC, respectively. At MAF $< 2 \times 10^{-4}$, the r^2 are 0.75, 0.64, and 0.48 for GEL, TOPMed and HRC, respectively. These results suggest

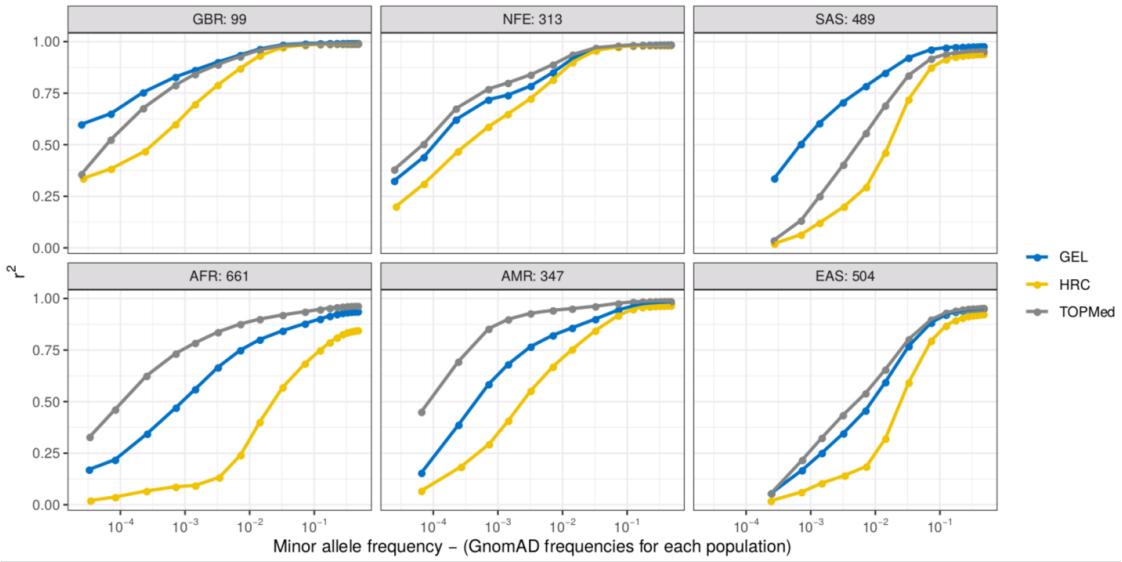


Figure 2.8: The x-axis shows non-reference allele frequency on a log scale, focusing in on rarer variants. The y-axis is imputation performance (r^2). The performance of the reference panels HRC (yellow), TOPMed (grey), GEL (blue) are shown as lines in each plot. The variants are stratified by GnomAD allele frequency (v3.3.1) of their corresponding population.

that the GEL reference panel is a well-matched reference panel to impute the UK Biobank samples, of which over 80% are White British or Irish, and for which South Asian is the second-largest ethnic group. The TOPMed panel out-performs GEL and HRC in African, American and marginally in East Asian samples.

2.7 Discussion

In this chapter, we built the GEL reference panel with 78,195 sample and 342 million autosomal variants. The GEL reference panel has a diverse population, with more than 49,641 White British and 5,283 South Asian individuals. We also identified the relatedness in the dataset, where 16% of samples are in a duo family and 44% in a trio family, leaving only 39% of samples unrelated. The GEL reference panel is 8.7 times more samples than the HRC and 1.1 times more samples than the TOPMed r2 reference panel. Nearly 200 million rare variants are not seen in either HRC or TOPMed (Figure 2.4). The relatedness of the GEL dataset allowed phasing of singletons by transmission. It resulted in GEL possessing slightly more variants than TOPMed, despite TOPMed having 24% more samples than GEL.

A major challenge of building the reference panel is to ensure good quality variant calls. We introduced the site level quality control filters which removed nearly half of the variants included in the raw aggVCF data, whilst covering the entire allele frequency spectrum. The Ts/Tv ratio was improved from 1.28 to 1.84. We then cross-examined the common variants captured by GEL to other popular reference panels: HRC, TOPMed, and UK10K. GEL contains over 1 million more common variants than HRC and a similar common variant count to UK10K. However, based on our findings, we have strong suspicion that UK10K has inflated their common variant count, which would cause the potential false positive GWAS results in the UKB imputed by HRC and UK10K panels. More analysis of this issue will be discussed in Section 4.2.3.

The reference panel size and matching population are important factors affecting the phasing and imputation accuracy. In the trio phasing experiment, the GEL reference panel outperformed HRC for all 1000 Genomes populations, except for the South American populations. By imputing the 1000 Genomes samples on the UK Biobank Axiom array, the White British imputation achieved $r^2 = 0.75$ at allele frequency of 2×10^{-4}.

The reference panel is made accessible for all GEL registered researchers. We plan in future to work on chromosome X phasing, once the aggregate VCF files are ready.

3

Imputing UK Biobank samples using the GEL reference panel

Contents

3.1 Overview 47
3.2 UK Biobank imputation pipeline 48
3.2.1 UK Biobank SNP array data quality control and phasing 48
3.2.2 UK Biobank imputation using GEL reference panel . . 48
3.3 GEL imputed UK Biobank data 49
3.3.1 Imputation quality 49
3.3.2 GEL Imputed UKB dataset 49
3.4 Discussion 52

3.1 Overview

In this chapter, we will describe the imputation of 488,315 UKB samples using the GEL reference panel. We will examine the imputation quality in terms of INFO score, and compare it to the previous UKB imputation, imputed by the HRC and UK10K reference panel. The imputed data are now available on the UKB platform under the field ID 21008.

3.2 UK Biobank imputation pipeline

3.2.1 UK Biobank SNP array data quality control and phasing

The UK Biobank SNP array data consist of 784,256 autosomal variants. We removed the set of 113,515 sites identified by the previous centralized UK Biobank as failing quality control[1], and also removed an additional set of 39,165 sites failing a test of the Hardy-Weinberg equilibrium in 409,703 White British samples, with the p-value threshold of 10^{-10}. The SNP array data were mapped using the LiftOver tool, GATK Picard LiftOver [84], from GRCh37 to GRCh38. Alleles with mismatching strand were flipped wherever possible. 495 variants were removed due to incompatibility between the two reference genomes, resulting in a final SNP array dataset with 631,081 autosomal variants, to be phased and imputed at GEL sites.

Haplotype estimation of the SNP array data is a prerequisite for imputation. Phasing was carried out one chromosome at a time using SHAPEIT4.2.2 without a reference panel. We ran SHAPEIT4 using its default 15 MCMC iterations and 30 threads. The runtime varied from 2 hours to 30 hours for each chromosome.

3.2.2 UK Biobank imputation using GEL reference panel

The imputation of normal filter set and lenient filter set were carried out independently to ensure the lenient filter phasing doesn't alter the result of the normal filter set. The autosomal imputation by GEL reference panel was carried out using IMPUTE5 (v1.1.4). The SNP array data were divided into 408 consecutive and overlapping chunks across the genome, and each chunk was further divided into 24 sample batches, with each batch containing 20,349 samples. IMPUTE5 was run on each of the 9,792 subsets using a single thread and default settings, on average below 4 minutes per genome, totalling about 1200 CPU days to impute all UK Biobank samples. Sample batches were merged using QCTOOL and the non-overlapping chunks were concatenated using cat-bgen [123].

3.3 GEL imputed UK Biobank data

3.3.1 Imputation quality

IMPUTE information (INFO) estimates the genotype imputation uncertainty via a score ranging from 0 to 1. A high INFO implies higher imputation quality, with a value of 1 indicating no uncertainty, and 0 indicating complete uncertainty about the genotype imputation. If an imputed variant on N samples has an INFO score α, it implies that the statistical power of association tests are approximately equivalent to αN perfectly observed genotype data. There is no single correct answer for which threshold to use. To perform GWAS on the UK Biobank data with 500,000 samples, it is typical to use the variants with INFO above 0.3, equivalent to >150,000 perfectly observed samples [1] (see Section 1.2.2 for the definition of INFO).

We compared GEL imputed UK Biobank INFO scores to those from the existing HRCUK10K imputed dataset[1]. The proportion of GEL imputed variants passing the INFO threshold of 0.3 are 8%, 78% and 98% for MAF $\leq 10^{-6}$, $10^{-6} <$ MAF $<= 10^{-5}$, and $10^{-5} <$ MAF $\leq 10^{-4}$, compared to 4%, 54%, and 78% for the HRC+UK10K imputed data (Figure 3.1). Among the 65 million variants imputed by both GEL and HRC+UK10K panels, 87% achieved better INFO score using the GEL panel (Figure 3.2). Furthermore, we also found there to be more imputation INFO improvement in INDELs than SNPs and more improvement in common variants (MAF>0.001) than less common variants. (Figure 3.3).

3.3.2 GEL Imputed UKB dataset

The GEL imputed UK Biobank data consists of 488,315 samples and 342,573,817 autosomal variants, which is 3 times more than its predecessor. Using effect prediction tool VEP (release 105) [124] to annotate the variants, we found 5 times more high impact consequence variants (Table 3.1). It is worth pointing out that GEL has 10-20 times more frameshift variants, due to better imputation quality of INDELs (Figure 3.3).

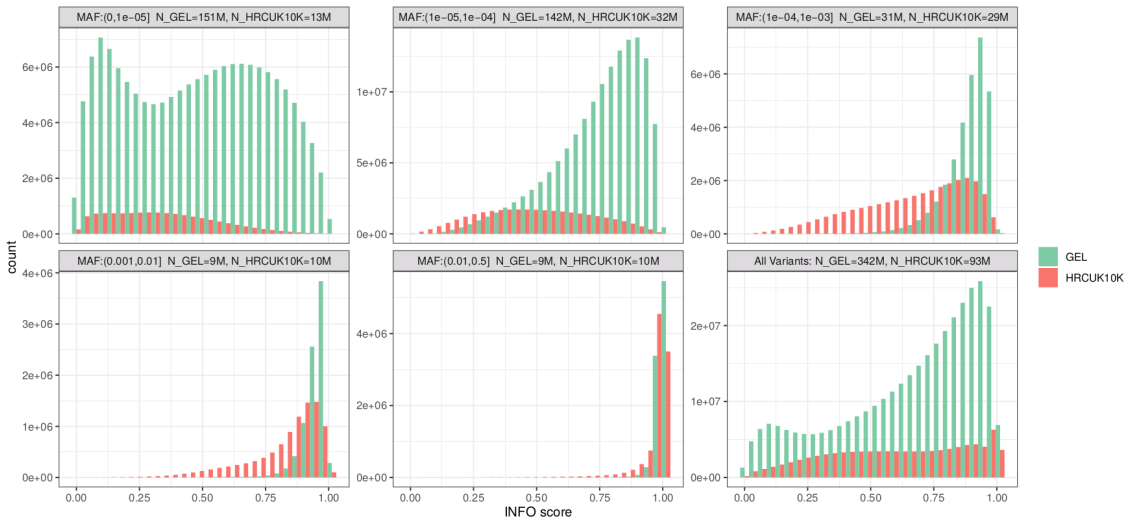


Figure 3.1: Imputation INFO score histogram comparison between GEL and HRCUK10K imputed UKB data. Each panel shows the distribution of INFO scores for GEL and HRCUK10K imputed variants in different MAF bins. The total number of variants in each bin is provided in the panel legend.

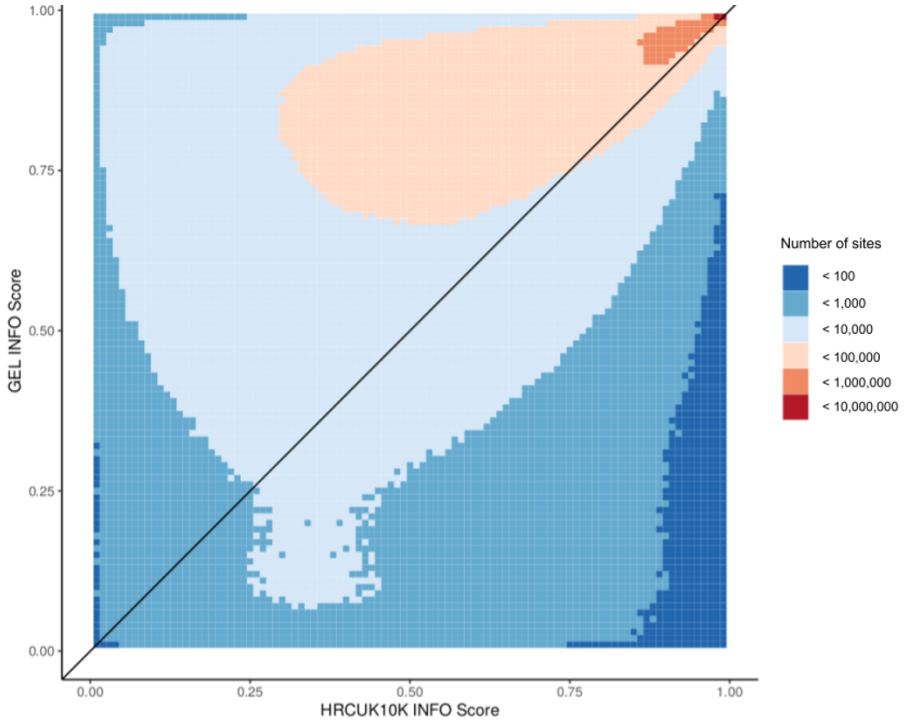


Figure 3.2: Comparison of INFO scores at sites that are in both GEL and HRCUK10K reference panels. A heatmap shows the number of sites in INFO score categories. In total, 65 million sites in both GEL and HRC+UK10K panels are examined.

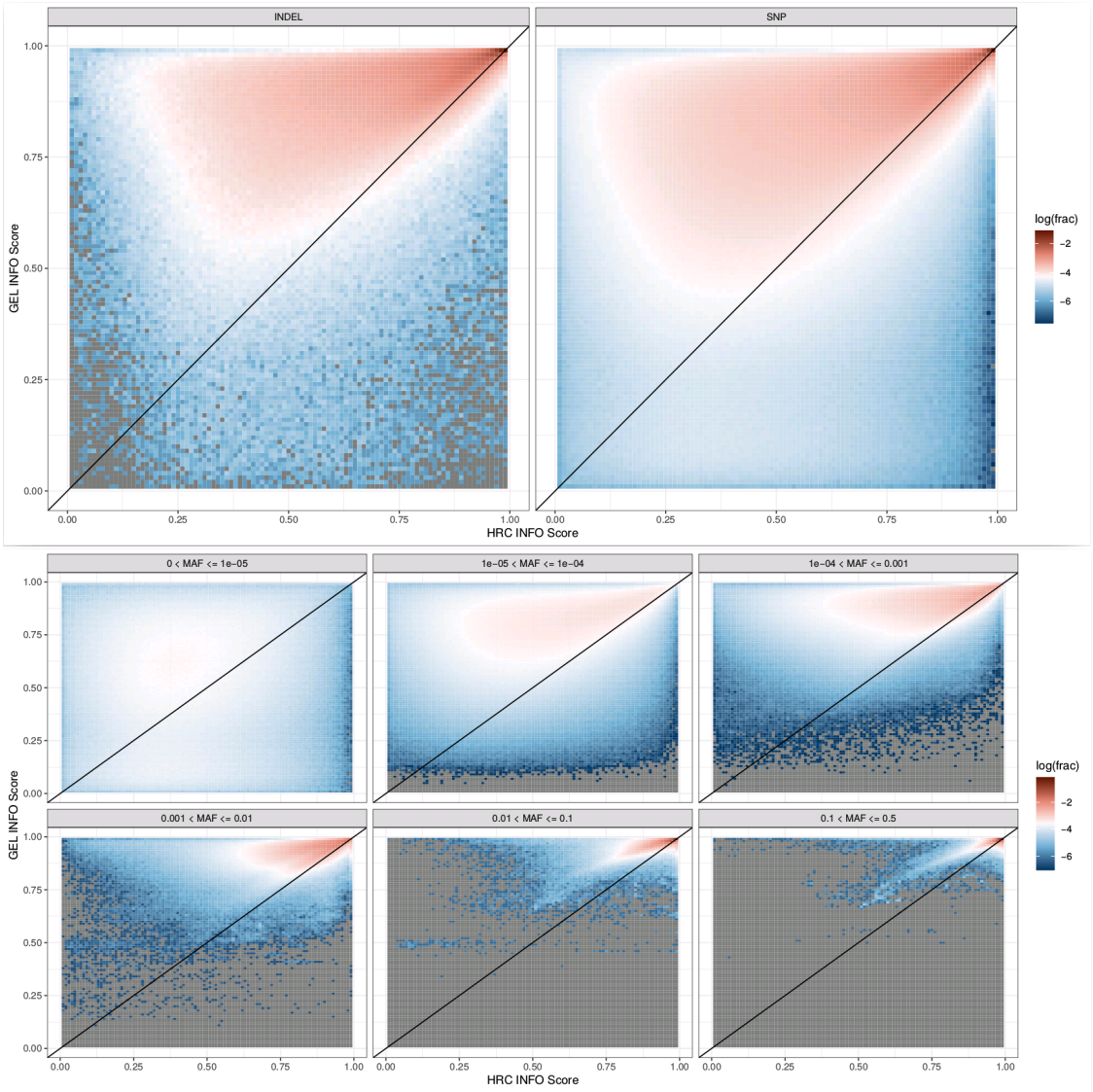


Figure 3.3: INFO score comparison of the common sites imputed by GEL and HRCUK10K, separating in SNPs and INDELs (above) and minor allele frequency bins (below). The pixel colour indicate the fraction of variant count in terms of the total variant count. The fraction is in a log scale.

The imputed data are stored in 22 8-bit zstd compressed BGEN files [123]. Variants are aligned to the GRCh38 reference genome. Each chromosome is stored in a separate BGEN file, with the filesize ranging from 19Gb to 103Gb, and totalling 1.2Tb. All variants are assigned with a unique ID, either using the rsid from dbSNP build 155 [1] wherever possible, or in the form of chr:pos_ref_alt, when the rsid is unavailable for the variant. Imputed genotypes are stored as genotype posterior

[1] https://ftp.ncbi.nih.gov/snp/latest_release/

High impact consequences	HRCUK10K (A)	GEL (A)	HRCUK10K (B)	GEL (B)
Splice acceptor	7,593	41,532	5,637	25,777
Splice donor	10,730	55,730	8,221	35,493
Stop gained	16,525	78,227	11,910	50,691
Frameshift	5,536	102,695	5,000	62,333
Stop lost	1,510	6,900	1,173	4,366
Start lost	2,406	10,541	1,803	6,730
Total	44,300	295,625	33,744	185,390
Missense	704,682	2,485,273	526,700	1,673,850

Table 3.1: The number of variant breaking down by high impact consequences predicted by VEP[124] (release 105). (A) indicates all variants in the imputed UKB data. (B) is a subset of GEL imputed UKB data with INFO > 0.3 and minor allele count > 5, which are common thresholds to be set for GWAS.

probabilities, consisting of three 8-bit floating numbers, representing the probability of the genotype being homozygote reference, heterozygote or homozygote alternate, respectively. Despite having more than 3 times the number of variants, the GEL imputed UKB BGEN files are nearly half of the size to its predecessor, HRCUK10K imputed BGEN files (2.2Tb). This is mainly due to the storage method, with the genotype dosages stored in 8-bit instead of 16-bit. The compression of the rare variants is highly efficient using the BGEN format.

The imputed data are made available on the UKB platform under the field ID 21008. IMPUTE5 allows output of the imputation data with the phase information. We stored the phase genotypes in VCF format. Due to the filesize, we only stored the rounded dosage as the phased genotypes. The phased imputed UKB data will be uploaded to the UKB platform soon.

3.4 Discussion

There are several existing UK Biobank imputation, including the imputation using the reference panel of HRCUK10K [1], TOPMed [37], and 200K UK Biobank [54]. The UK Biobank exome data has also been imputed using the 200K UK Biobank WES data [125]. The GEL imputed UK Biobank is an addition to the the imputation data resources.

The GEL imputed UKB data have 3 times more variants than HRCUK10K imputed UKB (H-UKB). In addition, they have 5 times more high consequence variants than HRCUK10K, with a very large 10-20 fold increase in frameshift variants, indicating a significant improvement in calling of INDELs (Figure 3.3). Comparing the variants in common between both imputed datasets, 87% of GEL imputed variants achieved better INFO score than HRCUK10K. The better common variant (MAF>0.001) imputation quality in Figure 3.3 also echoes the findings in Chapter 4. Well imputed rare as well as common variants are vital for downstream analyses.

Due to the delay of the chromosome X aggregated VCF files by GEL, we have been unable to produce the phased chromosome X phased data as yet. However, there are plans to work on the phasing of chromosome X as soon as the aggregated VCFs are ready.

4

UK Biobank imputed data genome-wide association studies and analysis

Contents

4.1 Overview 55
4.2 Genome-wide association studies 56
4.2.1 The genome-wide association studies for four quantitative traits 56
4.2.2 Rare variant associations 59
4.2.3 Common variant associations 63
4.3 Fine mapping rare variant associations 68
4.3.1 Conditional joint analysis 68
4.3.2 Results 69
4.4 Fine-mapping common variant associations 72
4.4.1 Bayesian fine-mapping 72
4.4.2 Results 74
4.5 Discussion 75

4.1 Overview

GEL imputed UK Biobank data provide a unique opportunity for the discovery of novel genetic associations. We performed exemplar genome-wide association studies (GWAS) on four quantitative UK Biobank traits, namely standing height, body mass index (BMI), systolic blood pressure (SBP) and diastolic blood pressure (DBP).

Compared to the HRCUK10K imputed UK Biobank data (H-UKB), the GEL imputed UK Biobank (G-UKB) data found 1,699 potential more rare associations (AF $<$ 0.05), including 38 rare functional associations. The GWAS using G-UKB discovered more than 70% of the EWAS findings [53], compared to 56% found by H-UKB.

Section 4.2.1 describes details of the GWAS pipeline. The rare variant GWAS results will be discussed in Section 4.2.2, focusing on the factors that affect the power of GWAS, including effect size, allele frequency, and sample size. In Section 4.2.3, we will explain why G-UKB found fewer common variant associations, presenting evidence that the H-UKB common variant calls are less reliable. G-UKB data also benefits other downstream analyses, such as conditional joint analysis (Section 4.3.2) and Bayesian fine-mapping (Section 4.4). We also list ten putative causal rare association findings by the conditional joint analysis.

4.2 Genome-wide association studies

4.2.1 The genome-wide association studies for four quantitative traits

We selected four quantitative traits to demonstrate the GWAS performance of difference between G-UKB and H-UKB. These traits are standing height (HEIGHT), body mass index (BMI), systolic (SBP) and diastolic (DBP) blood pressure. Samples from 40 to 70 years old are included. For all traits, outliers that are beyond ± 4 standard deviations to the mean value are removed.

SBP and DBP are based on the automated blood pressure readings. We calculated mean SBP and DBP values from two automated blood pressure readings (n = 418,755) or two manual readings (n = 25,888), when the automated readings are not available. For individuals with one manual and one automated blood pressure measurement (n = 13,521), the mean value is calculated between these two values. For individuals with only one available blood pressure reading (n = 413), we used this singular value. The mean blood pressure readings are then adjusted

based on the pressure-lowering medication (n=94,289). Adding 15 and 10 mmHg to SBP and DBP, respectively, if pressure-lowering medication is taken [126].

Variants with minor allele count lower than 5 are not included in the association tests. The trait measures were transformed using rank inverse normal transformation (RINT) to ensure the input phenotype to be normally distributed, in order to increase the power of GWAS and reduce false positives at rare variant due to outliers [127]. Suppose the number of sample in a GWAS is n,e RINT transformation of i^{th} sample's phenotype X_i is as follows,

$$\text{RINT}(X_i) = \Phi^{-1}\left(\frac{\text{rank}(X_i) - c}{n + 1 - 2c}\right) \tag{4.1}$$

with the constant $c = 3/8$ as the Blom offset. $\text{RINT}(\mathbf{X}) \sim \mathcal{N}(0, 1)$ [127].

The genome-wide association studies were carried out using REGENIE [14], which estimates a shrinkage parameter for each consecutive block of 1000 SNPs from the array data in a ridge regression (step 1), which can be treated as a prediction of local polygenic scores from that block. In step 2, REGENIE uses the step 1 prediction to perform a linear regression on quantitative traits using the "leave-one-chromosome-out" (LOCO) scheme [12, 14]. We used the UKB SNP array data for REGENIE step 1 and imputed data for step 2, accounting for sex, age, sex squared, sex $\times$ age, and 20 principal components as the covariates [128]. Tests are carried out on 429,460 White British samples. The association tests for G-UKB and H-UKB used the identical setup to facilitate comparison. H-UKB is called based on the reference genome GRCh37. In order to compare the results between G-UKB and H-UKB, the summary statistics of H-UKB are lifted over from GRCh37 to GRCh38, using Picard LiftOver [84].

In the following analysis, we split the findings into two categories, rare variant (MAF ≤ 0.05) and common variant (MAF > 0.05) associations. If unspecified, both categories used the same genome-wide significant P-value threshold, which is 5×10^{-8}. The INFO threshold is 0.8 for rare variants and 0.3 for common variants. Figure 4.2 shows that rare variant associations with low INFO tend to have inflated effect sizes, which justified our choice of higher INFO threshold for rare variants.

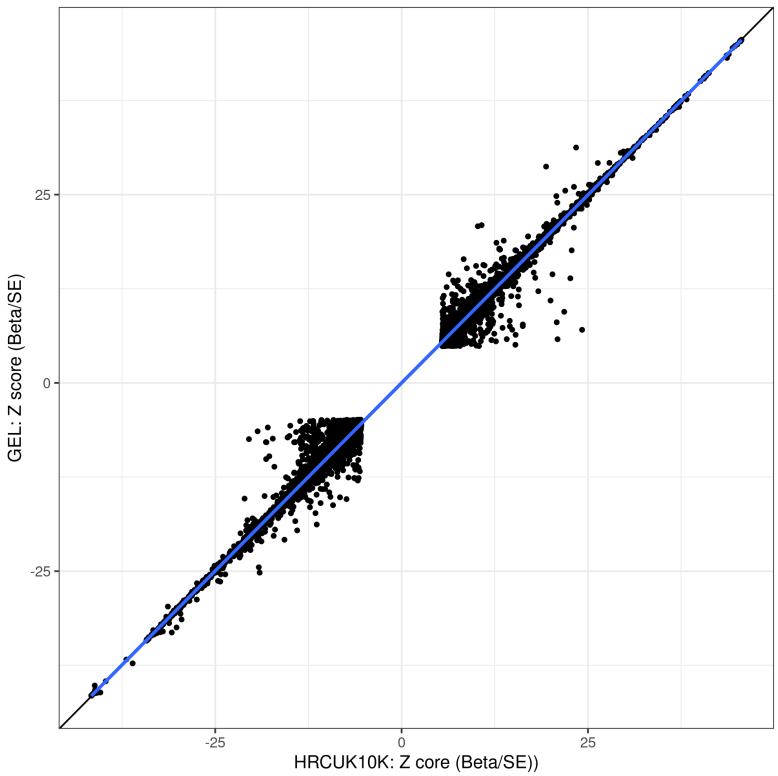


Figure 4.1: Comparing the Z-scores from G-UKB and H-UKB for the markers that were shown to be significant for height association (p-value $< 5 \times 10^{-8}$) with INFO $>$ 0.3. The blue line shows the linear regression line estimated on these data. The Pearson correlation r^2 is 0.99.

REGENIE tested, in total, 194 million and 67 million variants from G-UKB and H-UKB, respectively. G-UKB dataset had 102 million rare variants and 6.6 million common variants, passing the INFO threshold, whereas H-UKB had 14 million and 7.3 million such rare and common variants, respectively. Across all four traits, 31,699 rare associations and 455,392 common associations were found in G-UKB, when H-UKB found 30,711 and 494,215 rare and common associations, respectively (Figure 4.1). The detailed GWAS hits analysis results will be discussed in the next section.

The GWAS results using H-UKB and G-UKB are highly correlated. The variants that are present in both datasets showed a high correlation ($r^2 = 0.99$) (Figure 4.1). Over 90% of the genome-wide significant associations (GWAS hits) are common variants (MAF>0.05), as they are better powered. In the following sections, we will carry out more detailed analysis to show the advantage of using G-UKB to H-UKB, in terms of rare variant associations as well as common variant associations.

	P-value	INFO	G-UKB	H-UKB
Common associations	5×10^{-8}	0.3	455,392	494,215
Rare associations	5×10^{-8}	0.8	31,699	30,711
Common functional	5×10^{-8}	0.3	2,701 (0.59%)	2,747 (0.56%)
Rare functional	5×10^{-8}	0.8	511 (1.64%)	473(1.57%)
In common to EWAS	5×10^{-8}	/	26 (76%)	22 (65%)
In common to EWAS	2.18×10^{-11}	/	24 (70%)	19 (56%)

Table 4.1: Genome-wide significant association counts across all four traits. Common variants are variants with MAF > 0.05 and rare variants are those with MAF ≤ 0.05, excluding the HLA region in chromosome 6. The number of functionally important variants including high and moderate impact variant effect annotated by VEP are presented in the table, followed by the percentage found in genome-wide significant in the same category. We also showed the associations that are in common with UK Biobank EWAS results and its proportion in all the EWAS findings [53].

Exome-wide association studies (EWAS) examined 430,998 UKB individuals with European ancestry focusing on predicted loss of function variants, including stop-gain, frame-shift, stop-lost, start-lost and essential splice variants, using REGENIE RINT transformed data [53]. The genome-wide significant threshold is set at 2.18×10^{-11}. EWAS has reported 31, 0, 1, and 2 rare (MAF ≤ 0.053) genome-wide significant variant-trait associations across height, BMI, SBP, and DBP, respectively [53]. The GWAS procedure by Backman et al. (2021) was very similar to what we have done, though not necessarily identical [53]. Across all four traits, we discovered 70% of the EWAS findings can be found in G-UKB GWAS, whereas H-UKB only found 56%. If we relax the P-value threshold to 5×10^{-8}, G-UKB could discover 76% of these reported rare associations (Figure 4.1).

4.2.2 Rare variant associations

Statistical power of a regression analysis is influenced by effect size, allele frequency and the number of samples [129]. In a linear regression model for quantitative traits,

$$y = \mu + \beta X + \epsilon \tag{4.2}$$

where y is the phenotype measurement, β is the effect size, X is the genotype matrix, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The effect size estimate is $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2_{\hat{\beta}})$ and the squared z-score is $z^2 = (\hat{\beta}/\sigma_{\hat{\beta}})^2 \sim \chi^2_1(\lambda)$, with the non-centrality parameter $\lambda = (\beta/\sigma_{\hat{\beta}})^2$. Under

the assumption of the Hardy Weinberg equilibrium, the variance of a genotype is approximately $2f(1-f)$, where f is the allele frequency, and the standard error of the effect size estimate $\sigma_{\hat{\beta}}$ is

$$\sigma_{\hat{\beta}} = \frac{\sigma}{\sqrt{\text{Var}(X)N}} \approx \frac{\sigma}{\sqrt{2f(1-f)N}} \tag{4.3}$$

Hence, the non-centrality parameter is

$$\lambda = 2f(1-f)\beta^2 N/\sigma^2 \tag{4.4}$$

When the allele frequency decreases, the minimum effect size has to increase, in order to reach a particular significant level. Figure 4.2 shows that there were rarely any significant associations to be detected beyond the theoretical curve, where the statistical power is 0.5. Although real rare variant associations tend to have higher effect sizes, an inflated effect size still implies problems. This issue was well-captured by imputation INFO score. Figure 4.2 also shows rare variant associations with the allele frequency below 5×10^{-4} are exclusively detected by G-UKB. Similar results can be observed in the other three traits as well (Figure A.1).

We further validated the GWAS findings by examining the concordance between imputed variants and the 200,000 UK Biobank exome sequencing samples [53]. Among the 487,091 associations across all four traits, we found 18,740 associations corresponding to 15,362 associated markers in the exome sequencing data. In Figure 4.3, we demonstrate a high overall r^2 correlation between the GEL imputed dosages and the exome sequencing data at the genome-wide significant sites, with 97% of them having $r^2 > 0.8$. The proportion of variants surpassing the $r^2 > 0.8$ threshold are as follows: 97% ($AF > 10^{-1}$), 97% ($10^{-2} < AF \leq 10^{-1}$), 96% ($10^{-3} < AF \leq 10^{-2}$), 79% ($10^{-4} < AF \leq 10^{-3}$), and 51% ($AF \leq 10^{-4}$). Assuming that variants passing the $r^2 > 0.8$ threshold are reliable imputation for association tests, we can consider the GEL imputed UKB to be generally "trustworthy" for allele frequencies above 10^{-4}. We need to take caution for rare variant associations ($AF \leq 10^{-4}$).

In Figure 4.2 and Figure A.1, we observed a consistent trend towards negative (detrimental) effect sizes in rare variant associations. In Figure 4.3, we grouped

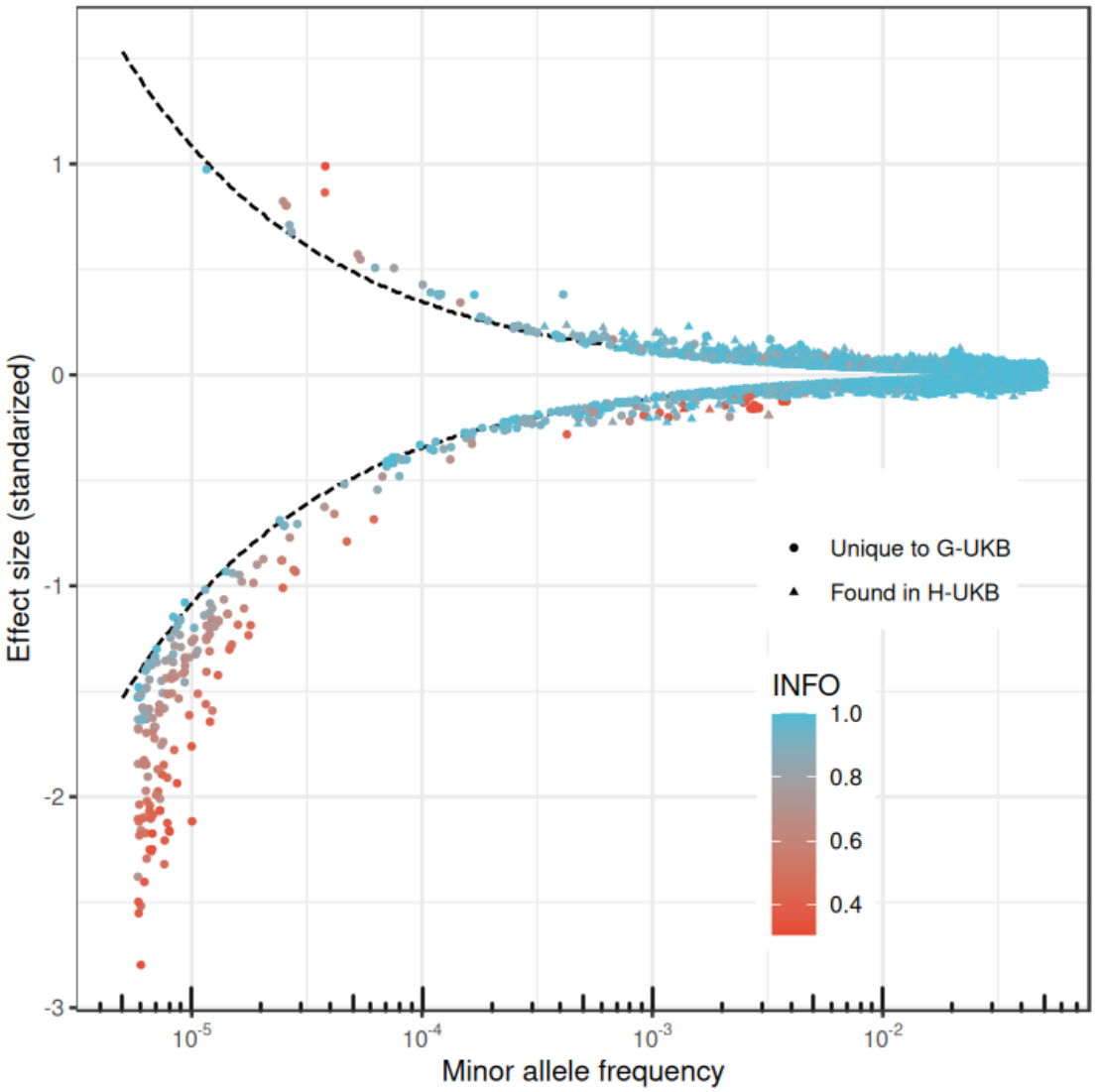


Figure 4.2: Genome-wide significant rare variants (MAF < 0.05) associations for standing height using G-UKB data shows rare associations require higher effect sizes to be detected. The triangles are G-UKB variants that can be found in H-UKB and the dots are those unique to G-UKB. The dotted line shows the theoretical minimum effect size for different allele frequency when the statistical power is 0.5, assuming $\alpha = 5 \times 10^{-8}$ and $N = 418,755$. The dots/triangles are coloured by the G-UKB imputation INFO score.

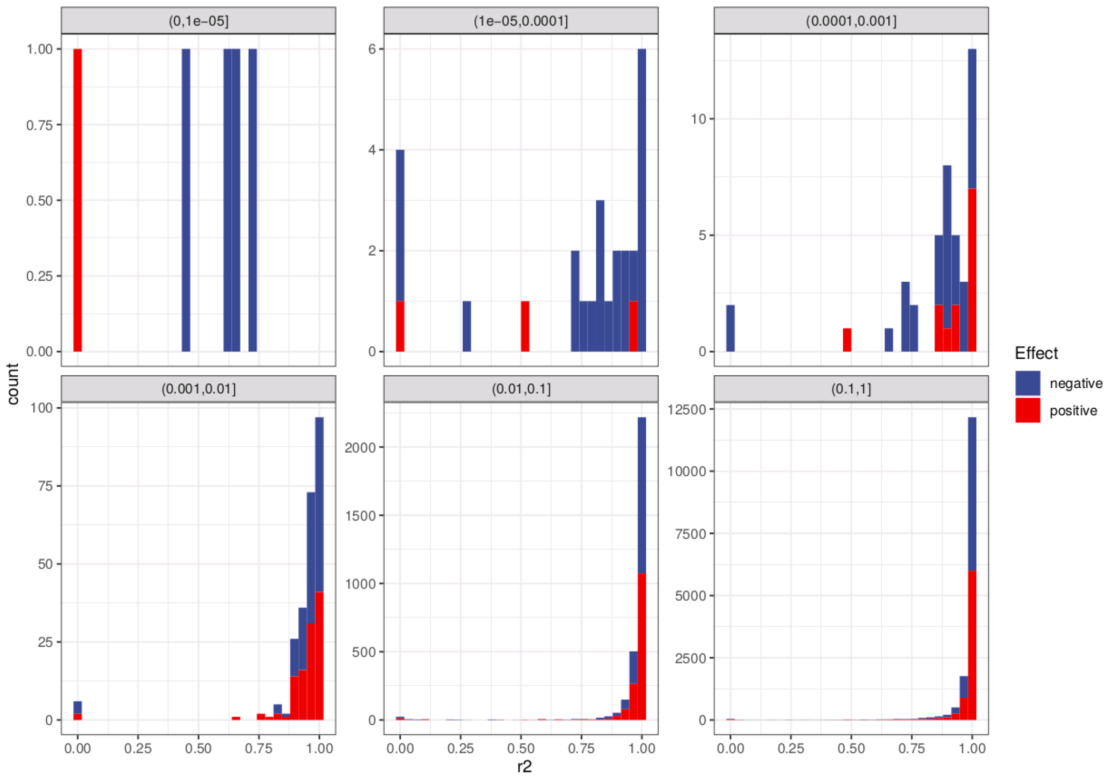


Figure 4.3: The histograms of squared correlation r^2 between the exome sequencing variants and GEL imputed variants in 200,000 UK Biobank samples that are available in the exome sequencing data, considering the GWAS significant variants with INFO > 0.3, stratified by allele frequencies. 97% examined GWAS hits are with $r^2 > 0.8$. The proportion of variants surpassing the $r^2 > 0.8$ threshold are as follows: 97% ($AF > 10^{-1}$), 97% ($10^{-2} < AF \leq 10^{-1}$), 96% ($10^{-3} < AF \leq 10^{-2}$), 79% ($10^{-4} < AF \leq 10^{-3}$), and 51% ($AF \leq 10^{-4}$). The bars in histograms are coloured by the effect size directions. We considered negative effect sizes in HEIGHT and positive effect sizes in BMI, SBP, and DBP as indicative of negative or detrimental effects.

the GWAS and exome overlapping variants based on positive and negative effects. We consider negative effect size in HEIGHT and positive effect size in BMI, SBP and DBP as negative or detrimental effects. In this analysis, we also observed higher proportion of negative (detrimental) effect associations with rare variants and most of these rare imputed variants have a considerably high imputation quality. We also found similar negative effect sizes in rare associations reported by the exome analysis [53].

A genetic mutation that causes the alteration of the molecular function of a protein is called a functional variant.They are particularly informative for association studies and fine-mapping [17]. By examining the functional variants with the

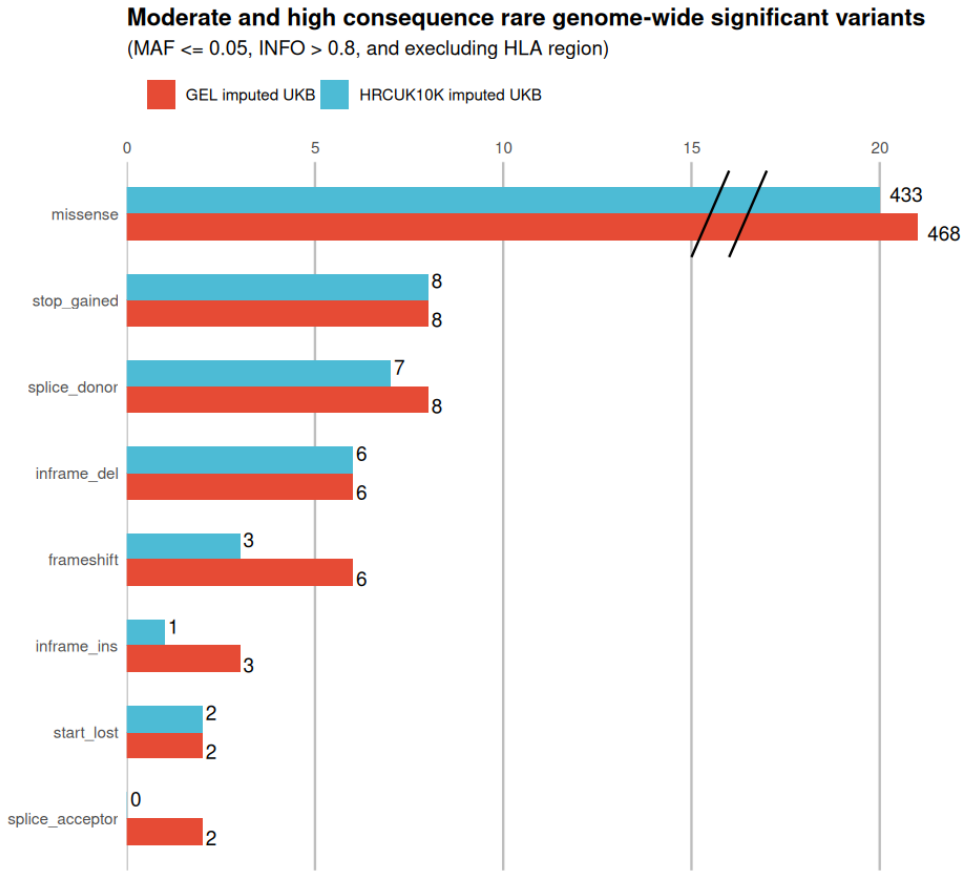


Figure 4.4: The number of significant rare associations captured by G-UKB and H-UKB data, stratified by variant effects with high and moderate consequences, defined by VEP.

following high and moderate predicted consequences: start lost, stop lost, stop gained, frame-shift, splice donor, splice acceptor, inframe insertion, inframe deletion, and missense, the G-UKB found more rare associations than in H-UKB in all these categories (Figure 4.4). Variants were annotated using VEP (release 105) [124].

4.2.3 Common variant associations

G-UKB detected 38,823 fewer common variant GWAS hits than H-UKB. Regarding concerns described in the previous chapter about potential false common variant calls in H-UKB, mainly caused by the excess of common variants in UK10K, we examine whether some of the excessive H-UKB GWAS hits could be false.

In total, there were 64,579 common significant associations that were detected by H-UKB, but failed to be detected by G-UKB, either because the variants were

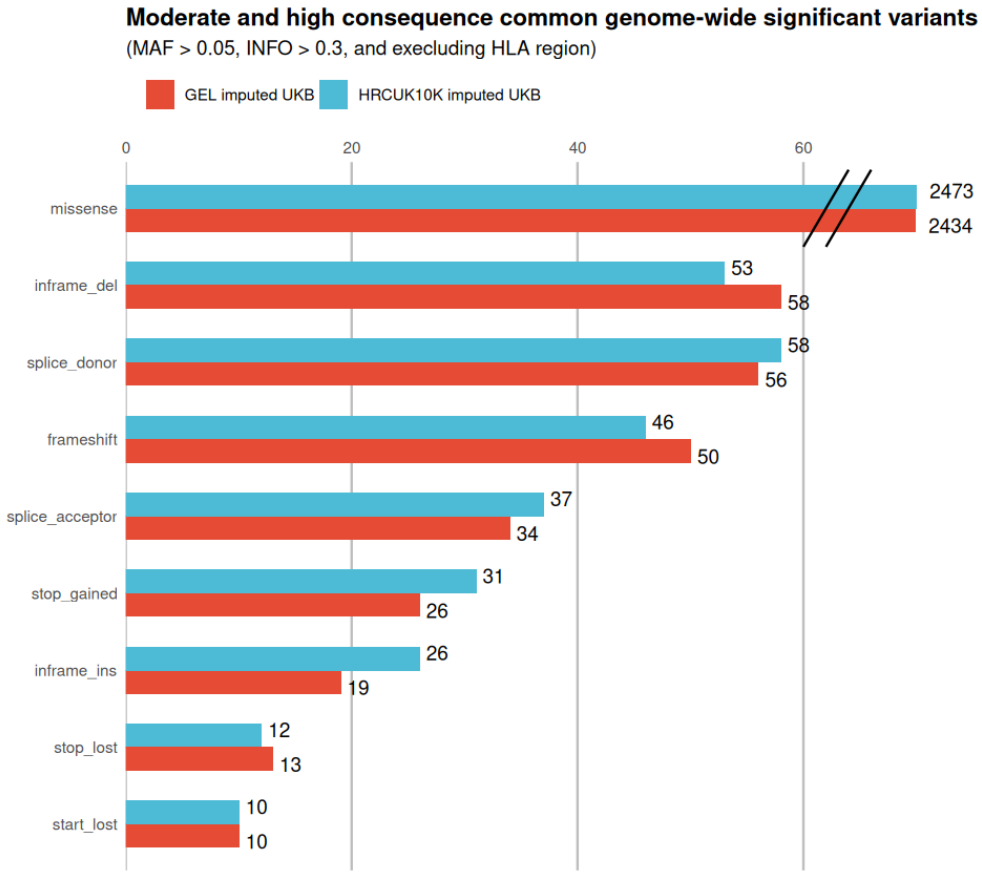


Figure 4.5: The number of significant common associations captured by G-UKB and H-UKB data, stratified by variant effects with high and moderate consequences, defined by VEP.

not present in H-UKB (35,823) or marked as not significant by the H-UKB GWAS. Among 22,549 (70%) of unique G-UKB GWAS hits, only 12,455 (19.2%) of them can be found in TOPMed r2. This suggests that the excessive H-UKB common variant associations might be compromised.

In Figure 4.6, we further compared the significant associations that are unique to each dataset, in terms of the allele frequency concordance to the TOPMed allele frequency. We found that the G-UKB unique GWAS hits match with the TOPMed allele frequency better than H-UKB unique GWAS hits. Many rare variants in TOPMed have an inflated allele frequency in H-UKB. In spite of having lower concordance in general, the allele frequency of H-UKB GWAS hits matched the TOPMed allele frequency quite well, indicating most of the H-UKB unique associations, when existed in TOPMed, are likely to be real. However, the variants

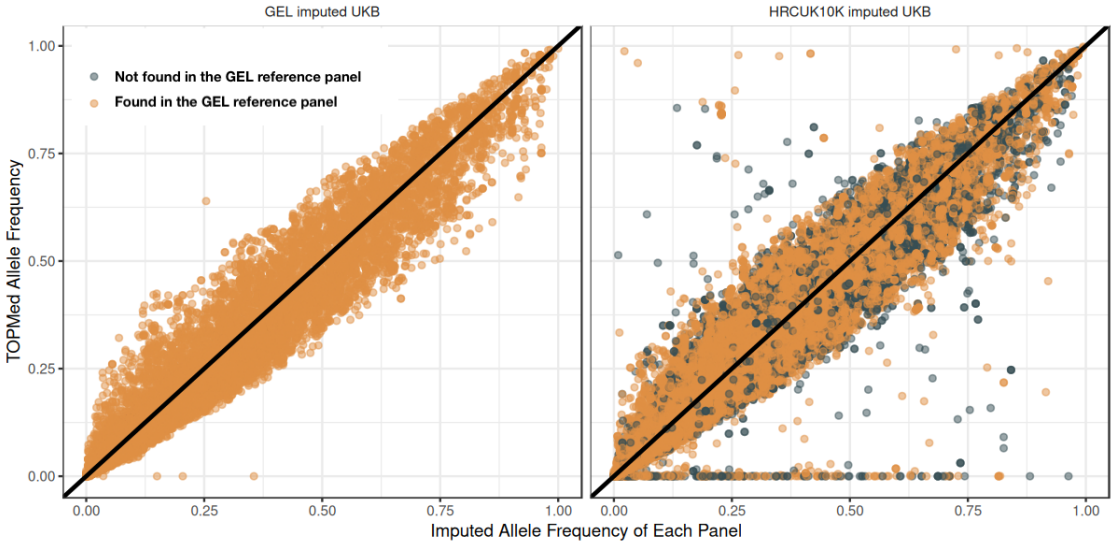


Figure 4.6: Comparing allele frequency concordance of GWAS significant sites to TOPMed between associations that are unique to G-UKB (left panel) and unique to H-UKB (right panel). Variants that don't locate within the HLA region and with INFO > 0.3 that are included in both datasets are included.

that are in common between the H-UKB unique GWAS hits and the TOPMed variants make up less than 20% of the H-UKB unique sites.

We further examined the imputed variant allele frequency concordance to TOPMed among associations, where a large P-value discrepancy was produced by between G-UKB and H-UKB. The large P-value discrepancy sites were defined as variants that were significant (P-value $< 5 \times 10^{-8}$) in one dataset but 1,000 times larger in the other dataset (P-value $> 5 \times 10^{-5}$). 14,898 GEL standard filter sites and 9,168 GEL lenient filter sites were identified. The GEL standard filter sites had a much better allele frequency concordance in both H-UKB P-value underestimation and overestimation cases (Figure 4.7). The GEL lenient filter sites in G-UKB showed potential for false negative GWAS hits, but no sign of false positives hits (Figure 4.8). These results strengthened our belief that, despite having less common variant associations, G-UKB associations are more reliable and less likely to be false positives. Regarding the standard filter sites, the G-UKB insignificant sites that are marked as significant in H-UKB are not likely to be false negative.

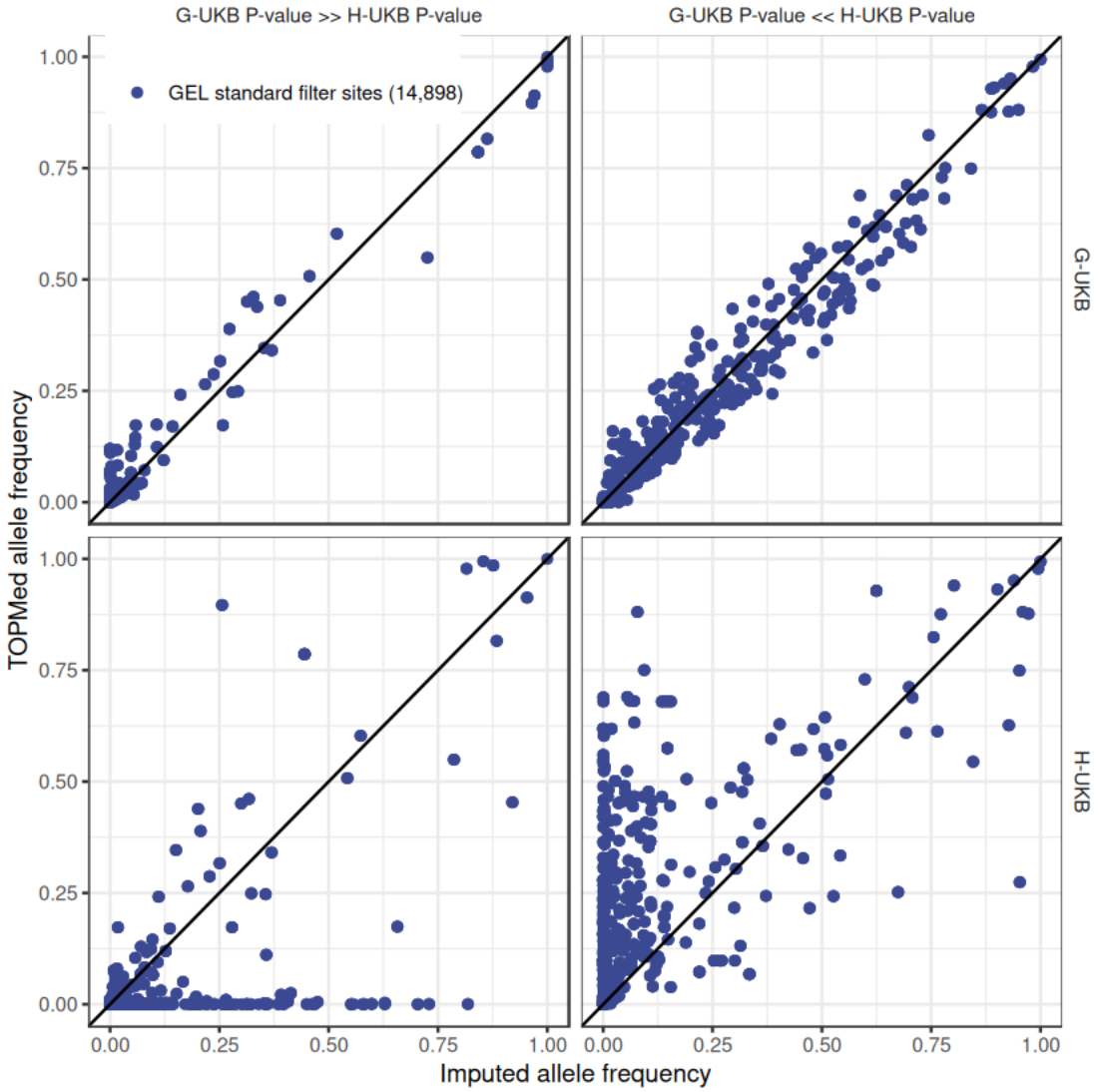


Figure 4.7: Comparing allele frequency concordance to the TOPMed allele frequency between large P-value discrepancy sites. Associations that are significant in H-UKB but with much bigger P-value ($> 5 \times 10^{-5}$) in G-UKB are allocated in the left column, separated by GEL normal filter sites and GEL lenient filter sites. Associations that are significant in G-UKB, but with much bigger P-Value ($> 5 \times 10^{-5}$) in H-UKB.

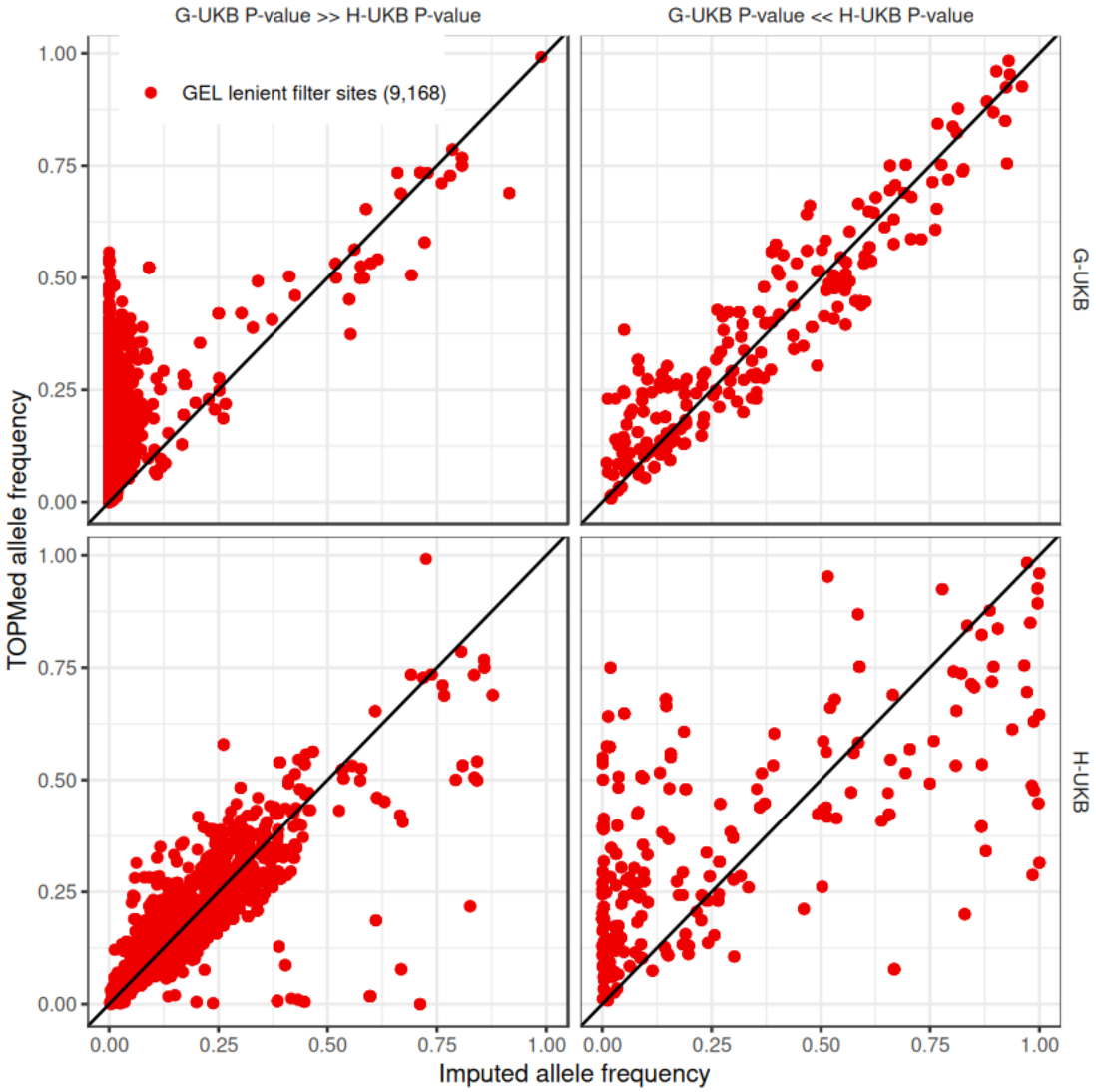


Figure 4.8: Comparing allele frequency concordance to TOPMed between large P-value discrepancy sites. Associations that are significant in H-UKB but with much bigger P-value ($> 5 \times 10^{-5}$) in G-UKB are allocated in the left column, separated by GEL normal filter sites and GEL lenient filter sites. Associations that are significant in G-UKB, but with much bigger P-Value ($> 5 \times 10^{-5}$) in H-UKB.

4.3 Fine mapping rare variant associations

4.3.1 Conditional joint analysis

A standard GWAS procedure uses marginal models, considering one variant at a time, in order to make the GWAS computationally tractable. Meanwhile, a joint model, considering all the selected variants and estimating their joint effect simultaneously, is more time-consuming. However, finding non-correlated variants are advantageous for fine-mapping of causal variants.

In order to use the joint model and compute in a reasonable timeframe, we carried out a conditional joint analysis using a step-wise forward selection procedure, one chromosome at a time. Our approach is similar to that was mentioned in [130]. First, we defined a set of genome-wide significant variants in one chromosome (P-value $< 5 \times 10^{-8}$) identified in the marginal regression by REGENIE, as $\mathbf{S}$. Use another set of variants $\mathbf{R}$ as the conditional variables in a joint regression, starting with the most significant variant in the marginal regression. For each step, we calculated the P-value for all the remaining variants in $\mathbf{S}$, one at a time, conditioned on $\mathbf{R}$. Move the variant with smallest P-value from set $\mathbf{S}$ to set $\mathbf{R}$, and then repeat until there are no more variants in set $\mathbf{S}$ remaining significant. All the above joint regression are conditioned on the covariates mentioned in Section 4.

The step-wise forward selection approach identifies a "smallest" subset of variants, for which the joint regression model explains the data so well that any additional variants would not lead to a better fit to the data. Variants that are included in the final selected set $\mathbf{R}$ should show incomplete LD among each other, since the colinearity among variants with high LD ensures only one in such a region is included in $\mathbf{R}$. Step-wise forward selection is often used to find regions for Bayesian fine-mapping [17]. To demonstrate the performance on finding rare causal variants using G-UKB in comparison to H-UKB, we will focus on the putative causal variants found by the step-wise forward selection approach.

rsid	Phenotype	MAF	Data	Effect size	$-\log_{10}$ P-value
rs535071161	BMI	0.0008360	G-UKB	0.969368	8.80021
rs948848696	BMI	0.0000895	G-UKB	3.416520	10.08210
rs1286160511	BMI	0.0000376	G-UKB	5.120000	9.47132
rs577620983	SBP	0.0041567	G-UKB	1.754600	8.01582
rs746206908	SBP	0.0000487	G-UKB	16.249800	7.56910
rs151036307	SBP	0.0020308	G-UKB	2.846980	9.46999
chr12:118537814_T_C	SBP	0.0000166	G-UKB	32.867900	9.45829
rs200999181	SBP	0.0011677	H-UKB	3.376680	8.61512
rs1327113379	SBP	0.0012055	G-UKB	3.404630	8.92087
rs374660188	SBP	0.0044004	G-UKB	-1.845340	9.59273
8:135602725_TCAC_T	SBP	0.0046057	H-UKB	-1.837860	9.60753
chr1:21479845_T_TGTGTGG	DBP	0.0000125	G-UKB	19.378300	7.90003
rs757561770	DBP	0.0000093	G-UKB	22.383400	8.85968
chr3:159719798_A_G	DBP	0.0000125	G-UKB	20.784000	9.36453
rs773828312	DBP	0.0020958	G-UKB	-1.482880	8.06360

Table 4.2: Ultra-low frequency associations flagged by step-wise regression (INFO > 0.8), see Figure 4.9

4.3.2 Results

We found a directional effect of ultra-rare variant associations (Figure A.1), for example, there are more associations being found with the effect of increasing the blood pressure than blood pressure lowing. By only looking at the independent variants flagged by conditional joint analysis (Figure 4.9), the directional trends are diminished compared to what we see in Figure A.1, but not fully eliminated in blood pressure traits and BMI. We list all the ultra-low frequency associations flagged by the step-wise regression in Table 4.2. Many of these associations are in or adjacent to genes that were found associated with the traits in the literature, hence plausible. The effect size negative trend still persists from the associations selected by the conditional joint analysis.

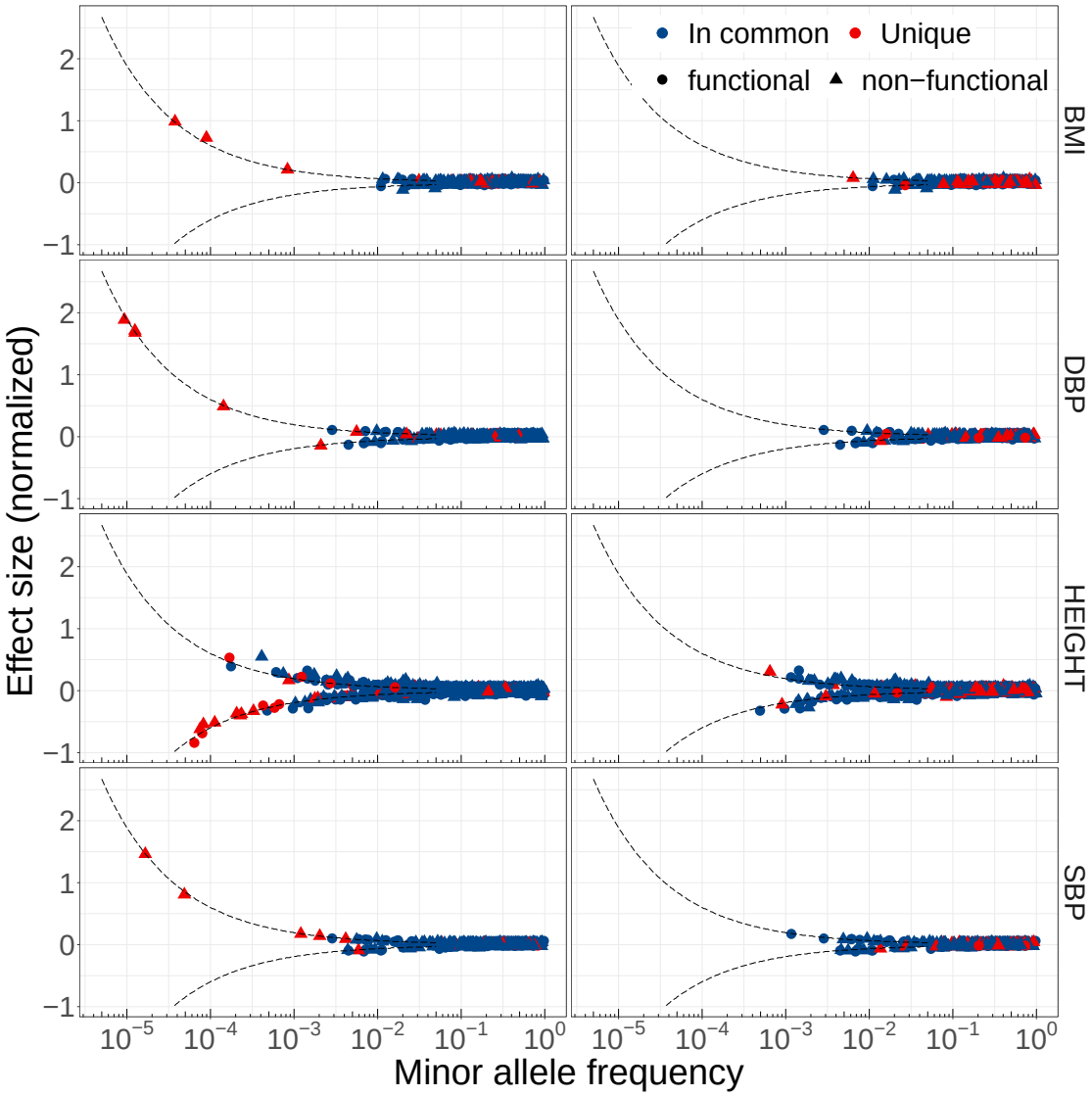


Figure 4.9: Effect size v.s. imputed allele frequency of independent genome-wide significant ($p < 5 \times 10^{-8}$) associations identified by the step-wise regressions (INFO > 0.8). The blue dots/triangles indicate the variants that were flagged by step-wise regressions by one dataset and at the same time a significant GWAS association in the other, both GEL-UKB (left column) and HRC-UKB (right column), and the red dots indicate unique variatns to each datasets. The shape of the data points indicates the consequences of the variants predicted by VEP, with dots indicating functional variants including, stop gained, stop lost, splice donor/acceptor, frameshift, inframe insertion/deletion and missense, and triangles indicating non-functional including intergenic, intronic and unknown consequences. The dotted line shows the theoretical minimum effect size detectable for corresponding allele frequencies assuming the power is 0.5.

Phenotype	rsid	MAF	Effect size	Consequence	Gene	Note
BMI	rs150391548	0.049	-0.128	intergenic	/	Near COX6A1, a gene known to be associated with BMI [131] (Figure A.3)
HEIGHT	rs554308136	0.034	0.2	intron	ZBTB14	(Figure A.2)
SBP	rs577620983	0.004	1.75	TF-binding site	/	Near PTP4A2 and KPNA6. (Figure A.4)
	rs9332745	0.028	-0.76	missense	KMT2A	KMT2A is associated with triglycerides level. [132, 133] (Figure A.5)
	rs151036307	0.002	2.84	synonymous	TNS2	SBP association reported by the UKB whole exome imputation paper. [125] (Figure A.6).
	rs17583877	0.006	-1.87	intron	ADRA1B	A deletion in ADRA1B gene, which is known to be associated with SBP and DBP [134]. (Figure A.7)
	rs575127487	0.023	-0.97	intron	KCNK9	KCNK9 is known to be controlling the potassium pathway (Figure A.8).
DBP	rs757561770	9.3×10^{-6}	22.38	intron	FGD4	FGD4 gene is known to be associated with systolic blood pressure [135] and hypertension [136]. (Figure 4.10)
	rs773828312	0.002	-1.48	intron	BZW2	BZW2 was found to be associated with diastolic blood pressure [137] (Figure A.9)
	rs368338720	0.022	0.47	intron	MTUS1	(Figure A.10).

Table 4.3: A list of 10 rare putative causal associations, flagged only by the G-UKB conditional joint analysis. These associations were only found G-UKB, i.e. they were either not significant in H-UKB or were not present in H-UKB. The Manhattan plots are provided in the appendix.

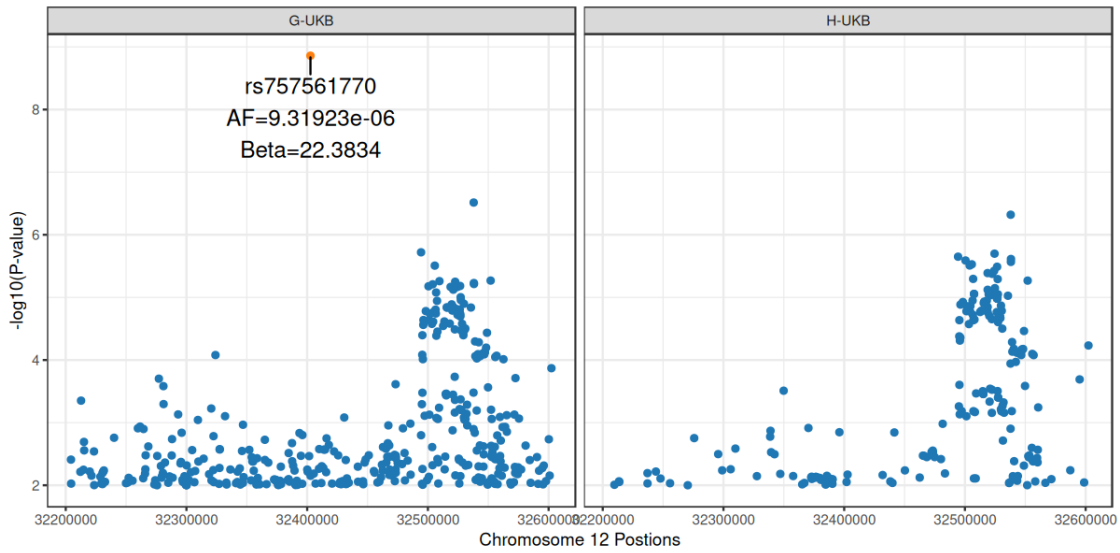


Figure 4.10: Manhattan plot comparing G-UKB and H-UKB results for DBP association. rs757561770 is an ultra-rare variant near FGD4, which known to be associated with systolic blood pressure [135] and hypertension [136].

We also hand-picked 10 LD-independent rare putative causal associations that are only significant in G-UKB, flagged the G-UKB conditional joint analysis. (Table 4.3). Some of these variants are selected because they are functionally important, and the others are selected because they are in or adjacent to genes that were found associated with the traits in the literature. The rarest association was rs757561770 (MAF=9.3×10^{-6}), an intron variant of FGD4 gene associated with DBP with an effect size of 22.38 mmHg (Figure 4.10), containing 9 heterozygote carriers in the entire imputed dataset, none of which were outliers. FGD4 gene is known to be associated with systolic blood pressure [135] and hypertension [136].

4.4 Fine-mapping common variant associations

4.4.1 Bayesian fine-mapping

Bayesian fine-mapping aims to find the posterior of probability of a model given the data. The model M is an indicator vector, where the i^{th} element is 1 when the i^{th} variant is causal and 0 when non-causal.

$$\Pr(M_i|\mathbf{X}) = \frac{\Pr(\mathbf{X}|M_i)\Pr(M_i)}{\sum_{M \in m} \Pr(\mathbf{X}|M)P(M)} \tag{4.5}$$

The Bayesian fine-mapping methods try to find the posterior inclusion probability (PIP_i), by summing up all model posteriors that include site i to be causal. We then rank the PIP*s* from high to low, and take the cumulative sum of PIPs until it is at least α, which we call the α-credible set.

The number of models increases exponentially with the increasing number of variants. The integral for the marginal likelihood and the posterior probability for many models posed huge challenges in computation [17].

For the purpose of demonstrating the performance of GEL imputed data, we used the method developed in Maller et al. 2012 [138]. Assuming only one causal variant within the defined region, the posterior is proportional to the Bayes factor, which can be subsequently be estimated using the asymptotic Bayes factor (ABF) described in Wakefield 2019 [139]. The benefit of using ABF is that it only requires the regression summary statistics.

Here is a brief sketch to show how the posteriors can be expressed as the Bayes factor. Suppose we define a region with k sites, the genotypes of the region forms a vector $\mathbf{X}$, and assume a uniform prior $\Pr(M_i|M) = \frac{1}{K}$. The Bayes factor (BF_i) comparing model M_i and M_0 (no causal variant) depends only on the genotype at site i.

$$\text{BF}_i = \frac{\Pr(\mathbf{X}|M_i)}{\Pr(\mathbf{X}|M_0)} = \frac{\Pr(\mathbf{X}_i|M_i)}{\Pr(\mathbf{X}_i|M_0)} \tag{4.6}$$

The posterior of having one causal variant is proportional to the causal variant's BF_i.

$$\begin{aligned}
\Pr(M_i|\mathbf{X}, M) &= \frac{\Pr(\mathbf{X}|M_i, M)\Pr(M_i|M)}{\Pr(\mathbf{X}|M)} \\
&= \frac{1}{k}\frac{\Pr(\mathbf{X}|M_i)}{\Pr(\mathbf{X}|M)} \\
&= \frac{1}{k}\frac{\Pr(\mathbf{X}|M_i)/\Pr(\mathbf{X}|M_0)}{\Pr(\mathbf{X}|M)/\Pr(\mathbf{X}|M_0)} \\
&= \frac{1}{k}\frac{\Pr(\mathbf{X}|M_i)/\Pr(\mathbf{X}|M_0)}{\frac{\sum_{i=1}^{k}\Pr(\mathbf{X}|M_i)\Pr(M_i|M)}{\Pr(\mathbf{X}|M_0)}}
\end{aligned} \tag{4.7}$$

Based on (4.6) and (4.7), the posterior for one causal variant can be expressed as the following,

	$W = 0.02^2$		$W = 0.2^2$	
Traits	HRCUK10K	GEL	HRCUK10K	GEL
Height	15 (0.2%)	15 (0.07%)	15 (0.19%)	15 (0.07%)
BMI	26 (0.29%)	25 (0.1%)	27 (0.3%)	30 (0.11%)
SBP	21 (0.25%)	21 (0.09%)	21 (0.25%)	21 (0.09%)
DBP	19 (0.2%)	20 (0.07%)	19 (0.26%)	21 (0.95%)

Table 4.4: The median number of variants included in the Bayesian fine-mapping 95% credible set and the the median proportion of the credible set to the number of sites contained in the region.

$$\Pr(M_i|\mathbf{X}, M) = \frac{BF_i}{\sum_{i=1}^{k} \mathrm{BF}_i} \approx \frac{\mathrm{ABF}_i}{\sum_{i=1}^{k} \mathrm{ABF}_i} \tag{4.8}$$

According to Wakefield 2019 [139], the asymptotic Bayes factor (ABF_i) in favour of having one causal variant, can be derived only using the regression summary statistics, where V_i the standard error, Z is the Z score. It assumes a normal prior $\sim \mathcal{N}(0, W)$, where W is the variance. Common choices of W are 0.2^2 and 0.02^2 [1].

$$\mathrm{ABF}_i = \left[\frac{\sqrt{V_i + W}}{W} \exp\Big(- \frac{z^2}{2} \frac{W}{V_i + W}\Big)\right]^{-1} \tag{4.9}$$

Since we assume there is only one causal variant within a region, the PIP_i is equivalent to the model posterior $\Pr(M_i|\mathbf{X})$.

We used GWAS summary statistics from Section 4, to compare G-UKB and H-UKB Bayesian fine-mapping performance. Using the genetic distance in the HapMap recombination map [140], the fine-mapping regions are defined as 0.125 centimorgans (cM) plus 25 kb each side of the H-UKB genome-wide significant marker (p-value $< 5 \times 10^{-8}$). Regions are merged into one if they overlap. With this method, we defined 1,660, 711, 505, 546, and 505 non-overlapping regions for HEIGHT, BMI, SBP and DBP, among which 60, 30, 33, 51 regions for each respective trait found no genome-wide significant markers in G-UKB, and were removed from this analysis.

4.4.2 Results

95% credible sets for each region were derived by the PIP of variants enclosed in the region using the summary statistics from (4.8). We found that the choices

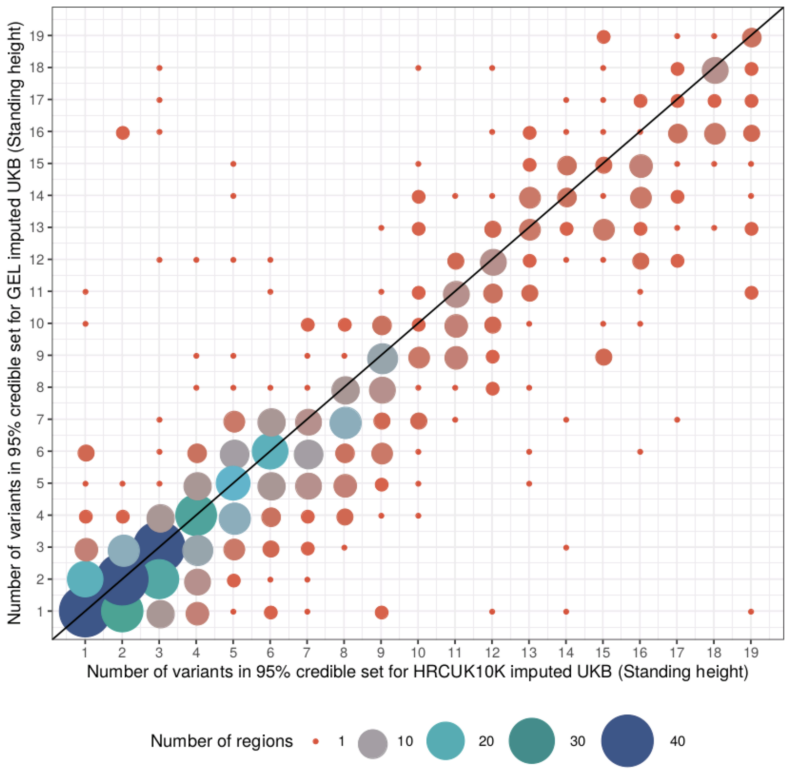


Figure 4.11: Compare 95% credible sets size ranging from 1 to 20 variants for standing height. This plot used the prior of $W = 0.02^2$. The size of the dots represents the number of fine-mapping regions.

of prior variance do not affect the results of H-UKB, but smaller priors (0.02^2) benefit the G-UKB results (Table 4.4). The median 95% credible set size are almost identical using G-UKB and H-UKB, but the proportion of variants in credible set of each region is around 3 times smaller in G-UKB than H-UKB (Table 4.4). Across all four traits, we found 44% of the regions have a lower 95% credible set size using G-UKB comparing to H-UKB, 30% remained the same and 25% H-UKB outperformed G-UKB (Figure 4.11).

4.5 Discussion

G-UKB found 1,699 more rare variant association and 38 rare functional associations than H-UKB. G-UKB found 70% of the associations found in EWAS [53], while H-UKB found only 56% of them. The main difficulty in detecting rare associations is that the GWAS approach has less power to detect ultra-rare associations with

moderate effect size, despite a drastic increase in testable rare variants in G-UKB (102 million), a seven fold increase to H-UKB (14 million). The low number of novel ultra-rare GWAS findings are also seen in UKB exome sequencing analysis [53, 141].

G-UKB detected less common variant associations than H-UKB. We conjectured that this is largely due to the inflation of false common variant calls in H-UKB, introduced by UK10K reference panel. We have the following evidence,

1. Over 1 million (11%) H-UKB common variants are not found in TOPMed (Table 2.4) and 16.75% of common variants are not found in GEL standard filter sites (Table 2.3).

2. We saw no inflation of common functional GWAS hits in H-UKB (Table 4.1).

3. 80% of common GWAS hits that are unique to H-UKB cannot be found in TOPMed.

4. The G-UKB standard filter sites tend to match the TOPMed allele frequency better, when H-UKB P-value estimates have large discrepancies to those estimated by G-UKB (Figure 4.7).

We demonstrated the imputation using large high coverage reference panel, such as GEL, could benefit the discovery of novel causal associations. Ten rare putative causal associations are flagged by the conditional joint analysis, which were not found to be significant in H-UKB. It also improved the Bayesian fine-mapping performance. Over 74% of the regions has achieved lower or equal credible set size, using G-UKB data.

We used four traits as a demonstration of the GEL imputed data. In terms of future work, we plan to expand the analysis to other UK Biobank phenotypes. We would also like to work on more rare variant analyses, including burden testing of genes.

5

Efficient Discrete time Wright Fisher Simulation

Contents

5.1 Overview 78
5.2 The ARGON2 DTWF simulator 78
5.2.1 The ancestral recombination graph 78
5.2.2 Simulation algorithm 79
5.2.3 The merging algorithm for coalescence events 81
5.2.4 Recombination and gene conversion 82
5.2.5 Mutation 84
5.3 ARGON2 validation and performance 85
5.3.1 Generating the ARG 85
5.3.2 Performance on a single constant sized population . . . 92
5.3.3 Performance on multiple populations with migration . . 94
5.4 ARG augmentation: simulating new samples on an existing ARG 94
5.4.1 New data structure: Genealogy and fixed blocks 95
5.4.2 Simulation heuristics 95
5.4.3 Simulation results on simulated ARGs 97
5.4.4 Simulation results on inferred ARG 99
5.5 Discussion 101

5.1 Overview

Genome simulation has served as an important tool in the development of population and statistical genetics during the past two decades [86–88]. With large genetic sequencing datasets becoming increasingly accessible, the demand for large-scale simulation in terms of sample size, genome length and complex demographic histories has been growing. In this chapter, we present a new implementation of a simulator sampling from the discrete time Wright Fisher (DTWF) process, called ARGON2. Simulating from the DTWF process is in general more computationally demanding than the coalescent simulation, However it leads to more accurate simulation results in terms of long range LD and recent generation relatedness [100–102, 142]. In addition, simulating specific aspects of genealogical histories, such as high migration rates or non-crossover gene conversion events, is cumbersome using standard coalescent simulation. In this chapter, we show that ARGON2 may be substantially faster than state-of-the-art coalescent simulation in some of these scenarios. We also develop new methodology to generate new samples conditional on existing simulated or inferred ARGs, which may be used to obtain increasingly realistic simulated data and used to aid study design.

The structure of this chapter is as follows: Section 5.2 describes the details of the ARGON2 simulation algorithm and its implementation. In Section 5.3, we assess simulated ARGs by comparing their topological properties, and benchmark the performance of several simulators. In Section 5.4, we describe a novel approach to perform data augmentation using inferred ARGs.

5.2 The ARGON2 DTWF simulator

5.2.1 The ancestral recombination graph

The term “ancestral recombination graph” (ARG) was originally used to redefine the coalescent process with recombination [114, 143, 144], referring to both the stochastic process and the encoding of the genealogy.

The term ARG has been increasingly used to solely refer to the graph-based encoding of a genealogical object (e.g., [145, 146]), while the process itself is most often referred to as the coalescent with recombination [81, 90, 91]. An ARG implicitly defines a genealogical tree for each position, which are often referred to as marginal tree or local tree. The topology of consecutive trees is progressively modified by the occurrence of recombination events. There is some variation in how nodes and edges in an ARG may be used (see e.g., [101, 146] or [95]). To store the simulated genealogy, ARGON2 uses a data structure similar to that used in ARGON, which we refer to as *SimpleARG*. This structure consists of two main elements, *Edges* represented by the parent node ID and child node ID, alongside with the corresponding genome range and the *Height* of each node, in number of generations. *SimpleARG* may be easily converted to the ARG format used in the ARG-Needle and tskit software packages.

Coalescent events are usually rare compared to the number of simulated individuals across generations, implying that the genetic material in the simulation is more likely transmitted from one generation to another without changes. Using an ARG node for each individual would therefore lead to the creation of a large number of unary nodes, where a parental node has only one child, leading to very large memory costs. However, recording unary nodes, may be desirable in some cases. By recording the unary nodes, an entire pedigree of past ancestors will be recorded. For instance it is possible to keep track of recombination events in the genealogy with a fully recorded pedigree. ARGON2 optionally allows retaining this kind of information, at the cost of a significant increase in required resources.

5.2.2 Simulation algorithm

ARGON2 proceeds backwards in time and processes one generation at a time, using an approach similar to that of ARGON. All present day samples are initialized as instances of the class *Individual*, which contains a list of *Block* objects, representing genetic materials travel along the genealogy from the present day samples. A *Block* stores the starting position, ending position, number of descendants, and pointers

to *Edges*. Throughout the simulation, *Block*s become increasingly fragmented and each segment become increasingly smaller due to coalescent and recombination events. A *Block* will only be freed from memory when it reaches the root. *Block* is the backbone of the DTWF simulation, whilst *Individual* may be used to enable additional optional features, such as pedigree tracking, diploid simulation, or sex-specific mutation and recombination maps. Diploid simulation mode allows an individual only to recombine in between two randomly chosen parental individuals, while in the haploid mode, an individual can have any number of parents.

ARGON2 allows the simulation of multiple populations (demes) with different sizes. Suppose there are K demes, each with population size $N_{e,k}$, with a generation specific pairwise-migration rate across demes. An *Individual* is stored in a 2-D array $\mathbf{C}^g_{k,i}$ for each generation g. The haploid DTWF simulator algorithm is described below.

Algorithm 1 Discrete time Wright Fisher haploid simulation algorithm with multiple populations, migration, and gene conversion.

Initialize: assign individuals of all K populations in $\mathbf{C}^0_{k,i}$ at generation $(g = 0)$
while existing blocks not reaching the root **do**
  $g \leftarrow g + 1$
  Initialize: $\mathbf{C}^g$ with an empty vector.
  for individual $\in \mathbf{C}^{(g-1)}$ **do**
    while crossover/non-crossover recombination **do**
      Process the recombinant (s).
    end while
    Draw a random integer $k \in [1, K]$ based on migration rate.
    Draw a random integer $i \in [0, N_{e,k}]$.
    if $\mathbf{C}^g_{i,k}$ is occupied **then**
      Coalesce the two individuals and record the events in *SimpleARG*.
      if any block reached the root **then** Free the block
      end if
    else
      $\mathbf{C}^g_{i,k} \leftarrow$ current individual.
    end if
  end for
end while

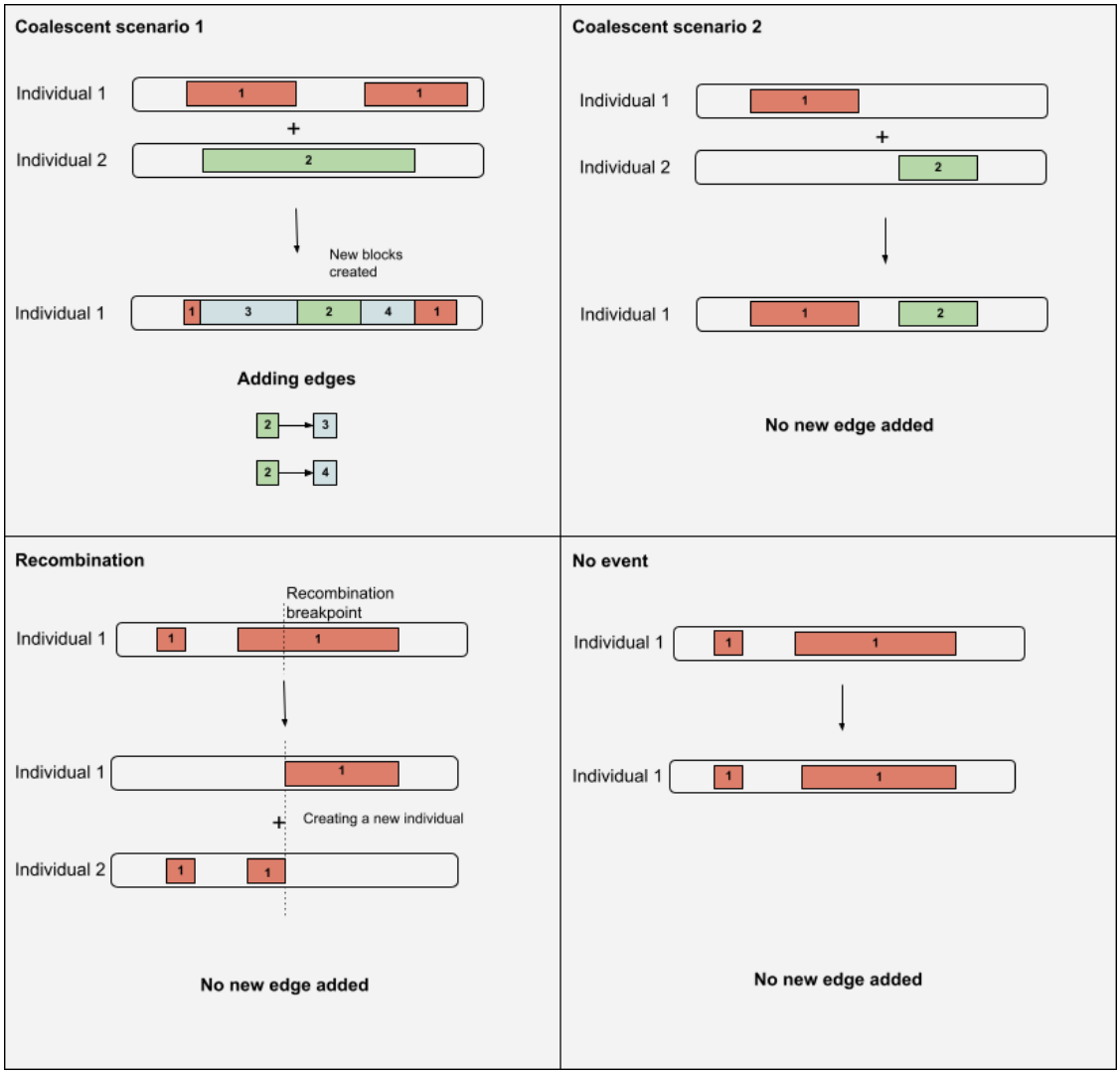


Figure 5.1: Schematic of coalescent and recombination implementation in ARGON2. Coalescent scenario 1 depicts two coalescing individuals with overlapping segments; coalescent scenario 2 shows two coalescing individuals without overlapping segments. Among all the cases, only coalescent with overlapping segments can create new edges.

5.2.3 The merging algorithm for coalescence events

When a coalescent event happens, the fragmented *Block*s of two coalesced individuals are merged. The merging operation is the most time-consuming part of the simulation, taking up more than 70% of the total run-time. For larger simulations, this proportion is expected to be even higher, as the fragmentation will be more severe. In Figure 5.2, we show that individuals have much more fragmented genome segments (i.e. more *Block*s) when the sample size, genome length or Ne are large.

ARGON2 uses the forward sweeping algorithm, used in ARGON [101], to merge

the coalesced blocks. When merging the *Block*s of two individuals, it scans the genome from left to right. *Block*s are stored in a linked list (data structure), which has $O(1)$ complexity for insertion and deletion of a given node. When merging two individuals with comparable numbers of blocks, the forward sweeping algorithm complexity is close to optimal and linear in terms of number of blocks. When individuals are more fragmented, i.e. high block count per individual, the forward sweeping algorithm is more efficient than binary tree structure (adopted e.g. in [95]), which require $\log(n)$ cost for insertions.

5.2.4 Recombination and gene conversion

Hudson's coalescent with recombination algorithm allows an individual genome to recombine multiple times and have more than two parents across two generations. In large-scale simulations, this may cause an inflation of lineages and distort the relatedness pattern [100–102]; in some extreme cases, this could make every pair of individuals second or third degree cousins. Nelson et al. (2020) [102] showed that the number of lineages in Hudson's coalescent algorithm tends to grow beyond the population size. In DTWF simulation, on the other hand, the number of lineages can never grow over the population size. DTWF models such as ARGON [101] and msprime-dtwf [102] increased the simulation accuracy by only allowing the recombination to happen back-and-forth between two ancestors in one generation. This approach is also adopted in ARGON2.

There are two types of DNA repair mechanisms during the meiotic recombination, namely, the cross-over (CO) and non-crossover (NCO) DNA repair. Both mechanisms are modelled in ARGON and ARGON2. Non-crossover recombination, often referred to as gene conversion, is not as commonly modelled, though it has been estimated to be an order of magnitude more likely to occur, compared to CO recombination [147]. The NCO gene conversion fills the double strand break (DSB) with a short homologous segment, usually from 50-1000 base pair [148], whereas CO recombination affects several mega-base long segments. NCO gene conversion could affect local LD [149], increase GC content over time [150].

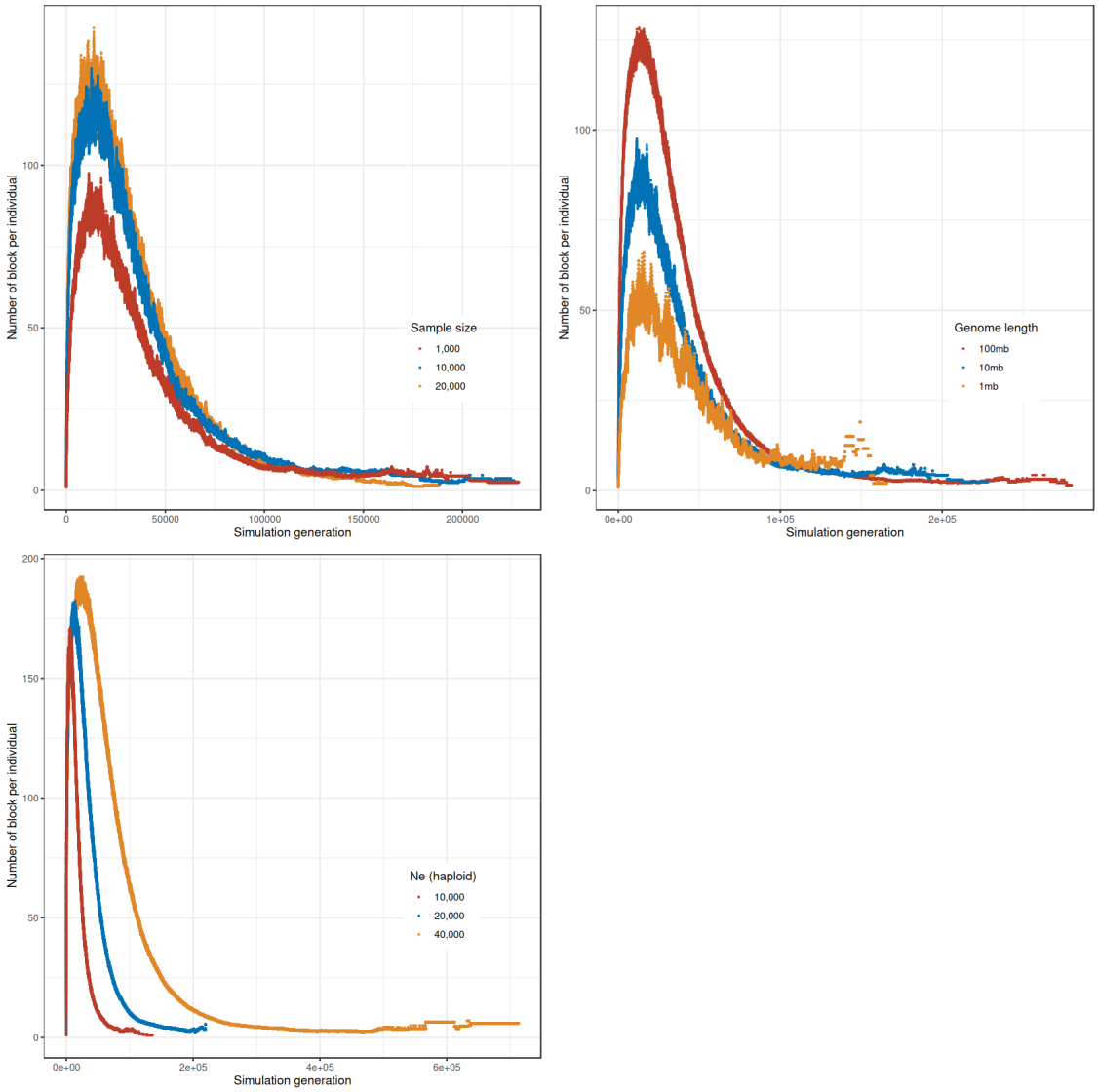


Figure 5.2: The fragmentation situation (Number of block per individual) using different parameters in each generation during the simulation. The upper left plot shows the simulation of a 10mb genome with varying sample sizes. The average number of blocks per individual increases increased sample size. The upper right plot shows the simulation of $1,000$ samples. By increasing the genome length, the average number of blocks per individual increases. The lower left plot shows the simulation of 100mb genome with $10,000$ samples, showing the increase of effective population size (Ne) can also increase the fragmentation.

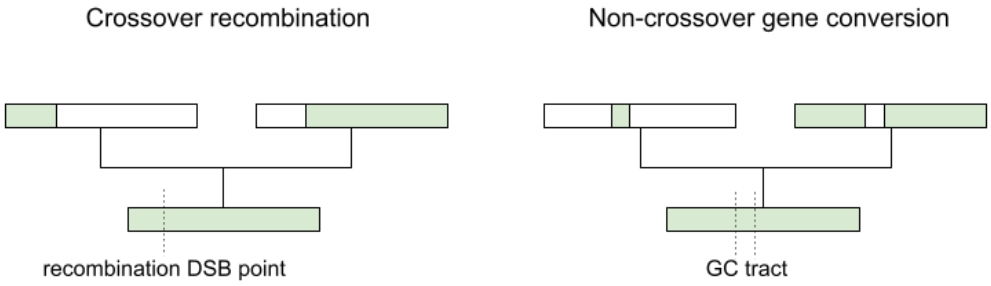


Figure 5.3: The schematic of crossover recombination and non-crossover gene conversion. DSB stands for double strand break.

To implement NCO gene conversion, we adopted an infinite sites approximation and assumed that the length of a NCO tract is exponentially distributed [151]. Suppose the per generation recombination rate is ρ and gene conversion rate is ρ_{GC}, to simulate the recombination with gene conversion, we first draw from an exponential distribution with the rate of $\rho + \rho_{GC}$ to determine the next break point for either CO and NCO recombination. Then we draw a sample from a Bernoulli distribution with the parameter of $\rho_{GC}/(\rho + \rho_{GC})$ to determine whether the "breaking" event is a gene conversion event. The operation of gene conversion is accomplished by two recombination events and a coalescent event back to the original individual, once the gene conversion tract length is determined (Figure 5.3). When the right boundary of the gene conversion tract exceeds the genome, the NCO gene conversion will be indistinguishable from the CO recombination event.

5.2.5 Mutation

Edges in an ARGON2 ARG represent ancestral relationships and allow retrieving the genomic extent (l) during which this relationship holds, as well as the height of the parent and child nodes involved. Suppose the per generation mutation rate is μ, we can simulate the number of mutation N_i on edge i using a Poisson distribution, $N_i \sim \text{Poi}(\mu l(h_p - h_c))$. Generating mutations on edges is a highly efficient operation, but determining which samples carry a mutation is time-consuming, and requires

significant optimization. A naive implementation requires the traversal using post-order tree traversal for every mutation. In ARGON2 we used the ARG traversal functions in ARG-Needle [146], which is based on the ARG traversal algorithm used in the ARGON simulator and enables efficiently computing the carriers of a set of mutations on the ARG.

5.3 ARGON2 validation and performance

5.3.1 Generating the ARG

We performed extensive simulations to verify the calibration of the ARGON2 simulator. To this end, we extracted a number of summary statistics from each simulation, including the number of segregating sites, heterozygosity, and average pairwise correlations. We used these statistics to test for significant deviations across different simulators, including ARGON [101], msprime v1.0 (Hudson algorithm) and msprime-dtwf v1.0 (DTWF ancestry simulation) [95]. We tested the following statistics (also see [101]):

- S: number of segregating sites in the sample.
- S_{rare}: number of polymorphic sites with minor allele frequency below 10% in the sample.
- S_{mid}: number of polymorphic sites with minor allele frequency between 10% and 40%.
- S_{common}: number of polymorphic sites with minor allele frequency higher than 40% in the sample.
- singleton: number of sites with one mutation in the samples.
- doubleton: number of sites with two mutations in the samples.
- θ heterozygosity: computed as in (5.1)
- r_{50}: average pairwise correlation of markers at distance between 0 and 50Kb.

- r_{100}: average pairwise correlation of markers at distance between 50 and 100Kb.
- r_{150}: average pairwise correlation of markers at distance between 100 and 150Kb.
- r^2_{50}: average pairwise squared correlation of markers at distance between 0 and 50Kb.
- r^2_{100}: average pairwise correlation of markers at distance between 50 and 100Kb.
- r^2_{150}: average pairwise correlation of markers at distance between 100 and 150Kb.

For n samples with the genome length of l, and allele frequency for position i is p_i, the heterozygosity θ is calculated as follows,

$$\theta = \frac{n}{(n-1)} \sum_{i=1}^{l} \frac{2p_i(1-p_i)}{l} \tag{5.1}$$

The number of segregating sites and heterozygosity were computed using bcftools [121] and the correlation was calculated using PLINK 1.9 [11], accounting for all variants above minor allele frequency 1%. We used the aforementioned four simulators to perform $1,000$ independent simulations of 10 Mb genome with 50 haploid samples. The recombination rate was set to 10^{-8} per basepair, per generation and the mutation rate was set to 2×10^{-8} per basepair, per generation. The results are shown in the tables below, where we report the mean value, standard error (S.E.), and Z-score of each comparison. Z-scores ($-1.94 < Z < 1.94$) that are corresponding to p-value lower than 0.05 are marked as red for caution, though multiple testing should be taken into account when evaluating these results.

The tests were carried out for four demographic models, including a constant population size model, exponential expansion model, bottleneck model, and island model [101]. We chose a relatively small sample size, where the Hudson's coalescent algorithm [97] is a reasonable approximation of the discrete-time Wright Fisher

process [89], so that all simulators should produce comparable results regardless of the model they adopt, as less distortion would be made due to polytomy [96]. The tables below show ARGON2 results match all the other three simulators.

Measurement	ARGON2 (S.E.)	argon1 (S.E. | Z score)	msprime (S.E. | Z score)	msprime-dtwf (S.E. | Z score)
S	4.14×10^3 (5.94)	4.15×10^3 (6.11|-0.2)	4.13×10^3 (5.66|1.69)	4.14×10^3 (5.86|0.16)
θ	7.97×10^{-4} (1.67×10^{-6})	7.97×10^{-4} (1.72×10^{-6}|-0.2)	7.94×10^{-4} (1.66×10^{-6}|1.25)	7.97×10^{-4} (1.64×10^{-6}|-0.13)
singleton	8×10^2 (2.27)	8.01×10^2 (2.19|-0.48)	7.99×10^2 (2.17|0.19)	8.01×10^2 (2.28|-0.37)
doubleton	3.98×10^2 (1.53)	4.01×10^2 (1.6|-1.28)	3.96×10^2 (1.56|1.01)	3.97×10^2 (1.55|0.3)
S_{rare}	2.34×10^3 (4.11)	2.34×10^3 (4.1|-0.24)	2.33×10^3 (3.83|1.07)	2.34×10^3 (4.08|0.43)
S_{mid}	1.46×10^3 (3.78)	1.46×10^3 (3.68|-0.37)	1.46×10^3 (3.53|0.82)	1.46×10^3 (3.74|-0.37)
S_{common}	3.1×10^2 (1.59)	3.08×10^2 (1.64|0.62)	3.06×10^2 (1.59|1.35)	3.08×10^2 (1.54|0.52)
r_{50}	6.34×10^{-2} (3.29×10^{-4})	6.39×10^{-2} (1.13×10^{-4}|-1.57)	6.42×10^{-2} (6.02×10^{-4}|-1.13)	6.39×10^{-2} (1.36×10^{-4}|-1.4)
r_{100}	3.6×10^{-2} (1.73×10^{-4})	3.59×10^{-2} (5.77×10^{-5}|0.25)	3.6×10^{-2} (3.1×10^{-4}|-0.24)	3.6×10^{-2} (7.03×10^{-5}|-0.25)
r_{150}	3.08×10^{-2} (1.5×10^{-4})	3.07×10^{-2} (4.77×10^{-5}|0.73)	3.08×10^{-2} (2.49×10^{-4}|0.05)	3.08×10^{-2} (5.48×10^{-5}|0.23)
r_{50}^2	5.18×10^{-1} (2.78×10^{-3})	5.14×10^{-1} (9.41×10^{-4}|1.55)	5.01×10^{-1} (4.43×10^{-3}|3.21**)	5.12×10^{-1} (1.09×10^{-3}|1.96)
r_{100}^2	5.31×10^{-1} (2.32×10^{-3})	5.27×10^{-1} (8.4×10^{-4}|1.87)	5.21×10^{-1} (3.94×10^{-3}|2.22*)	5.26×10^{-1} (1.02×10^{-3}|2.12*)
r_{150}^2	5.54×10^{-1} (2.24×10^{-3})	5.51×10^{-1} (7.28×10^{-4}|1.46)	5.46×10^{-1} (4.08×10^{-3}|1.67)	5.49×10^{-1} (9.24×10^{-4}|2.08*)

Table 5.1: Simulations of 50 haploid samples in a constant population, with the effective population size of 20,000. The simulation is carried out with 50 haploid samples on 5 Mb genome for 1,000 times. The recombination rate is 10^{-8} recombinations per bp per generation, and the mutation rate is 2×10^{-8} mutations per bp per generation. Z score corresponding P-value lower than 0.05 and 0.0045 (Bonferroni correction) are marked as red with one asterisk and two asterisks, respectively.

Measurement	ARGON2 (S.E.)	argon1 (se|Z)	msprime (S.E.|Z)	msprime-dtwf (S.E.|Z)
S	2.99×10^3 (5.66)	2.98×10^3 (5.91|0.47)	2.99×10^3 (5.9|0.13)	2.98×10^3 (5.93|0.53)
θ	7.39×10^{-4} (1.64×10^{-6})	7.4×10^{-4} (1.72×10^{-6}|-0.22)	7.41×10^{-4} (1.74×10^{-6}|-0.54)	7.38×10^{-4} (1.72×10^{-6}|0.47)
singleton	2.32×10^2 (1.92)	2.3×10^2 (1.91|0.81)	2.31×10^2 (1.88|0.44)	2.29×10^2 (1.95|1.07)
doubleton	1.72×10^2 (1.7)	1.74×10^2 (1.72|-0.54)	1.71×10^2 (1.67|0.71)	1.74×10^2 (1.75|-0.73)
S_{rare}	1.22×10^3 (4.04)	1.21×10^3 (4.16|1.21)	1.22×10^3 (4.16|0.69)	1.22×10^3 (4.06|0.02)
S_{mid}	1.43×10^3 (3.87)	1.43×10^3 (4.08|0.2)	1.43×10^3 (4.02|0.35)	1.42×10^3 (3.93|1.49)
S_{common}	3.03×10^2 (1.64)	3.08×10^2 (1.69|-2.04*)	3.08×10^2 (1.66|-2.02*)	3.07×10^2 (1.66|-1.56)
r_{50}	1.18×10^{-1} (3.95×10^{-4})	1.19×10^{-1} (3.17×10^{-4}|-1.03)	1.19×10^{-1} (6.73×10^{-4}|-0.3)	1.19×10^{-1} (4.63×10^{-4}|-0.24)
r_{100}	7.39×10^{-2} (2.9×10^{-4})	7.44×10^{-2} (2.26×10^{-4}|-1.29)	7.42×10^{-2} (5.02×10^{-4}|-0.42)	7.39×10^{-2} (3.49×10^{-4}|0.07)
r_{150}	6.43×10^{-2} (2.58×10^{-4})	6.44×10^{-2} (2.03×10^{-4}|-0.22)	6.43×10^{-2} (4.8×10^{-4}|0.05)	6.4×10^{-2} (3.09×10^{-4}|0.85)
r_{50}^2	4.72×10^{-1} (1.48×10^{-3})	4.71×10^{-1} (1.18×10^{-3}|0.42)	4.72×10^{-1} (2.74×10^{-3}|-0.08)	4.71×10^{-1} (1.74×10^{-3}|0.34)
r_{100}^2	4.26×10^{-1} (1.67×10^{-3})	4.26×10^{-1} (1.3×10^{-3}|0.27)	4.26×10^{-1} (2.92×10^{-3}|0.08)	4.25×10^{-1} (2.02×10^{-3}|0.43)
r_{150}^2	4.31×10^{-1} (1.69×10^{-3})	4.3×10^{-1} (1.39×10^{-3}|0.47)	4.31×10^{-1} (3.1×10^{-3}|0.01)	4.3×10^{-1} (2.11×10^{-3}|0.25)

Table 5.2: Simulations of 50 haploid samples with a population bottleneck. The effective population size shrank from generation 500 to a present day size of 2,000, starting from 20,000 haploid individuals. The simulation is carried out with 50 haploid samples on 5 Mb genome for 1,000 times. The recombination rate is 10^{-8} recombinations per bp per generation, and the mutation rate is 2×10^{-8} mutations per bp per generation. Z score corresponding P-value lower than 0.05 and 0.0045 (Bonferroni correction) are marked as red with one asterisk and two asterisks, respectively.

Measurement	ARGON2 (S.E.)	argon1 (S.E. | Z score)	msprime (S.E. | Z score)	msprime-dtwf (S.E. | Z score)
S	4.7×10^3 (5.76)	4.71×10^3 (5.91|-1.08)	4.69×10^3 (5.8|1.27)	4.69×10^3 (5.65|1.37)
θ	8.07×10^{-4} (1.7×10^{-6})	8.11×10^{-4} (1.63×10^{-6}|-1.39)	8.07×10^{-4} (1.71×10^{-6}|0.23)	8.06×10^{-4} (1.65×10^{-6}|0.53)
singleton	1.31×10^3 (2.21)	1.31×10^3 (2.28|-1.43)	1.31×10^3 (2.15|1.38)	1.3×10^3 (2.2|1.87)
doubleton	4.37×10^2 (1.35)	4.36×10^2 (1.32|0.28)	4.37×10^2 (1.36|-0.23)	4.38×10^2 (1.36|-0.65)
S_{rare}	2.89×10^3 (3.79)	2.9×10^3 (4.07|-0.5)	2.89×10^3 (3.87|1.53)	2.89×10^3 (3.84|1.36)
S_{mid}	1.46×10^3 (3.8)	1.47×10^3 (3.65|-0.34)	1.46×10^3 (3.72|0.86)	1.46×10^3 (3.55|0.83)
S_{common}	3.06×10^2 (1.62)	3.1×10^2 (1.55|-1.84)	3.08×10^2 (1.66|-0.97)	3.06×10^2 (1.57|-0.24)
r_{50}	5.17×10^{-2} (5.51×10^{-4})	5.14×10^{-2} (7.65×10^{-5}|0.62)	5.12×10^{-2} (2.17×10^{-4}|0.84)	5.17×10^{-2} (5.5×10^{-4}|0.07)
r_{100}	2.88×10^{-2} (2.01×10^{-4})	2.88×10^{-2} (3.13×10^{-5}|0.07)	2.87×10^{-2} (9.58×10^{-5}|0.17)	2.88×10^{-2} (2.54×10^{-4}|-0.12)
r_{150}	2.47×10^{-2} (1.25×10^{-4})	2.48×10^{-2} (2.33×10^{-5}|-0.43)	2.48×10^{-2} (6.88×10^{-5}|-0.6)	2.49×10^{-2} (2.23×10^{-4}|-0.56)
r^2_{50}	5.39×10^{-1} (5.14×10^{-3})	5.29×10^{-1} (8.24×10^{-4}|1.87)	5.32×10^{-1} (2.63×10^{-3}|1.14)	5.27×10^{-1} (7.56×10^{-3}|1.3)
r^2_{100}	5.73×10^{-1} (4×10^{-3})	5.68×10^{-1} (6.45×10^{-4}|1.19)	5.68×10^{-1} (2.2×10^{-3}|1.05)	5.68×10^{-1} (5.13×10^{-3}|0.85)
r^2_{150}	5.98×10^{-1} (2.26×10^{-3})	5.97×10^{-1} (4.94×10^{-4}|0.25)	5.97×10^{-1} (1.59×10^{-3}|0.29)	5.93×10^{-1} (5.29×10^{-3}|0.79)

Table 5.3: Simulations of 50 haploid samples in an exponentially expanding population, with the effective population size expanding from 20,000 to 200,000, from the 500^{th}generation to a present day. The simulation is carried out with 50 haploid samples on 5 Mb genome for 1,000 times. The recombination rate is 10^{-8} recombinations per bp per generation, and the mutation rate is 2×10^{-8} mutations per bp per generation. Z score corresponding P-value lower than 0.05 and 0.0045 (Bonferroni correction) are marked as red with one asterisk and two asterisks, respectively..

Measurement	ARGON2 (S.E.)	argon1 (S.E. | Z score)	msprime (S.E. | Z score)	msprime-dtwf (S.E. | Z score)
S	4.03×10^3 (5.92)	4.04×10^3 (5.85|-0.97)	4.03×10^3 (6.2|0.24)	4.03×10^3 (5.92|0.25)
θ	8.09×10^{-4} (1.65×10^{-6})	8.1×10^{-4} (1.74×10^{-6}|-0.34)	8.09×10^{-4} (1.72×10^{-6}|-0.09)	8.09×10^{-4} (1.72×10^{-6}|0.07)
singleton	8.13×10^2 (2.38)	8.16×10^2 (2.32|-0.93)	8.11×10^2 (2.44|0.34)	8.07×10^2 (2.32|1.8)
doubleton	4.07×10^2 (1.62)	4.07×10^2 (1.58|-0.06)	4.07×10^2 (1.63|-0.03)	4.09×10^2 (1.67|-1.11)
S_{rare}	2.19×10^3 (4.13)	2.19×10^3 (3.77|-1.27)	2.18×10^3 (4.23|0.17)	2.18×10^3 (4|0.63)
S_{mid}	1.5×10^3 (3.71)	1.5×10^3 (3.68|-0.35)	1.49×10^3 (3.7|0.43)	1.5×10^3 (3.75|-0.96)
S_{common}	3.1×10^2 (1.65)	3.09×10^2 (1.62|0.62)	3.11×10^2 (1.69|-0.25)	3.07×10^2 (1.61|1.55)
r_{50}	7.09×10^{-2} (6.51×10^{-4})	7.12×10^{-2} (1.19×10^{-4}|-0.5)	7.18×10^{-2} (8.32×10^{-4}|-0.85)	7.16×10^{-2} (2.33×10^{-4}|-1.02)
r_{100}	4.2×10^{-2} (3.52×10^{-4})	4.16×10^{-2} (6.28×10^{-5}|1.15)	4.19×10^{-2} (3.38×10^{-4}|0.23)	4.19×10^{-2} (1.23×10^{-4}|0.27)
r_{150}	3.66×10^{-2} (2.76×10^{-4})	3.62×10^{-2} (5.35×10^{-5}|1.46)	3.64×10^{-2} (3.99×10^{-4}|0.24)	3.64×10^{-2} (1.03×10^{-4}|0.44)
r_{50}^2	5.01×10^{-1} (4.71×10^{-3})	5.02×10^{-1} (9.18×10^{-4}|-0.23)	4.91×10^{-1} (9.03×10^{-3}|0.94)	5.02×10^{-1} (1.75×10^{-3}|-0.12)
r_{100}^2	5.1×10^{-1} (4.59×10^{-3})	5.08×10^{-1} (9.05×10^{-4}|0.33)	5.06×10^{-1} (8.89×10^{-3}|0.38)	5.07×10^{-1} (1.64×10^{-3}|0.5)
r_{150}^2	5.29×10^{-1} (4.16×10^{-3})	5.29×10^{-1} (8.38×10^{-4}|-0.22)	5.28×10^{-1} (8.77×10^{-3}|0.03)	5.3×10^{-1} (1.43×10^{-3}|-0.23)

Table 5.4: Simulations of 80 haploid samples in an Island model consists of four demems with symetric pairwise migration rate of 0.001 individuals per generation. Each deme was inisialized with 20 samples in the present day. The simulation is carried out with 50 haploid samples on 5 Mb genome for 1,000 times. The recombination rate is 10^{-8} recombinations per bp per generation, and the mutation rate is 2×10^{-8} mutations per bp per generation. Z score corresponding P-value lower than 0.05 and 0.0045 (Bonferroni correction) are marked as red with one asterisk and two asterisks, respectively.

5.3.2 Performance on a single constant sized population

Like ARGON, ARGON2 adopted a two-stage simulation approach, consisting of ARG building and mutation generation. We compared ARGON2 ARG building, mutation + I/O, and the total run-time (Figure 5.4) against ARGON1 (version 0.1.160415 with OpenJDK 11.0.15), msprime (v1.2.0) and its DTWF module msprime-dtwf, with varying genome length and effective population size.

All benchmark experiments were run on a single Intel Xeon E5-2690 CPU. In particular, for running ARGON1, the Java flag "-XX:+UseSerialGC" was used to ensure a single processor simulation. The total run-time and memory were measured using the Bash command *time*. To test the ARG building time, *simulate_ancestry* function in ARGON1, *generate_ARG* function in ARGON2 and *sim_ancestry* function in msprime were used for respective simulators. tskit provides two approaches to generate mutation data in VCF format, a Python function *write_vcf* or the tskit command line tool. Although it requires writing the out "TreeSequence" output, the latter is at least 10 times faster, was therefore adopted it in our benchmarks. Since the msprime mutation generation and I/O processes are inseparable, we measured the mutation generation and I/O together. In particular for msprime, we measured mutation and I/O in the following way.

```
# Simulating the mutation
mts = msprime.sim_mutations(ts, rate=mu)
mts.dump(output + ".tsk")
cmd = "python3 -m tskit vcf " + output + ".tsk | bgzip -c > " +
    output
os.system(cmd)
```

The performance of ARGON2 is comparable to that of msprime when simulating a small population size and short genome length. For larger genome size, ARGON2 still shows better performance than msprime-dtwf and ARGON1. For much larger simulation, ARGON2 is slower than msprime-dtwf, but still around 2 times faster than ARGON1. With increased the genome length and effective population size, the number of *Block*s increases exponentially, namely, the genomes become more fragmented, likely causing the drop in performance. In future work, we plan to implement defragmentation operations, which are likely to lead to an improvement

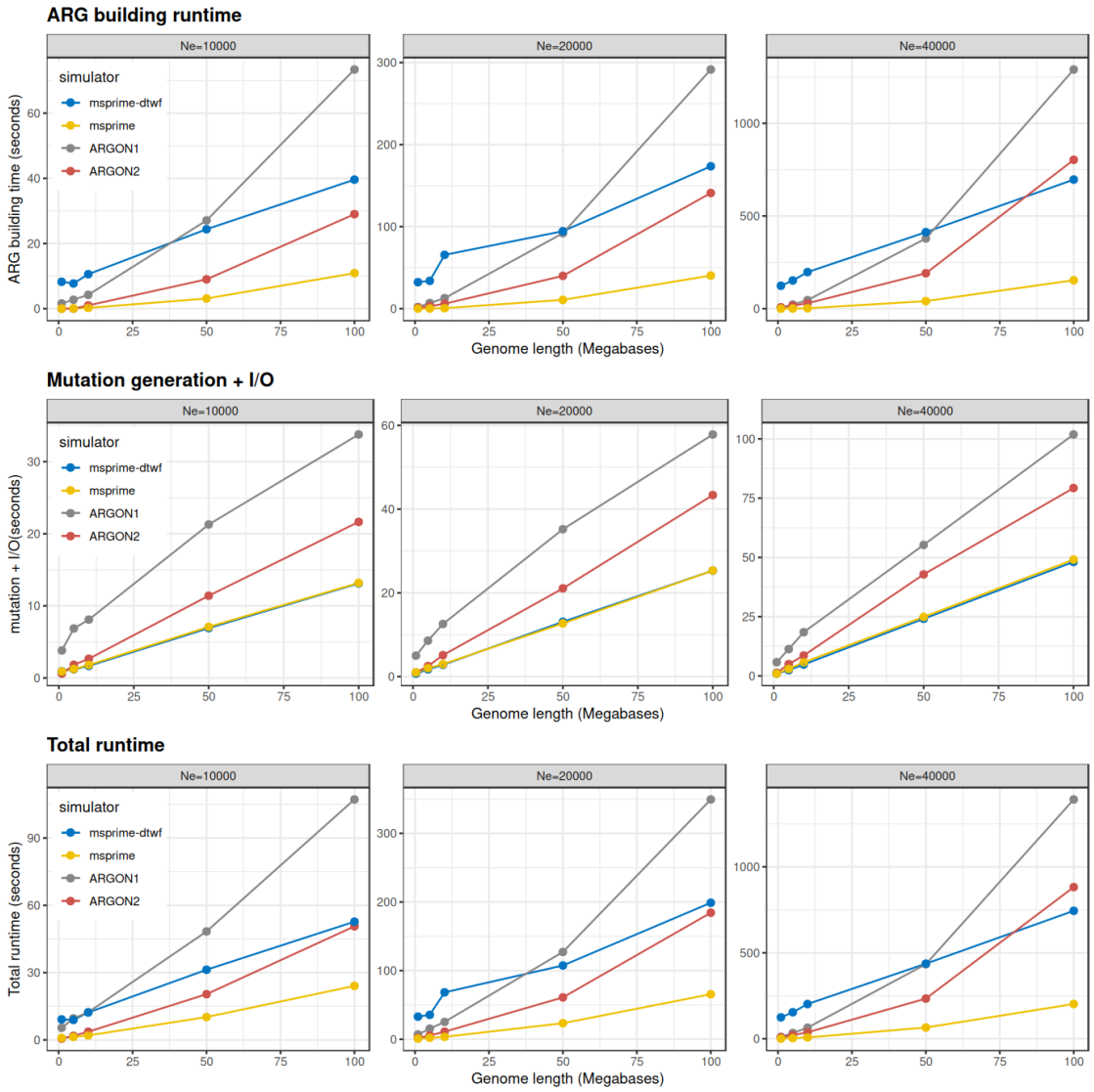


Figure 5.4: The runtime of ARGON1, ARGON2, msprime, and msprime-dtwf, breaking down in ARG building (the first row), mutation generation + I/O (the second row) and the total runtime (the third row). The performance is measured by simulating 1,000 haploid samples with a varying genome length (x-axis) and effective population size (columns). The population size Ne is specified on the top of each panel.

in performance. The mutation performance of ARGON and ARGON2 is currently slower than msprime and msprime-dtwf (Figure 5.4), likely due to the same fragmentation issue. In terms of the total run-time, ARGON2 is faster than msprime-dtwf for small simulations and has comparable performance for larger simulations.

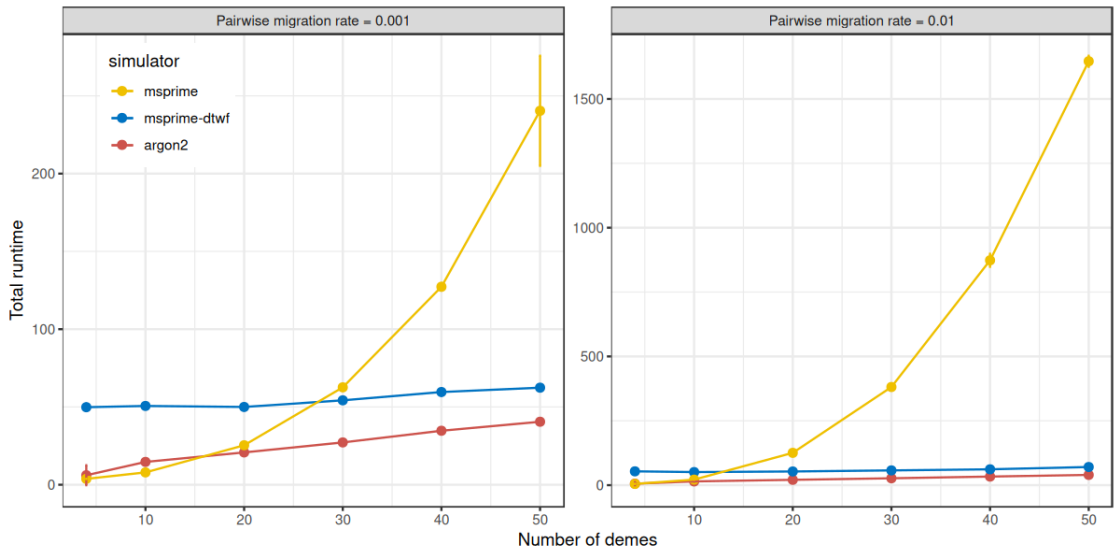


Figure 5.5: The performance (total runtime) comparison of ARGON2, msprime coalescent model, and msprime discrete time Wright Fisher model (msprime-dtwf) in the Island model simulation up to 50 demes. The left plot shows a low pairwise migration rate and the right plot shows a high pairwise migration rate.

5.3.3 Performance on multiple populations with migration

Unlike in the case of Hudson's coalescent algorithm, simulation of migration events under the DTWFs creates very little computational overhead. The difference in performance becomes noticeable when the number of demes and migration rates in the simulated demographic model increase. We simulated an island model, where we allowed all-to-all pairwise migration, with a varying number of demes. We set the total effective population size as 20,000 (haploid), equally divided among all the demes, and sampled 100 individuals from each deme. In this setting, when the number of demes is increased, the total sample size also increases. Figure 5.5 shows that the runtime of the Hudson's algorithm (msprime) rapidly increases with the number of demes, while that of DTWF stays constant. As a result, msprime becomes extremely inefficient compared to DTWF models in complex demographic models that involve multiple demes (e.g. > 10) and high migration rates.

5.4 ARG augmentation: simulating new samples on an existing ARG

5.4.1 New data structure: Genealogy and fixed blocks

In order to facilitate generating new samples on an existing ARG, we developed a new ARG data structure, referred to as *Genealogy* in the remainder of this chapter. It contains a collection of interconnected *Node*s at discrete generations. In the ARG augmentation algorithm, numerous operations of querying lineages over a given segment were required.

Such operation is efficiently achieved using an interval tree, which requires $O(\log n)$ computation for queries of this kind. We therefore used interval tree to store parent and children information within nodes. *Genealogy* can be converted from and to the ARGON2 *SimpleARGs* format, as well as ARG-Needle and tskit formats. The conversion overhead is an order of magnitude smaller than the simulation algorithm, and therefore negligible. The *Genealogy* data structure was used to store the existing ARG as a reference, which is not modified during the process of ARG augmentation simulations. *Genealogy* also implements a built-in ARG traversal algorithm and a naive mutation generation function, in contrast to the more efficient ARG-Needle mutation function used in previously described simulations.

In our ARG augmentation algorithm, it is important to differentiate between genetic material that is already present in the existing ARG and new material that belongs to new samples to be added to the ARG. We refer to these former type of genetic materials as fixed *Block*s, while free *Block* refers to novel simulated material. A free *Block* is simulated following the DTWF procedure (Algorithm 1), while a fixed block is only allowed to follow the lineages specified in the existing ARG.

Whenever a free *Block* coalesces with a fixed block, it becomes fixed. From that time on, during the simulation, such a block will track the lineages of the fixed *Block*, until it reaches a root of the existing ARG.

5.4.2 Simulation heuristics

In this section, we will address four major issues that we faced, when simulating new samples over an augmented ARG, namely, recovering hidden recombination events, the presence of incompatible lineages, information loss regarding individuals,

and the conversion between continuous and discrete timescale. We develop heuristic strategies to efficiently address each of these issues. These heuristics may introduce biases to the ARG augmentation procedure, compared to a standard DTWF simulation. The main application of ARG augmentation, however, is in the simulation of new samples starting from an ARG inferred from real data. In this context, the biases introduced by these heuristic strategies are largely overshadowed by the biases introduced when an ARG is inferred from data (Discussed in 5.4.4).

Hidden recombinations

The exact time of recombination events is often not recorded in simulated ARGs [95, 101] or in inferred ARGs [63, 146, 152], as knowledge of the time of recombination events does not influence downstream operations such as mutation generations. In ARGON, ARGON2, and ARG-Needle inferred ARGs, recombination events are implicitly assumed to occur between neighbouring recombination and coalescent nodes. Specifically, when an ARG contains a recombination node (where multiple parents are present), one can only ascertain that such a recombination occurred within an interval of time $(t_a, t_b]$, where t_a is the time of the recombination node, and t_b is the time of the closest coalescence node backwards in time. In such cases, the recombination time t_{rec} can be assumed to be uniformly distributed within the region of $(t_a, t_b]$, i.e. $P(t_{\text{rec}}|\text{recombination} \in (t_a, t_b]) \sim \text{unif}(t_a, t_b)$. Note that this approach ignores the case of lineages recombining and coalescing multiple times, which are expected to occur frequently [106, 107], as this information is permanently lost in these encodings of an ARG.

Incompatible lineage merging

The second issue is that random coalescent processes could cause the merging of nodes that are incompatible to the reference genealogy. Specifically, during a simulation sampling from the DTWF backwards in time, two *Individual*s who select the same parent may be merged with no restrictions. However, during ARG augmentation there are additional constraints to consider, in order to make sure that the existing ARG (composed of fixed *Block*s) is preserved. Specifically, the algorithm

should not allow coalescing *Individual*s in a way that would lead to the merging of fixed lineages that are incompatible with the existing ARG. To ensure this, we adopted a heuristic approach where all incompatible blocks are shifted into a new *Individual* whenever a conflict arises. These individuals are chosen at random; if the choice of individual continues to lead to a conflict, the random sampling is repeated until no incompatible blocks are found. This approach is equivalent to introducing recombination events at the beginning and the end of each conflicting block.

No Individual information

In an inferred ARG, genetic segments do not come with the Individual information, namely we don't have the information on how the inferred genetic segments travel together within Individuals of an underlying population pedigree. In order to enable performing data augmentation using an inferred ARG, we needed to somehow assign ancestral genetic blocks to Individuals. We achieved this by assigning inferred genetic segments to a unique Individual when initializing the simulation. In cases where the number of inferred segments exceeds the number of effective Individuals in an assumed demographic model, we randomly coalesce the Individuals using the coalescent paradigm discussed above for the case of incompatible lineage merging.

Continuous scale vs discrete scale

Algorithms for genealogical inference produce ARGs using a continuous timescale, while in the DTWF process, Individuals and Blocks belong to discrete generations. To perform data augmentation using inferred ARGs, we therefore need to perform a discretization step. This discretization rounds the time assignment of each block, pushing up and down the nodes to match a discrete timescale. In some cases, this leads to the formation of polytomies (coalescence events involving more than 2 Individuals).

5.4.3 Simulation results on simulated ARGs

Unlike in the case of an inferred ARG, simulated ARG sampled using ARGON2 are on a discrete timescale and preserve the "Individual" structure. We evaluated the

	ARGON2 (S.E.)	ARG augmentation (S.E.)	Z-score
S	4.5×10^3 (2.82×10^1)	4.52×10^3 (2.23×10^1)	-0.57
θ	6.56×10^{-4} (6.78×10^{-6})	6.67×10^{-4} (5.97×10^{-6})	-1.19
Singleton	6.67×10^2 (7.52)	6.75×10^2 (7.94)	-0.76
Doubleton	3.34×10^2 (5.92)	3.24×10^2 (6.41)	1.18
S_{rare}	3.06×10^3 (2.14×10^1)	3.05×10^3 $(1.64 \times 10^1$	0.37
S_{mid}	1.18×10^3 (1.49×10^1)	1.2×10^3 (1.28×10^1)	-0.99
S_{common}	2.59×10^2 (5.81)	2.7×10^2 (6.2)	-1.34
R_{50}	4.58×10^{-2} (3.76×10^{-4})	4.65×10^{-2} (3.24×10^{-4})	-1.3
R_{100}	1.74×10^{-2} (1.82×10^{-4})	1.75×10^{-2} (1.56×10^{-4})	-0.23
R_{150}	1.27×10^{-2} (1.17×10^{-4})	1.26×10^{-2} (1.52×10^{-4})	0.63
R^2_{50}	5.44×10^{-1} (3.47×10^{-3})	5.42×10^{-1} (2.68×10^{-3})	0.43
R^2_{100}	5.72×10^{-1} (1.8×10^{-3})	5.68×10^{-1} (2.15×10^{-3})	1.53
R^2_{150}	5.69×10^{-1} (2.04×10^{-3})	5.66×10^{-1} (2.7×10^{-3})	0.68

Table 5.5: Verifying the ARG augmentation algorithm to ARGON2 simulations. The simulated ARG had 500 samples and 10 new samples were added to it. The ARG augmentation simulation results were compared to an independent simulation of ARGON2 with 510 samples (ARGON2). Z-scores that are corresponding to P-value lower than 0.05 are marked as red for caution, otherwise as green. Both simulations used a constant population size of 20,000 haploid samples.

	ARGON2 (S.E.)	ARG augmentation (S.E.)	Z-score
S	4.94×10^4 (6.13×10^1)	5.25×10^4 (300)	-10.3
θ	6.58×10^{-4} (1.65×10^{-6})	6.58×10^{-4} $(1.45 \times 10^{-6}$	-0.24
Singleton	6.64×10^3 (2.43×10^1)	8.03×10^3 (1.76×10^2)	-7.85
Doubleton	3.31×10^3 (1.5×10^1)	4.24×10^3 (7.6×10^1)	-12.04
S_{rare}	3.49×10^4 (5.93×10^1)	3.81×10^4 (3.1×10^2)	-9.99
S_{mid}	1.18×10^4 (3.14×10^1)	1.18×10^4 (3.6×10^1)	-0.11
S_{common}	2.66×10^3 (1.64×10^1)	2.65×10^3 (1.44×10^1)	0.11
R_{50}	4.28×10^{-2} (1.06×10^{-4})	3.92×10^{-2} (1.76×10^{-4})	17.44
R_{100}	1.54×10^{-2} (3.98×10^{-5})	1.27×10^{-2} (6.58×10^{-5})	35.16
R_{150}	1.07×10^{-2} (2.7×10^{-5})	8.67×10^{-3} (4.95×10^{-5})	36.5
R^2_{50}	5.45×10^{-1} (9.37×10^{-4})	5.46×10^{-1} (1.51×10^{-3})	-0.36
R^2_{50}	5.77×10^{-1} (3.75×10^{-4})	5.72×10^{-1} (8.34×10^{-4})	5.15
R^2_{150}	5.72×10^{-1} (4.74×10^{-4})	5.65×10^{-1} (8.49×10^{-4})	6.83

Table 5.6: Verifying the ARG augmentation algorithm by comparing metrics to ARGON2 simulations, where the simulated ARG has 500 samples and 500 new samples are added to it (ARG augmentation). These metrics are compared to an independent simulation of ARGON2 with 1000 samples (ARGON2). Z-scores that are corresponding to P-value lower than 0.05 are marked as red for caution, otherwise as green. Both simulations used a constant population size of 20,000 haploid samples.

performance of our ARG augmentation algorithm by sampling new individuals on top of an existing simulated ARG output by ARGON2. The simulated ARG still presents issues linked to the problems of incompatible lineage merging and missing recombination timings. We refer to the simulated ARG as the "old ARG" on top of which we aim to simulate new samples. Ideally, the augmented ARG should have the same characteristics, measured using summary statistics, as an ARG of the same total size simulated in one go using the same parameters.

We created the "old ARG" with 500 samples, using ARGON2. We devised two ARG augmentation scenarios: in the first simulation, we added 10 samples to the reference ARG (Table 5.5) and in the second scenario, we added 500 samples (Table 5.6). The ARG augmentation with small new sample sizes has little distortion of the property of simulated ARG, whereas the distortion is significant when the new sample size is big in relation to the "old ARG". However, in the next section, we will show that the biases introduced by the heuristics are still secondary, in comparison to the biases introduced by the ARG inferences.

5.4.4 Simulation results on inferred ARG

Rather than simulating on a true ARG, in this section, we aim to simulate new samples on an inferred ARG. First, we generated synthetic sequencing data and then inferred the ARG using the ARG-Needle algorithm [146]. Specifically, we simulated $1,000$ haploid samples using msprime with a constant population size ($20,000$ haploid samples), to be used as a reference panel. We aim to compare the performance when simulating $1,000$ additional samples using HAPGEN2 and ARGON2's ARG augmentation based on an ARG-Needle inferred ARG. To infer the ARG using ARG-Needle we proceeds as follows,

(1) We built a set of decoding quantities (used by the ASMC algorithm [153] within ARG-Needle) assuming a constant demographic model of size $N_e = 20,000$ and providing the reference panel encoded in a VCF file. (2) We inferred the ARG using ARG-Needle in sequencing mode. (3) We performed "ARG normalization" of the inferred ARG [146].

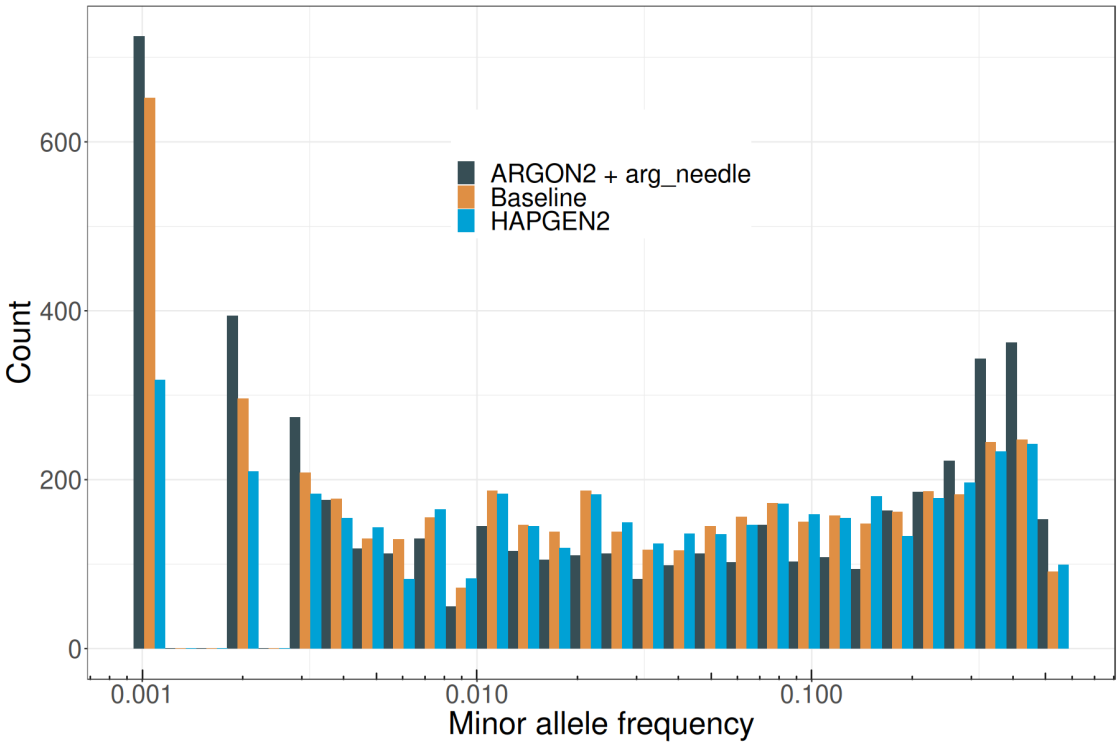


Figure 5.6: The simulated allele frequency spectrum comparison between ARGON2 ARG augmentation and HAPGEN2. The reference panel consists of 1,000 haploid samples was created using msprime. The inferred reference panel ARG used by ARGON2 is produced by ARG-Needle. The allele frequency was obtained by 1,000 additional samples simulated by ARGON2 and HAPGEN2, respectively. The baseline allele frequency is the reference panel allele frequency.

Figure 5.6 shows ARGON2 using ARG-Needle inferred ARG had better matching to the reference allele frequency for rare allele frequency bins compared to HAPGEN2, though with a sizeable distortion for common variants. Copying rare variants using the Li-Stephens model is often challenging, it is expected that ARGON2 approach is superior to the Li-Stephens approach, especially simulating with larger sample size where more rare variants could emerge. We also noticed that that the distortion of the common variant count is largely due to the bias introduced by ARG inference. To demonstrate this, we simulated directly on the ARG-Needle inferred reference ARG, by randomly assigning mutations to existing edges. In Figure 5.7, we show the excess common variants is an artefact without ARG augmentation.

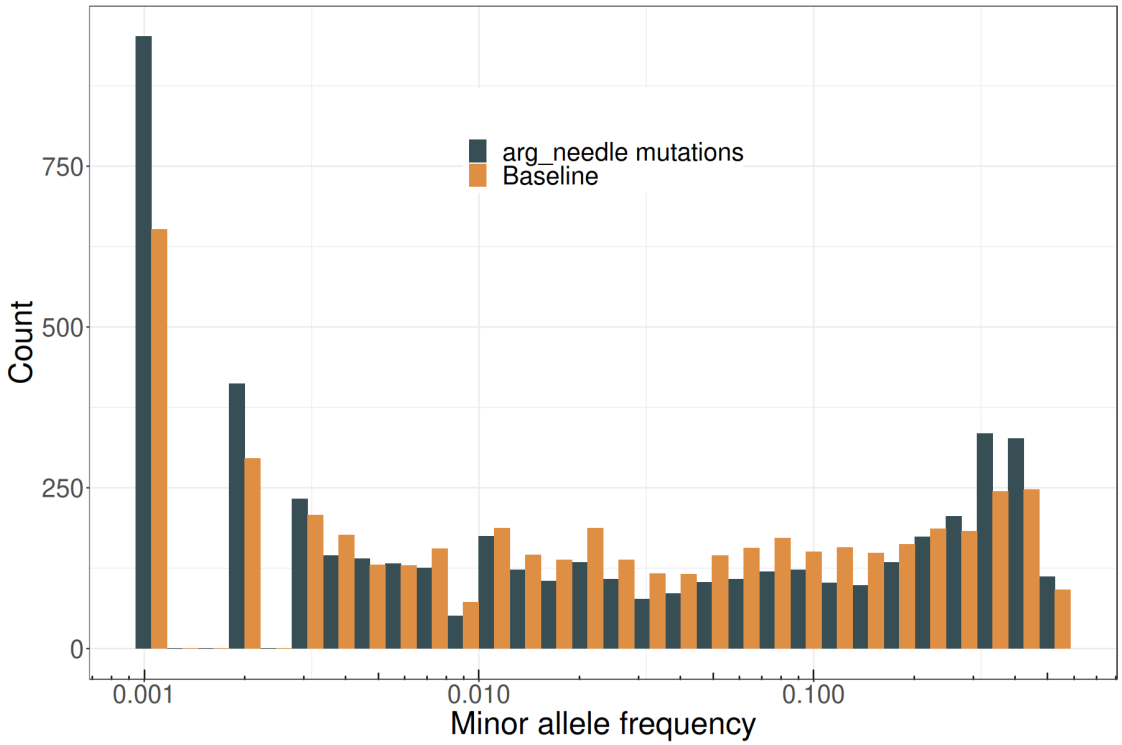


Figure 5.7: Allele frequency spectrum of the mutation simulation of ARG inferred by ARG-Needle. The reference panel is created using msprime and the baseline allele frequency is derived from the reference panel. The ARG-Needle mutations are randomly simulated on the ARG inference of the reference panel.

5.5 Discussion

In this chapter, we described ARGON2, a new implementation of the ARGON discrete time Wright Fisher simulator. We compared ARGON2 to existing simulators, including ARGON1, msprime, and msprime-dtwf, using various metrics such as the number of segregating sites, heterozygosity, and LD, to verify that the topology and the mutation process of ARGON2 is well aligned with other existing simulators. The ARGON2 implementation is consistently faster than its predecessor, ARGON, and more than 2 times faster than msprime-dtwf for smaller simulations ($N_e = 10,000$, genome length = 50 Mb) for ARG building. Despite the similar performance for larger simulations, ARGON2 has a worse scaling property than msprime-dtwf, which is likely due to the substantial increase in fragmentation in these regimes. We have also shown that DTWF is much faster than Hudson's coalescent algorithm when simulating large number of demes on a fully connected Island model, especially

when there is a high migration rate among the demes; we observed a 40x speed-up when simulating 50 demes with high migration rate (Figure 5.5).

We also explored a new simulation strategy, which we refer to as ARG augmentation, where we simulate new samples on an existing simulated or inferred ARG. We showed that, using this approach to simulate from an inferred ARG, ARGON2 is able to mimic features of samples in the reference panel, producing a more realistic rare allele frequency spectrum than other popular methods that are based on the Li-Stephens copying model, such as HAPGEN2 [87]. We expect this approach to improve as methods for ARG inference continue to be developed. ARGON2 is currently a work in progress, and we plan to take a number of steps to improve its scalability. One immediate such step is the implementation of a strategy to reduce fragmentation, which is likely to benefit simulation as well as a number of downstream applications. This may be facilitated by the adoption of a better representation of Block objects that can better support the merging of adjacent blocks, that done in ARG-Needle [146]. In addition, we have observed ARGON2 to be more efficient than coalescent based models for the simulation of complex models, but complex demographic models are difficult to work with without a user-friendly interface.We therefore intend to work on a better Python interface for the specification of demography models in the future. Finally, the distortion of Hudson's coalescent algorithm affects the recent generation disproportionally compared to the distant past. The combination of coalescent and DTWF simulation, adopted in [100] and [102], may therefore lead to improved performance.

6
Conclusion

This thesis revolved around the construction of data resources and infrastructure to facilitate large-scale genomic analyses, including genome-wide association and simulation. One of the major contributions of this thesis is the development of the GEL reference panel with hundreds of millions of novel rare variants unique to the GEL populations. The reference panel enables high quality rare variant imputation for White British and South Asian populations, which benefits the discovery of novel rare variants associations and other downstream analyses. The phasing reference panel is available for all Genomics England registered users. A second contribution is the delivery of a GEL-imputed version of the UK Biobank data set, which contains over three times more variants than a previous version imputed using HRC+UK10K [1]. This unique data resource will allow thousands of UK Biobank researchers to detect associations to rare and common genomic variation in hundreds of heritable traits. The third contribution is the development of an efficient discrete-time Wright-Fisher simulator, ARGON2. ARGON2 allows accurate simulation of large genomic data sets, as well as data augmentation, which may be used to test novel methodology and experimental designs.

In Chapter 2.3, we set out to build a reference panel using the Genomics England 100,000 Genome project samples. The project has sequenced 120,000 genomes from over 85,000 rare disease and cancer patients alongside their close relatives. The

use of high coverage sequencing produced an almost complete characterization of genetic variation in this cohort, one of the largest human genetic variation resources ever collected in the UK. The dataset consists of a diverse set of ancestries with percentages self reporting as White, Asian, Black and Mixed ancestry of 70%, 9%, 2.3% and 1.85% respectively, and contains a large number of related individuals. For example, 44% and 16% of the individuals have at least one and two first degree relative also in the study, respectively, which greatly aids genotype phasing.

We designed a set of quality control filters to remove more than half of the variants, resulting in a reference panel of 342 million autosomal variants. The overall Ts/Tv ratio has been improved from 1.1 to 1.8, with the removed variant Ts/Tv ratio to be 0.54. We developed a multi-stage phasing strategy to phase the reference panel, which required 6500 CPU days. The reference panel phasing accuracy was assessed using 1,178 parents from 589 father-mother-child trio families in the 1000 Genomes project, for which the phasing ground truth can be directly estimated. The mean switch error rate was 0.18% for the CEU population. Most populations achieved better switch error rate than those obtained using the HRC panel, especially Southern Asian populations, which saw a nearly 3-fold reduction in switch error rate (Figure 2.7). In the imputation experiment, we compared the imputation accuracy results of 2,405 1000 Genomes project samples across three reference panels: TOPMed, HRC and GEL. We did so by subsetting the data to emulate the UKB Axiom SNP array. Samples were imputed using the reference panels and the results were compared in terms of the correlation between imputed dosages and sequencing ground truth. GEL-imputed data resulted in a better imputation r^2 than HRC-imputed data in all populations and across the entire allele frequency spectrum. GEL has also demonstrated an advantage in imputing rare variants of White British and South Asian samples compared to the TOPMed reference panel. For example, at MAF $= 10^{-4}$ and in White British samples, the r^2 is 0.6 using GEL and 0.3 using TOPMed. The South Asian samples imputation at MAF $= 10^{-4}$ had GEL $r^2 = 0.32$ and TOPMed $r^2 = 0.02$ (Figure 2.8).

The imputation of 448,315 UK Biobank samples, described in Chapter 3, was carried out using a fast imputation method, IMPUTE5, taking 1,200 CPU days to accomplish. We benchmarked imputation quality in terms of IMPUTE INFO. Using the GEL reference panel, over 87% of existing markers in the HRCUK10K imputed UKBiobank data (H-UKB) achieved a higher imputation INFO (Figure 3.2). In comparison to the H-UKB data, the number of variants imputed using GEL has tripled, while the number of functional variants of likely high impact was nearly 5 times higher (Table 3.1).

In Chapter 4, we demonstrated how the G-UKB dataset may be used to facilitate GWAS and other downstream analyses. We conducted a GWAS of four quantitative traits, standing height, BMI, systolic and diastolic blood pressure using both G-UKB and H-UKB. The rare associations are very hard to find and usually under-powered as discussed in 4.2.2. G-UKB rare association results are reliable for $AF > 10^{-4}$ and over 50% of the associations below the allele frequency of 10^{-4} are reliable by comparing the imputed markers and WES data (Figure: 4.3). G-UKB found 1,699 more rare associations and 38 functional variants, including stop-loss, stop-gain, frameshift, inframe insertion/deletion, missense, and etc., identified by VEP [124] than H-UKB across all four traits (see 4.2.2). The functional importance is informative for fine-mapping the causal variants in GWAS [17]. Over 70% of associations identified by exome-wide association studies [53] were found using G-UKB, while H-UKB only found 56% of them. We used conditional joint analysis (step-wise regression) to identify 10 novel rare variants associations that are independent of the LD with other variants, including one missense, one TF-binding site and one synonymous variants. Using G-UKB also benefited Bayesian fine-mapping.

Finally, in Chapter 5, we described the implementation of a discrete time Wright Fisher (DTWF) simulator, ARGON2, which enables accurate large-scale simulation of genomic data. DTWF simulators can produce more realistic patterns of genomic variation than coalescent simulators when the sample size is very large, particularly in terms of long range LD and the distribution of rare genomic

variation [98–102]. Although coalescent simulation is generally very efficient, our simulations show that ARGON2 is significantly faster than state-of-the-art coalescent simulation under complex demographic models. ARGON2 also enables generating new synthetic genomes using an existing reference panel, by simulating new samples on a genealogy inferred from existing data. Similar to imputation, this strategy enables augmenting the data with novel variation. This approach can be used to create large synthetic data sets for populations for which extensive genotyping or sequencing reference panels have not yet been built, enabling the testing of experimental designs and novel methodology.

Appendices

A
Additional Figures

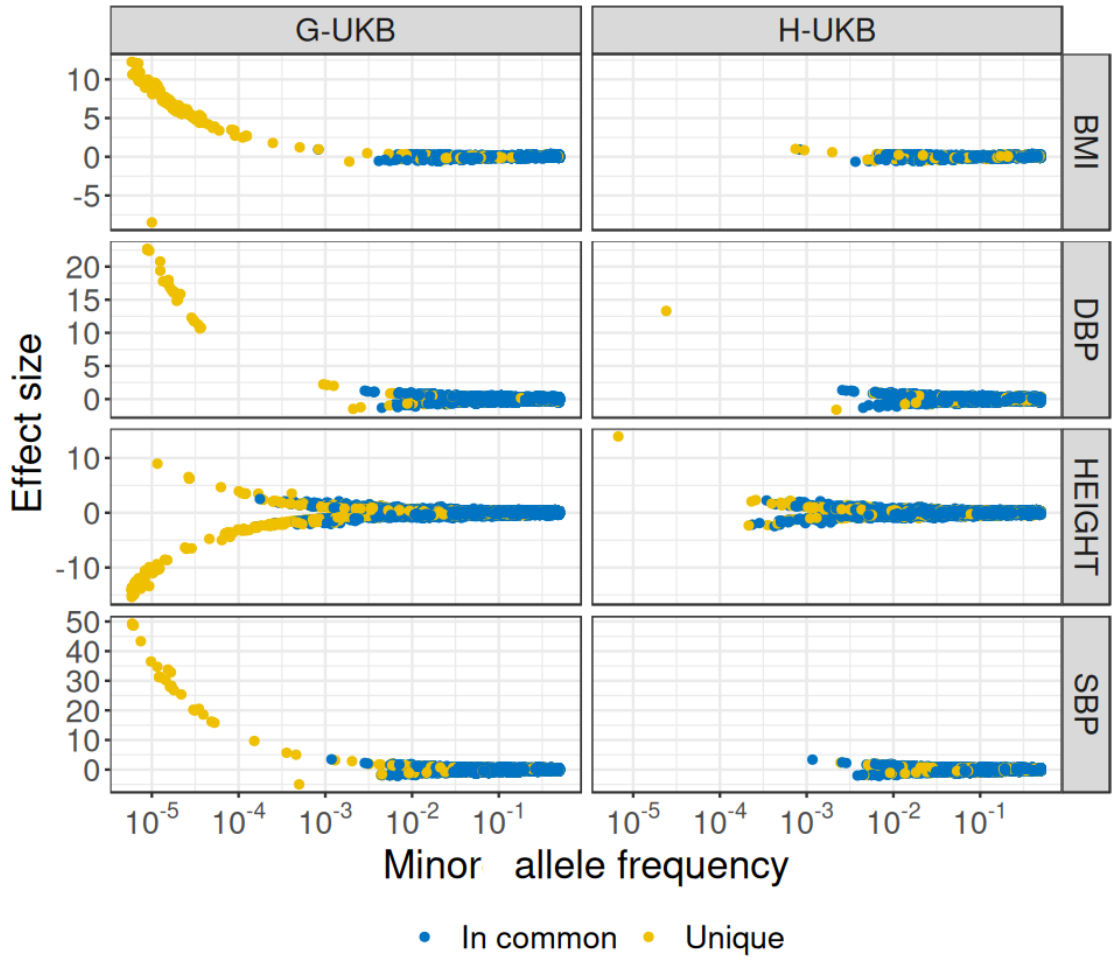


Figure A.1: Genome-wide significant associations (p-value $< 5 \times 10^{-8}$) with INFO $>$ 0.8 are plotted along the imputed MAF, showing with the decreasing of the MAF, the associations captured requires higher effect size, due to the declining of power. Blue dots are associations found both in GEL (G-UKB) and HRCUK10K (H-UKB) imputed UKB data, and yellow dots are those unique to each dataset.

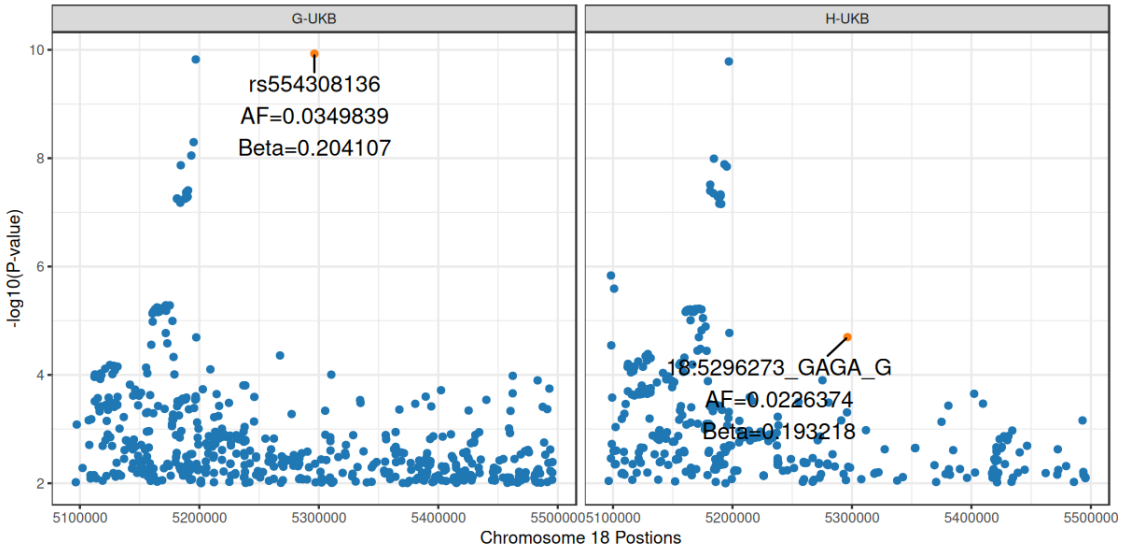


Figure A.2: Manhattan plot comparing G-UKB and H-UKB results for standing height association rs554308136 is a variant in ZBTB14 gene.

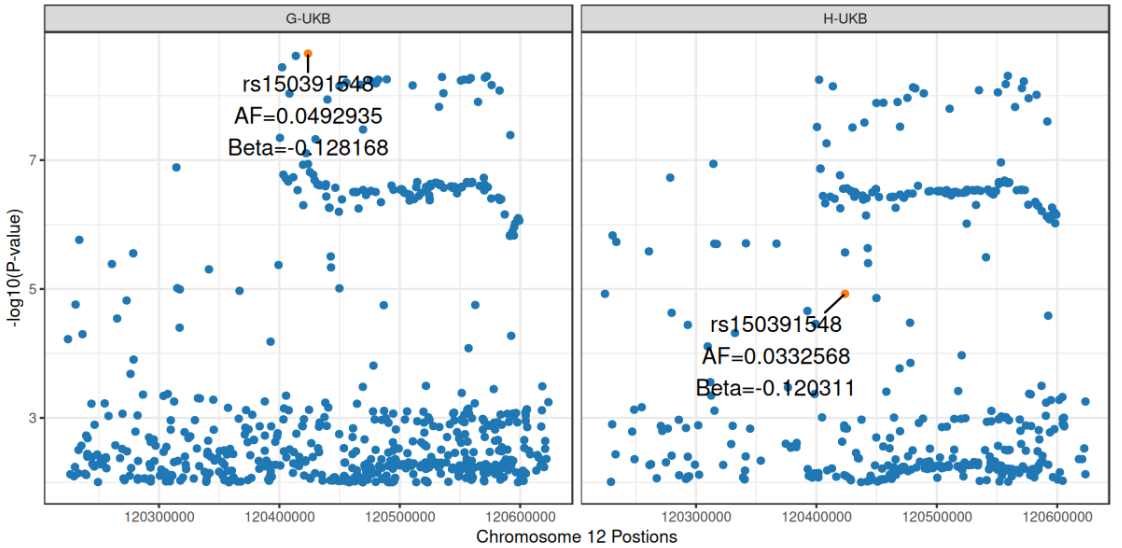


Figure A.3: Manhattan plot comparing G-UKB and H-UKB results for BMI association. rs150391548, near COX6A1 Gene which was found to be associated with BMI in [131].

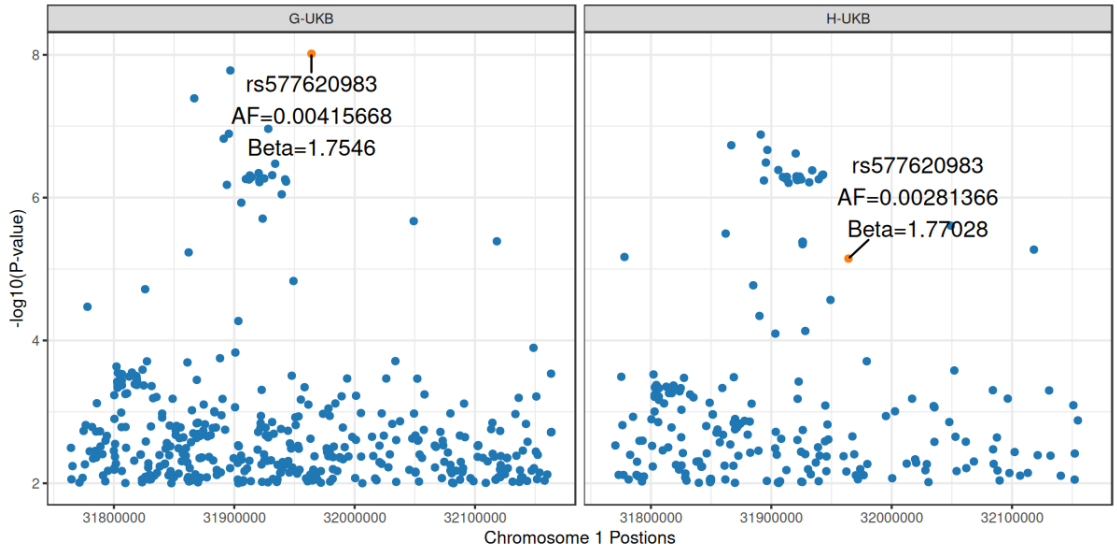


Figure A.4: Manhattan plot comparing G-UKB and H-UKB results for SBP association. rs577620983 is a TF-binding site near PTP4A2 and KPNA6 genes.

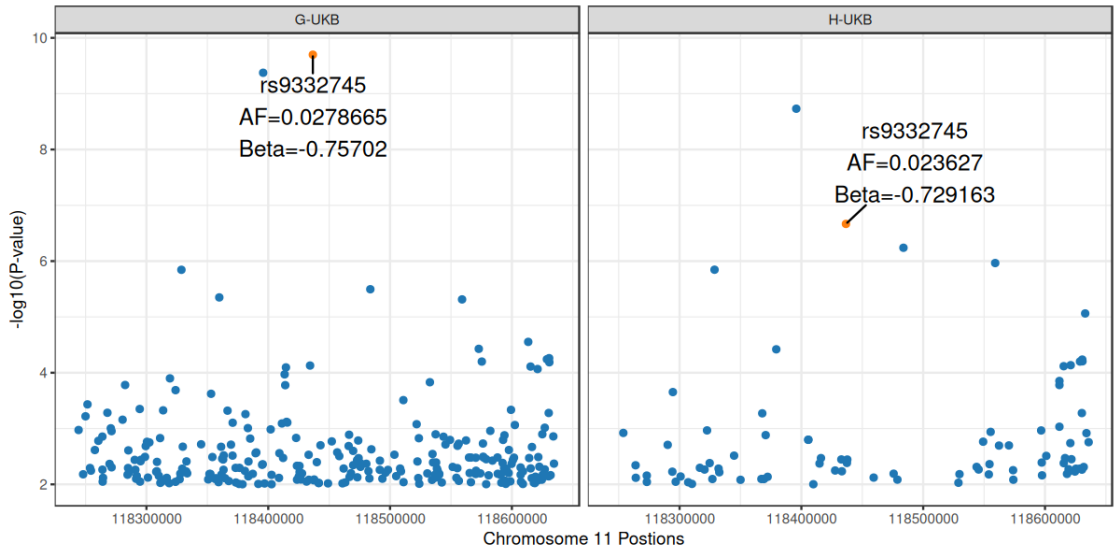


Figure A.5: Manhattan plot comparing G-UKB and H-UKB results for SBP association. rs9332745 is a missense variant in KMT2A gene, which is known to be associated with triglycerides level. [132, 133].

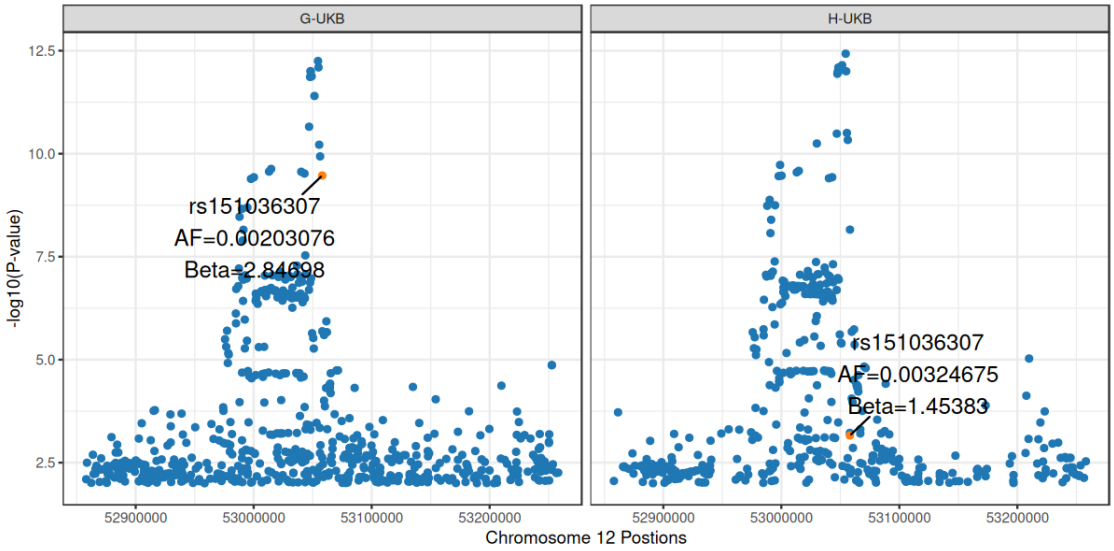


Figure A.6: Manhattan plot comparing G-UKB and H-UKB results for SBP association. rs151036307 is a synonymous variant in TNS2, also found significant in UKB whole exome imputation, associated with SBP [125].

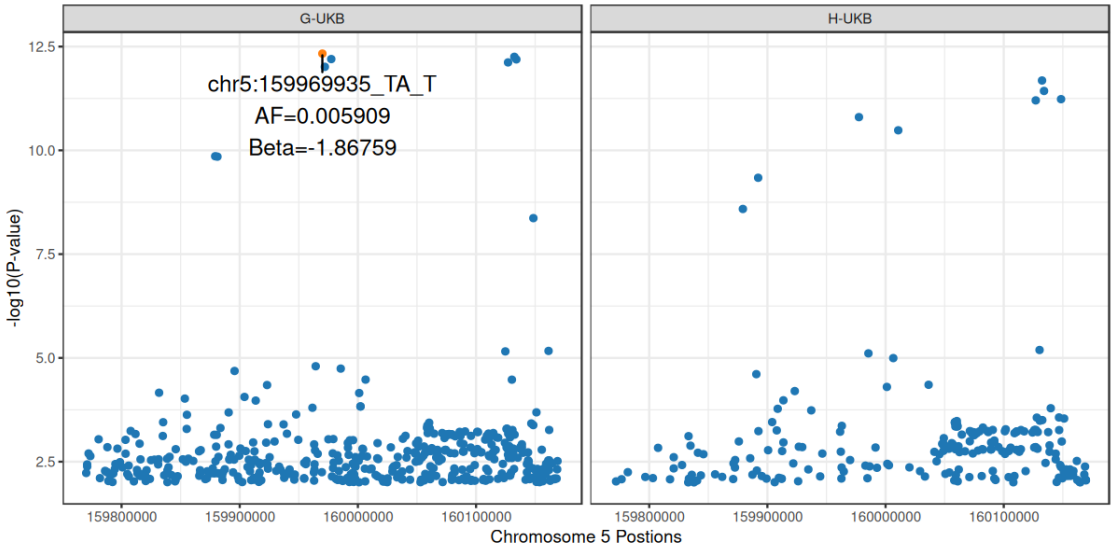


Figure A.7: Manhattan plot comparing G-UKB and H-UKB results for SBP association. rs1755838777 is a variant belong to ADRA1B gene. The gene is found to be associated with SBP and DBP in the literature [134].

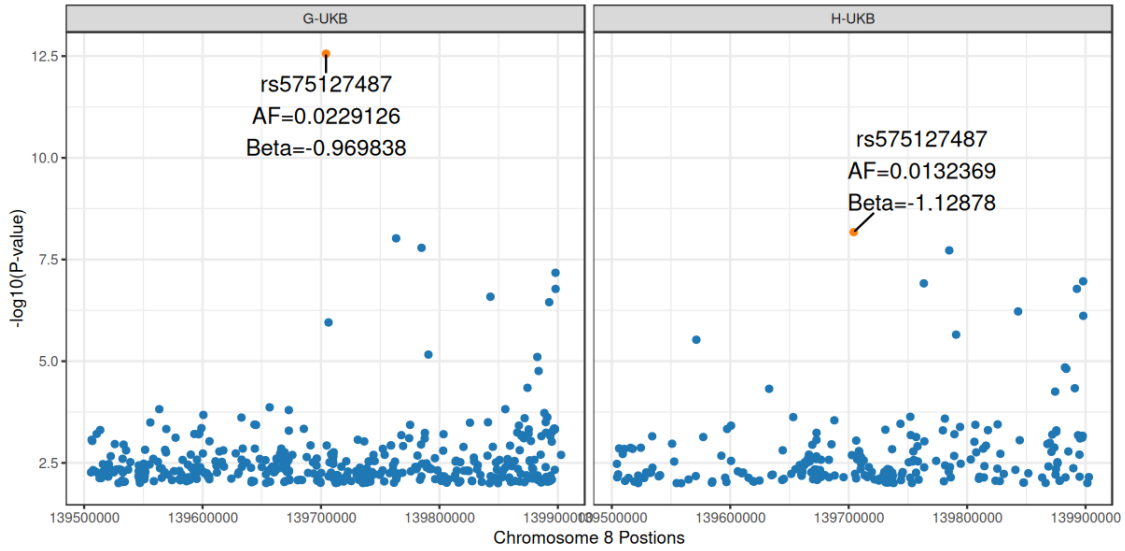


Figure A.8: Manhattan plot comparing G-UKB and H-UKB results for SBP association. rs575127487 is an intron variant of KCNK9, which controls the potassium pathway.

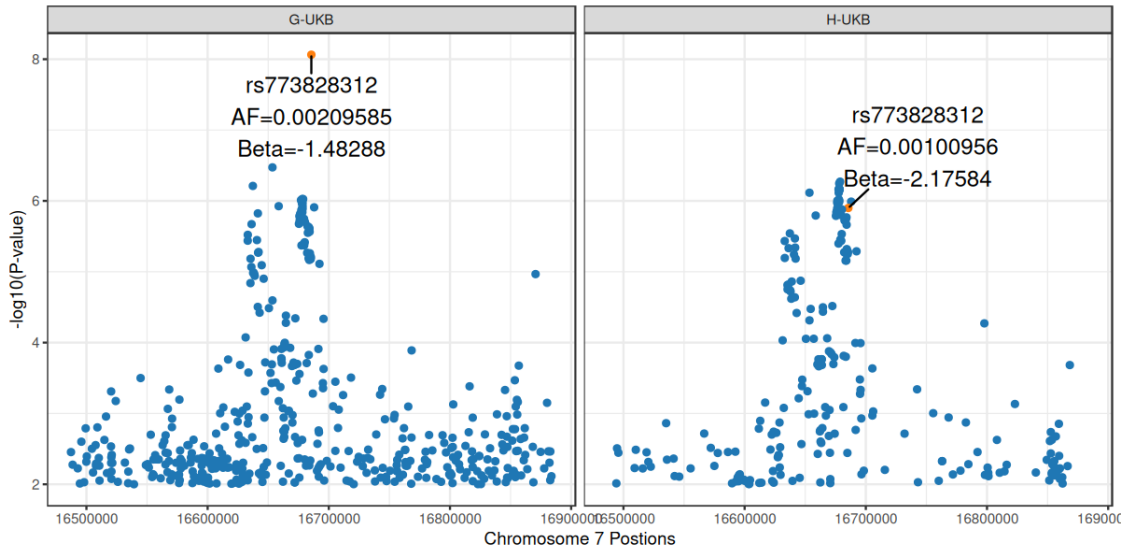


Figure A.9: Manhattan plot comparing G-UKB and H-UKB results for DBP association. rs773828312 belongs to BZW2 gene, which was found to be associated with diastolic blood pressure [137].

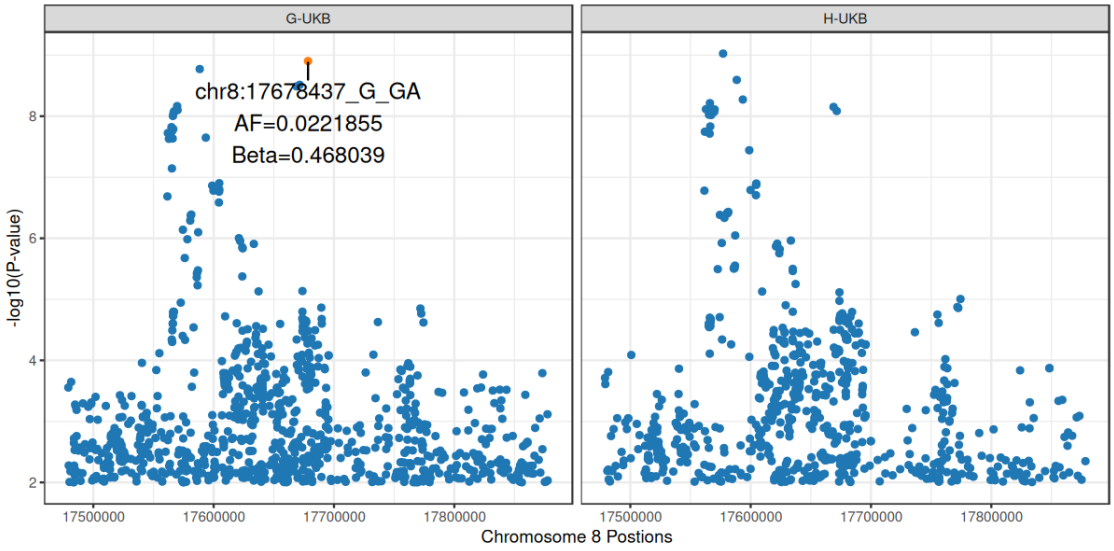


Figure A.10: Manhattan plot comparing G-UKB and H-UKB results for DBP association. The deletion is known as rs368338720, in MTUS1 gene.

Bibliography

[1] Clare Bycroft et al. "The UK Biobank Resource with Deep Phenotyping and Genomic Data". In: *Nature* 562.7726 (7726 Oct. 2018), pp. 203–209. URL: https://www.nature.com/articles/s41586-018-0579-z (visited on 01/08/2021).

[2] J. D. Watson and F. H. C. Crick. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356 (4356 Apr. 1953), pp. 737–738. URL: https://www.nature.com/articles/171737a0.

[3] Peter M Visscher and Michael E Goddard. "From R.A. Fisher's 1918 Paper to GWAS a Century Later". In: *Genetics* 211.4 (Apr. 1, 2019), pp. 1125–1130. URL: https://doi.org/10.1534/genetics.118.301594.

[4] Roland Fisher. "The Correlation between Relatives on the Supposition of Mendelian Inheritance". In: *Transactions of the Royal Society of Edinburgh* 52 (1918), pp. 339–433.

[5] Melina Claussnitzer et al. "A Brief History of Human Disease Genetics". In: *Nature* 577.7789 (7789 Jan. 2020), pp. 179–189. URL: https://www.nature.com/articles/s41586-019-1879-7 (visited on 09/24/2021).

[6] Richard A. Gibbs et al. "The International HapMap Project". In: *Nature* 426.6968 (6968 Dec. 2003), pp. 789–796. URL: https://www.nature.com/articles/nature02168 (visited on 09/24/2021).

[7] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *American Journal of Human Genetics* 101.1 (July 6, 2017), pp. 5–22. pmid: 28686856. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501872/ (visited on 09/24/2021).

[8] John Novembre et al. "Addressing the Challenges of Polygenic Scores in Human Genetic Research". In: *The American Journal of Human Genetics* 109.12 (Dec. 1, 2022), pp. 2095–2100. pmid: 36459976. URL: https://www.cell.com/ajhg/abstract/S0002-9297(22)00460-8 (visited on 12/22/2022).

[9] Peter Dornbos et al. "A Combined Polygenic Score of 21,293 Rare and 22 Common Variants Improves Diabetes Diagnosis Based on Hemoglobin A1C Levels". In: *Nature Genetics* 54.11 (11 Nov. 2022), pp. 1609–1614. URL: https://www.nature.com/articles/s41588-022-01200-1 (visited on 12/22/2022).

[10] Jonathan Marchini and Bryan Howie. "Genotype Imputation for Genome-Wide Association Studies". In: *Nature Reviews Genetics* 11.7 (July 2010), pp. 499–511. URL: https://www.nature.com/nrg/journal/v11/n7/full/nrg2796.html (visited on 08/01/2017).

[11] Shaun Purcell et al. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. pmid: 17701901. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/ (visited on 12/21/2021).

[12] Po-Ru Loh et al. "Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts". In: *Nature Genetics* 47.3 (3 Mar. 2015), pp. 284–290. URL: https://www.nature.com/articles/ng.3190 (visited on 10/05/2022).

[13] Longda Jiang et al. "A Resource-Efficient Tool for Mixed Model Association Analysis of Large-Scale Data". In: *Nature Genetics* 51.12 (12 Dec. 2019), pp. 1749–1755. URL: https://www.nature.com/articles/s41588-019-0530-8 (visited on 01/05/2023).

[14] Joelle Mbatchou et al. "Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits". In: *Nature Genetics* 53.7 (7 July 2021), pp. 1097–1103. URL: https://www.nature.com/articles/s41588-021-00870-7 (visited on 09/26/2021).

[15] Jeffrey D. Wall and Jonathan K. Pritchard. "Haplotype Blocks and Linkage Disequilibrium in the Human Genome". In: *Nature Reviews Genetics* 4.8 (8 Aug. 2003), pp. 587–597. URL: https://www.nature.com/articles/nrg1123 (visited on 06/03/2023).

[16] Montgomery Slatkin. "Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future". In: *Nature Reviews Genetics* 9.6 (6 June 2008), pp. 477–485. URL: https://www.nature.com/articles/nrg2361 (visited on 01/08/2023).

[17] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. "From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping". In: *Nature Reviews Genetics* 19.8 (8 Aug. 2018), pp. 491–504. URL: https://www.nature.com/articles/s41576-018-0016-z (visited on 10/13/2021).

[18] Daniel O. Stram. "Tag SNP Selection for Association Studies". In: *Genetic Epidemiology* 27.4 (2004), pp. 365–374. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20028 (visited on 06/03/2023).

[19] Keyue Ding and Iftikhar J. Kullo. "Methods for the Selection of Tagging SNPs: A Comparison of Tagging Efficiency and Performance". In: *European Journal of Human Genetics* 15.2 (2 Feb. 2007), pp. 228–236. URL: https://www.nature.com/articles/5201755 (visited on 06/03/2023).

[20] Jian Yang et al. "Conditional and Joint Multiple-SNP Analysis of GWAS Summary Statistics Identifies Additional Variants Influencing Complex Traits". In: *Nature Genetics* 44.4 (Mar. 18, 2012), 369–375, S1–3. pmid: 22426310.

[21] Zhihong Zhu et al. "Causal Associations between Risk Factors and Common Diseases Inferred from GWAS Summary Data". In: *Nature Communications* 9.1 (Jan. 15, 2018), p. 224. pmid: 29335400.

[22] Kristin L. Ayers and Heather J. Cordell. "SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression". In: *Genetic Epidemiology* 34.8 (2010), pp. 879–891. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20543 (visited on 06/03/2023).

[23] Matthew Stephens. "A Unified Framework for Association Analysis with Multiple Related Phenotypes". In: *PLOS ONE* 8.7 (July 5, 2013), e65245. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0065245 (visited on 06/03/2023).

[24] Farhad Hormozdiari et al. "Identifying Causal Variants at Loci with Multiple Signals of Association". In: *Genetics* 198.2 (Oct. 1, 2014), pp. 497–508. URL: https://doi.org/10.1534/genetics.114.167908 (visited on 06/03/2023).

[25] Wenan Chen et al. "Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics". In: *Genetics* 200.3 (July 1, 2015), pp. 719–736. URL: https://doi.org/10.1534/genetics.115.176107 (visited on 06/03/2023).

[26] Christian Benner et al. "FINEMAP: Efficient Variable Selection Using Summary Data from Genome-Wide Association Studies". In: *Bioinformatics* 32.10 (May 15, 2016), pp. 1493–1501. URL: https://doi.org/10.1093/bioinformatics/btw018 (visited on 06/03/2023).

[27] Joseph K. Pickrell. "Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits". In: *The American Journal of Human Genetics* 94.4 (Apr. 3, 2014), pp. 559–573. URL: https://www.sciencedirect.com/science/article/pii/S0002929714001062 (visited on 06/03/2023).

[28] Gleb Kichaev et al. "Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies". In: *PLOS Genetics* 10.10 (Oct. 30, 2014), e1004722. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004722 (visited on 06/03/2023).

[29] Wenan Chen et al. "Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics". In: *Genetics* 204.3 (Nov. 1, 2016), pp. 933–958. URL: https://doi.org/10.1534/genetics.116.188953 (visited on 06/03/2023).

[30] Peter M. Visscher, William G. Hill, and Naomi R. Wray. "Heritability in the Genomics Era — Concepts and Misconceptions". In: *Nature Reviews Genetics* 9.4 (4 Apr. 2008), pp. 255–266. URL: https://www.nature.com/articles/nrg2322 (visited on 06/03/2023).

[31] Noah Zaitlen and Peter Kraft. "Heritability in the Genome-Wide Association Era". In: *Human Genetics* 131.10 (Oct. 1, 2012), pp. 1655–1664. URL: https://doi.org/10.1007/s00439-012-1199-6 (visited on 06/03/2023).

[32] Albert Tenesa and Chris S. Haley. "The Heritability of Human Disease: Estimation, Uses and Abuses". In: *Nature Reviews Genetics* 14.2 (2 Feb. 2013), pp. 139–149. URL: https://www.nature.com/articles/nrg3377 (visited on 06/03/2023).

[33] Brendan Maher. "Personal Genomes: The Case of the Missing Heritability". In: *Nature* 456.7218 (7218 Nov. 1, 2008), pp. 18–21. URL: https://www.nature.com/articles/456018a.

[34] Greg Gibson. "Rare and Common Variants: Twenty Arguments". In: *Nature Reviews Genetics* 13.2 (2 Feb. 2012), pp. 135–145. URL: https://www.nature.com/articles/nrg3118 (visited on 12/14/2022).

[35] Loïc Yengo et al. "A Saturated Map of Common Genetic Variants Associated with Human Height". In: *Nature* 610.7933 (7933 Oct. 2022), pp. 704–712. URL: https://www.nature.com/articles/s41586-022-05275-y.

[36] Teri A. Manolio et al. "Finding the Missing Heritability of Complex Diseases". In: *Nature* 461.7265 (7265 Oct. 2009), pp. 747–753. URL: https://www.nature.com/articles/nature08494.

[37] Daniel Taliun et al. "Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program". In: *Nature* 590.7845 (7845 Feb. 2021), pp. 290–299. URL: https://www.nature.com/articles/s41586-021-03205-y (visited on 07/14/2023).

[38] Marta Byrska-Bishop et al. "High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios". In: *bioRxiv* (Feb. 7, 2021), p. 2021.02.06.430068. URL: https://www.biorxiv.org/content/10.1101/2021.02.06.430068v1 (visited on 05/15/2021).

[39] Gil A. McVean et al. "An Integrated Map of Genetic Variation from 1,092 Human Genomes". In: *Nature* 491.7422 (7422 Nov. 2012), pp. 56–65. URL: https://www.nature.com/articles/nature11632 (visited on 01/05/2023).

[40] Iain Mathieson and Gil McVean. "Differential Confounding of Rare and Common Variants in Spatially Structured Populations". In: *Nature Genetics* 44.3 (3 Mar. 2012), pp. 243–246. URL: https://www.nature.com/articles/ng.1074 (visited on 01/05/2023).

[41] Eric S. Lander et al. "Initial Sequencing and Analysis of the Human Genome". In: *Nature* 409.6822 (6822 Feb. 2001), pp. 860–921. URL: https://www.nature.com/articles/35057062 (visited on 12/14/2022).

[42] Taishan Hu et al. "Next-Generation Sequencing Technologies: An Overview". In: *Human Immunology*. Next Generation Sequencing and Its Application to Medical Laboratory Immunology 82.11 (Nov. 1, 2021), pp. 801–811. URL: https://www.sciencedirect.com/science/article/pii/S0198885921000628 (visited on 12/14/2022).

[43] Kris Wetterstrand. *The Cost of Sequencing a Human Genome*. genome.gov. 2021. URL: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (visited on 12/14/2022).

[44] Jesse J. Salk, Michael W. Schmitt, and Lawrence A. Loeb. "Enhancing the Accuracy of Next-Generation Sequencing for Detecting Rare and Subclonal Mutations". In: *Nature Reviews Genetics* 19.5 (5 May 2018), pp. 269–285. URL: https://www.nature.com/articles/nrg.2017.117 (visited on 12/16/2022).

[45] Lisenka E. L. M. Vissers et al. "Array-Based Comparative Genomic Hybridization for the Genomewide Detection of Submicroscopic Chromosomal Abnormalities". In: *The American Journal of Human Genetics* 73.6 (Dec. 1, 2003), pp. 1261–1270. pmid: 14628292. URL: https://www.cell.com/ajhg/abstract/S0002-9297(07)63979-2 (visited on 12/16/2022).

[46] Roger Bumgarner. "Overview of DNA Microarrays: Types, Applications, and Their Future". In: *Current Protocols in Molecular Biology* 101.1 (2013), pp. 22.1.1–22.1.11. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb2201s101 (visited on 12/16/2022).

[47] David Altshuler, Peter Donnelly, and The International HapMap Consortium. "A Haplotype Map of the Human Genome". In: *Nature* 437.7063 (7063 Oct. 2005), pp. 1299–1320. URL: https://www.nature.com/articles/nature04226 (visited on 12/16/2022).

[48] Pardis C. Sabeti et al. "Genome-Wide Detection and Characterization of Positive Selection in Human Populations". In: *Nature* 449.7164 (7164 Oct. 2007), pp. 913–918. URL: https://www.nature.com/articles/nature06250 (visited on 12/16/2022).

[49] Paul R. Burton et al. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls". In: *Nature* 447.7145 (7145 June 2007), pp. 661–678. URL: https://www.nature.com/articles/nature05911.

[50] Sayantan Das et al. "Next-Generation Genotype Imputation Service and Methods". In: *Nature Genetics* 48.10 (10 Oct. 2016), pp. 1284–1287. URL: https://www.nature.com/articles/ng.3656 (visited on 05/24/2021).

[51] Adam Auton et al. "A Global Reference for Human Genetic Variation". In: *Nature* 526.7571 (7571 Oct. 2015), pp. 68–74. URL: https://www.nature.com/articles/nature15393 (visited on 05/26/2021).

[52] Konrad J. Karczewski et al. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans". In: *Nature* 581.7809 (7809 May 2020), pp. 434–443. URL: https://www.nature.com/articles/s41586-020-2308-7 (visited on 05/02/2021).

[53] Joshua D. Backman et al. "Exome Sequencing and Analysis of 454,787 UK Biobank Participants". In: *Nature* 599.7886 (7886 Nov. 2021), pp. 628–634. URL: https://www.nature.com/articles/s41586-021-04103-z (visited on 02/10/2022).

[54] Bjarni V. Halldorsson et al. "The Sequences of 150,119 Genomes in the UK Biobank". In: *Nature* 607.7920 (7920 July 2022), pp. 732–740. URL: https://www.nature.com/articles/s41586-022-04965-x (visited on 12/16/2022).

[55] Mitja I. Kurki et al. “FinnGen Provides Genetic Insights from a Well-Phenotyped Isolated Population”. In: *Nature* 613.7944 (7944 Jan. 2023), pp. 508–518. URL: https://www.nature.com/articles/s41586-022-05473-8 (visited on 06/03/2023).

[56] Zhengming Chen et al. “China Kadoorie Biobank of 0.5 Million People: Survey Methods, Baseline Characteristics and Long-Term Follow-Up”. In: *International Journal of Epidemiology* 40.6 (Dec. 1, 2011), pp. 1652–1666. URL: https://doi.org/10.1093/ije/dyr120 (visited on 06/03/2023).

[57] Akiko Nagai et al. “Overview of the BioBank Japan Project: Study Design and Profile”. In: *Journal of Epidemiology* 27 (3 Suppl Feb. 8, 2017), S2–S8. pmid: 28189464. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5350590/ (visited on 06/03/2023).

[58] R. Holliday. “Recombination and Meiosis”. In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 277.955 (Mar. 21, 1977), pp. 359–370. pmid: 16297.

[59] Jonathan Marchini. “Haplotype Estimation and Genotype Imputation”. In: *Handbook of Statistical Genomics*. John Wiley & Sons, Ltd, 2019, pp. 87–114. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119487845.ch3 (visited on 12/30/2022).

[60] Ryan Tewhey et al. “The Importance of Phase Information for Human Genomics”. In: *Nature Reviews Genetics* 12.3 (Mar. 2011), pp. 215–223. URL: https://www.nature.com/articles/nrg2950 (visited on 10/07/2018).

[61] Simon Myers et al. “A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome”. In: *Science* 310.5746 (Oct. 14, 2005), pp. 321–324. URL: https://www.science.org/doi/10.1126/science.1117196 (visited on 01/01/2023).

[62] Garrett Hellenthal et al. “A Genetic Atlas of Human Admixture History”. In: *Science* 343.6172 (Feb. 14, 2014), pp. 747–751. URL: https://www.science.org/doi/10.1126/science.1243518 (visited on 01/04/2023).

[63] Jerome Kelleher et al. “Inferring Whole-Genome Histories in Large Population Datasets”. In: *Nature Genetics* 51.9 (9 Sept. 2019), pp. 1330–1338. URL: https://www.nature.com/articles/s41588-019-0483-y (visited on 09/26/2021).

[64] Leo Speidel. “Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies”. In: (2021).

[65] Sharon R. Browning and Brian L. Browning. “Haplotype Phasing: Existing Methods and New Developments”. In: *Nature Reviews Genetics* 12.10 (Oct. 2011), p. 703. URL: https://www.nature.com/articles/nrg3054 (visited on 12/08/2017).

[66] Matthew Stephens, Nicholas J. Smith, and Peter Donnelly. “A New Statistical Method for Haplotype Reconstruction from Population Data”. In: *The American Journal of Human Genetics* 68.4 (Apr. 1, 2001), pp. 978–989. pmid: 11254454. URL: https://www.cell.com/ajhg/abstract/S0002-9297(07)61424-4 (visited on 01/04/2023).

[67] Po-Ru Loh et al. "Reference-Based Phasing Using the Haplotype Reference Consortium Panel". In: *Nature Genetics* 48.11 (Nov. 2016), p. 1443. URL: https://www.nature.com/articles/ng.3679 (visited on 01/15/2018).

[68] Brian L. Browning et al. "Fast Two-Stage Phasing of Large-Scale Sequence Data". In: *The American Journal of Human Genetics* 108.10 (Oct. 7, 2021), pp. 1880–1890. pmid: 34478634. URL: https://www.cell.com/ajhg/abstract/S0002-9297(21)00304-9 (visited on 01/04/2023).

[69] Olivier Delaneau et al. "Accurate, Scalable and Integrative Haplotype Estimation". In: *Nature Communications* 10.1 (1 Nov. 28, 2019), p. 5436. URL: https://www.nature.com/articles/s41467-019-13225-y (visited on 05/04/2021).

[70] S. Rubinacci. "Genotype Imputation Methods for next Generation Datasets". http://purl.org/dc/dcmitype/Text. University of Oxford. URL: https://ora.ox.ac.uk/objects/uuid:04c9363c-dd91-44df-9306-1932a69fedce (visited on 04/28/2021).

[71] S. T. Sherry et al. "dbSNP: The NCBI Database of Genetic Variation". In: *Nucleic Acids Research* 29.1 (Jan. 1, 2001), pp. 308–311. pmid: 11125122.

[72] Brian L. Browning, Ying Zhou, and Sharon R. Browning. "A One-Penny Imputed Genome from Next-Generation Reference Panels". In: *The American Journal of Human Genetics* 103.3 (Sept. 6, 2018), pp. 338–348. pmid: 30100085. URL: https://www.cell.com/ajhg/abstract/S0002-9297(18)30242-8 (visited on 01/04/2023).

[73] Robert W. Davies et al. "Rapid Genotype Imputation from Sequence with Reference Panels". In: *Nature Genetics* 53.7 (July 2021), pp. 1104–1111. pmid: 34083788.

[74] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. "Genotype Imputation Using the Positional Burrows Wheeler Transform". In: *PLOS Genetics* 16.11 (Nov. 16, 2020), e1009049. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009049 (visited on 05/24/2021).

[75] Yichen Si, Brett Vanderwerff, and Sebastian Zöllner. "Why Are Rare Variants Hard to Impute? Coalescent Models Reveal Theoretical Limits in Existing Algorithms". In: *Genetics* 217.4 (Apr. 1, 2021), iyab011. URL: https://doi.org/10.1093/genetics/iyab011 (visited on 01/08/2023).

[76] Na Li and Matthew Stephens. "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data". In: *Genetics* 165.4 (Dec. 2003), pp. 2213–2233. pmid: 14704198.

[77] Jonathan Marchini et al. "A Comparison of Phasing Algorithms for Trios and Unrelated Individuals". In: *The American Journal of Human Genetics* 78.3 (Mar. 1, 2006), pp. 437–450. URL: http://www.sciencedirect.com/science/article/pii/S0002929707623830 (visited on 12/08/2017).

[78] Brian L. Browning and Sharon R. Browning. "Statistical Phasing of 150,119 Sequenced Genomes in the UK Biobank". In: *The American Journal of Human Genetics* 110.1 (Jan. 5, 2023), pp. 161–165. pmid: 36450278. URL: https://www.cell.com/ajhg/abstract/S0002-9297(22)00499-2 (visited on 03/09/2024).

[79] Simone Rubinacci et al. "Efficient Phasing and Imputation of Low-Coverage Sequencing Data Using Large Reference Panels". In: *Nature Genetics* 53.1 (1 Jan. 2021), pp. 120–126. URL: https://www.nature.com/articles/s41588-020-00756-0 (visited on 05/26/2021).

[80] L. R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In: *Proceedings of the IEEE* 77.2 (Feb. 1989), pp. 257–286.

[81] J. F. C. Kingman. "The Coalescent". In: *Stochastic Processes and their Applications* 13.3 (Sept. 1, 1982), pp. 235–248. URL: https://www.sciencedirect.com/science/article/pii/0304414982900114 (visited on 11/19/2022).

[82] Olivier Delaneau, Jean-Francois Zagury, and Jonathan Marchini. "Improved Whole-Chromosome Phasing for Disease and Population Genetic Studies". In: *Nature Methods* 10.1 (Jan. 2013), pp. 5–6. URL: https://www.nature.com/articles/nmeth.2307 (visited on 05/01/2024).

[83] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. "A Linear Complexity Phasing Method for Thousands of Genomes". In: *Nature Methods* 9.2 (Feb. 2012), pp. 179–181. URL: http://www.nature.com/nmeth/journal/v9/n2/abs/nmeth.1785.html (visited on 06/23/2017).

[84] Mark A. DePristo et al. "A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data". In: *Nature Genetics* 43.5 (5 May 2011), pp. 491–498. URL: https://www.nature.com/articles/ng.806 (visited on 05/25/2021).

[85] Olivier Delaneau and Jonathan Marchini. "Integrating Sequence and Array Data to Create an Improved 1000 Genomes Project Haplotype Reference Panel". In: *Nature Communications* 5.1 (1 June 13, 2014), p. 3934. URL: https://www.nature.com/articles/ncomms4934 (visited on 12/30/2022).

[86] Stephen F. Schaffner et al. "Calibrating a Coalescent Simulation of Human Genome Sequence Variation". In: *Genome Research* 15.11 (Jan. 11, 2005), pp. 1576–1583. pmid: 16251467. URL: https://genome.cshlp.org/content/15/11/1576 (visited on 01/11/2023).

[87] Zhan Su, Jonathan Marchini, and Peter Donnelly. "HAPGEN2: Simulation of Multiple Disease SNPs". In: *Bioinformatics* 27.16 (Aug. 15, 2011), pp. 2304–2305. URL: https://doi.org/10.1093/bioinformatics/btr341 (visited on 09/30/2021).

[88] Alicia R. Martin et al. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations”. In: *The American Journal of Human Genetics* 100.4 (Apr. 6, 2017), pp. 635–649. pmid: 28366442. URL: https://www.cell.com/ajhg/abstract/S0002-9297(17)30107-6 (visited on 01/11/2023).

[89] John Wakely. *Coalescent Theory: An Introduction*. Macmillan Learning, Apr. 22, 2016. 352 pp. Google Books: x30RAgAACAAJ.

[90] Richard R. Hudson. “Properties of a Neutral Allele Model with Intragenic Recombination”. In: *Theoretical Population Biology* 23.2 (Apr. 1, 1983), pp. 183–201. URL: https://www.sciencedirect.com/science/article/pii/0040580983900138 (visited on 11/20/2022).

[91] R. R. Hudson. “Gene Genealogies and the Coalescent Process.” In: *Oxford Surveys in Evolutionary Biology* 7 (1990), pp. 1–44. URL: https://www.cabdirect.org/cabdirect/abstract/19910191040 (visited on 11/21/2022).

[92] P. Donnelly and S. Tavaré. “Coalescents and Genealogical Structure under Neutrality”. In: *Annual Review of Genetics* 29 (1995), pp. 401–421. pmid: 8825481.

[93] G. A. Watterson. “On the Number of Segregating Sites in Genetical Models without Recombination”. In: *Theoretical Population Biology* 7.2 (Apr. 1975), pp. 256–276. pmid: 1145509.

[94] Richard R. Hudson. “Generating Samples under a Wright–Fisher Neutral Model of Genetic Variation”. In: *Bioinformatics* 18.2 (Feb. 1, 2002), pp. 337–338. URL: https://doi.org/10.1093/bioinformatics/18.2.337 (visited on 11/21/2022).

[95] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. “Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes”. In: *PLOS Computational Biology* 12.5 (May 4, 2016), e1004842. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004842 (visited on 11/21/2022).

[96] Franz Baumdicker et al. “Efficient Ancestry and Mutation Simulation with Msprime 1.0”. In: *Genetics* 220.3 (Mar. 1, 2022), iyab229. URL: https://doi.org/10.1093/genetics/iyab229 (visited on 11/21/2022).

[97] Jotun Hein, Mikkel Schierup, and Carsten Wuif. *Gene Genealogies, Variation and Evolution - a Primer in Coalescent Theory*. Oxford University Press, 2005.

[98] Yun-Xin Fu. “Exact Coalescent for the Wright–Fisher Model”. In: *Theoretical Population Biology* 69.4 (June 1, 2006), pp. 385–394. URL: https://www.sciencedirect.com/science/article/pii/S0040580905001498 (visited on 12/07/2022).

[99] John Wakeley et al. “Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman’s Coalescent”. In: *Genetics* 190.4 (Apr. 1, 2012), pp. 1433–1445. URL: https://doi.org/10.1534/genetics.111.135574 (visited on 11/21/2022).

[100] Anand Bhaskar, Andrew G. Clark, and Yun S. Song. "Distortion of Genealogical Properties When the Sample Is Very Large". In: *Proceedings of the National Academy of Sciences* 111.6 (Feb. 11, 2014), pp. 2385–2390. URL: https://www.pnas.org/doi/full/10.1073/pnas.1322709111 (visited on 11/21/2022).

[101] Pier Francesco Palamara. "ARGON: Fast, Whole-Genome Simulation of the Discrete Time Wright-fisher Process". In: *Bioinformatics* 32.19 (Oct. 1, 2016), pp. 3032–3034. URL: https://doi.org/10.1093/bioinformatics/btw355 (visited on 05/25/2021).

[102] Dominic Nelson et al. "Accounting for Long-Range Correlations in Genome-Wide Simulations of Large Cohorts". In: *PLOS Genetics* 16.5 (May 5, 2020), e1008619. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008619 (visited on 09/22/2021).

[103] Guillaume Laval and Laurent Excoffier. "SIMCOAL 2.0: A Program to Simulate Genomic Diversity over Large Recombining Regions in a Subdivided Population with a Complex History". In: *Bioinformatics* 20.15 (Oct. 12, 2004), pp. 2485–2487. URL: https://doi.org/10.1093/bioinformatics/bth264 (visited on 12/18/2022).

[104] Carsten Wiuf and Jotun Hein. "Recombination as a Point Process along Sequences". In: *Theoretical Population Biology* 55.3 (June 1, 1999), pp. 248–259. URL: https://www.sciencedirect.com/science/article/pii/S0040580998914034 (visited on 01/11/2023).

[105] Gilean A.T McVean and Niall J Cardin. "Approximating the Coalescent with Recombination". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459 (July 29, 2005), pp. 1387–1393. URL: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2005.1673 (visited on 06/06/2020).

[106] Paul Marjoram and Jeff D. Wall. "Fast "Coalescent" Simulation". In: *BMC Genetics* 7.1 (Mar. 15, 2006), p. 16. URL: https://doi.org/10.1186/1471-2156-7-16 (visited on 01/11/2023).

[107] Gary K. Chen, Paul Marjoram, and Jeffrey D. Wall. "Fast and Flexible Simulation of DNA Sequence Data". In: *Genome Research* 19.1 (Jan. 1, 2009), pp. 136–142. pmid: 19029539. URL: https://genome.cshlp.org/content/19/1/136 (visited on 01/11/2023).

[108] Laurent Excoffier and Matthieu Foll. "Fastsimcoal: A Continuous-Time Coalescent Simulator of Genomic Diversity under Arbitrarily Complex Evolutionary Scenarios". In: *Bioinformatics* 27.9 (May 1, 2011), pp. 1332–1334. URL: https://doi.org/10.1093/bioinformatics/btr124 (visited on 11/21/2022).

[109] Laurent Excoffier et al. "Fastsimcoal2: Demographic Inference under Complex Evolutionary Scenarios". In: *Bioinformatics* 37.24 (Dec. 15, 2021), pp. 4882–4885. URL: https://doi.org/10.1093/bioinformatics/btab468 (visited on 01/11/2023).

[110] Paul R. Staab et al. "Scrm: Efficiently Simulating Long Sequences Using the Approximated Coalescent with Recombination". In: *Bioinformatics* 31.10 (May 15, 2015), pp. 1680–1682. URL: https://doi.org/10.1093/bioinformatics/btu861 (visited on 11/21/2022).

[111] Garrett Hellenthal and Matthew Stephens. "msHOT: Modifying Hudson's Ms Simulator to Incorporate Crossover and Gene Conversion Hotspots". In: *Bioinformatics* 23.4 (Feb. 15, 2007), pp. 520–521. URL: https://doi.org/10.1093/bioinformatics/btl622 (visited on 01/11/2023).

[112] Ilya Shlyakhter, Pardis C. Sabeti, and Stephen F. Schaffner. "Cosi2: An Efficient Simulator of Exact and Approximate Coalescent with Selection". In: *Bioinformatics* 30.23 (Dec. 1, 2014), pp. 3427–3429. URL: https://doi.org/10.1093/bioinformatics/btu562 (visited on 12/20/2022).

[113] Liming Liang, Sebastian Zöllner, and Gonçalo R. Abecasis. "GENOME: A Rapid Coalescent-Based Whole Genome Simulator". In: *Bioinformatics* 23.12 (June 15, 2007), pp. 1565–1567. URL: https://doi.org/10.1093/bioinformatics/btm138 (visited on 11/21/2022).

[114] Robert C. Griffiths and Paul Marjoram. "An Ancestral Recombination Graph". In: *Progress in Population Genetics and Human Evolution*. Progress in Population Genetics and Human Evolution. Springer, 1997, pp. 257–270.

[115] Peter Ralph, Kevin Thornton, and Jerome Kelleher. "Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes". In: *Genetics* 215.3 (July 1, 2020), pp. 779–797. URL: https://doi.org/10.1534/genetics.120.303253 (visited on 01/11/2023).

[116] "100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report". In: *New England Journal of Medicine* 385.20 (Nov. 11, 2021), pp. 1868–1880. URL: https://doi.org/10.1056/NEJMoa2035790 (visited on 12/20/2021).

[117] Alkes L. Price et al. "Long-Range LD Can Confound Genome Scans in Admixed Populations". In: *American Journal of Human Genetics* 83.1 (July 11, 2008), pp. 132–135. pmid: 18606306. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2443852/ (visited on 09/30/2021).

[118] Shane McCarthy et al. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1279–1283. URL: https://www.nature.com/articles/ng.3643 (visited on 05/05/2018).

[119] Jing Wang et al. "Genome Measures Used for Quality Control Are Dependent on Gene Function and Ancestry". In: *Bioinformatics* 31.3 (Feb. 1, 2015), pp. 318–323. pmid: 25297068. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4308666/ (visited on 12/21/2021).

[120] Klaudia Walter et al. "The UK10K Project Identifies Rare Variants in Health and Disease". In: *Nature* 526.7571 (7571 Oct. 2015), pp. 82–90. URL: https://www.nature.com/articles/nature14962 (visited on 01/02/2023).

[121] Heng Li. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data". In: *Bioinformatics* 27.21 (Nov. 1, 2011), pp. 2987–2993. pmid: 21903627. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198575/ (visited on 10/02/2022).

[122] Christian Fuchsberger, Gonçalo R. Abecasis, and David A. Hinds. "Minimac2: Faster Genotype Imputation". In: *Bioinformatics* 31.5 (Mar. 1, 2015), pp. 782–784. URL: https://doi.org/10.1093/bioinformatics/btu704 (visited on 05/24/2021).

[123] Gavin Band and Jonathan Marchini. "BGEN: A Binary File Format for Imputed Genotype and Haplotype Data". In: (May 2, 2018), p. 308296. URL: https://www.biorxiv.org/content/10.1101/308296v2 (visited on 12/05/2021).

[124] William McLaren et al. "Deriving the Consequences of Genomic Variants with the Ensembl API and SNP Effect Predictor". In: *Bioinformatics (Oxford, England)* 26.16 (Aug. 15, 2010), pp. 2069–2070. pmid: 20562413.

[125] Alison R. Barton et al. "Whole-Exome Imputation within UK Biobank Powers Rare Coding Variant Association and Fine-Mapping Analyses". In: *Nature Genetics* 53.8 (8 Aug. 2021), pp. 1260–1269. URL: https://www.nature.com/articles/s41588-021-00892-1 (visited on 10/02/2021).

[126] Martin D. Tobin et al. "Adjusting for Treatment Effects in Studies of Quantitative Traits: Antihypertensive Therapy and Systolic Blood Pressure". In: *Statistics in Medicine* 24.19 (Oct. 15, 2005), pp. 2911–2935. pmid: 16152135.

[127] Zachary R. McCaw et al. "Operating Characteristics of the Rank-Based Inverse Normal Transformation for Quantitative Trait Analysis in Genome-Wide Association Studies". In: *Biometrics* 76.4 (2020), pp. 1262–1272. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13214 (visited on 10/04/2022).

[128] Alkes L. Price et al. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies". In: *Nature Genetics* 38.8 (8 Aug. 2006), pp. 904–909. URL: https://www.nature.com/articles/ng1847 (visited on 06/10/2020).

[129] Pak C. Sham and Shaun M. Purcell. "Statistical Power and Significance Testing in Large-Scale Genetic Studies". In: *Nature Reviews Genetics* 15.5 (5 May 2014), pp. 335–346. URL: https://www.nature.com/articles/nrg3706 (visited on 10/25/2022).

[130] Jian Yang et al. "Conditional and Joint Multiple-SNP Analysis of GWAS Summary Statistics Identifies Additional Variants Influencing Complex Traits". In: *Nature Genetics* 44.4 (4 Apr. 2012), pp. 369–375. URL: https://www.nature.com/articles/ng.2213 (visited on 07/15/2022).

[131] Zhaozhong Zhu et al. "Shared Genetic and Experimental Links between Obesity-Related Traits and Asthma Subtypes in UK Biobank". In: *The Journal of Allergy and Clinical Immunology* 145.2 (Feb. 2020), pp. 537–549. pmid: 31669095.

[132] Tom G. Richardson et al. "Evaluating the Relationship between Circulating Lipoprotein Lipids and Apolipoproteins with Risk of Coronary Heart Disease: A Multivariable Mendelian Randomisation Analysis". In: *PLoS medicine* 17.3 (Mar. 2020), e1003062. pmid: 32203549.

[133] Saori Sakaue et al. "A Cross-Population Atlas of Genetic Associations for 220 Human Phenotypes". In: *Nature Genetics* 53.10 (10 Oct. 2021), pp. 1415–1424. URL: https://www.nature.com/articles/s41588-021-00931-x (visited on 10/11/2021).

[134] Thomas J. Hoffmann et al. "Genome-Wide Association Analyses Using Electronic Health Records Identify New Loci Influencing Blood Pressure Variation". In: *Nature Genetics* 49.1 (Jan. 2017), pp. 54–64. pmid: 27841878.

[135] Gleb Kichaev et al. "Leveraging Polygenic Functional Enrichment to Improve GWAS Power". In: *American Journal of Human Genetics* 104.1 (Jan. 3, 2019), pp. 65–75. pmid: 30595370.

[136] Fumihiko Takeuchi et al. "Interethnic Analyses of Blood Pressure Loci in Populations of East Asian and European Descent". In: *Nature Communications* 9.1 (Nov. 28, 2018), p. 5052. pmid: 30487518.

[137] Denis Plotnikov et al. "High Blood Pressure and Intraocular Pressure: A Mendelian Randomization Study". In: *Investigative Ophthalmology & Visual Science* 63.6 (June 1, 2022), p. 29. pmid: 35762941.

[138] Julian B. Maller et al. "Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases". In: *Nature Genetics* 44.12 (12 Dec. 2012), pp. 1294–1301. URL: https://www.nature.com/articles/ng.2435 (visited on 10/18/2022).

[139] Jon Wakefield. "Bayes Factors for Genome-Wide Association Studies: Comparison with P-values". In: *Genetic Epidemiology* 33.1 (2009), pp. 79–86. pmid: 18642345.

[140] Kelly A. Frazer et al. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs". In: *Nature* 449.7164 (7164 Oct. 2007), pp. 851–861. URL: https://www.nature.com/articles/nature06258 (visited on 05/26/2021).

[141] Quanli Wang et al. "Rare Variant Contribution to Human Disease in 281,104 UK Biobank Exomes". In: *Nature* (Aug. 10, 2021), pp. 1–6. URL: https://www.nature.com/articles/s41586-021-03855-y (visited on 09/11/2021).

[142] John Wakeley and Tsuyoshi Takahashi. "Gene Genealogies When the Sample Size Exceeds the Effective Size of the Population". In: *Molecular Biology and Evolution* 20.2 (Feb. 1, 2003), pp. 208–213. URL: https://doi.org/10.1093/molbev/msg024 (visited on 01/11/2023).

[143] R. C. Griffiths. "The Two-Locus Ancestral Graph". In: *Lecture Notes-Monograph Series* 18 (1991), pp. 100–117. JSTOR: 4355649. URL: https://www.jstor.org/stable/4355649 (visited on 11/21/2022).

[144] S. N. Ethier and R. C. Griffiths. "On the Two-Locus Sampling Distribution". In: *Journal of Mathematical Biology* 29.2 (Dec. 1, 1990), pp. 131–159. URL: https://doi.org/10.1007/BF00168175 (visited on 11/21/2022).

[145] Matthew D. Rasmussen et al. "Genome-Wide Inference of Ancestral Recombination Graphs". In: *PLOS Genetics* 10.5 (May 15, 2014), e1004342. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004342 (visited on 11/21/2022).

[146] Brian C. Zhang, Arjun Biddanda, and Pier Francesco Palamara. *Biobank-Scale Inference of Ancestral Recombination Graphs Enables Genealogy-Based Mixed Model Association of Complex Traits*. Nov. 4, 2021. URL: https://www.biorxiv.org/content/10.1101/2021.11.03.466843v1 (visited on 11/30/2022). preprint.

[147] J Gay, S Myers, and G McVean. "Estimating Meiotic Gene Conversion Rates From Population Genetic Data". In: *Genetics* 177.2 (Oct. 1, 2007), pp. 881–894. URL: https://doi.org/10.1534/genetics.107.078907 (visited on 12/16/2022).

[148] Alec J. Jeffreys and Celia A. May. "Intense and Highly Localized Gene Conversion Activity in Human Meiotic Crossover Hot Spots". In: *Nature Genetics* 36.2 (2 Feb. 2004), pp. 151–156. URL: https://www.nature.com/articles/ng1287 (visited on 12/16/2022).

[149] Molly Przeworski and Jeffrey D. Wall. "Why Is There so Little Intragenic Linkage Disequilibrium in Humans ?" In: *Genetics Research* 77.2 (Jan. 2001), pp. 143–151. URL: https://www.cambridge.org/core/journals/genetics-research/article/why-is-there-so-little-intragenic-linkage-disequilibrium-in-humans/024A2AF4D33BAC26571FE69642FBFB87 (visited on 12/16/2022).

[150] Amy L Williams et al. "Non-Crossover Gene Conversions Show Strong GC Bias and Unexpected Clustering in Humans". In: *eLife* 4 (Mar. 25, 2015). Ed. by Bernard de Massy, e04637. URL: https://doi.org/10.7554/eLife.04637 (visited on 11/28/2022).

[151] C. Wiuf and J. Hein. "The Coalescent with Gene Conversion". In: *Genetics* 155.1 (May 2000), pp. 451–462. pmid: 10790416.

[152] Leo Speidel et al. "A Method for Genome-Wide Genealogy Estimation for Thousands of Samples". In: *Nature Genetics* 51.9 (9 Sept. 2019), pp. 1321–1329. URL: https://www.nature.com/articles/s41588-019-0484-x (visited on 09/26/2021).

[153] Pier Francesco Palamara et al. "High-Throughput Inference of Pairwise Coalescence Times Identifies Signals of Selection and Enriched Disease Heritability". In: *Nature Genetics* 50.9 (9 Sept. 2018), pp. 1311–1317. URL: https://www.nature.com/articles/s41588-018-0177-x (visited on 01/11/2023).