The Impact of Genetic Drift on the Distribution and Uses of Linkage Disequilibrium

Marcus Tutert

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2021

To my Mom and Dad, without whose never-ending support, love and selflessness, I would not be where I am today. I am eternally grateful. Thank you for being the giants whose shoulders I was able to stand upon.

Acknowledgements

I would like to thank first and foremost my supervisors Gil McVean and Robbie Davies. Thank you Gil, for all your mentorship over the years. I have learnt an incredible amount from you and I have become a more rigorous scientist because of your efforts and support. Thank you Robbie for your incredible help in the past year. Bringing on another supervisor midway through a project is never easy, but you were an incredible boon to my research efforts, and always willing and more than enthusiastic to help improve the quality of my work. Thank you Robert Hinch for your help in this project and mentorship for my first few years at Oxford. To all the members of the McVean lab, thank you for your continued help and for answering any questions that I may have had, however small. I offer a sincere thanks to Kiran, who was a constant source of friendship and help. Thank you to my friends Chris, Xanita and Steph for always being up for an adventure. A heartfelt thanks goes to my sister, Larissa. Your support and curiosity in my work was always appreciated, and I look up to you every single day more than you know. Thank you to all others who have helped throughout the course of my degree including Luke for his keen eye in visualizing data, Andrew for his programming assistance and Jason for his assistance in statistics. Special thanks to my friends at St Cross College including Calum, Gabriel, Joe and Will for always making sure I don't take myself too seriously. Thank you to all my previous scientific mentors, in particular Philip Awadalla and members of the Awadalla lab including Armande, Elyssa, Kim and Heather in inspiring me to pursue scientific research. Thank you to all my friends back home. Last, but not least, thank you Izzy for your patience with me and my work over the last few years both at Oxford and away. Your encouragement and company has been invaluable - thank you for believing that I could do it.

Abstract

With the advent of Genome Wide Association Studies (GWAS), researchers have gained unprecedented access to data in order to examine associations between variants and complex traits and diseases. To analyze these studies, researchers have typically utilized measurements of Linkage Disequilibrium (LD) from the GWAS population or from a reference population which may be genetically drifted relative to the GWAS. However, these measurements are only point estimates and thus, do not capture the stochastic nature of forces such as recombination and genetic drift. Moreover, point estimates from reference populations do not take into consideration errors and inconsistencies in the ancestral composition between the reference and GWAS population. Therefore, in this thesis, we instead model LD as a distribution and demonstrate the extent to which statistical genetic applications can be improved in terms of their robustness and calibration by modelling uncertainty in LD.

First, we characterise the distribution of LD and demonstrate that genetic drift between populations can result in a high degree of variation of the measurements of LD. Additionally, we demonstrate that this variation leads to variation downstream within the accuracy of statistical genetic applications including summary statistic imputation and fine-mapping.

Secondly, to generate LD measurements in a genetically drifted population, we use a reference population and weight the individual reference haplotypes. We demonstrate how repeatedly sampling these weights provides a process to obtain the distribution of LD in the drifted population and how the weights can be updated in a marginal fashion.

Thirdly, using GWAS summary statistics and a reference population, we derive a Maximum Likelihood Estimate (MLE) which infers both the level of noise in the GWAS population as well as the drift between the GWAS and the reference population.

Lastly, we show that by repeatedly updating the weights on our reference haplotypes while incorporating GWAS summary statistic data leads to a method to infer the in-sample distributions of LD from a GWAS. The benefits of our approach over using LD obtained from a reference population is illustrated in the context of fine-mapping admixed populations.

We end by discussing more sophisticated methods for generating our weighted reference haplotypes, the ability to leverage distributions of LD in statistical genetic applications and the effective and efficient data sharing of LD distributions.

Contents

List of Abbreviations

1000G . . . . . 1000 Genomes Project

AFR . . . . . African

AMR . . . . . American

ARG . . . . . Ancestral Recombination Graph

CEU . . . . . Utah residents with Northern and Western European ancestry from the CEPH collection

CHB . . . . . Han Chinese in Beijing, China

DAF . . . . . Distribution of Allele Frequency

EAS . . . . . East Asian

EUR . . . . . European

GWAS . . . . . Genome Wide Association Study

HMM . . . . . Hidden Markov Model

IQR . . . . . Interquartile Range

LD . . . . . Linkage Disequilibrium

MAF . . . . . Minor Allele Frequency

Mb . . . . . Mega-base

MLE . . . . . Maximum Likelihood Estimate

MRCA . . . . Most Recent Common Ancestor

MVN . . . . . Multivariate Normal Distribution

OOA . . . . . Out Of Africa

PIP . . . . . Posterior Inclusion Probability

SAS . . . . . South Asian

SNP . . . . . Single Nucleotide Polymorphism

SSS . . . . . Shotgun Stochastic Search

UKBB . . . . . UK Biobank

WBA . . . . . White British Ancestry

YRI . . . . . Yoruba in Ibadan, Nigeria

1 Introduction

Contents

The non-random association of alleles at different loci, known as linkage disequilibrium (LD) is a sensitive indicator of the population genetic forces that structure genomic variation within a population (Slatkin 2008). LD has been researched in the context of selection (Hill and Robertson 1968), inbreeding (Weir and Cockerham 1969), gene conversion (Wiehe et al. 2000), among many other

areas, which have produced powerful insights within the field of human genetics.

In recent years, many large genome-wide association studies (GWAS) have been conducted that have generated results that have expanded our knowledge about the biology of complex traits (Visscher et al. 2017). In these studies, various statistical genetic methods are employed that either implicitly or explicitly use LD, for example imputation of missing genotypes (Y. Li et al. 2009) and fine mapping to disentangle correlated summary statistics into a set of causal variants (Wallace et al. 2015). However, in these analyses, LD is often treated as a fixed constant, rather than as a measure of a population with inherent sampling variation. Further, the influence that genetic drift plays on the differences in LD between a reference and GWAS population is often overlooked (Benner et al. 2017). Such differences are important to consider in settings when the individual level data of the GWAS population is unavailable - and GWAS summary statistic analyses along with reference population data are required.

1.1 Thesis Aims

In this thesis, we aim to characterise the distribution of LD in populations and to investigate to what extent this distribution is impacted by the result of genetic drift and differences in LD measures among subpopulations or subsamples from the same population. We hope to demonstrate that relying on point estimates of measures of LD from reference populations is not suitable for downstream statistical genetic analyses for two reasons. First, point estimates do not capture the stochastic nature of forces, such as recombination and genetic drift which shape patterns of LD across the human genome and lead to differences among populations (or subpopulations) in haplotpye structure. Secondly, point estimates from reference populations do not take into consideration errors and inconsistencies in the ancestral composition between the reference and GWAS population.

Consequently, we hope that the work in this thesis produces a shift in the analysis approach, such that researchers who intend to carry out GWAS summary statistic analyses will no longer routinely rely on point estimates of measures of LD from reference populations. We hope that researchers will recognize the existence and importance of the distribution of LD and begin to work on leveraging properties of this distribution for downstream applications.

We begin this chapter by establishing how LD materialises within a population and the measurements that are used to describe this property. We describe how genetic drift generates LD and conversely, how forces like recombination act to break down LD between loci. To characterize the stochastic nature of genetic drift and its impact on measures of LD, we describe a Wright-Fisher model which considers the evolution of a population forwards in time. Similarly, we also describe the construction of the genealogical history of a set of present-day samples backwards in time, in a process referred to as the *coalescent*. To better understand the impact of genetic drift on the distribution of LD, we focus on a common, yet often misunderstood and mis-estimated measure of population differentiation, Wright's F_{st} statistic. Next, we highlight common statistical genetic applications that rely on LD such as summary statistic imputation and fine-mapping. In particular, we highlight the ramifications of misspecifying LD in statistical genetic applications, and how a lack of genetic diversity within existing collections of reference population data exacerbates these issues.

1.2 Linkage Disequilibrium in Population Genetics

1.2.1 Properties of Linkage Disequilibrium

Linkage disequilibrium arises in a population when a new mutation, which gives rise to a new allele at a locus, occurs on the background of an allele at another

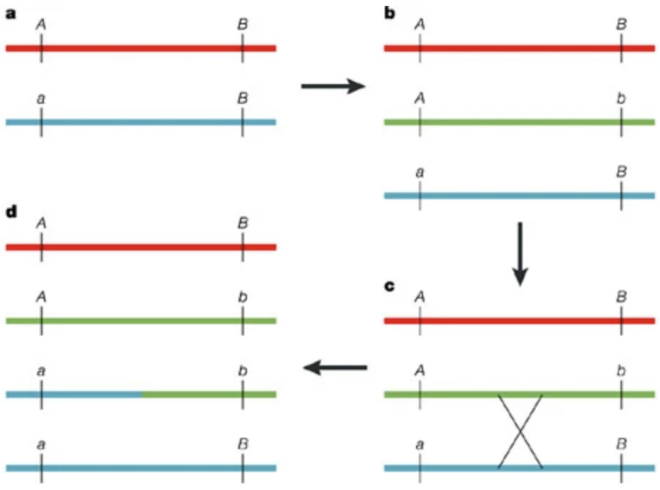

Figure 1.1: Schematic illustrating how LD can arise in a population of haplotypes. a) At the outset, there is a polymorphic locus with alleles A and a. b) When a mutation occurs at a nearby locus, changing an allele B to b, this occurs on a single chromosome bearing either allele A or a at the first locus (A in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The b allele will always be found on a chromosome with the A allele at the adjacent locus. c) The association between alleles at the two loci will gradually be disrupted by recombination between the loci. d) This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (a, b) increases in frequency. Source: Ardlie et al. 2002.

locus (Figure 1.1 a and b). Early on in the lifetime of this new mutation[1], the allele at one locus will be found on the same chromosome with an allele at a second locus at a greater frequency than expected from random assortment (ie. if the loci were segregating independently within the population) (Ardlie et al. 2002).

The most common and fundamental measure for LD is the coefficient of LD D, which measures the covariance between two alleles A and B at two loci (Hill and Robertson 1968) given as:

$$D_{AB} = \pi_{AB} - \pi_A \pi_B, \tag{1.1}$$

where π_{AB} refers to the frequency of haplotypes possessing both alleles A and B, and π_A and π_B correspond to the allele frequencies at the respective loci. The

[1]Prior to recombination occurring.

allele pair AB is commonly referred to as a haplotype, and π_{AB} describes the haplotype frequency. The sign of D is arbitrary and depends on which allele one defines as the reference.

An alternative measure for LD is the correlation coefficient between alleles A and B at the two loci of interest (Weir 1979) given as:

$$r_{AB} = \frac{D_{AB}}{\sqrt{\pi_A(1-\pi_A)\pi_B(1-\pi_B)}}. \tag{1.2}$$

Linkage disequilibrium is a property of the population and therefore can be estimated from the allele and haplotype frequencies in a sample from the population. Along with these estimates, the confidence intervals of measures such as D and r can also be obtained (Jones et al. 2016). Research in this area has focused, in part, on accounting for uncertainty in LD measurements to accurately estimate other population genetic parameters such as the effective population size (Hill and Robertson 1968). These analyses have usually been predicated on measures of LD approaching a steady state equilibrium, balanced by random drift and mutation against recombination within a population (Hill and Robertson 1968, Ohta and Kimura 1969). For instance, Hill and Weir described the variance-covariance structure of LD measures by using moment estimators (Hill and Weir 1988). Hudson, using a Monte Carlo method, studied the joint distribution of LD and allele frequencies (Hudson 1985). Briefly, these investigations have suggested that measures of LD can have high variances and covariances (Hill and Weir 1988). However, these analyses assume properties of the sample of haplotypes such as a very large effective population size, which in real world settings is often not the case (Schiffels and Durbin 2014).

1.2.2 Patterns of Linkage Disequilibrium Arising from Population Genetic Forces

To better understand the structure of LD across the human genome, we describe four population genetic forces: recombination, drift, selection and population structure,

and describe their effect on the patterns and the magnitude of LD.

Firstly, the process of recombination, which is the exchange of genetic material between different organisms will break down the associations between loci (Coop and M. Przeworski 2007). This process will decrease the magnitude of LD from one generation to the next according to:

$$D_{AB}(t+1) = (1-c)D_{AB}(t), \tag{1.3}$$

where t is the index of the generation and c is the recombination rate[2] (Jennings 1917) (Figure 1.1 c and d). The breakdown of LD for pairs of loci which are far apart will occur much quicker than those separated by a short distance (Pritchard and Przeworski 2001). Recombination will act more strongly for loci separated by a large distance, as these loci are more likely to be separated onto different chromatids during chromosomal crossover during sexual reproduction (Foss et al. 1993). Therefore, we refer to loci which are relatively close together on a chromosome as being linked.

Secondly, genetic drift which leads to changes in allele and haplotypes frequencies in a population over time due to sampling variation (Lynch et al. 2016) can generate LD. Consider a population of haplotypes in which two closely linked loci are in linkage equilibrium. Sampling a subset of the population will generate observed non-zero measures of LD, even if the expectation of LD is zero across the entire population (Slatkin 2008). The increase in drift within a stable (non-growing) and small population will also increase LD, as haplotypes gradually become lost from the population over time (Ardlie et al. 2002).

Thirdly, LD can also be influenced by selection which is the differential survival of individuals due to their genetics (Hurst 2009). Balancing selection, which describes the active maintenance of alleles segregating within a population, will permit LD to exist indefinitely in a population (Charlesworth 2006). However, there may also

[2]Where c can not be greater than 0.5.

be suppression of LD around a balanced locus, as in the case of sex chromosome systems (Charlesworth et al. 1997) and therefore balancing selection can lead to a reduction in LD (Schierup et al. 2000). A more complex example of selection influencing patterns of LD exists within selective sweeps. Selective sweeps occur when a beneficial mutation is swept to fixation through natural selection (Smith and Haigh 1974). A selective sweep impacts patterns of linked genetic variation through what is known as the hitch-hiking effect in which the beneficial mutation alters frequencies of alleles at closely linked loci. Interestingly, depending on the relative position of the selected and nearby neutral loci, sweeps can increase, decrease or even eliminate LD entirely (McVean 2007, Stephan et al. 2006).

Lastly, population structure which arises from demographic events such as population subdivision, changes in population size, and the exchange of individuals among populations through migration will affect LD across the entire genome (Slatkin 2008, Goldstein and Weale 2001).

Gene flow, as a result of migration of individuals between subpopulations, will generate LD in each subpopulation when allele frequencies differ between the subpopulations (Slatkin 2008). This result was developed by Nei, who proposed that when isolated populations begin to exchange genes by migration, LD may increase temporarily within the population (Nei and Li 1973).

Changes in the size of a population may also lead to an increase or decrease in LD (Rogers 2014). For instance, the rapid reduction of population size known as a bottleneck (Chan et al. 2006) will result in the loss of haplotypes in the population and increase LD (Ardlie et al. 2002). After the bottleneck has occurred, a reduced population size will also increase the effects of genetic drift, thereby, further increasing the levels of LD within the population (Ohta 1982). The identification of higher levels of genome-wide LD within a population can indicate past bottlenecks in the history of the population (Schmegner et al. 2005, Zhang et al. 2004, Thornton and Andolfatto 2006). For instance, it has been argued that long-distance LD in

non-African human populations is a consequence of a bottleneck which occurred early in human history (Slatkin 2008). Conversely, population growth will lead to a decline in LD due to a larger population size, thereby reducing the effects of genetic drift (Rogers 2014).

1.2.3 Wright-Fisher Model

To better understand the impact of genetic drift on the distribution of LD, we now characterize how a population evolves forward in time through a Wright-Fisher process. (Wright 1931, Fisher 1931). This model has proven to be a very useful tool to mathematically describe the stochastic nature of genetic drift.

The Wright-Fisher model considers a haploid population with constant size of N individuals and a single locus at which two alleles (A and a) are observed. Time is measured in non-overlapping and discrete generations where all individuals within the population die in the current generation when the next generation of individuals are born. We ignore the effects of mutation and selection, and as such, alleles are considered neutral. This results in equal probability for each individual to produce offspring. Therefore, reproduction is considered a random sampling process whereby alleles are inherited by the successive generation drawn with replacement from the gene pool of the current generation.

Each draw of this random sampling process results in either allele A or a. This process describes the binomial sampling of alleles at every generation. Let X_t be the number of alleles A in generation t. If i individuals in the population carry the allele A, then the probability of sampling the A allele is equal to its frequency in the current generation t, $\pi_i = \frac{i}{N}$. In the subsequent generation, $t+1$, the probability of observing j copies of the allele, $X_{t+1} = j$ is binomially distributed as:

$$P(j|i) = \binom{N}{j} \pi_i^j (1-\pi_i)^{N-j}. \tag{1.4}$$

Using this formulation, we can also calculate the expected number of alleles in generation $t+1$ given the number of alleles in generation t as:

$$\mathbb{E}[X_{t+1}|X_t] = N\pi_i = N\frac{X_t}{N} = X_t. \tag{1.5}$$

Similarly, the variance of alleles in generation $t+1$ given the number of alleles in generation t is:

$$Var[X_{t+1}|X_t] = N\pi_i(1-\pi_i) = X_t(1-\frac{X_t}{N}). \tag{1.6}$$

Over time, due to the random sampling of alleles at each generation, allele frequencies will *drift* as a consequence of the underlying stochastic nature of reproduction. At each generation, the fixation or loss of an allele may occur which results in the number of copies of the allele remaining at N or 0 respectively for all future generations. Furthermore, from the properties of the binomial distribution, as the population size increases, the variance of the allele frequency at each generation will decrease. This implies that genetic drift will have a smaller effect when a population is large, but a larger effect when the population is small (Figure 1.2).

Carrying out statistical inference in a Wright-Fisher model involves analysing properties of the distribution of the allele frequency (DAF) (Tataru et al. 2015). However, there is no tractable analytical form for this distribution (Ewens 2004) and therefore researchers have had to rely on numerical and analytical approximations. One particular approach involves using moment-based approximations, which aim to fit the first moments (ie. the mean and variance) of the true DAF. These approximations typically use either the beta distribution (Balding and Nichols 1995) or the normal distribution (Nicholson et al. 2002). These distributions are motivated by the diffusion limit of the Wright-Fisher model. For instance, the beta distribution is the stationary DAF under neutral evolution with recurrent mutation (Wright 1945).

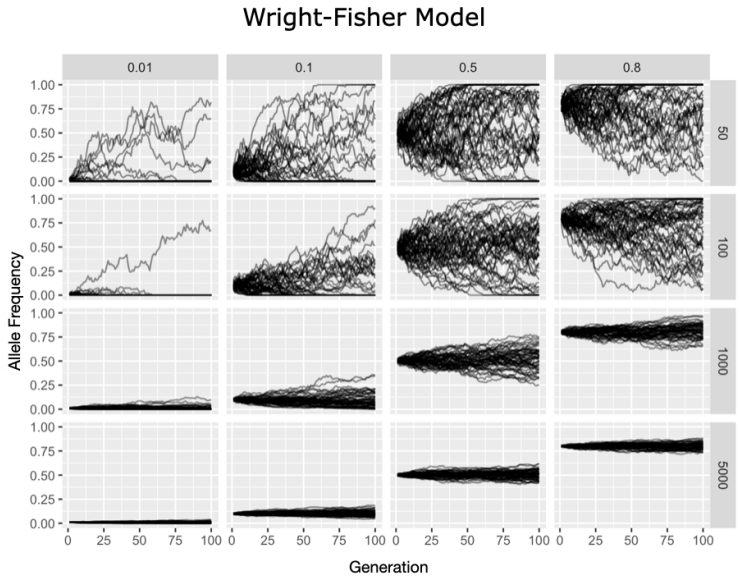

Figure 1.2: Simulation of a Wright-Fisher model to illustrate the dynamics of genetic drift and the resulting change of allele frequencies over time. We ran each replicate simulation 100 generations forwards in time beginning each of the replicate simulations at a given allele frequency. The titles on the columns of the subplots denote the initial starting allele frequency of each of the simulations. The title on the rows of the subplots denotes the haploid population size (N) in each replicate simulation. Each graph displays the allele frequency on the y-axis and the generation time on the x-axis. Each line corresponds to a single a simulation, and we ran 50 replicate simulations in each setting. Code generated using template adapted from `https://stephens999.github.io/fiveMinuteStats/wright_fisher_model.html`

Further, various extensions to the Wright-Fisher model have been introduced which allow for a more realistic characterisation of human genetic variation. For instance, incorporating mutations at the alleles (Dangerfield et al. 2012), considering more than one loci (Ethier and Nagylaki 1989) and introducing recombination (Griffiths, Jenkins, et al. 2016). The two locus model with recombination (Griffiths 1981) is a particularly useful extension to the Wright-Fisher model in the context of this thesis. Under this model, we are able to examine measures of LD over time, and assess the effect of recombination and drift on these measures. In Chapter 2, we will outline and simulate this model in detail.

The Wright-Fisher model also deepens our understanding of the causes of LD outlined in the previous section in two ways. Firstly, this model elucidates how

measures of LD are by definition, affected by genetic drift as a consequence of the change in allele frequencies in a population over time. Secondly, under this model, we can obtain expressions that relate the size of the population to the level of genetic drift within the population.

1.2.4 Genealogical Interpretation of Linkage Disequilibrium

In a Wright-Fisher model, we consider each successive generation formed by the random sampling of alleles from the gene pool of the current generation. In other words, we discussed the forward in time evolution of a population. However, given a set of present-day genetic samples, one can also construct a backwards in time approach which characterizes the genealogical relationship of the present-day samples.

Tracing the ancestral lineages backwards in time from the set of present-day samples is performed through a process known as the *coalescent*, whose statistical properties were originally developed by Kingman (Kingman 1982).

Under the coalescent, ancestral relationships between individuals are represented as lineages in a genealogical tree. Given our set of present-day samples, at each generation, individuals independently choose one ancestor at random. If by chance, two individuals were to choose the same ancestor at some point backwards in time, we join the lineages in what is known as a coalescence event. The process of coalescing lineages continues until we have only one remaining lineage which we define as belonging to the most recent common ancestor (MRCA) of our present-day samples.

In the coalescent, mutations are considered neutral, and therefore, independent of the genealogical process (Wakeley 2009). Mutational events are added to the coalescent by superimposing mutations on branches proportional to their length. The population-scaled mutation rate (for a haploid sequence) is given by

$$\theta = 2N_e\mu, \tag{1.7}$$

where μ defines a constant rate of mutation per site per generation and N_e is the effective population size (Wakeley 2009). θ is assumed to be constant as $N_e \to \infty$. The factor of two is added for historical reasons and is related to the expected number of pairwise differences between two haploid sequences (Tajima 1993, Wakeley 2009).

Recombination along the sequence of a sample can also be incorporated within the coalescent, an idea first introduced by Hudson (Hudson 1983). Unlike mutational events, recombination will substantially effect the structure of the genealogy (ie. the topology). Recombination will alter the genealogical relationship between different genetic segments, such that the two chromosomes may be closely related at one segment, but distantly related at another segment. One way to visualize the effect of recombination on the effect of the structure of the genealogy across the sequence is by noting that recombination will introduce a sequence of marginal trees of different topologies (Figure 1.3). This representation is most commonly shown through an ancestral recombination graph (ARG) (Hudson 1995).

Coalescent models with recombination are traditionally based on the *diploid* Wright-Fisher model (Tran et al. 2013). Populations in this model will contain $2N$ gene copies at each locus. The dynamics of the population depend on the population-scaled recombination parameter $\rho = 4N_e r$, where r is probability of recombination per locus per generation, as well as the population scaled mutation rate[3] (Equation 1.7)

The characterization of the genealogical relationship between a set of present-day samples is important within the context of LD for several reasons. For instance, the levels of LD between loci in a population is a consequence of the genealogical relationship of the samples within the population. (Nordborg and Tavaré 2002). Consider tracing the path of a mutation on a tree through its genealogical history to the MRCA that introduced this mutation into the population. Haplotypes that

[3]Multiplying by a constant of 4 rather than 2, to consider a diploid rather than haploid sequence.

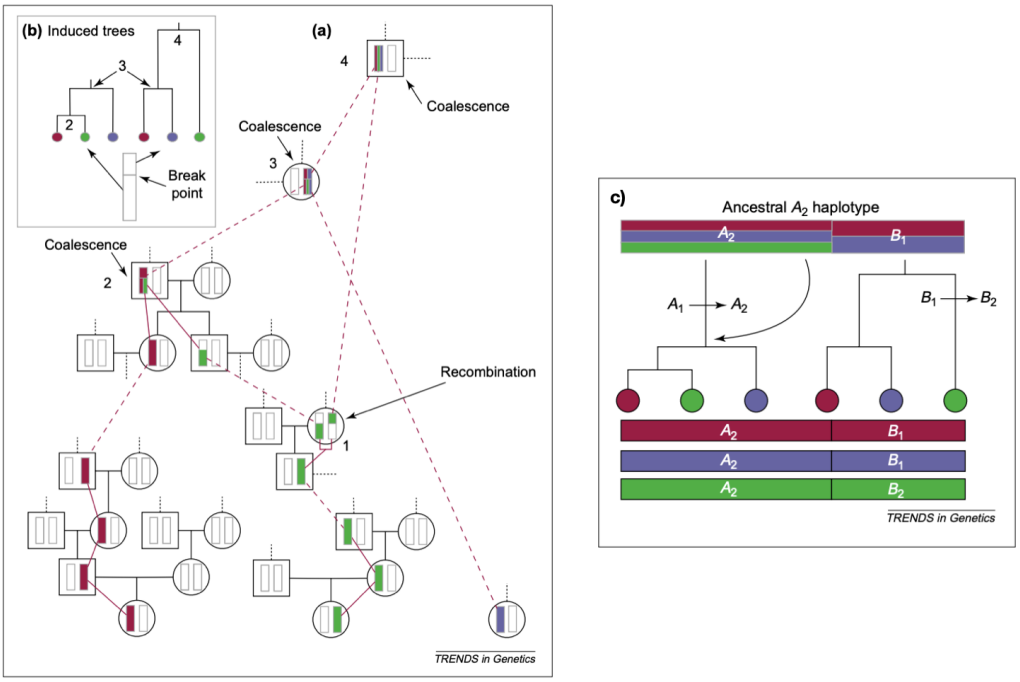

Figure 1.3: A genealogy for three copies of a short chromosomal segment (left) and displaying the genetic polymorphism as the result of mutations along the branches of this genealogical tree (right). (a) Tracing the segmental lineages back in time, we observe the following events: (1) the 'green' lineage undergoes recombination and splits into two lineages, which are then traced separately; (2) one of the resulting green lineages coalesces with the 'red' lineage, creating a segment, part of which is ancestral to both green and red, part of which is ancestral to red only; (3) the 'blue' lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and red, partially ancestral to all three colors; (4) the 'other' part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colors in its entirety. (b) The recombination event induces different genealogical trees on either side of the break. Genetic polymorphism is the result of mutations along the branches of genealogical trees. The genealogical trees for linked chromosomal positions will not, in general, be statistically independent. Therefore, neither will the allelic states of linked loci be statistically independent, that is, there will be LD between the loci. This is illustrated here by adding two unique mutations to the trees from Fig. 1. The mutation at locus B occurred earlier than the mutation at locus A. The A2 allele therefore arose in a population that was polymorphic with respect to locus B. The most recent common ancestor of the A2 alleles carried a B1 allele. Two of the three haplotypes shown here still do; the third recombined with a haplotype carrying a B2 allele (as depicted as in the genealogy on the left). Source: Nordborg and Tavaré 2002.

inherit this mutation must also inherit from the MRCA, a stretch of genetic material in the surrounding area (Slatkin 1994, Nordborg and Tavaré 2002). This implies that closely linked loci will be highly correlated while distantly linked loci will be roughly independent of one another (Nordborg and Tavaré 2002). Therefore, the allelic states (generated through mutation) of pairs of closely linked loci will implicitly be correlated and in LD with each other (Figure 1.3).

Additionally, an approximation for the expected measures[4] of LD for a given population, σ_d^2, can be calculated by the covariance of coalescence times for pairs of samples in the population (McVean 2002). Therefore, the behaviour of LD under different demographic events such as geographical subdivisions (Kruglyak 1999), bottlenecks (Reich et al. 2001) and population growth (Slatkin 1994) can be interpreted in terms of the underlying genealogy of the population.

A similar construction to derive the genealogical relationship of a set of present-day samples considers starting with a single ancestral lineage (the MRCA) and moving forwards in time by splitting the ancestral lineage repeatedly, in what is sometimes informally referred to as the forwards in time coalescent[5] (Griffiths and Tavaré 1998, Griffiths and Tavaré 1994).

This construction helps to describe properties of the genealogy such as the number of descendants conditional on a single lineage and provides us with the distribution of the allele frequencies in a sample of haplotypes. We motivate the derivation of this distribution by noting that measures of LD such as D and r explicitly rely on the allele frequencies at pairs of loci. Therefore, characterising the distribution of allele frequencies is a step towards characterising the distribution of LD within a population.

[4]This is an expression for the ratio of expectations of $\mathbb{E}[D_{AB}^2]$ and $\mathbb{E}[\pi_A(1-\pi_A)\pi_B(1-\pi_B)]$, developed by Ohta and Kimura (Ohta and Kimura 1971).

[5]Note that this is different than the Wright-Fisher model which describes a population evolving forwards in time as a random sampling process.

In this construction, we begin by counting the number of alleles on a genealogical tree. To do so, we trace backwards in time through the tree until we observe a mutational event. This event indicates that an ancestor which carries the mutation will have passed the allele generated by this mutation to its descendants. As a result, ancestral lineages can either die through mutation or through coalescence. This process will be repeated until we are left with the final ancestral lineage which represents some final type of allele. The number of lineages in the construction of the forwards in time coalescent will decrease monotonically from N to 1 (Figure 1.4).

To understand the statistical properties of the forwards in time coalescent and its relationship to the distribution of allele frequencies of the sample haplotypes, we use an urn model which offers an efficient simulation framework in population genetic and probability theory (Kotz and Balakrishnan 1997). The Hoppe urn model (Donnelly 1986, Hoppe 1987) defines the distribution of the allelic types and frequencies for N haplotypes under an infinite alleles model (Kimura and Crow 1964)[6]. In this urn model, balls of different colours represent different alleles. The number of balls within the urn of colour (ie. allele type) i is denoted as α_i. Furthermore, the urn contains an extra white *mutation* ball which possesses a mass θ relative to the other coloured balls of mass one. While j coloured balls remain in the urn, we sample from the $j+1$ balls in the urn with probability proportional to the mass of the ball. If we have drawn the white ball, we return this ball to the urn and introduce one new coloured ball (of a colour not yet seen) into the urn. If we have drawn the coloured ball, we return the ball to the urn and add in an additional ball of the same colour. We stop this process when there are N non-white balls remaining in the urn. The connection between drawing balls from the Hoppe urn and the construction of the forwards in time coalescent is described in Figure 1.4.

[6]The infinitely-many-alleles model specifies that every mutation makes a new type of allele, never seen before in the population.

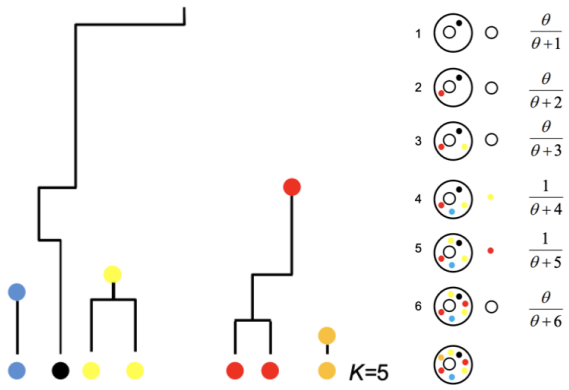

Figure 1.4: Genealogical construction of the Hoppe urn model. Different coloured circles on the tree indicate different allelic types. The drawing of the urn balls is shown on the right with probabilities indicated. Note how the probabilities of the numbered events are identical in both the urn and the genealogical tree and that no formulation of this model currently exists with recombination between the samples. Figure taken with permission from Simon Myers' notes on the coalescent: `http://www.stats.ox.ac.uk/~myers/mathgenteaching.html`.

If we define a random variable X_n as the allele type (ie. colour) of the n^{th} additional ball added to the urn, then:

$$P(X_{n+1} = i \mid X_1, X_2, ..., X_n) = \frac{\alpha_i + v_i(n)}{\theta + n}, \tag{1.8}$$

where $v_i(n)$ is the number of copies of allele i after we have drawn n balls from the urn (Hoppe 1987). As the limit of our draws approaches infinity we have:

$$\lim_{n\to\infty} \frac{v_i(n)}{n} = \pi_i, \tag{1.9}$$

where π_i is the i^{th} allele frequency. The partitions $(\pi_1, \pi_2, ..., \pi_K)$ which define the allelic frequencies of all K allele types within our sample of haplotypes will be Dirichlet distributed as $Dir(\alpha_1, ..., \alpha_K)$ (Watterson 1976, Donnelly 1986, Hoppe 1987).

1.3 F_{st} as a Measure of Genetic Differentiation

Earlier, we highlighted the role of genetic drift in generating LD within a population. As a result of genetic drift, populations that are no longer freely mixing will

become differentiated over time (Holsinger et al. 2009). A popular measure of this differentiation is F_{st}.

F_{st} was originally developed by Wright (Wright 1951) and was related to the probability of identity by descent. Therefore, F_{st} is sometimes known as the co-ancestry coefficient. Wright defined F_{st} as it relates to an underlying property of the genealogical process, rather than as a statistic that one can calculate from the data. F_{st} can be also formulated in terms of the progress of the fixation of an allele and is therefore sometimes referred to as a fixation index.

A natural interpretation of F_{st} suggested by Wright involves the ratio of variances of the allele frequencies, akin to a classical F-statistic. Cockerham (Cockerham 1969) formalized this definition as:

$$F_{st} = \frac{Var(\pi_i)}{\pi(1-\pi)}, \tag{1.10}$$

where $Var(\pi_i)$ is the variance of the allele frequency across subpopulations and π is the mean allele frequency of the entire population, assuming a single bi-allelic locus. Another useful estimator for F_{st} was suggested by Hudson (Bhatia et al. 2013). This estimator is independent of sample size even when F_{st} is not identical across populations and corresponds to a simple average of the population specific F_{st} estimates.

Confusingly, F_{st} can be interpreted both as a parameter of the underlying genealogical process (Weir and Cockerham 1984), or as a statistic of the samples drawn from a population (Takahata and Nei 1984). Traditionally, F_{st} is estimated across an entire region as the information contained at a single allele is usually not informative (Holsinger et al. 2009). Thus, it is often times necessary to combine information in a principled manner, either across SNPs or across different subpopulations. Estimates of F_{st} can also change due to the presence or absence of rare SNPs. Consequently, the appropriate choice of the SNPs used in the analysis

is an important consideration to avoid ascertainment bias (Bhatia et al. 2013) as differences in these estimates of F_{st} have the potential to alter inference regarding demographic events in the history of a population (Holsinger et al. 2009).

Measures of population differentiation can also be derived in the context of specific demographic events (Meirmans and Hedrick 2011). For instance, consider an ancestral population which has split some time ago t (in generations) in the past and have formed two isolated randomly mating populations. The estimates of population split times and the levels of genetic differentiation between the diverged populations have played a substantial role in the understanding of human population history (Slatkin 1987). A notable result surrounding this demographic event was developed by Nielsen (Nielsen et al. 1998) who derived a MLE for the divergence time between the two diverged populations under a coalescent framework. Nielsen derived that the estimated divergence time $\hat{T}$ between the pair of populations is given as:

$$\hat{T} = -log(1 - \hat{F_{st}}), \tag{1.11}$$

where $\hat{F_{st}}$ is an estimate of F_{st} within the population. Note that the divergence time T is scaled such that $T = \frac{t}{2N_e}$ where t is the number of generations since the two populations diverged and N_e is the effective population size of the diverged populations.

One final result we highlight to relate our discussions of genetic drift, population differentiation and measures of F_{st} is the Balding-Nichols model of genetic drift (Balding and Nichols 1995). Here, the beta model is used as an approximation to the DAF across identical and derived subpoulations separated by Wrights F_{st} from an ancestral population with allele frequency π as:

$$Beta(\lambda\pi, 1 - \lambda\pi), \tag{1.12}$$

where $\lambda = \frac{1-F_{st}}{F_{st}}$ is parametrised such that this distribution possesses the same the mean and variance as the Wright-Fisher DAF.

1.4 Applications of Linkage Disequilibrium in Statistical Genetics

The ability to measure and leverage the properties of LD across the human genome has played an enormous role in statistical genetic applications. We will focus on reviewing the role of LD in two common usages: performing genotype and summary statistic imputation and improving the detection of causal variants in a GWAS through fine-mapping.

1.4.1 Genotype and Summary Statistic Imputation

Genotype imputation refers to the prediction of variants within a sample of individuals that were not originally measured (Y. Li et al. 2009). Genotype imputation is most commonly used to increase power of a GWAS study but can also aid in combining results across different GWAS cohorts that may rely on different genotyping platforms (Marchini and Howie 2010).

The process of genotype imputation typically involves a GWAS dataset generated with a genotyping array. This is a relatively cost-efficient method designed to interrogate common genetic variation at markers across the genome (Tam et al. 2019). The set of genotyped markers in the GWAS is compared to a reference population of haplotypes which contain a much higher number of markers that have been usually constructed through the more expensive process of whole-genome sequencing (Schwarze et al. 2018). Several reference populations are available for this purpose including the Haplotype Reference Consortium (The Haplotype Reference Consortium 2016), the HapMap project (The International HapMap Consortium 2005) and the 1000G project (The 1000G Project Consortium 2015). Genotype imputation is performed by identifying stretches of shared haplotype sequences and the missing genotypes in the study samples are filled in by copying the alleles in the matching reference haplotypes (Figure 1.5) (Howie et al. 2009).

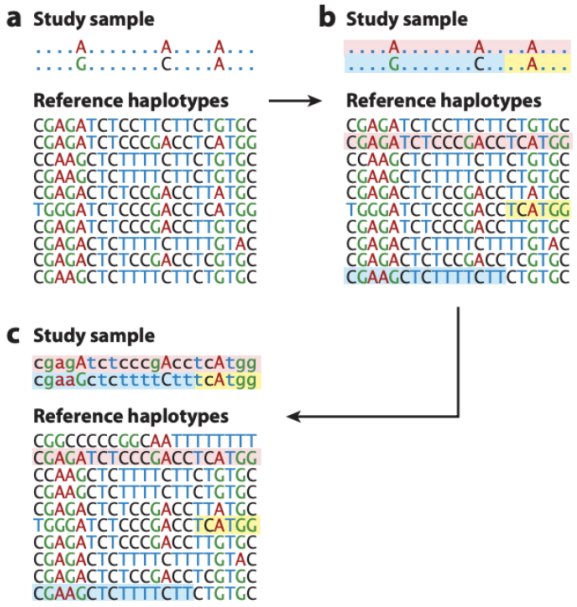

Figure 1.5: Genotype imputation in a sample of apparently unrelated individuals. (a) The observed data, which consists of genotypes at a modest number of genetic markers in each sample being studied and, in addition, of detailed information on genotypes (or haplotypes) for a reference sample. (b) The process of identifying regions of chromosome shared between a study sample and individuals in the reference populations. (c) Observed genotypes and haplotype sharing information have been combined to fill in a series of unobserved genotypes in the study sample. Source: Y. Li et al. 2009.

Although genotype imputation is a powerful tool, it has several drawbacks. For instance, the process is computationally intensive and requires the individual level genotypes from the GWAS study as input (Das et al. 2016). Therefore, an alternative to genotype imputation is summary statistic imputation, which directly imputes the GWAS summary statistics from missing variants. Summary statistic imputation only requires the GWAS summary statistics and the LD measures between variants (Wu et al. 2020). Briefly, this process uses the z-scores within the summary statistics which are assumed to be normally distributed under the null model of no association. Therefore, under the null, the vector of z-scores Z at SNPs in a locus are approximately distributed in a multivariate normal fashion as $Z \sim \text{MVN}(0, \Sigma)$ where Σ is the correlation matrix corresponding to the LD between the SNPs (Pasaniuc, Zaitlen, et al. 2014). In Chapter 2, we will fully

derive the steps necessary to impute summary statistics given the typed summary statistics and a LD panel from a reference population.

In recent years, summary statistic imputation has become a popular tool to re-analyze existing GWAS summary statistics deposited online to search for associated variants to a phenotype that were previously untyped, or were not imputed. For example, Pasaniuc used summary statistics from a large meta-analysis of four lipid traits to re-release the resulting imputed summary statistics at 1000G variants (Pasaniuc, Zaitlen, et al. 2014). Summary statistic imputation has been shown to produce novel associations using existing GWAS summary statistics. The GIANT project, a GWAS which examined the genetics underpinning of height (Allen et al. 2010), was re-analyzed with summary statistic imputation and 34 novel loci were identified, 19 of which were replicated in the UK Biobank (Rüeger et al. 2018).

1.4.2 Fine-mapping

The presence of LD between variants in a GWAS both helps and hinders statistical genetic analyses. On one hand, the LD between variants helps to find additional regions of association across the genome. Conversely, the LD between the the causal and non-causal variants in the region complicates and hinders the challenge of identifying the causal variants themselves. To overcome this obstacle, methods known as fine-mapping were developed in order to statistically infer the underlying causal variant(s) by accounting for LD between variants (Wallace 2020).

After the collection of sequencing data, GWAS association results are produced. These associations are typically performed using univariate analyses, for instance using linear or logistic regression and include covariates to correct for population structure in the GWAS population (Visscher et al. 2017). These results are then segmented into any number of regions of interest, typically based on the strength of association passing above a significance threshold[7] (Fadista et al. 2016).

[7] Traditionally $-\log_{10}(p_{\text{value}}) \geq 1e-8$.

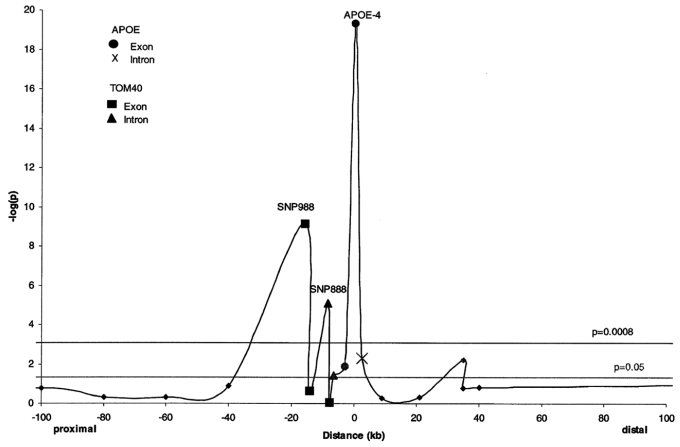

Figure 1.6: Plot of $-\log_{10}(p_{\text{value}})$ for case-control test for allelic association with Alzheimer's Disease, for SNPs immediately surrounding APOE (<100 kb). Source: Scott et al. 2000.

Unfortunately, due to the presence of LD, selecting the SNP with the highest association (known as the lead SNP) or naively ranking the associations in a sorted order is an inexact process to infer causal variants as the tagged variant may actually be correlated with the true and potentially unmeasured causal variant. Inference is further complicated when the causal variant has been imputed, thereby resulting in a potential loss of power to detect the true level of association of the imputed variant (Faye et al. 2013).

Furthermore, LD does not necessary decrease in a monotonic pattern according to distance from the lead SNP. A prime example of this phenomenon is given by the *APOE* locus on chromosome 19 and its association to Alzheimer's disease. In this region, the p-values from a GWAS do not follow a monotonic decreasing pattern due to the large and complex patterns of LD (Figure 1.6) (Scott et al. 2000).

One possible approach to disentangle the set of causal variants is the use of conditioning the association signals in a region (Spain and Barrett 2015). This process involves an iterative procedure in which one conditions the association of SNPs using the LD measures of the lead SNP until no SNPs remain above a significance threshold. Unfortunately, while these approaches are able to generate

the most likely sources and number of signals in a region, they lack the probabilistic rigour of assigning the probabilities of SNPs being causal in a region. Therefore, Bayesian methods have been developed to establish a probabilistic framework for fine-mapping analyses (Maller et al. 2012).

Traditionally, fine-mapping has used the full genotype-phenotype data-set. However, recently, for computational storage and genetic privacy reasons, methods have been developed which instead use the GWAS summary statistics (typically z-scores and/or p-values), along with LD measures from an ancestrally similar reference population. Several methods including: PAINTOR (Kichaev et al. 2014), CAVIAR (Chen et al. 2016) and SuSie (Wang et al. 2020), were all developed in the context of performing fine-mapping using GWAS summary statistics.

One novel development in the fine-mapping software field is FINEMAP (C. Benner et al. 2016). This algorithm uses a shotgun stochastic search (SSS) algorithm which quickly explores the space of causal configurations by focusing computational power on causal configurations with non-negligible probabilities (Hans et al. 2007). FINEMAP is thousands of times faster than competing methods and scales much more efficiently with the number of causal variants in the region.

FINEMAP along with other Bayesian based fine-mapping methods focus on the set of configurations of SNPs most likely to be causal in a region (Guan and M. Stephens 2011, Wilson et al. 2010). This is done by computing the posterior probability of a specific model M, conditional on the observed data D. Typically, the results of these Bayesian methods are given as the posterior probability of inclusion (PIP) for each respective SNP. The PIP for a given SNP j, is computed by the sum of the posteriors within the model where the SNP j is causal ($c_j = 1$). Formally:

$$PIP_j = P(c_j = 1|D) = \sum_{M, c_j=1} P(M|D). \tag{1.13}$$

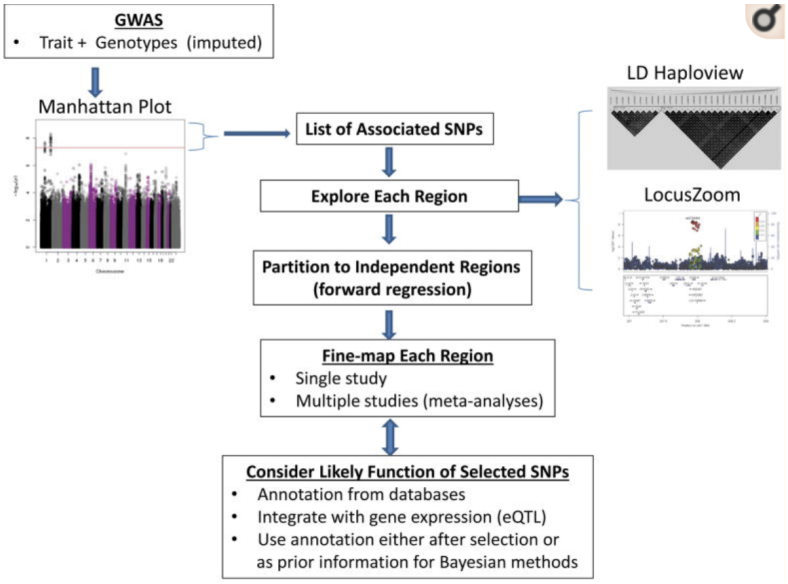

Figure 1.7: Workflow of a typical GWAS analysis from the initial sequencing and data collection to annotation of SNPs selected from fine-mapping analyses. Source: Schaid et al. 2018.

Bayesian fine-mapping methods usually output an α credible set. A typical construction of the credible set is performed by ranking SNPs (from highest to lowest PIPs) and including SNPs within the credible set until the cumulative sum of the PIPs reach α, usually set at 95 percent.

Fine-mapping can also be performed across ancestrally diverse populations in a process known as trans-ethnic fine-mapping (Asimit et al. 2016). This technique has the potential to uncover associations of variants that are consistent across populations, with similar direction and magnitude of effect size (Shi et al. 2020). The premise of trans-ethnic studies is that differences in LD patterns can more accurately deconvolute signals of association and variant tagging within a GWAS. In particular, the inclusion of samples with African ancestry have shown great promise in reducing the size of the resulting credible set (Hutchinson et al. 2020). This is due to the low levels of LD as well as an overall higher level of genetic diversity in the African population relative to European or East Asian populations (Li and Yu 2014).

The process of fine-mapping aims to identify variants that can be followed-up with functional experiments in the lab, often times to aid in the development of

therapeutics to treat disease (Broekema et al. 2021). For instance, as a result of GWAS analyses, loss-of-function mutations in *SLC30A8* which encodes a zinc transporter expressed in pancreatic islets was shown to be protective for Type 2 Diabetes (Flannick et al. 2014). This has led to the development of drugs that act as antagonists for these transport mechanisms (Adulcikas et al. 2019).

1.5 Consequences of Misspecified Linkage Disequilibrium

The application of LD in statistical genetics has led to increases in the power to detect associations of a variant with a phenotype through imputation, development of fine-mapping methods to identify causal variants and many other novel applications. However, it is not always the case that researchers can use the individual level data from the GWAS population and consequently, in-sample measures of LD. Researchers will therefore instead use reference populations as a substitute. Figure 1.8 describes a typical workflow in a GWAS setting where either the in-sample and optimal LD measures can be used or the LD measures from a reference population are used to perform downstream fine-mapping and other statistical genetic applications (Benner et al. 2017). In principle, these reference populations will closely match the GWAS population in terms of their ancestral composition, quality control pipeline and many other salient properties.

However, heterogeneity often exists between data garnered from GWAS and reference populations. For instance, differences in ancestry, genotyping platform and analysis pipelines (Zuvich et al. 2011) have the potential to generate differential associations. Populations with relatively similar ancestries, such as the European subpopulations in the 1000G, still have marked differences in LD patterns (Novembre et al. 2008) which can lead to inaccurate downstream analyses (Benner et al. 2017). Furthermore, reference populations may contain errors from genotyping or

imputation that are not reflected within the GWAS summary statistics (Anderson-Trocmé et al. 2020). Conversely, the GWAS itself may contain errors of a similar nature (Laurie et al. 2010).

Further, to choose the appropriate reference population, the size of the GWAS population must also be taken into consideration. Typically, a well powered GWAS study now scales into the orders of 100,000 cases and controls or more (Visscher et al. 2017). However, the most widely available and used reference populations for statistical genetic applications, the 1000G project, contains only 2504 samples. Consequently, this makes estimating and substituting the LD measures in GWAS populations using this reference population challenging due to the lower resolution of the LD obtained from the reference population relative to the GWAS LD.

A possible solution to this problem of misspecification involves the regularization of the LD panel generated from the reference population (Pasaniuc, Zaitlen, et al. 2014). This can be accomplished in several manners. The most simple involves setting all correlations between SNPs greater than some distance apart to be zero or to adjust the correlation structure based on a recombination map (Wen and M. Stephens 2010). Alternatively, independent LD blocks could be generated from the reference population data (Berisa and Pickrell 2016).

Misspecification of the GWAS LD measures using a reference population can confound downstream fine-mapping analysis in many problematic ways (Benner et al. 2017, Schaid et al. 2018). For example, a GWAS meta-analysis for lipid levels found inconsistencies between the LD measures from the EUR population in the 1000G and the available z-scores from the summary statistics (Chen et al. 2016). This study highlighted a pair of SNPs, rs2144300 and rs4846914 which were in perfect LD measures ($D = 1$) in the EUR population. However, the z-scores from the summary statistics were 9.16 and 9.442. After fine-mapping, this resulted in rs2144300 with a PIP of 0.05, much lower than the PIP of 0.61 for rs4846914. This result is inconsistent with the correlation from the EUR population suggesting that

these two SNPs should have the same (or at the least, very similar) PIPs. The authors suggested two possible reasons for the source of this inconsistency. First, the LD pattern in the EUR panel was not reflective of the true in-sample GWAS LD patterns. Secondly, and most likely to be the source, meta-analyzed z-scores of the imputed genotypes were not computed for the same set of individuals. Such an effect often happens when a quality control threshold (such as the imputed INFO-score) is employed in meta-analyses (Zhu and M. Stephens 2018).

1.5.1 Sharing and Adapting Linkage Disequilibrium Information

To avoid the ramifications of using mis-specified LD panels from reference populations, it has been suggested to *share* LD data (Zheng et al. 2017). This would have the benefit of no longer needing to use the same reference populations (eg: 1000G European population) for every GWAS but instead one would use the GWAS population located on some online repository. There are two advantages of this approach. First, the dimensions of the LD panel do not scale with the number of individuals in the GWAS making data exchange easier. Secondly, measures of LD do not contain any identifying information about the participants in the study thus protecting genetic privacy (Erlich and Narayanan 2014).

However, many drawbacks with this approach still exist. Firstly, sharing LD information places the onus on researchers to deposit these LD panels onto online repositories. Secondly, GWAS data is currently located across the internet on various public repositories, therefore, collating one meta repository of LD information is extremely challenging (Benner et al. 2017). Thirdly, sharing LD panels is done after quality control has been performed on the GWAS data. Consequently, these panels will not present a full and holistic picture of which variants and samples were filtered at which steps in the quality control pipeline. Such information

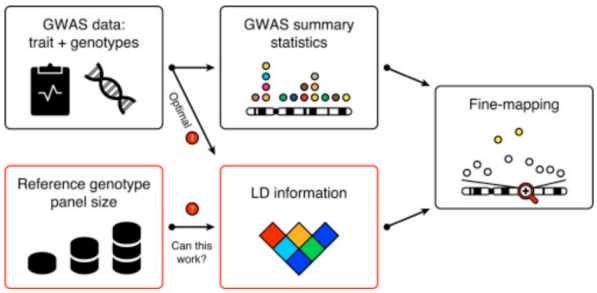

Figure 1.8: Schematic of fine-mapping causal variants in trait-associated genomic regions by Using GWAS summary statistics and LD information. Source: Benner et al. 2017.

is important for downstream interpretation of GWAS analyses, and still will be absent when LD information is shared.

Instead of directly sharing LD information, substantial work has recently been conducted regarding the development of tools to use existing reference populations to adapt to the specific GWAS population at hand. Two commonly used methods that use this approach are DISTMIX (D. Lee et al. 2015 and Adapt-mix (Park et al. 2015). These two tools have a statistical framework that places weights onto different subpopulations in a reference population to more closely match the ancestral composition of the GWAS. However, such an approach has many drawbacks. For instance, these methods are developped solely for the purpose of obtaining imputed summary statistics. As such, these methods do not infer the GWAS in-sample LD measures. Therefore, the use of such a tool for other statistical genetic analyses (eg. fine-mapping) is not possible. In addition, these tools do not incorporate error within the GWAS summary statistics or within the reference population. Furthermore, these tools are not applicable for data-sets with admixed individuals due to the high variability of ancestral proportions across the genome. Lastly, these methods apply weights to the subpopulations in the reference population (eg. 1000G EUR, AFR, etc.) rather than onto individual haplotypes. This leads to the tools being reliant on the accurate labelling of the subpopulations.

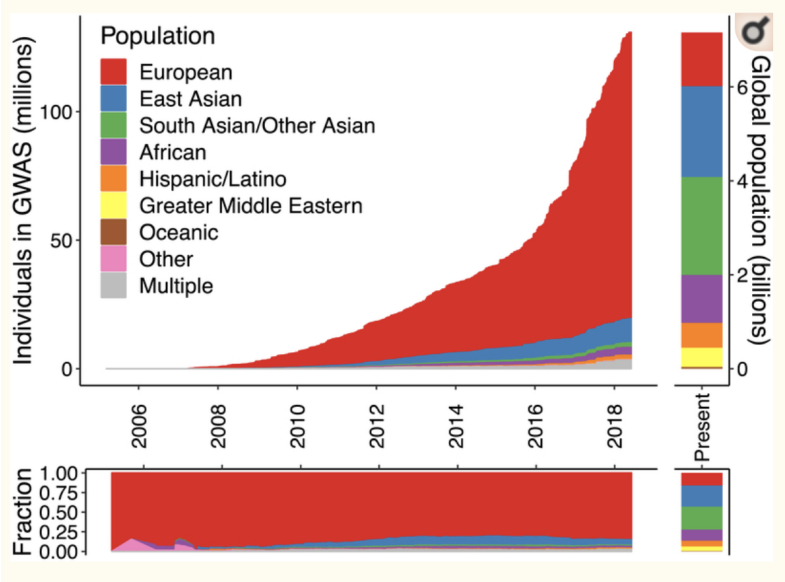

Figure 1.9: Ancestry of GWAS participants over time compared to the global population. Cumulative data as reported by the GWAS catalog. Individuals whose ancestry is "not reported" are not shown. Source: Martin et al. 2019.

1.5.2 Data Resources and the Lack of Genetic Diversity

The use of genetic data from existing reference populations also present a significant challenge as the collection of existing genetic data for both reference and GWAS populations are overwhelmingly European, yet, these populations make up only 16 percent of the global population (Figure 1.9) (Martin et al. 2019).

Although the sample sizes of GWAS studies in European populations have grown, non-European population GWAS sample sizes have remained relatively stagnant. This trend is even more concerning when considering that the relative population sizes of African and Hispanic populations, among other underrepresented populations, have grown in recent years (Morales et al. 2018).

It should be highlighted that genetic data from these underrepresented populations are considered the "lowest hanging fruit" (Martin et al. 2019) in the effort to better understand the genetic underpinnings of disease and complex traits. Genetic variation in non-European populations may be rare or absent in European populations and even with the largest sample sizes, these variants would not be

discovered in an European GWAS. One example of this involves the *SLC16A11* association with Type 2 diabetes in Latino populations (Mercader and Florez 2017). The lack of genetic diversity also presents a challenge for imputation, as existing imputation methods will have trouble imputing variants which are rare in non-European populations and are therefore biased by the ancestral composition of the reference population (Bai et al. 2020).

The lack of diversity within existing genetic data further highlights the difficulty in appropriately choosing a suitable reference population for summary statistic analyses. As methods to analyze admixed GWAS populations grow more powerful, improving the diversity of reference populations to use in summary statistic complex trait analyses will become all the more important.

1.6 Conclusion: From Point Measures to Distributions

In this chapter, we have described the causes and consequences of LD. In particular, we have focused on the stochastic nature of population genetic forces such as genetic drift that shape patterns of LD. We have done so to motivate investigations to better understand how drift and finite sampling influences the distribution of LD and the extent which statistical genetic applications can be improved (in terms of robustness and calibration) by modelling uncertainty in LD arising through such processes.

The rest of this thesis describes the investigations arising from these motivations. In Chapter 2, our primary aim is to characterize and summarize the properties of the distribution of LD and assess the impact of genetic drift on this distribution. Chapter 3 describes a method to generate the distribution of LD that corresponds to populations under varying levels of genetic drift relative to a reference population. In Chapter 4, we derive a MLE which estimates the level of drift between a GWAS population and a reference population. Critically, we derive our MLE using solely the GWAS summary statistics, rather than the individual level-data and

extend our method to account for underlying error that may be present within the GWAS population. Lastly, in Chapter 5, we integrate our investigations into the properties of the distribution of LD and the impact of genetic drift, the ability of using a reference population to generate this distribution, and our estimates of genetic drift between a GWAS and reference population using GWAS summary statistics. Accordingly, we develop a method that aims to infer the distribution of LD within the GWAS population. Importantly, we do so using only the GWAS summary statistics and data from a reference population that has drifted relative to the GWAS population.

2

Modelling the Impact of Genetic Drift on Linkage Disequilibrium

Contents

In this chapter, we model the distribution of LD in a population under genetic drift. Historically, researchers have been more concerned with treating measurements of LD as point estimates, rather than understanding the distribution of LD and its relationship to various population genetic forces. Therefore, much of the work presented in this chapter concerning the properties of the distribution of LD are surprisingly simple but suggest wide-ranging implications for various statistical genetic applications.

We begin by simulating a two-locus Wright-Fisher model to understand how measurements of LD change over time due to genetic drift. Next, we extend this model to allow for recombination between loci. Using this model, we illustrate the stochastic natures of recombination and drift and their effect on measurements of LD. By simulating a series of populations from identical starting conditions, we describe how over time the measurements of LD will tend to become divergent between populations.

Motivated by the results of these two-locus Wright-Fisher simulations, we characterize the distribution of LD by simulating populations under the coalescent. We demonstrate that sampling from the distribution results in point estimates of LD measurements which possess a high degree of variance among subpopulations or subsamples from the same population. This variance is especially pronounced when high levels of genetic drift exist between subpopulations.

We finish by incorporating the distribution of LD within two statistical genetic applications, specifically, summary statistic imputation and fine-mapping. We show that the process of sampling LD panels from this distribution enables us to incorporate the underlying variation of the point estimates. Importantly, we demonstrate that the accuracy within these applications will substantially vary across LD panels sampled from the distribution. Further, sampling LD panels from increasingly drifted reference populations relative to the GWAS population leads to increased variability of the accuracy in these applications.

2.1 Linkage Disequilibrium under a Two-Locus Wright-Fisher Model

We begin by outlining the construction of a two-locus Wright-Fisher model which helps us to understand how stochastic population genetic forces play a role in influencing measurements of LD over time.

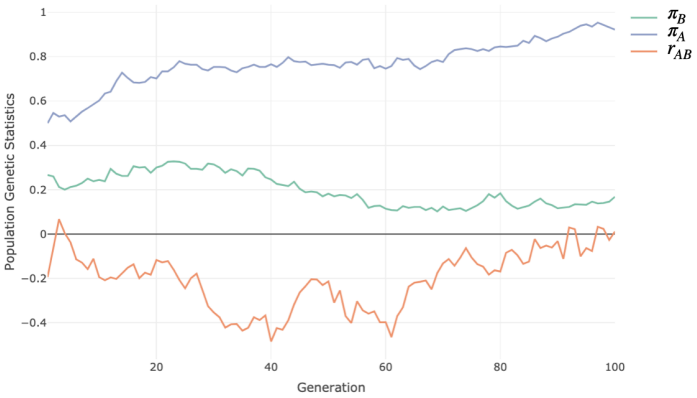

Figure 2.1: Population genetic statistics (π_A, π_B and r_{AB}) from a two-locus Wright-Fisher simulation over 100 generations. Initial haplotype frequencies π_{01}, π_{00}, π_{10} and π_{11} were all set to 30, 20, 40 and 10% in a population of 1000 haplotypes.

Consider a haploid population of constant size N. Within this population exists two segregating bi-allelic loci labelled A and B. The presence or the absence of the alternative allele at each of the two loci is denoted as 1 and 0 respectively. Therefore, the four haplotypes within the population are denoted as: h_{10}, h_{00}, h_{11} and h_{01}. We define the initial frequencies of the four haplotypes at generation one as: π_{10}, π_{00}, π_{11} and π_{01}. Moving forwards in time at each generation, alleles are inherited by the next generation drawn with replacement from the gene pool of the population at the current generation. Therefore, genetic drift is the product of random sampling of alleles at each generation. At each generation in this model, we measure both the allele frequencies at the two loci (π_A and π_B) as well as the LD (r_{AB}) between the loci (Figure 2.1). The stochastic nature of these measurements illustrates that genetic drift does not only generate changes in allele frequencies but also in measurements of LD over time.

Next, we introduce recombination into the Wright-Fisher model to more closely mimic real world genetic variation. At each generation, we incorporate the probability that a pair of haplotypes at loci A and B will recombine and exchange alleles. Therefore, haplotypes h_{10} and h_{01} may recombine and form haplotypes h_{11}

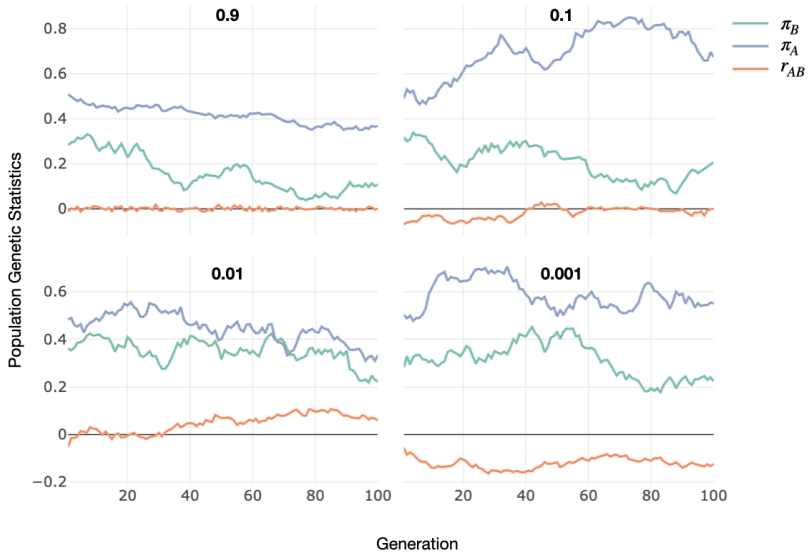

Figure 2.2: Population genetic statistics (π_A, π_B and r_{AB}) over 100 generations from a two-locus Wright-Fisher simulation with recombination. Each population began with initial haplotype frequencies π_{01}, π_{00}, π_{10} and π_{11} were set to 30, 20, 40 and 10% in a population of 1000 haplotypes. The titles on the subplots indicate the recombination rate set within the simulation.

and h_{00}[1] (See Figure 1.1 for an illustration of recombination). The recombination rate defines the expected number of cross-over events per generation per haplotype pair. We simulate a Wright-Fisher model across a series of recombination rates using the same initial haplotype frequencies as in Figure 2.1 and again measure π_A, π_B and r_{AB} at each generation (Figure 2.2).

As the recombination rate increases, there is a corresponding decrease in the magnitude of LD. At the highest recombination rate (0.9), r_{AB} fluctuates around zero across the entire time span of the simulation. Over a long enough timescale, we expect that the presence of any recombination in the Wright-Fisher model to gradually break down the LD between the alleles, thereby reducing r_{AB} to zero (Barton and Otto 2005).

By re-running the Wright-Fisher simulations across replicate populations, we investigate the stochastic nature of recombination and genetic drift and its impact on measurements of LD between populations. We run the Wright-Fisher model

[1]This pairing of haplotypes is the only combination that will effectively result in recombination.

M times, representing M different populations, all starting with the same initial haplotype frequencies. These replicate populations each evolve forwards in time according to the unique reproduction and recombination events that occur at each generation. We set $M = 1000$ and run the simulation across a series of different recombination rates (Figure 2.3).

When the recombination rate is low, measurements of LD (r_{AB}) display high levels of variance between the replicate populations at each generation as indicated by the divergent trajectories of the measurements of LD over time. This variance is a direct consequence of the stochastic nature of recombination which varies the level of coupling between different genealogies[2] for each replicate population. As previously observed in Figure 2.2, when the recombination rate is very high (0.9), recombination almost entirely breaks up the association between alleles resulting in minimal divergence in the measurements of LD between the replicate populations.

2.2 Joint Distribution of Linkage Disequilibrium and Allele Frequencies

By using two-locus Wright-Fisher models, we showed how measurements of LD between populations vary over time as a result of the stochastic nature of random genetic drift and recombination. However, instead of analyzing measurements of LD over time, we can analyze the distribution of the measurements of LD among a set of samples from either one population or a set of closely related populations. The variation in measurements of LD between present-day populations is a result of both the finite sampling of individuals that form these populations and the stochastic nature of population genetic forces in the history of these populations (Figure 2.3). Here, we intend to characterize the shape and properties of the distribution of LD that result from this variation. Further, we are motivated to investigate the *joint*

[2]More generally, with the introduction of recombination events, we generate different *ARGs* as recombination events alter the genealogical structure between loci and induce marginal trees along the sequence.

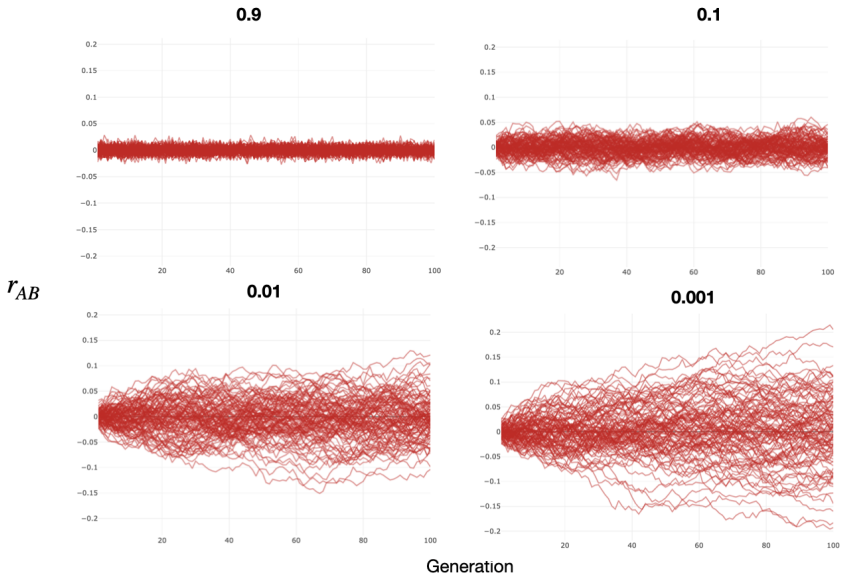

Figure 2.3: Measurements of LD (r_{AB}) in 1000 replicate populations simulated from a two-locus Wright-Fisher model with recombination over 100 generations. Each population began with identical initial haplotype frequencies as in the simulations in Figure 2.2. Each line corresponds to a single replicate population, and each subplot corresponds to the recombination rate (0.9, 0.1, 0.01, 0.001) set within the simulation as denoted in the title of the subplots.

distribution of LD and the allele frequencies, as measures of LD are explicitly linked to the allele frequencies at the alleles under consideration (Equations 1.2 and 1.1).

Consider a genealogical tree[3] with a single recombination hotspot. The sample space of all possible allele frequencies at each allele and LD measurements between every pair of alleles is both astronomically large and complex, even when considering relatively few mutations and samples. Therefore, characterizing the joint distribution of LD and allele frequencies is computationally daunting and difficult to visualize in 2-dimensional space. However, these problems can be circumvented by conditioning the joint distribution on a specific pair of SNPs which we define as the *conditional SNP pair*. Conditioning greatly reduces the dimensionality of the distribution as it is equivalent to slicing through the multi-dimensional surface that represents the joint distribution.

[3]Unless otherwise specified, the genealogical tree refers to the ARG and we implicitly consider all marginal trees induced by recombination. Further, the genealogical tree refers to both the structure of the genealogy, as well as the mutational events which have been superimposed onto the branches of the genealogical tree.

We implement this idea using simulated genetic data under the coalescent from *msprime* (Kelleher, Etheridge, et al. 2016). Using *msprime*, we simulate a large set of P haplotypes with a sequence length of 1Mb under a human level mutation rate of 1e-8 with an effective population size of 5,000. Because we do not simulate any demographic events, we consider these haplotypes to be drawn from a panmictic population. Additionally, to investigate the impact of recombination on these joint distributions, we use a recombination map with a 1000 base pair hotspot in the middle of the 1Mb region. The strength of this hotspot is given by $4N_e r$ where r is the per generation per base recombination rate. We randomly split these P samples into M populations with each group containing N samples[4]. Without loss of generality, we define the first sampled population as the *target* population and the remaining $N-1$ populations as the *peripheral* populations. Conditional SNP pairs are selected by randomly choosing a pair of segregating SNPs within the target population that fall on either side of the recombination hotspot. We specifically select SNPs on either side of the hotspot in order to examine how the distribution of LD changes when the correlation between the marginal trees carrying either SNP is varied. The SNPs chosen are defined as A and B. We record the value of LD as measured by r between the two SNPs as r_{AB}^{Target}. We additionally record the geometric mean of the allele frequencies of the SNPs A and B in the target population π_A^{Target} and π_B^{Target} as $\sqrt[2]{\pi_A^{\text{Target}}\pi_B^{\text{Target}}}$. The geometric mean produces a single metric that can be useful to summarize the pair of allele frequencies in the conditional SNP pair. For instance, we can utilize this metric to assess if the pair of SNPs is more aptly defined as: rare-rare, rare-common or common-common. This entire process is illustrated by Steps 1-3 in Figure 2.5.

For three separate *msprime* simulations and conditional SNP pair, we plot the distribution of measurements of LD in the peripheral populations conditional on the conditional SNP pair chosen in each simulation (Figure 2.4).

[4]This implies that we have $P = NM$ samples across all populations.

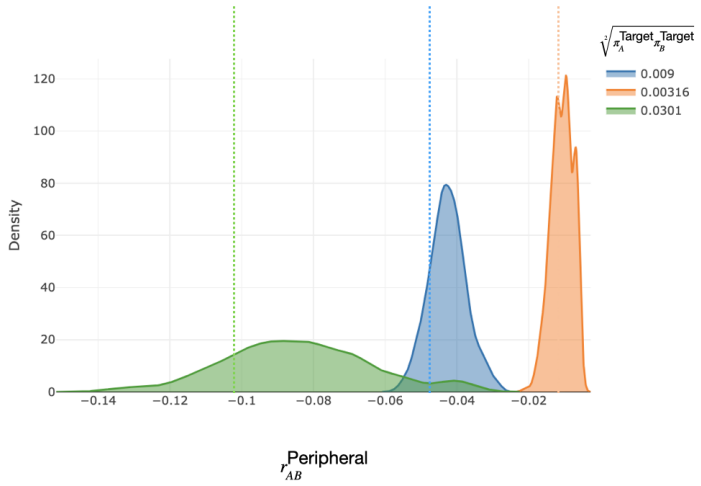

Figure 2.4: Distribution of measurements of LD in peripheral populations for three independent *msprime* simulations for a conditional SNP pair in the target population. The number of samples in each population M and the number of populations N were both set to 1000. The recombination hotspot was set to $4N_e r = 1$ in each of these simulations. Each of the three distributions across peripheral populations corresponds to a single *msprime* simulation. Subsequent conditioning is described within Figure 2.5, Steps 1-3. For each respective simulation, the vertical coloured dotted lines correspond to the value of r_{AB}^{Target}. The legend describes the geometric mean of the allele frequency in the target population for the conditional SNP pair for each of the three simulations.

We observe that the measurements of LD for a conditional SNP pair in the target population are roughly normally distributed across the peripheral populations . Moreover, the variances of these measurements are significantly different in all three simulations. These differences suggest that the variation of measurements of LD are dependent on either the properties of the genealogies generated by *msprime* or the choice of the conditional SNP pair.

The variance of these distributions in Figure 2.4 reflects the extent to which the peripheral populations and the target population share a similar measurement of LD at the conditional SNP pair. For example, the orange density plot in Figure 2.4 displays a relatively small variance, which indicates that it is more likely that sampling a measurement of LD from a peripheral population will result in a similar measurement of LD from the target population. Likewise, the green density plot displays a larger variance of these measurements of LD which indicates that it is

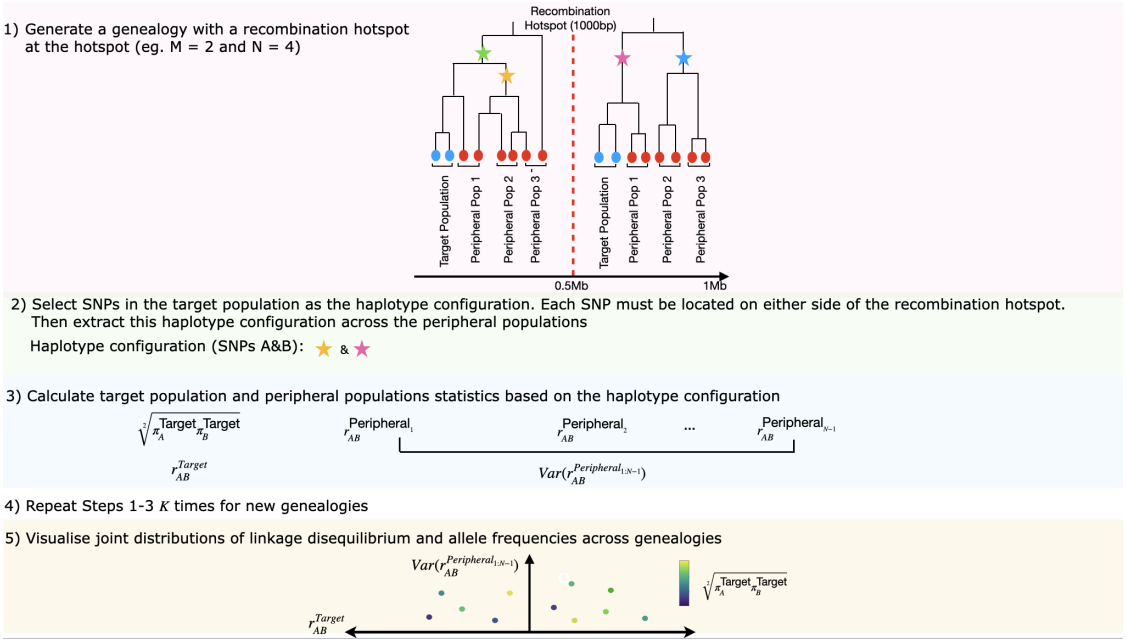

Figure 2.5: Steps to visualise the joint distribution of measurements of LD and the allele frequencies for a conditional SNP pair in a panmictic population. Step 1: Simulate a large genealogy with a recombination hotspot. The length of the genetic sequence is represented by the rightward arrow while the recombination hotspot is denoted by the middle dotted red line. Sample haplotypes are represented by the coloured circles at the bottom of the tree. The samples are separated into a set of M blue samples representing the target population and a set of M red samples in each of the $N-1$ peripheral populations. Distinct mutations correspond to the coloured stars on these genealogies. Step 2: Choose a conditional SNP pair (eg. yellow and fuchsia mutations) amongst the mutations that fall on either side of the recombination hotspot. Step 3: Calculate metrics that describe the joint distribution in target and peripheral populations. Step 4: Repeat Steps 1-3 K times for new genealogies. Step 5: Visualize the relationship between these metrics. Each point in the visualization corresponds to a single simulated genealogy.

more likely that sampling a measurement of LD from a peripheral population will result in a more dissimilar measurement of LD from the target population.

To gain a deeper understanding of the characteristics of the joint distribution of measurements of LD and allele frequencies, we repeat the process above K times, each time simulating a new genealogy along with a randomly chosen conditional SNP pair (Figure 2.5). For each simulation, we obtain the measurement of LD and the geometric mean of the allele frequency at the conditional SNP pair in the target population. Further, we obtain the variance in the measurements of the LD in the peripheral populations at this conditional SNP pair. These metrics summarise the

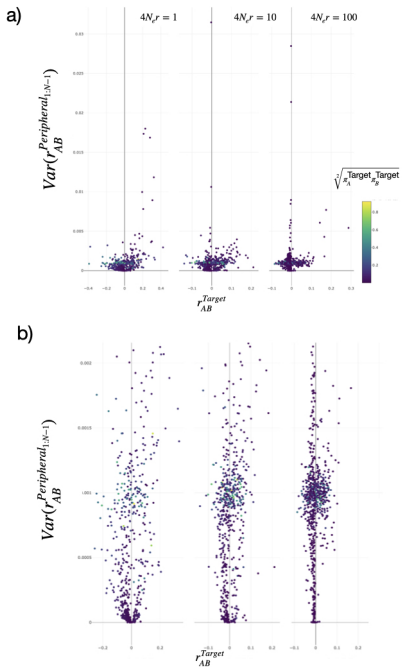

Figure 2.6: The joint distribution of LD and the allele frequencies for a conditional SNP pair in samples drawn from a panmictic population. a) Each point corresponds to a unique simulated genealogy with $M = 1000$ samples per population and $N = 1000$ populations. We repeatedly simulated genealogies $K = 1000$ times. Each subplot from left to right corresponds to the hotspot strength within the sequence as denoted in the title of the subplots. b) Is identical to a) however we only consider points where $Var(r_{AB}^{Peripheral_{1:N-1}}) < 0.003$.

salient properties of the joint distribution. We investigate the effect of the hotspot recombination rate on the properties of the joint distribution by simulating each set of the K genealogies under a different value for $4N_e r$.

Figure 2.6 illustrates that the measurements of LD at the conditional SNP pair across the peripheral populations exhibit relatively low values of variance for the majority of genealogies. However, a small fraction of the genealogies result in relatively high and outlying variances. Interestingly, these outlying variances correspond to genealogies where the conditional SNP pair has a low geometric mean allele frequency. In other words, both SNPs are rare. In addition, we observe that these outlying variances are more evident when r_{AB}^{Target} is greater than zero.

When we focus on genealogies that exhibit low (and non-outlying) values of

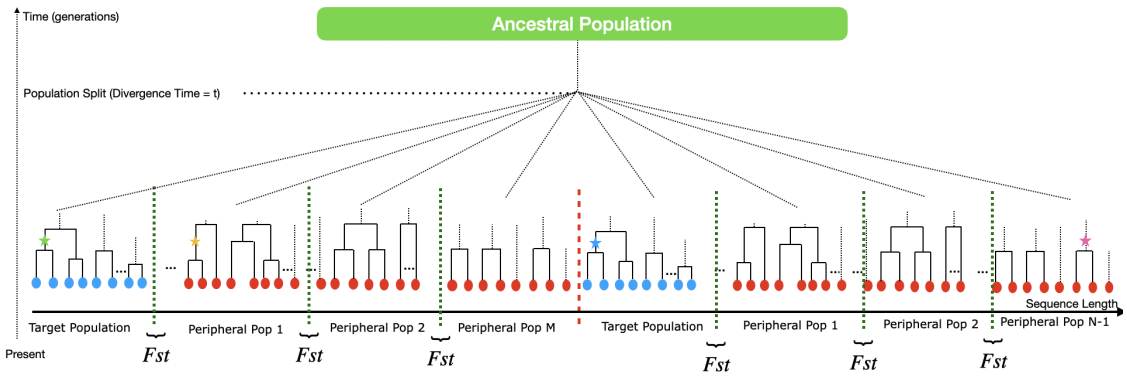

Figure 2.7: An ancestral population splits t generations in the past and forms the set of $N-1$ peripheral populations and the target population in the present. This split generates genetic drift as measured by F_{st} between each of the present-day populations. All other labels are consistent with Figure 2.5.

$Var(r_{AB}^{Peripheral_{1:N-1}})$ (Figure 2.6 b), these variances are scattered in a *fountain-like* pattern. This pattern reflects a lower bound for $Var(r_{AB}^{Peripheral_{1:N-1}})$ that increases as the value of $|r_{AB}^{Target}|$ also increases. Further, we observe that changing the strength of the recombination hotstpot results in the magnitude of r_{AB}^{Target} which decreases as the recombination rate around the hotspot increases. This result again demonstrates how recombination can act to break down LD between SNPs. Interestingly, as the hotspot increases to $4N_e r = 100$ we notice that the variance of the peripheral population LD measurements appears to approach a steady state value of 0.001.

2.2.1 Genetic Drift

In the previous section, we formed the peripheral and target populations drawing sample haplotypes from a panmictic population within *msprime*. Thus, these populations are not structured in any meaningful way and little to no genetic drift is found between them. In contrast, in this section we generate genetic drift between the N populations to investigate the impact of drift on the joint distribution of the measurements of LD and the allele frequencies. In our model, an ancestral population splits t generations in the past to form peripheral and target populations, resulting in genetic drift between the populations (Figure 2.7).

A population split is a suitable choice for the demographic event to generate drift between the populations due to the following reasons: this demographic event is easily implemented in *msprime*, the theoretical underpinning of genetic drift under a population split is well documented (Holsinger et al. 2009) and a single parameter t controls the level of drift between the diverged populations. In *msprime*, we set the effective population size of the ancestral and diverged populations to 5,000. Note that this simulation framework will also induce a coalescence event at the time of population split. We note that the unique case of a population split $t = 0$ generations ago is equivalent to forming the peripheral and target populations using samples drawn from a panmictic population as we did previously.

Therefore, the distribution of measures of LD in the peripheral populations for a conditional SNP pair under this demographic model can be obtained from a single *msprime* simulation. We visualize this distribution in the same fashion as Figure 2.4. However, we now generate three genealogies for three different population split times. As before, we simulate all genealogies using a recombination hotspot set to $4N_e r = 1$. These distributions are shown in Figure 2.8.

The measurements of LD across the peripheral populations for each of these genealogies appear to be once again roughly normally distributed. Therefore, we speculate that the the shape of these distributions is independent of the underlying demographic model used within these simulations though will be influenced by the additional drift.

As in the previous section, we gain a deeper understanding of the characteristics of the joint distribution of measurements of LD and the allele frequencies by repeating the process above K times, each time simulating a new genealogy along with a randomly chosen conditional SNP pair. We again visualize these joint conditional distributions across a series of recombination rates at the hotspot, as well as across a series of population split times (Figure 2.9).

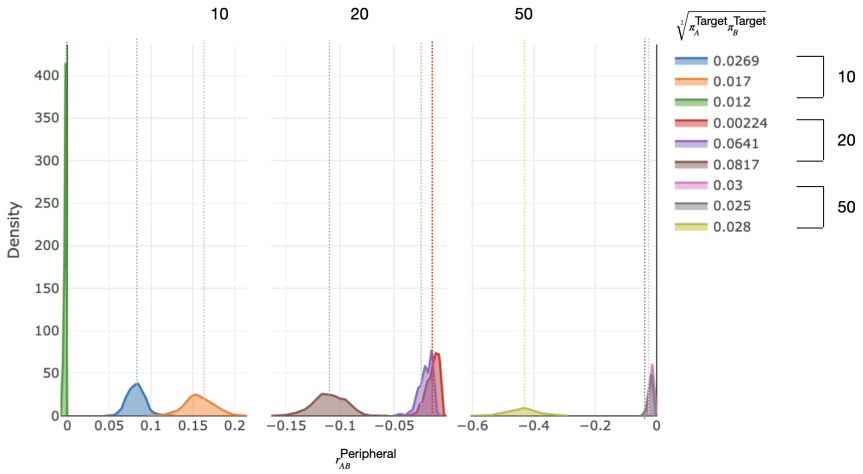

Figure 2.8: Distribution of measurements of LD in peripheral populations under genetic drift for a conditional SNP pair. Each of the distributions correspond to a single *msprime* simulation and subsequent conditioning as described within Figure 2.5, Steps 1-3. The number of samples in each population M and the number of populations N were both set to 1000. The strength of the recombination hotspot was set to $4N_e r = 1$ in each of these simulations. For each respective simulation, the vertical coloured dotted lines correspond to the value of r_{AB}^{Target}. The legend describes the geometric mean of the allele frequency in the target population for the conditional SNP pair for each of the simulations and is sorted by the population split time. The title of the subplots indicate the time of the population split set within the demographic model (Figure 2.7).

In Figure 2.9, we observe the same structure of the joint distribution when no genetic drift was generated between the populations (Figure 2.6). Most of the genealogies result in values of $Var(r_{AB}^{Peripheral_{1:N-1}})$ that are relatively low but a small fraction of the genealogies generates variances that are substantial outliers. As illustrated previously in the panmictic setting, we observe that these outliers are most commonly found in genealogies when r_{AB}^{Target} is greater than zero. Further, these outlying points correspond to conditional SNP pairs of rare variants. Therefore, we speculate these points relate to a rare variant which arises on the background of another rare variant leading to coupling linkage and increased variance (Good 2020). In addition, the magnitude of these outlying points are roughly the same as in the panmictic setting. A notable feature of these plots is that as the time of the population split increases further into the past and the genetic drift between the populations grows larger, the non-outlying points begin to turn inwards (Figure 2.9

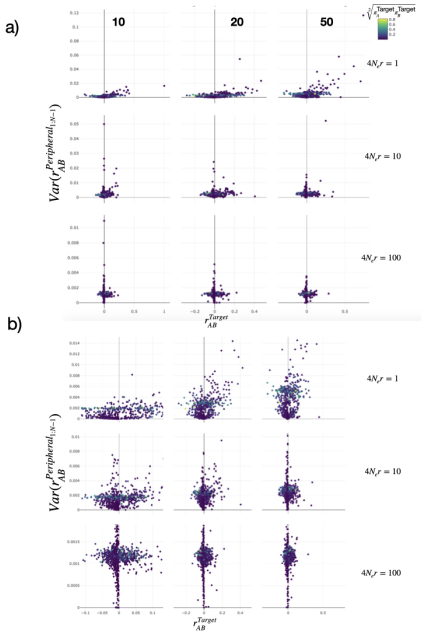

Figure 2.9: The joint distribution of LD and the allele frequencies for conditional SNP pairs in populations under genetic drift. a) Each point corresponds to a simulated genealogy with $M = 1000$ samples and $N = 1000$ populations. Each column of subplots corresponds to the time of the population split (10, 20 and 50 generations ago) which forms the N divergent populations. Each row of the subplots indicates the strength of the recombination hotspot ($4N_e r$). b) is identical to a) but focused on only the centre of mass of the distribution and the non-outlying variances in each experiment.

b). As a result, the lower bound of $Var(r_{AB}^{Peripheral_{1:N-1}})$ for a given value of r_{AB}^{Target} increases under higher levels of genetic drift relative to lower values of genetic drift. We speculate that increasing the genetic drift between the populations results in increasing the variance of the allele frequencies between populations. Consequently, the measurements of LD may possess increased variance because measurements of LD explicitly rely on the allele frequencies. Lastly, as previously shown in the panmictic setting, we observe that the magnitude of r_{AB}^{Target} decreases as the recombination rate around the hotspot increases. Furthermore, similar to the panmictic simulations, we note that the variance of the peripheral population LD measurements appears to approach a steady state value of 0.001 under high recombination.

2.3 Impact of Genetic Drift on Linkage Disequilibrium on Downstream Applications

In the previous section, we characterized the joint distribution of LD and the allele frequencies in populations under genetic drift. Given the variation in measurements of LD that we observed across populations, we aim to investigate the extent to which this variation impacts downstream statistical genetic applications. We focus on two common methods that use LD, specifically, summary statistic imputation and fine-mapping. Because these methods rely on point estimates of the measurements of LD within a population, they do not directly assess how variation of these measurements impact downstream results. To overcome this obstacle, we repeatedly sample LD panels from the distribution of LD and utilize the point estimates of the measurements of LD in each panel. Through sampling LD panels, we describe how variation in the measurements of LD across the sampled panels generates variation across the results of these applications. We implement this process using LD panels obtained from replicate populations in *msprime*. Each replicate population generated with *msprime* is drawn from the underlying sample space of all possible genealogies consistent with a given demographic model. This process is similar to the steps described in Figure 2.5 but we now consider SNPs across an entire region rather then choosing a conditional SNP pair.

2.3.1 Summary Statistic Imputation

Summary statistic imputation is an attractive tool for estimating associations at untyped SNPs in a GWAS study. These methods are fast, scalable and can provide accurate results almost on par with genotype imputation (Pasaniuc, Zaitlen, et al. 2014).

Typically, summary statistic imputation involves the imputation of z-scores from a GWAS. Z-scores are defined at each SNP as $\frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$ where $\hat{\beta}$ is the effect size and $\sigma_{\hat{\beta}}$ is

the standard error of the effect size. LD between SNPs induces a covariance between their observed z-scores, according to the correlation between the two SNPs (Pasaniuc, Zaitlen, et al. 2014). Under the null model of no association, the distribution of the z-scores are approximated by a multivariate normal distribution with mean μ and correlation Σ. Here, Σ_{ij} denotes the correlation (ie. Pearson r) between SNPs i and j.

To describe the process of summary statistic imputation we partition a vector of the z-scores Z into two components Z_t (of size m) and Z_i (of size $m - n$). These components correspond to the z-scores at the typed and imputed SNPs respectively. Additionally, we partition the mean vector and variance covariance matrix into $(\mu_t, \mu_i)^T$ which corresponds to the mean at typed and imputed SNPs, covariances among imputed SNPs ($\Sigma_{i,i}$), covariances among typed and imputed SNPs ($\Sigma_{i,t}$) and covariance among typed SNPs ($\Sigma_{t,t}$) (Pasaniuc, Zaitlen, et al. 2014). Therefore, $Z_i|Z_t$ is Gaussian distributed with mean $\mu_{Z_i|Z_t} = \mu_i + \Sigma_{i,t}\Sigma_{t,t}^{-1}(Z_t - \mu_t)$ and variance $\Sigma_{i|t} = \Sigma_{i,i} - \Sigma_{i,t}\Sigma_{t,t}^{-1}\Sigma_{i,t}^T$.

We can represent the imputed z-scores at SNPs i $Z_{i|t}$ as a linear combination of the typed z-scores Z_t using weights $W = \Sigma_{i,t}\Sigma_{t,t}^{-1}$. Let A be the variance-covariance matrix among typed SNPs. From the multivariate normal approximation under the null model we assume that $Z_t \sim \mathcal{N}(0, A)$. Therefore $Z_{i|t}$ has variance WAW^T and we use $\frac{Z_{i|t}}{\sqrt{WAW^T}}$ as the imputed z-scores at SNPs i. (Pasaniuc, Zaitlen, et al. 2014).

We assess the accuracy of summary statistic imputation using simulated GWAS data. We first simulate both a source population[5] and a reference population that has drifted relative to the source population. As in the previous section, we simulate these two populations under the demographic event of an ancestral population which has split some time t generations[6] ago in the past to form the source and reference population. We use *msprime* to simulate the source and reference population, each with a sample size of 1000. We set the effective population size of the ancestral,

[5]This is different than the GWAS population which contains the case-control haplotypes simulated from HAPGEN2.

[6]In *msprime*, units of time are measured in non-coalescent units

source and reference population each to 5000. These populations are simulated under a uniform recombination and mutation rate of 1e-8 across 1Mb regions. We jointly remove all non-segregating SNPs in both the source and reference population, as well as SNPs below 1 percent minor allele frequency (MAF).

Using the haplotypes in the source population generated by *msprime*, we simulate GWAS case-control haplotypes using HAPGEN2. HAPGEN2 simulates case control data-sets at SNP markers (Su et al. 2011). We use HAPGEN2 to simulate a binary case-control GWAS study with 10,000 cases and controls respectively under the null model of no association. The output of HAPGEN2, a set of case and control simulated haplotypes, can be tested for association with the binary trait using standard logistic regression methods. We use SNPTEST to generate the GWAS summary statistics using a frequentist test with a logistic regression framework under a additive model to test for association between the SNPs and the mock trait (J. Marchini et al. 2007). These association tests provide us with a series of GWAS effect sizes and standard errors across SNPs that are converted into z-scores. Next, we mask and impute 50% of the summary statistics chosen at random among the SNPs in the GWAS region. Finally, we perform summary statistic imputation using the LD panel obtained from the GWAS case-control haplotypes (from HAPGEN2) or from the drifted reference population. We assess the accuracy by calculating the r^2 correlation coefficient between the true and imputed z-scores in two groups by sorting the SNPs in the GWAS as common ($\geq$ 5% MAF) or rare ($<$ 5% MAF). The accuracy of summary statistic imputation under the null model for a series of replicate regions across LD panels obtained from different populations is described in Figure 2.10.

Similarly, we also use HAPGEN2 to simulate GWAS regions that contain three causal variants. These three variants each possess heterozygous and homozygous log odds ratios for disease risk of 1.5 and 2.25 respectively. We choose these high log odds ratios to improve the accuracy of downstream fine-mapping and summary

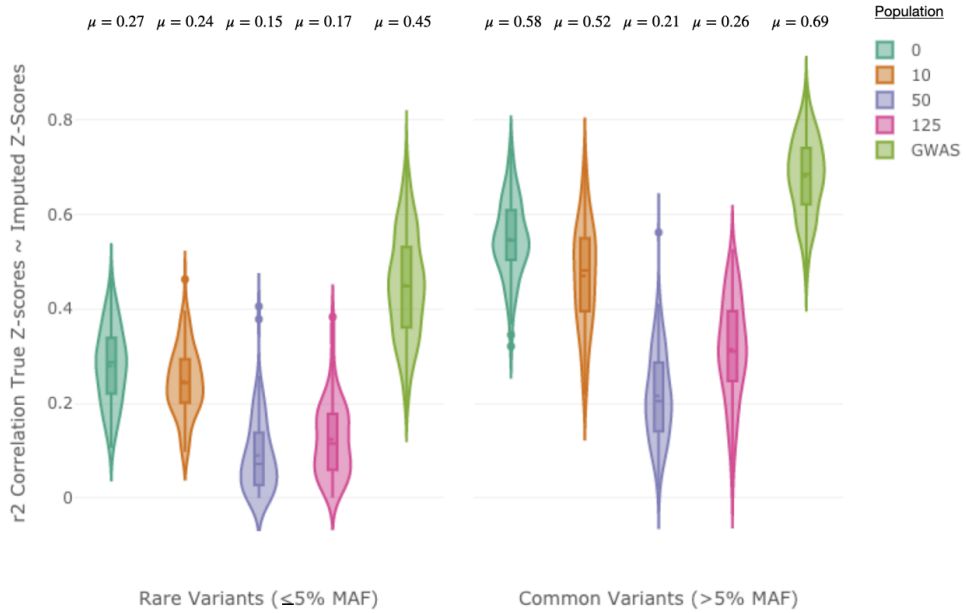

Figure 2.10: Distribution of the accuracy of summary statistic imputation under the null across 1000 replicate regions. Each violin plot displays the density of the r^2 correlation of summary statistic imputation across all replicate regions between the true and imputed z-scores. Violin plots on the left include variants at $\leq$ 5% MAF and $\geq$ 1% in the GWAS population and the violin plots on the right include variants at $>$ 5% MAF in the GWAS population. Each colour corresponds to the population in which the LD panel was obtained for use in summary statistic imputation. These populations included the GWAS case-control haplotypes and four reference populations that split 0, 10, 50 and 125 generations ago from the GWAS population. The statistic μ above each violin plot represents the mean of the r^2 correlation across the 1000 replicates for each of the different populations.

statistic imputation in order to obtain more noticeable differences between the LD panels. Performing fine-mapping with three causal variants is a difficult problem (C. Benner et al. 2016, Benner et al. 2017) and simulating data under these high odds ratios should improve the overall accuracy of FINEMAP. We run HAPGEN2 with the same parameters we used previously under the null model. Again, we generate summary statistics with SNPTEST and mask and impute half of these summary statistics at random. The accuracy of summary statistic imputation for both common ($>$ 5% MAF) or rare ($\leq$ 5% MAF) GWAS SNPs under the causal model for a series of replicate regions across LD panels obtained from different populations is described in Figure 2.11.

Overall, the LD panel obtained from the GWAS population provides the highest mean accuracy in both the null and causal models of association. The LD from the GWAS case-control haplotypes represents the gold standard for accuracy as the correlation between z-scores are nearly identical to the correlation in the LD panel between the GWAS SNPs (Benner et al. 2017). Importantly, although the in-sample LD information represents our gold-standard accuracy, the process of summary statistic imputation through solving a system of linear equations results in a degree of inaccuracy[7] that is introduced into our imputed summary statistics. Therefore, we do not see perfect correlation between the true and the imputed z-scores. Moreover, through running the analysis across these replicate regions, we demonstrate that the accuracy of summary statistic imputation is highly variable from one region to the next. These results suggest that the variation across LD panels leads to substantial variation in the accuracy of summary statistic imputation.

Typically, using LD panels from reference populations that have more recently split in the past from the GWAS population lead to higher mean accuracy within summary statistic imputation across replicate regions. In comparison, using LD panels from reference populations that have split further back in the past from the GWAS population produce lower mean accuracy within summary statistic imputation. As genetic drift between the reference and GWAS population increases, measurements of LD in the reference and GWAS population become increasingly diverged (Novembre et al. 2008). Consequently, this leads to decreased accuracy within summary statistic imputation as the correlations between the SNPs in the reference populations no longer match the correlation between the z-scores in the GWAS summary statistics. However, one exception to this trend involves the reference populations that have split 50 and 125 generations ago from the GWAS population. Here, in both the causal and null model of association, the LD panels obtained from the reference population that split 125 generations ago outperforms

[7]For instance, when using ridge-regression to invert the LD panel

the reference populations that split 50 generations ago. In our simulation, we might expect that population pairs which possess higher levels of genetic drift may artificially obtain a higher accuracy relative to population pairs with lower levels of genetic drift as a consequence of removing more low frequency variants when the populations are highly diverged. Further, it appears that as the reference populations become increasingly diverged from the GWAS population we observe increasingly variable accuracy across replicate regions. For instance, the interquartile range (IQR) for the summary statistic imputation accuracy for the rare variants (Figure 2.10) using the population that diverged 0 generations ago is 0.12 compared to 0.21 using the population that diverged 125 generations ago.

We generally obtain a much higher mean accuracy of summary statistic imputation across the replicate regions for common variants as compared to the rare variants. This result is consistent across each of the populations used to obtain the LD panels. We note that rare variants are traditionally difficult to impute, even using genotype imputation methods (Y. Li et al. 2009) and the lower level of accuracy for the summary statistic imputation is in line with these results (Rüeger et al. 2018). Interestingly, with the exception of the LD panel obtained from the GWAS population, the variability of the accuracy of summary statistic imputation was greater when imputing common SNPs compared to imputing rare variants.

Finally, we observe that the accuracy of summary statistic imputation is appreciably higher when summary statistics were generated under a causal model of association rather than the null model. In the causal model, there is much greater variation across the z-scores which is entirely driven by LD to the causal variant(s). Given that summary statistic imputation (performed under the null model of association) aims to identify and leverage the presence of LD to perform imputation, relying on these signals of association may make it easier to identify the underlying patterns of LD in a region.

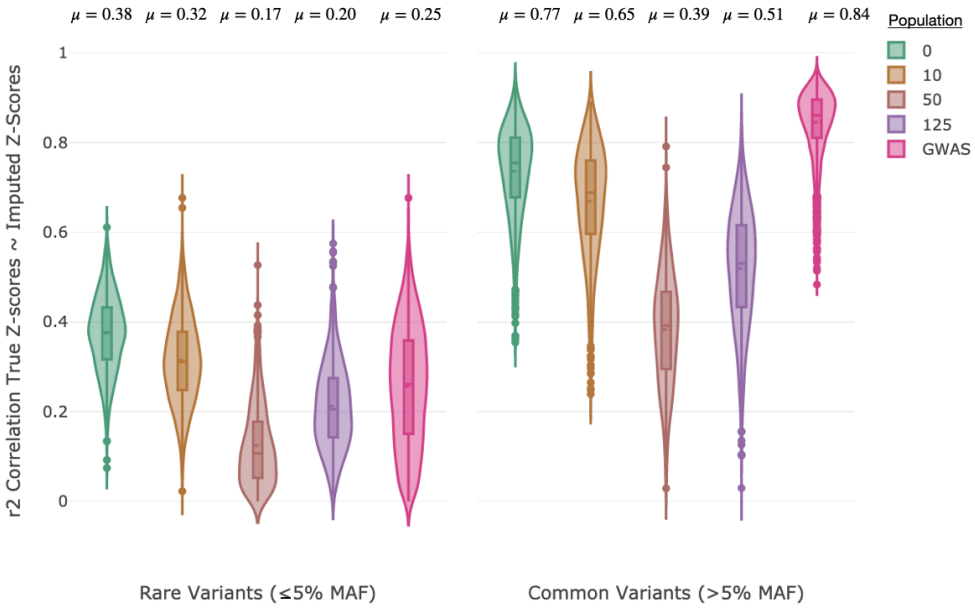

Figure 2.11: Distribution of the accuracy of summary statistic imputation under a model of association with three causal variants across 1000 replicate regions. Each violin plot displays the density of the r^2 correlation of summary statistic imputation between the true and imputed z-scores across LD panels from different populations. Violin plots on the left include variants at $\leq 5\%$ MAF in the GWAS population and the violin plots on the right include variants at $> 5\%$ MAF in the GWAS population. Each colour corresponds to the population in which the LD panel was obtained for use in summary statistic imputation. These populations included the GWAS population and four reference panels that split 0, 10, 50 and 125 generations ago from the GWAS population. The statistic μ above each violin plot represents the mean of the r^2 correlation across the 1000 replicates for each of the different populations.

2.3.2 Fine-mapping

LD is also used to fine-map signals of association at GWAS loci (Wallace et al. 2015). Within fine-mapping, LD panels are used to disentangle which SNPs are correlated with the trait of interest and which SNPs are instead tagged alongside the true causal SNPs due to LD. Several tools have been developed to perform fine-mapping. In this chapter, we use the tool FINEMAP (C. Benner et al. 2016).

To assess the accuracy of fine-mapping, we simulate GWAS summary statistics and genetic data in a similar manner as we previously did when assessing the accuracy of summary statistic imputation. However, in the case of fine-mapping

signals of association, it is only necessary to consider simulating GWAS data under a causal model.

We assess the accuracy of the fine-mapping results using two metrics after running FINEMAP on a set of GWAS summary statistics. We calculate both metrics using the 95% credible set of the plausible causal variants. To obtain the credible set, we rank the set of all configurations outputted by FINEMAP by their posterior probability. We filter this set to retain the configurations with the highest posterior probabilities which cumulatively contain 95 percent of the posterior probability and define the SNPs in these configurations as the 95 percent credible set. For each configuration of SNPs within the credible set, we calculate the fraction of true causal variants that appear within each configuration ($\frac{1}{3}$,$\frac{2}{3}$, or $\frac{3}{3}$) and multiply this fraction by the PIP of the respective configuration. Summing over all configurations results in the first metric used to assess accuracy, which is defined as the *weighted PIP*. The second metric examines how often the 95% credible set contains the true causal variants. For each of the three causal variants, we record if the causal variant is contained within the 95 percent credible set. Thus, the second metric can result in a score of 0, 1, 2 or 3.

We investigate how genetic drift between the reference and GWAS population affects fine-mapping performance by running FINEMAP using LD panels obtained from the GWAS population as well as from the drifted reference populations. We consider the setting when SNPTEST associations summary statistics are converted into z-scores and are directly read into FINEMAP to perform fine-mapping. The accuracy of FINEMAP for a series of replicate GWAS regions across LD panels from different populations is shown in Figure 2.12.

Within FINEMAP we set the flag **n-causal-snps** equal to three which is equal to the number of causal variants in the simulated region. However, in reality, it may not be known *a priori* how many causal variants are in a region, as opposed to the simulations presented in this chapter where HAPGEN2 is used to directly set

the number of causal variants. Therefore, we also considered running FINEMAP with the flag **n-causal-snps** with greater than three causal variants. Running these analyses resulted in lower weighted PIPs and less regions where a higher number of the true causal variants were contained in the credible set. However, changing this flag did not effect the relative performance of fine-mapping using LD panels from different populations. Therefore, we decided to set the number of causal variants within this flag in FINEMAP equal to three and have presented the results with this flag setting below. We note that previous evaluations of FINEMAP fine-mapping accuracy (Benner et al. 2017) have performed analyses in a similar manner and have set this flag equal to the number of causal variants that a region was simulated underneath. In the conclusion of this chapter we discuss other metrics (eg. proportion of false positives) which may be effected by changing the value of this flag and their usefulness within the context of our experiments.

In statistical genetic applications, it is common to combine summary statistic imputation and fine-mapping into one set of analyses (Rüeger et al. 2018, Pasaniuc, Zaitlen, et al. 2014). This process involves the following: curating the association statistics from a GWAS study, performing summary statistic imputation to impute the summary statistics at untyped variants and fine-mapping the region with the expanded set of variants.

In a similar process, we simulate the GWAS population and summary statistics and reference populations as previously described with three causal variants in the region. We mask and impute half of the summary statistics chosen at random within the region. Summary statistic imputation is performed using either the LD panel obtained from the reference populations, or the LD panel from the GWAS case-control haplotypes. We perform fine-mapping on a GWAS region with FINEMAP using the imputed summary statistics, along with the remaining typed summary statistics from SNPTEST. For each set of analyses, the LD panel provided to FINEMAP is the same as the LD panel used in summary statistic

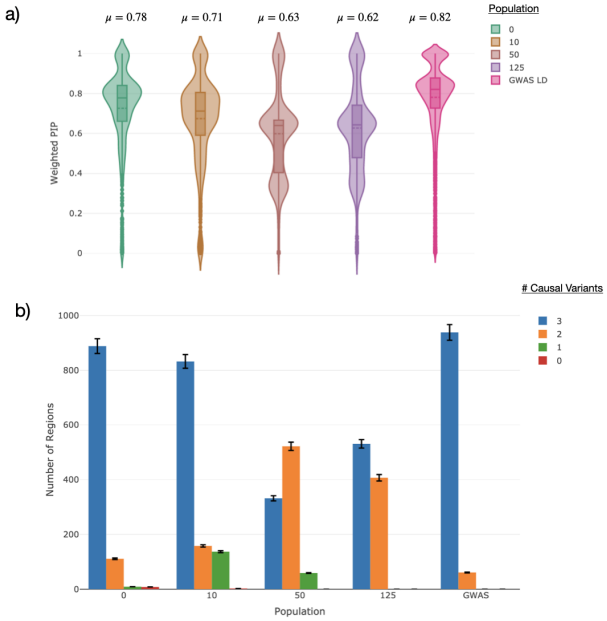

Figure 2.12: Accuracy of fine-mapping simulated GWAS data with the true summary statistics using the LD panels obtained from different populations across 1000 replicate regions. The true summary statistics were obtained directly from SNPTEST. The populations labelled 0, 10, 50, and 125 refer to the time (in generations) of the population split resulting in the drifted reference populations relative to the GWAS population. The GWAS label refers to the LD obtained from the case-control haplotypes simulated from HAPGEN2. a) Distribution of the weighted PIPs for a series of populations used to obtain the LD panels to perform fine-mapping with FINEMAP. The value μ above the distribution indicates the mean across the replicates for a given population. FINEMAP was run with flag **–n-causal-snps 3**. b) The number of regions in which the 95 percent credible set marginally contained 0, 1, 2 or 3 causal variants across 1000 replicate regions using the LD panels obtained from different populations. Error bars gives $\pm$ 1.96 times the standard error of the mean.

imputation. The consistency of using the same LD panel is reflective of real-world GWAS analyses and pipelines. The accuracy of FINEMAP according to these metrics using the imputed summary statistics for a series of replicate regions and different populations is shown in Figure 2.13.

We observe that when FINEMAP is provided with the true GWAS summary statistics directly from SNPTEST, the LD panel obtained from the GWAS case-control haplotypes still performs the best across the two fine-mapping metrics. However, as previously demonstrated using LD panels from different populations within summary statistic imputation, substantial variation exists across the accuracy

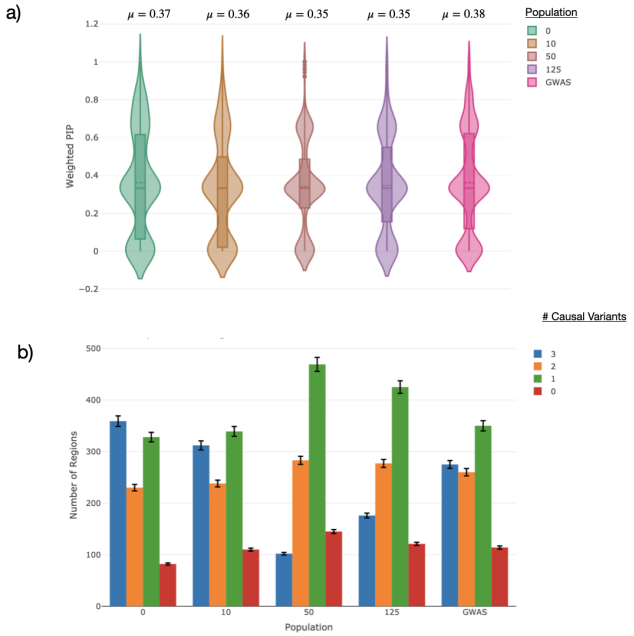

Figure 2.13: Accuracy of fine-mapping with imputed summary statistics using LD panels obtained from different populations across 1000 replicate regions. The populations labelled 0, 10, 50, and 125 refer to the time (in generations) of the population split resulting in the drifted reference populations relative to the GWAS population. The GWAS label refers to the the case-control haplotypes simulated from HAPGEN2. a) Distribution of the weighted PIPs across replicate regions. The value μ above the distribution indicates the mean across the replicates for a given population. FINEMAP was provided the imputed summary statistics and run with flag –**n-causal-snps 3**. b) The number of regions across the population in which the 95 percent credible set marginally contained 0, 1, 2 or 3 causal variants. Error bars gives ± 1.96 times the standard error of the mean.

of fine-mapping in different regions. Furthermore, fine-mapping accuracy decreases as the reference population used to obtain the LD panel becomes increasingly genetically drifted to the GWAS population. This decrease in accuracy is discernible in both metrics. First, using LD panels obtained from drifted reference populations results in lower weighted PIPs (Figure 2.13 a) relative to using LD obtained from the GWAS population. Secondly, the LD panels from the drifted reference populations display an increase in the proportion of regions where only one or two of the true causal variants were identified within the credible set (Figure 2.13 b). Interestingly, the weighted PIP of using FINEMAP with LD panels obtained from the reference population which split 125 generations ago, outperforms the

reference population which split 50 generations ago. As before, we speculate this result is due to filtering rare SNPs in highly drifted populations. Further, we observe that as the reference population becomes increasingly drifted relative to the GWAS population, the variability of the weighted PIPs across replicate regions increase. The increase in the variance of the weighted PIPs is likely a consequence of FINEMAP identifying more regions where only $\frac{2}{3}$ and $\frac{1}{3}$ of the true causal variants are found in the credible set (Figure 2.12 a).

We now consider the fine-mapping results when providing FINEMAP with the imputed summary statistics (Figure 2.13). We observe lower levels of accuracy for both fine-mapping metrics compared to the results when providing FINEMAP with the true GWAS summary statistics across all populations. Additionally, we observe extremely wide variation of the weighted PIPs for each of the populations (Figure 2.13 a). Interestingly, providing FINEMAP with the imputed summary statistics and the LD obtained from the GWAS population no longer out-performs the accuracy of using the LD obtained from a reference panel population that split 0 generations ago from the GWAS. Furthermore, LD panels obtained across all drifted reference populations perform equally poorly. Because the weighted PIP score is centered around roughly 0.33 for most of these distributions, this suggests that FINEMAP has been able to only correctly identify one of the three causal variants with any degree of certainty in these analyses. Finally, we observe many more regions where only one or none of the true causal variants are contained in the credible set relative to providing FINEMAP with the true GWAS summary statistics in each respective population setting (Figure 2.13 b).

2.4 Conclusion

In this chapter, we aimed to answer two fundamental questions involving how LD is measured in a population. Firstly, to what extent do measurements of LD

differ among subpopulations or subsamples from the same population? Secondly, does the variation across measurements of LD affect the downstream performance of statistical genetic applications?

We first described simulation results using a two-locus Wright-Fisher model with recombination to convey the motivation behind these questions. These Wright-Fisher models illustrated how point estimates of LD can be modelled as samples drawn from a distribution. Using a simulation-based approach, we described the variation in measurements of LD across populations. Importantly, in real-world settings, a single-point estimate of LD from a population does not capture this variation in any meaningful way. We observed this variation in measurements of LD across populations, even when the samples which formed these populations were drawn from a panmictic population. Furthermore, as we introduced genetic drift between the populations, we observed that the variation of these measurements of LD grew larger. The addition of further layers of population structure within the simulated populations would be a worthwhile endeavour for future investigation to assess the impact of other demographic events on the variation in measurements of LD.

However, the simulated approach used to describe the distribution of LD may not present an unbiased perspective of the variation of LD measurements for two reasons. Firstly, we exclude the measurement of LD when the conditional SNP pair is non-segregating in a peripheral population resulting in undefined measurements of LD (Equation 1.2). Filtering out these peripheral populations may artificially decrease the variance of the measurements of LD. Secondly, multiple genealogies produce the same conditional SNP pair in the target population. An interesting direction for future work would involve integrating over these multiple genealogies in a principled manner.

Having characterized the variation across point estimates of LD between populations, we investigated if this variation has any substantial effect downstream on the accuracy of statistical genetic applications. By generating simulated GWAS data

with *msprime*, we showed that variation in LD panels from replicate populations resulted in substantial variation of the accuracy within statistical genetic applications. We further demonstrated that these applications perform best when using LD panels obtained directly from the GWAS population. However, the significant variation in the accuracy of downstream results suggests that caution should be warranted when relying on these results in real-world settings. The incorporation of properties of the distribution of LD rather than using point estimates of measurements of LD in these statistical genetic applications will more accurately capture the inherent sampling variation of the LD panels. One limitation regarding these fine-mapping simulations and analyses is the absence of other potentially relevant fine-mapping metrics to evaluate the accuracy of FINEMAP using point estimates of LD from genetically drifted populations. For instance, future work in this area could involve analysing other metrics such as the Receiver Operating Characteristic (ROC) curve, or evaluating the magnitude of Type 2 error. Including these additional sets of metrics may provide a more comprehensive overview as to the strengths and weaknesses of fine-mapping under the different conditions considered in this chapter.

Previously, when we derived the summary statistic imputation method, we implicitly imputed the *expected* z-score under the multivariate normal approximation. However, a more nuanced approach in line with the results of this chapter would involve the incorporation of the distribution of LD within summary statistic imputation. Using this approach, we may be able to generate confidence intervals for the imputed summary statistics, rather than just a single point estimate of the expected z-score. A similar idea could be also be developed in the context of fine-mapping.

In conclusion, the results in this chapter suggest that statistical genetic applications can be improved by modelling the uncertainty in LD. Modelling uncertainty would prove to be particularly advantageous in settings where a genetically drifted reference population is used as a substitute for a GWAS population for the purposes

of summary statistic imputation and fine-mapping. In the next chapter, we develop a method that aims to model this distribution.

3

Generating Distributions of Linkage Disequilibrium Under Genetic Drift

Contents

In recent years, researchers have become increasingly reliant on using reference population data as a substitute for population genetic measurements such as LD and allele frequencies in a GWAS population. However, the reliance on reference population data leads to two major issues that are often overlooked. Firstly, a reference population provides only a single point estimate of the measurements of LD within a population. In Chapter 2, we demonstrated that these point estimates

of LD exhibit substantial variation between populations which may lead to variation downstream within the accuracy of statistical genetic applications. Secondly, a reference population does not often possess the same ancestral composition as the GWAS population, resulting in genetic drift between the two populations. Consequently, this drift leads to population genetic measurements in the reference population that may not be a suitable substitute for the measurements within the GWAS population.

In this chapter, we aim to address the above issues by modelling the distribution of measurements of LD in a GWAS population that is genetically drifted from a reference population. In doing so, the robustness and calibration of downstream statistical genetic applications can be improved.

Our method utilizes a reference population and weights the reference haplotypes to generate population genetic measurements including LD in a genetically drifted population. Crucially, we perform this process by repeatedly sampling weights from a distribution which allows for variation in the population genetic measurements. Using a genealogical interpretation, we model the distribution of allele frequencies in the GWAS population by using the reference population as an approximation of the ancestors to the GWAS population and modelling drift between the two. Further, we explain how genetic drift between the reference and GWAS population can be modelled by sampling the weights from a Gamma distribution.

Under different demographic histories using populations simulated under the coalescent along with a study of real GWAS data from the UK Biobank, we assess the calibration of our model and in what settings our model succeeds in generating realistic drifted population genetic measurements. Lastly, we construct an alternative manner to samples the weights through repeatedly updating the weights of the reference haplotypes. We end by briefly discussing how this process will lead to a method that better approximates the LD in the GWAS population and in a manner that accounts for the ascertainment process (See Chapter 5).

3.1 Modelling GWAS Populations under Genetic Drift

Consider both a reference population R that contains M_{ref} haplotypes and N variants, (See Table 3.1 for a description of all notation used in this chapter) and a GWAS population G which is genetically drifted relative to the reference panel and contains the same set of variants as R. Note that in real-world settings, the populations R and G may contain private variants which may constitute a large fraction of trait variance although in GWAS analyses these variants can be difficult to ascertain within a population (Tam et al. 2019). R and G are both binary encoded matrices with 0 and 1 denoting the reference or alternate allele for a given haplotype i and variant j respectively. Given the level of genetic drift between the two populations, either known in advance or inferred in some manner from the data[1], we aim to sample from the distribution of population genetic measurements, such as the allele frequencies and LD in G.

We begin by presenting a generative model to simulate population genetic measurements in the drifted population G using R. These statistics include both the drifted population allele frequencies π^G, as well as the drifted population LD measurements r^G. Importantly, under the generative model, we do not directly simulate G. Instead, we first simulate the correspondence between R and G using the haplotype weighting scheme described below. Next, using these weights along with R, we calculate the drifted population genetic measurements that would normally be obtained directly from G in real-world settings.

3.1.1 Haplotype Weighting Scheme

The reference population R is customarily treated as a static and rigid entity. However, as we are interested in using R to sample drifted population genetic

[1]See Chapter 4 for details on a method to infer F_{st} between a reference and GWAS population using summary statistics.

Table 3.1: A description of all notation and constants used in the modelling described in Chapter 3.

Symbol	Definition
R	Reference population (matrix) with dimensions (M_{ref} , N)
G	GWAS population (matrix) with dimensions (M_{GWAS} , N)
M_{ref}	Number of haplotypes in R
M_{GWAS}	Number of haplotypes in G
N	Number of variants in the reference population R
i	Haplotype index
j,$j^{'}$	SNP index
l	Weight matrix update index
π^G	Drifted allele frequencies
π^R	Reference allele frequencies
r^R	LD measure r in the reference population
r^G	Drifted LD measure r
W	Weight matrix with dimensions (M_{ref}, N)
$W^\star$	normalized weight matrix with dimensions (M_{ref}, N)
$w^{'}$	Proposed weight vector with length N
K	Number of weight states in $\mathbb{W}$
k	Index of weight state in $\mathbb{W}$
$\mathbb{W}$	Set of Gamma distributed weights
N_e	Effective population size
F_{st}	Genetic drift parameter within Gamma distributed weights
$\hat{F_{st}}$	Estimated genetic drift
$F_{st}^\star$	Effective genetic drift
c	Balding-Nichols genetic drift parameter ($\frac{1-F_{st}}{F_{st}}$)
t	Time (in generations)
α	Dirichlet concentration parameter

measurements from a distribution, rather than using R to obtain a single point estimate for these measurements, it is more appropriate to consider R as *adaptable* and the basis for a prior for the GWAS population characteristics. This adaptability comes about by not considering reference haplotypes in the population as equals but rather, over-weighting or under-weighting the reference haplotypes. Through this *haplotype weighting scheme*, we obtain a weighted version of R whose population genetic measurements can now more closely resemble their drifted counterpart measurements in G. Therefore, repeatedly generating weighted versions of R provides us with the ability to sample drifted population genetic measurements

from a distribution.

In order to formally implement the above scheme, we first generate a haplotype weight matrix W that possesses the same dimensions as R. Each element w_{ij} in W is a positive-valued weight from the set $\mathbb{W}$ which corresponds to the i^{th} haplotype and the j^{th} variant in the reference population. The weights at each respective haplotype (ie. along each row in W) remain constant across each SNP such that $w_{i1} = w_{ij} \forall j$. In Chapter 5, we relax this constraint and describe how the weights change on a per-SNP basis across the region.

These weights can be transformed into normalized proportions at each column in W, resulting in a normalized weight matrix denoted as $W^\star$. These proportions are defined for each element in $W^\star$ as $w^\star_{ij} = \frac{w_{ij}}{\sum_{l=1}^{M_{ref}} w_{lj}}$ which results in $\sum_{i=1}^{M_{ref}} w^\star_{ij} = 1$ for all variants j. By treating the weights as proportions, we produce more natural correspondences with various population genetic measurements such as the allele frequencies.

3.1.2 Generating Drifted Population Statistics

We use $W^\star$ and R to obtain the allele frequency at variant j in the drifted population G by calculating the mean weight carried by the haplotypes with the alternate allele at SNP j described as:

$$\pi^G_j = \sum_{i=1}^{M_{ref}} (R_{ij} w^\star_{ij}). \tag{3.1}$$

Notice that setting identical weights across $W^\star$ results in the reference allele frequencies π^R equalling the drifted allele frequencies π^G.

In a similar approach, the proportional weights in $W^\star$ and the reference population R are used to calculate drifted measurements of LD. Because the weights do not change along the region, each row in W can be thought of as being proportional to the number of copies of a given reference haplotype in the GWAS population. Therefore, the r measurement of LD between two SNPs j

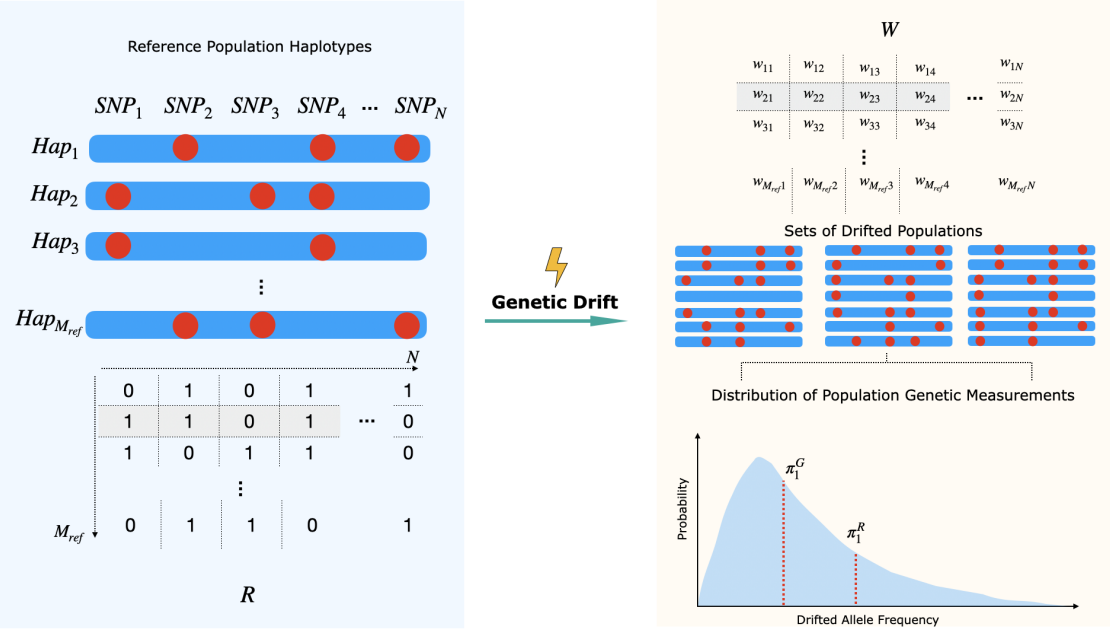

Figure 3.1: Haplotype weighting scheme used to generate a series of genetically drifted populations from a reference population R and a weight matrix W. Generating genetic drift through the weights is done in a number of different ways, each of which generates a different drifted population. These populations can then be used to obtain downstream population genetic measurements, such as the allele frequencies. The distribution of the drifted allele frequency of an arbitrary SNP across these populations is illustrated by the blue-shaded distribution in the right pane. The allele frequency in the drifted population G is denoted as π_1^G and the allele frequency of this SNP in the reference population is denoted by π_1^R. Both of these allele frequencies are indicated by the red dotted lines on this distribution.

and $j^{'}$ in the drifted population G is described as a weighted Pearson correlation where the weighted covariance is:

$$cov(G_j, G_{j'}) = \sum_{i=1}^{M_{ref}} w_{ij}^{\star}(R_{ij} - \pi_j^G)(R_{ij'} - \pi_{j'}^G) \tag{3.2}$$

and therefore the weighted correlation using this expression is:

$$corr(G_j, G_{j'}) = \frac{cov(G_j, G_j^{'})}{\sqrt{cov(G_j, G_j)cov(G_j^{'}, G_j^{'})}}. \tag{3.3}$$

Notice that when all the weights in $W^{\star}$ are identical, the weighted Pearson correlation simplifies to measuring r between pairs of SNPs (Equation 1.2).

3.1.3 Genealogical Interpretation

How should we model the distribution of W? To answer this question we consider a genealogical interpretation of the haplotype weighting scheme where the haplotypes in the reference population acts as the ancestors to their descendant haplotypes in the GWAS population. To illustrate and motivate this interpretation, we describe the coalescent tree without recombination showing the genealogical relationship between the two populations R and G (Figure 3.2).

In this genealogical model, we make two assumptions. First, the GWAS sample size is much greater than the reference sample size[2]. Secondly, these two populations are the result of an ancestral population split which occurred a relatively recent time ago t in the past, which induces genetic drift between the two populations. If the population split occurred fairly recently in the past, this implies that the coalescent events in the past beyond the split will largely occur between two GWAS samples, or a GWAS sample and a reference sample (Wakeley 2009). As a result, few reference samples coalesce with other reference samples prior to coalescing with a lineage with GWAS population descendants. Since the coalescence is largely occurring within the GWAS population this means that if we consider the genealogical relationships before the split t, ancestral lineages to both the reference and GWAS population are exchangeable. The exchangeability of these lineages means that the two trees depicted in Figure 3.2 b), which describe the genealogy are topologically equivalent. Consequently, under these assumptions, samples within the reference population can be used to approximate the distribution of the ancestors to the GWAS population.

Importantly, these approximations do not hold if there has been a high degree of coalescence in the reference population relative to the coalescence in the GWAS population. For instance, if the reference population experienced a bottleneck, these approximations would not hold (DeGiorgio et al. 2011). A prominent example of

[2]Ideally, $Ne_R > Ne_G$ as well.

this exception is the case of the Finnish population which experienced a bottleneck roughly 4000 years ago (Sajantila et al. 1996). Consequently, using a Finnish reference population in order to model a European GWAS population, such as the UK Biobank would result in a very poor approximation for the distribution of the ancestors to the GWAS population.

Using this genealogical interpretation, we can model the distribution of allele frequencies in the GWAS population. We begin by constructing the GWAS descendants from the reference population ancestors through a forwards in time coalescent process without recombination (Donnelly and Tavare 1995). This process was described in detail in Chapter 1 using a Hoppe Urn model (Figure 1.4)(Hoppe 1987). For a large descendant population generated under neutral evolution, it has been shown that the stationary distribution of allele frequencies of the descendants follows a symmetric Dirichlet distribution parametrised by α as $Dir(\alpha)$ (Kingman 1982, Watterson 1975, Hoppe 1987). When α is large ($\alpha \gg 1$), this models each descendant individual as having allele copies originating from all ancestors in equal proportions. Conversely, when α is small ($\alpha \ll 1$), each descendant individual has allele copies originating mostly from a single ancestor, with all ancestors being equally likely.

Therefore, within W, we assume the normalized weights are Dirichlet distributed, and that allows us to model the distribution of the GWAS allele frequencies using the haplotype weighting scheme. By modelling the distribution of allele frequencies with W, we implicitly obtain the distribution of the LD measurements in the GWAS population. Our model is similar to the STRUCTURE model (Pritchard, M. Stephens, et al. 2000) which follows the suggestion of Nichols and Balding (Balding and Nichols 1995) in using a Dirichlet distribution to model the allele frequencies at each locus in a population.

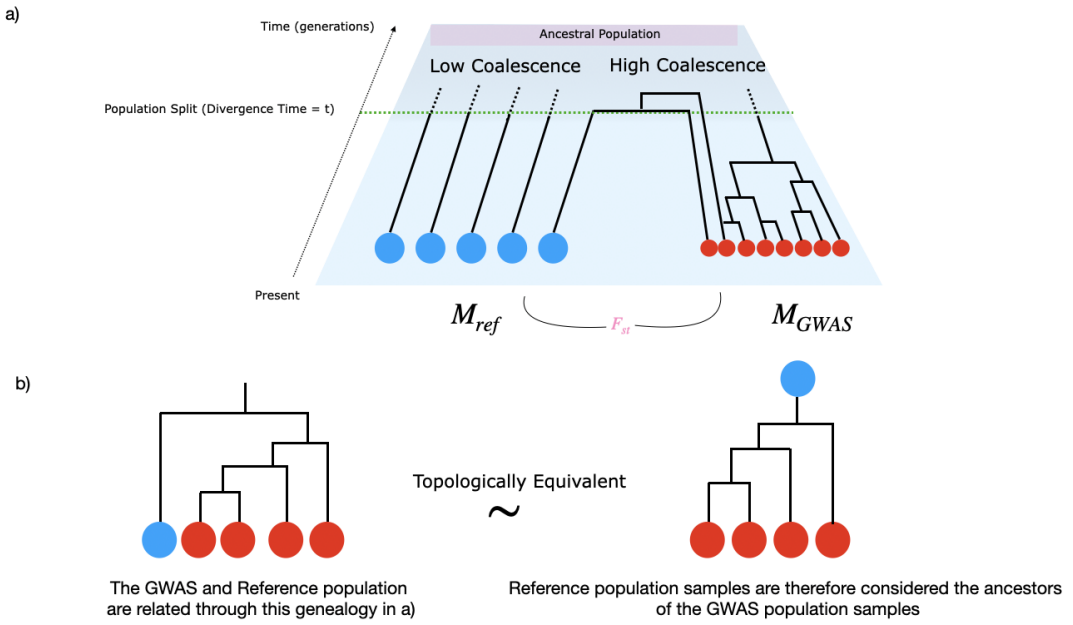

Figure 3.2: Genealogical interpretation of the haplotype weighting scheme under the coalescent without recombination. a) Genealogical relationship between the samples in the reference and GWAS population. The two populations are coloured as red and blue which correspond to the GWAS and reference population respectively. The number of samples in the GWAS population is much greater than the number of samples in the reference population. These two populations have split from an ancestral population some time t generations ago in the past. We assume the time of the split occurred recently in the past, resulting in low coalescence in the reference population as compared to the high coalescence in the GWAS population. b) Two topologically equivalent trees. The tree on the left hand side shows the relationship of a sample of the GWAS population relating to the other M_{ref} samples. Under the assumptions in our model, this tree is topologically equivalent to the tree on the right hand side. This equivalency indicates how we represent the reference samples as ancestors to the samples within the GWAS population.

3.1.4 Gamma Distributed Weights

For convenience, an alternative - and equivalent - procedure is to model the unnormalized weights, which can be achieved using the relationship between the Gamma and Dirichlet distributions. If the weights across the reference haplotypes are independently Gamma distributed with shape parameter (α_1,α_2,...,$\alpha_{M_{ref}}$), the weights become Dirichlet distributed $Dir(\alpha_1, \alpha_2, ..., \alpha_{M_{ref}})$ after normalizing (Wong 1998). In other words, this transformation gives us a process to generate Dirichlet distributed normalized weights by normalizing a set of Gamma distributed unnormalized weights. In the previous section, we demonstrated that this Dirichlet

distribution was symmetric and therefore parametrised for all reference haplotypes with a single parameter α. Therefore, using the relationship between the Gamma and Dirichlet distribution, the Gamma distribution is equivalently parametrised across all reference haplotypes using a single shape[3] parameter α as:

$$f(x;\alpha) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x}. \tag{3.4}$$

We set the shape parameter α equal to $\frac{1}{M_{ref}c}$ where $c = \frac{1-F_{st}}{F_{st}}$. To motivate this parametrisation, we consider a reference population of M_{ref} haplotypes in which i of these haplotypes encode the alternate allele at SNP j. We sample Gamma distributed weights using this parametrisation, populate W independently across reference haplotypes and normalize the weights to obtain $W^{\star}$. As a result of this process, the expectation of the allele frequency at SNP j is:

$$\mathbb{E}[\pi_j^G] = \frac{i}{M_{ref}} = \pi_j^R \tag{3.5}$$

and the variance is:

$$Var[\pi_j^G] = F_{st}\frac{i}{M_{ref}}(1 - \frac{i}{Mref}) = F_{st}\pi_j^R(1-\pi_j^R). \tag{3.6}$$

Next, recall that the reference population represents the ancestors to the GWAS population. Therefore, the marginal drifted allele frequency at each SNP in the descendant GWAS population is distributed according to the the Balding-Nichols beta model (Balding and Nichols 1995) (Equation 1.12). Consequently, setting $\alpha = \frac{1}{M_{ref}c}$ preserves the relationship between the haplotype weighting scheme and the Balding-Nichols model for genetic drift. Note that when we have more reference haplotypes, we require the Gamma distributed weights to be more skewed to achieve the same level of allele frequency drift as compared to fewer reference haplotypes.

[3]We additionally set the rate parameter equal to the shape parameter for convenience.

3.2 Assessing Model Coverage under Genetic Drift

3.2.1 Effective Genetic Drift

Importantly, observe that between R and G there exists two sources of variation. The first source of variation is genetic drift, and we have previously described modelling this variation using the haplotype weighting scheme under the Balding-Nichols model. The second source of variation is the finite sampling of the population. In order to model variation which arises from finite sampling, we implement an extension to the Bayesian bootstrap (Rubin 1981). Traditionally, the Bayesian bootstrap accounts for finite sampling by generating distributions of repeated samples within the *same* population. However, under the genealogical interpretation of the haplotype weighting scheme, the reference population is used to approximate the distribution of the ancestors to the GWAS population. Therefore, we extend the Bayesian bootstrap to instead model a population that may be somewhat diverged from the reference using repeated sampling with *drift*. In order to do so, we repeatedly sample weights from the Dirichlet distribution $Dir(\alpha)$ derived in the previous section and populate W in an independent and identically distributed manner across each of the reference haplotypes.

We now discuss how to appropriately set the correct level of F_{st} within Equation 3.4 to generate the set of gamma distributed weights under the Bayesian bootstrap with drift. Let Y be the allele count in a population. Suppose we have drawn M_{ref} samples from the reference population and M_{GWAS} samples from the drifted GWAS population. Consider $\frac{Y_{GWAS}}{M_{GWAS}}$ given $\frac{Y_{ref}}{M_{Ref}}$ which represents the distribution of the empirical frequency in the GWAS sample given the empirical frequency in the reference sample. The expectation of $\frac{Y_{GWAS}}{M_{GWAS}} - \frac{Y_{ref}}{M_{Ref}}$ is zero. Therefore, the reference sample frequency is an unbiased estimator of the GWAS sample frequency irrespective of drift.

Similarly, the variance of $\frac{Y_{GWAS}}{M_{GWAS}} - \frac{Y_{ref}}{M_{Ref}}$ dictates how well the allele frequency of the reference sample models the allele frequency of the GWAS sample. Assuming that the frequency of the variant in the ancestral population is x, under the Balding-Nichols beta-binomial model it follows that:

$$Var[\frac{Y_{GWAS}}{M_{GWAS}} - \frac{Y_{ref}}{M_{Ref}}] = x(1-x)(2F_{st} + \frac{1}{M_{ref}} + \frac{1}{M_{GWAS}}). \tag{3.7}$$

The variance of the Balding-Nichols model is $x(1-x)F_{st}$. However, Equation 3.7 implies that under the Bayesian bootstrap with drift we are required to set an *effective* level of drift ($F^{\star}_{st}$) as:

$$F^{\star}_{st} = 2F_{st} + \frac{1}{M_{ref}} + \frac{1}{M_{GWAS}}. \tag{3.8}$$

Interestingly, if $F_{st} = 0$, $F^{\star}_{st}$ is approximately $\frac{1}{min(M_{ref}, M_{GWAS})}$. This suggests that we are able to stochastically re-weight the reference population even in the absence of drift to obtain a better estimate of the distribution of the allele frequencies and LD measurements and account for the finite sampling of the population.

Within the set of gamma distributed weights $\mathbb{W}$ generated using Equation 3.4, we set the level of $F^{\star}_{st}$[4] that corresponds to the effective genetic drift under the Bayesian bootstrap with drift. Estimates for the level of drift ($\hat{F_{st}}$) can be obtained in several ways. For instance, drift can be estimated from the individual level GWAS data or inferred from the GWAS summary statistics as described in Chapter 4. In this chapter, we use the MLE developed by Nielsen (Nielsen et al. 1998) which was specifically derived for estimating genetic drift arising from a population split (Equation 1.11).

[4] $F^{\star}_{st}$ refers to the parameter used to generate the gamma distributed weights in $\mathbb{W}$, whereas $\hat{F_{st}}$ refers to the estimated level of genetic drift between R and G.

3.2.2 Coalescent Models

We now assess the calibration of the distributions of the population genetic measurements generated from the haplotype weighting scheme using simulated genetic data. If the true drifted measurements in G are found uniformly across the quantiles of the distribution, our model is well calibrated (Gentle 2009).

We simulate genetic data from *msprime* (Kelleher, Etheridge, et al. 2016) under the demographic event of a population split in which an ancestral population splits some time ago in the past to form the reference and GWAS population. In this demographic event (Figure 3.3), we vary the genetic drift between the GWAS and reference population by increasing or decreasing the time of the population split (Holsinger et al. 2009). In *msprime*, we set a mutation rate of 1e-8, an effective population size of the ancestral population of 10,000, a descendant population effective population size of 5,0000, and a sequence length of 1Mb. Importantly, we do not simulate recombination in the region. We draw 1000 and 5000 haplotypes within the two descendant populations in the demographic model to form the reference and GWAS samples respectively. We jointly filter out non-segregating SNPs in either the GWAS or the reference population, resulting in approximately 500 SNPs in a *msprime* region. Therefore, R and G both contain the same set of SNPs.

Next, we sample M_{ref} weights uniformly from $\mathbb{W}$ in an independent and identically distributed manner to populate W. Using R and W, we calculate the drifted allele frequencies and LD measurements (Equation 3.1 and 3.3). By repeatedly sampling the set of M_{ref} weights P times, we generate a distribution for the drifted allele frequencies and LD measurements under a value of $F_{st}^{\star}$ set in $\mathbb{W}$. Next, we split the distributions of each of the respective measurements of the drifted allele frequencies and LD into 100 quantiles. Finally, we record in which quantile of the distribution the true allele frequency and the LD measurements from G falls into respectively. To ensure robustness, we simulate Q replicate pairs

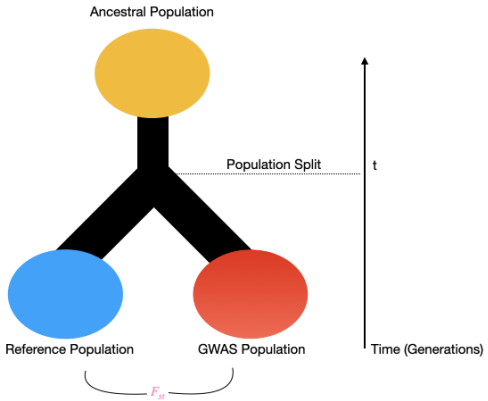

Figure 3.3: Demographic history of a population split from an ancestral population some time t ago in the past which forms the GWAS and reference population. These two present-day populations are drifted with respect to one another by a level of F_{st}.

of G and R, each time performing the process described above for each SNP and SNP-SNP pair within each simulation and pooling the results across replicates.

We simulate the replicate GWAS and reference samples across a series of divergence times corresponding to the population split between the reference and GWAS population and across different values of $F_{st}^{\star}$ set in $\mathbb{W}$. The calibration of our model in these different settings is shown in Figure 3.4.

First, focusing on the calibration of the allele frequency distributions (Figure 3.4 a), we observe that when the value of $F_{st}^{\star}$ is set too high relative to the split time (ie. subplots above the diagonal) we simulate too much drift in our model. Therefore, most of the drifted allele frequencies have too low of a minor allele frequency relative to their measurement in G. This results in a departure from the theoretical distribution. For instance, the distribution in the upper right corner in Figure 3.4 a) with $F_{st}^{\star} = 0.0506$ and the time of the population split set to 0 generations in the past exhibits this feature. Conversely, setting $F_{st}^{\star}$ too low relative to the split time (ie. subplots below the diagonal) simulates too little drift. Therefore, the drifted allele frequencies match the reference panel frequencies more than expected given the level of drift. This results in these calibration plots resulting in a departure from the theoretical distribution. For instance, the distribution in the lower left

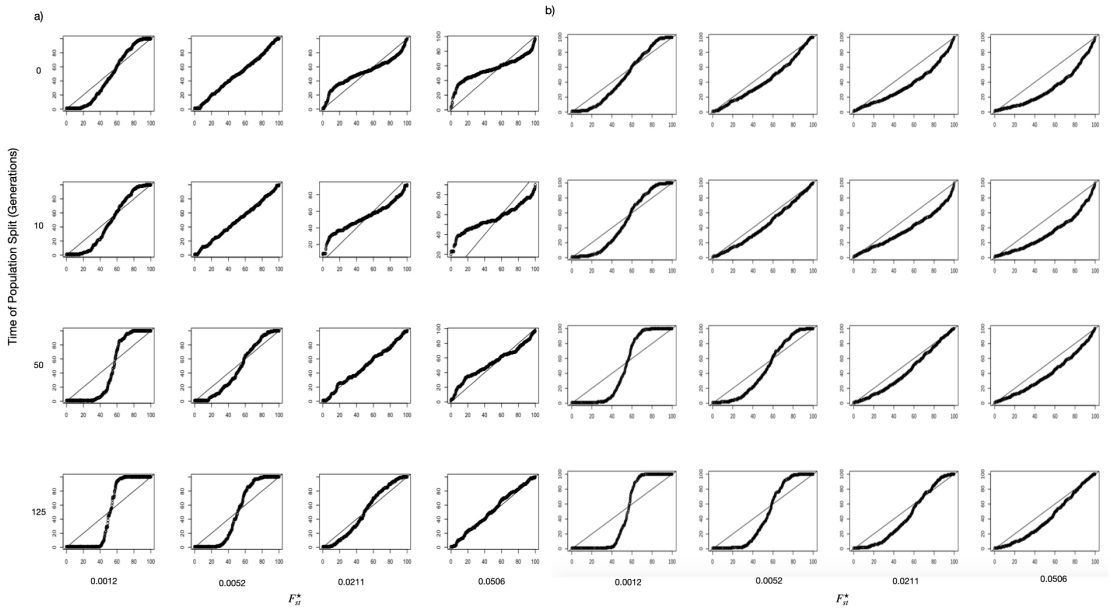

Figure 3.4: QQ plots showing the calibration of the distribution of a) the allele frequencies and b) LD measurements (r) across a series of diverged GWAS and reference populations from an ancestral population as a result of a population split for different values of $F_{st}^{\star}$. Each calibration plot corresponds to $Q = 1000$ replicate *msprime* population pairs. For a single *msprime* population, weight matrices W were repeatedly sampled $P = 1000$ times to generate the distribution of the drifted measurements. Each distribution was broken into 100 quantiles and for each allele frequency and LD measurement in the GWAS population we recorded the quantile in the distribution where the true value was located. This process was repeated across every SNP and SNP-SNP pair in the *msprime* population, generating a series of quantile counts. These counts were then pooled across the replicate *msprime* simulations. The x-axis denotes the theoretical quantiles from a Uniform distribution and the y-axis denotes the distribution of the observed quantiles of the counts. The black line displays the relationship where the observed and theoretical quantiles are perfectly aligned indicating perfect calibration. Each row of subplots corresponds to the time of the population split (in generations) of the GWAS and the reference population simulated in *msprime*. Each column of subplots denotes the $F_{st}^{\star}$ value used to generate $\mathbb{W}$. The levels of $F_{st}^{\star}$ were calculated using Equation 1.11 to estimate the level of F_{st}, and then adjusted to obtain the effective drift using Equation 3.7.

corner in Figure 3.4 a) with $F_{st}^{\star} = 0.0012$ and the time of the population split set to 125 generations in the past exhibits this feature.

Importantly, we observe that our model is well calibrated when $F_{st}^{\star}$ matches the level of genetic drift between the reference and GWAS population (ie. subplots along the diagonal). We further observe our model becomes less calibrated as the difference between $F_{st}^{\star}$ and the true genetic drift between the reference and

GWAS population increases. Interestingly, when the population split occurred 0 generations ago (ie. at the present) our model appears best calibrated when setting $F_{st}^{\star}$ to 0.0052 corresponding to a split time 10 generations ago. This result implies that in settings where the genetic drift between R and G is expected to be zero, we still observe excess variability between these populations even relative to re-sampling the weights without drift.

The calibration of the distributions of the LD measurements displays a similar trend as compared to the calibration of the distributions of the allele frequencies. However, a right skew in the distributions appears even when the value of $F_{st}^{\star}$ is set at the appropriate value. Therefore, this results suggests it is more challenging for the haplotype weighting scheme to model drift in the measurements of LD than it is to model drift in the allele frequencies.

3.2.3 Out of Africa Model

In order to simulate more realistic genetic variation within the reference and GWAS population, we simulate these populations under the Out of Africa (OOA) model (Figure 3.5). The OOA model refers to the human expansion out of Africa and the settlement of the New World. The demographic history of this model was inferred by Gutenkunst (Gutenkunst et al. 2009). This model describes the ancestral human population in Africa, the out of Africa event and the subsequent European-Asian population split which forms three present-day populations (CHB, CEU and YRI). These results are derived from the HapMap populations (The International HapMap Consortium 2005) and refer to a Chinese (CHB), European (CEU) and sub-Saharan Yoruba (YRI) population.

Under an OOA model, we simulate present-day CHB, CEU and YRI populations. We generate all possible pairings[5] of the GWAS and reference population for a total of three pairings. As before, we simulate the GWAS and reference populations using

[5]Labelling the GWAS and the reference population without loss of generality.

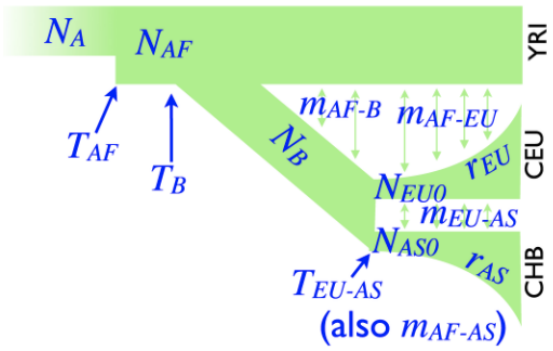

Figure 3.5: Out of Africa model described in Gutenkunst et al. 2009. We construct the GWAS and reference populations using pairs of populations from the CHB, CEU and YRI present-day populations in this model. Model parameters and notation are described in Table 1 of Gutenkunst et al. 2009. Source: Gutenkunst et al. 2009

msprime under a human level mutation rate and no recombination in a 1 Mb region. Each of the GWAS and reference populations contain 5000 haplotypes sampled from the OOA present-day populations for each respective population pairing. As in the previous section, we jointly remove any non-segregating variants in either the reference or GWAS population which results in approximately 500 SNPs in a region. Note that in the following analyses, we refer to the pairing of the OOA populations for R and G as "GWAS-Reference". In other words, the first population represents the GWAS and the second population represents the reference.

We again assess the calibration of the distribution of the allele frequencies and the measurements of LD across a series of replicate regions as we did previously under a demographic history of a population split. However, we no longer implement the MLE (Equation 1.11) to set the level of $F_{st}^{\star}$. This MLE is derived in the context of a single population split but the OOA model contains other combinations of demographic events. Therefore, we instead we rely on the results of Bhatia et al. 2013 which previously measured the levels of genetic drift between the CHB, CEU and YRI populations. Guided by these results, we set values of $F_{st}^{\star}$ as 0.15, 0.25 and 0.35. Although these values are not exact, they provide us with an appropriate range of values to assess the calibration of our model.

In these simulations, the minimum MAF in the GWAS population is $\frac{1}{M_{GWAS}}$ or 0.02%. Under high levels of $F_{st}^{\star}$, we may generate drifted allele frequencies that fall below this threshold. Consequently, when generating the weight matrices W, we only include the resulting drifted allele frequencies in the distribution if they possess a MAF greater than $\frac{1}{M_{GWAS}}$ to fit the constraints of our model. Similarly, the drifted measurements of LD across each pair of SNPs requires each of the SNPs to be above the MAF threshold. We repeatedly sample W until we have recorded 1000 drifted allele frequency measurements and 1000 drifted LD measurements that pass the MAF threshold for each SNP and each SNP-SNP pair in the population pairing. This filtering is performed marginally and results in a set of samples of W for SNP j which may be different than the set of samples for SNP $j^{'}$.

The calibration of the distributions generated by repeatedly sampling W for pairs of populations in the OOA model and across different values of $F_{st}^{\star}$ is shown in Figure 3.6.

We begin by examining the calibration of the distribution of the allele frequencies. The CHB-CEU population pairing appears to be best calibrated under a $F_{st}^{\star}$ of 0.25. Whereas, the YRI-CEU pairing appears best calibrated under a $F_{st}^{\star}$ of 0.25 and 0.35. Moreover, the CHB-YRI pairing also appears best calibrated under a $F_{st}^{\star}$ of 0.25 or 0.35. In all three cases, the distributions are not as well calibrated as compared to the demographic event of a population split and contain a rightward skew (Figure 3.4 a). One feature of these calibration plots that we note in particular are the "elbows" in the bottom left and upper right corners of the plot. These elbows correspond to when the true allele frequencies across a high number of SNPs is found only in the highest and lowest quantile in the distribution of the drifted allele frequencies we generate in our model. This results in the observed quantiles that we compare to the theoretical uniform quantiles possessing a "U" shape.

This elbow is most prominent in the case of the CHB-YRI pairing under a $F_{st}^{\star}$ of 0.15. Similar to the interpretation of the plots below the diagonal in Figure

3.4 a), one possible interpretation of this feature is that we do not simulate high enough levels of F_{st} to capture the drift in allele frequencies in these populations. Simulating additional drift will shift the quantiles of the distributions towards zero and one and therefore may better capture allele frequencies that drift towards rare minor allele frequencies in G. However, the the values of F_{st} are already in-line with the previous values estimated for the OOA population pairs (Bhatia et al. 2013) and this elbow remains in the calibration plots across each of the population pairings for different levels of F_{st}. These observations suggest that incorrectly settings the levels of F_{st} within our model is not the source of this feature. Another interpretation for this feature is the inability of our model to generate realistic patterns of genetic drift given the demographic events within the OOA model. For instance, migration between populations (eg. African and European) decreases the levels of drift (Bodmer and Cavalli-Sforza 1968). Additionally, the demographic event in the OOA model of a population expansion also decreases levels of drift (Keinan et al. 2007). Therefore, we may overestimate drift due to the fact that we do not incorporate these demographic events in our model. In contrast, we may also underestimate genetic drift generated from the OOA model in the haplotype weighting scheme. Our model incorporates only a single population split whereas the OOA model incorporates two population splits which form the three present-day populations. The haplotype weighting scheme does not account for the increase in genetic drift in the OOA model that may be generated from these additional population splits (Pickrell and Pritchard 2012).

Therefore, the weaker calibration of these distributions using simulated data from the OOA populations (ie. the appearance of the rightward skew and the elbows) relative to the calibration using data simulated from a population split is likely a consequence of our model being unable to generate realistic genetic drift under these other demographic events.

Next, we examine the calibration of the distribution of LD measurements. Here, we observe that in all three population pairings that our model is best calibrated under the highest level of $F_{st}^{\star}$ at 0.35. Interestingly, unlike the demographic event of a simple population split (Figure 3.4 b), the values of $F_{st}^{\star}$ which result in the highest calibration under our model for each of the population pairings does not always match between the LD measurements (Figure 3.6 a) and the allele frequencies (Figure 3.6 b).

3.2.4 UK Biobank

Finally, we assess the calibration of our model using real GWAS and reference population data. We employ the individual level genotyped data provided by the UK Biobank (Bycroft et al. 2018). Although the UK Biobank is composed of a largely European population, individuals of non-British and admixed ancestry are present in the data-set. Consequently, we select only individuals with assigned white-British ancestry (WBA) and those passing a PCA filter (See `https://github.com/Nealelab/UK_Biobank_GWAS` for details).

Each of the 1000G super-populations (EUR, SAS, AFR, EAS and AMR) are employed as the respective reference populations. Guided by previous estimates of genetic drift between the 1000G populations (Elhaik 2012), we use a range of $F_{st}^{\star}$ values (0.0001,0.05,0.1 and 0.25) to assess the calibration of our model. Again, as in the case of the OOA model, the values of $F_{st}^{\star}$ are not exact but provide an appropriate range to assess the calibration of the distributions under different population pairings. Implicitly within these values of $F_{st}^{\star}$, we assume that the UK Biobank and 1000G EUR populations are roughly equivalent, and use this assumption to obtain the estimates of drift of the UK Biobank population relative to the other non-European 1000G super populations.

Using the WBA filtered samples of the UK Biobank, we select ten 1Mb regions on Chromosome 22 and extract matching regions within each of the 1000G reference

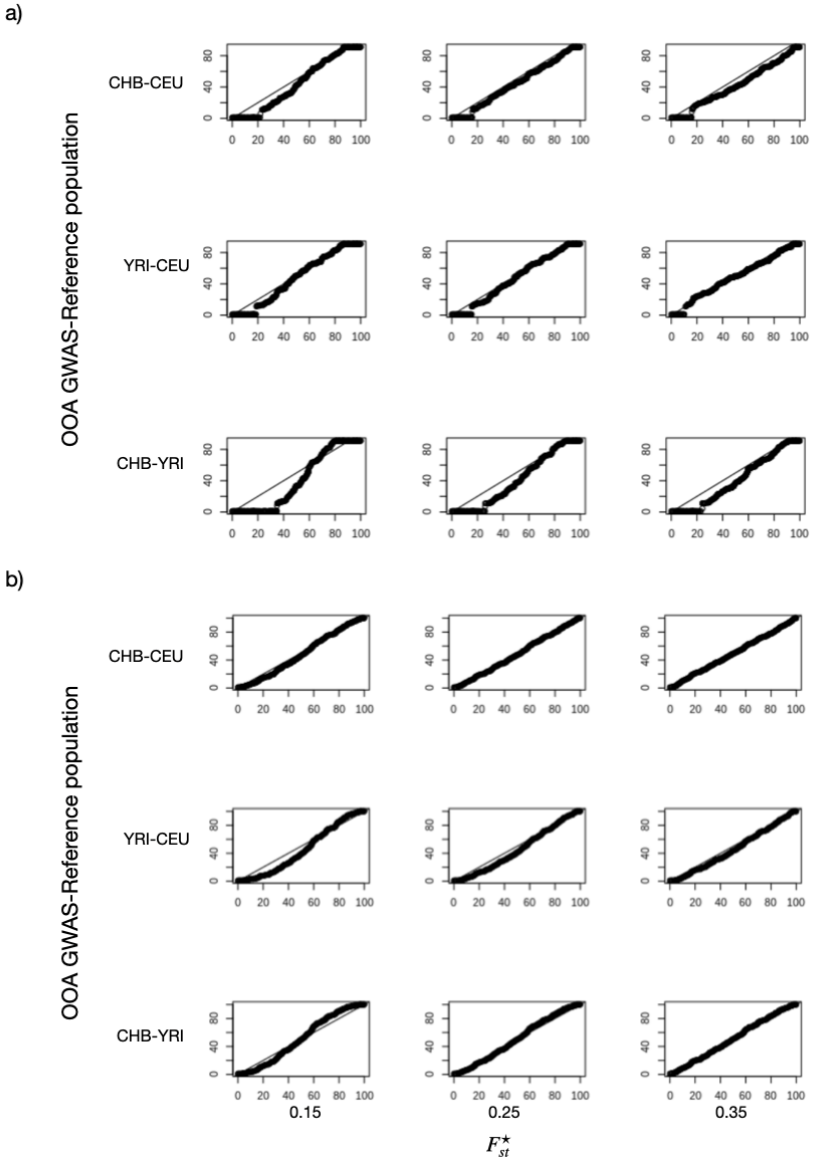

Figure 3.6: QQ plots showing the calibration of the distribution of a) the allele frequencies and b) LD measurements across a series of OOA GWAS and reference population pairings for different values of $F_{st}^{\star}$. Each calibration plot corresponds to $Q = 1000$ replicate *msprime* population pairs and weight matrices W were repeatedly sampled $P = 1000$ times to generate the distribution of the measurements. Each distribution was broken into 100 quantiles and for each allele frequency and LD measurement in the GWAS population we recorded the quantile in the distribution where the true value was located. This process was repeated across every SNP and SNP-SNP pair in the *msprime* population, generating a series of quantile counts. These counts were then pooled across the replicate *msprime* simulations. The x-axis denotes the theoretical quantiles from a Uniform distribution and the y-axis denotes the distribution of the observed quantiles of the counts. The black line displays the relationship where the observed and theoretical quantiles are perfectly aligned indicating perfect calibration. Each column of subplots corresponds to a particular OOA GWAS and reference population pairing. Each column of subplots denotes the $F_{st}^{\star}$ value used to generate $\mathbb{W}$.

populations. For each region, we jointly filter out SNPs in the GWAS and reference population that fall below 1% MAF in any of the reference populations. Further, we jointly filter out non-segregating SNPs in the GWAS and reference populations. Therefore, we always compare the same set of SNPs across different GWAS and reference population pairings. As in the previous section, we generate the distribution of the drifted population genetic measurements by repeatedly sampling weights using W. As previously, we aim to set a MAF threshold of $\frac{1}{M_{GWAS}}$ at approximately 0.0003%, to filter out the drifted allele frequencies and drifted LD measurements within the distributions. However, SNPs with very low allele frequencies are likely to have been filtered out in the process of quality control and treated as an error within sequencing or at some other point in data collection (Marees et al. 2018). Therefore, it is incredibly unlikely that we would observe a SNP with a MAF in this low range in the GWAS population. None the less, we still use a threshold of $\frac{1}{M_{GWAS}}$ due to the extensive computational run-time necessary to employ a more restrictive but realistic threshold such as 1% MAF. The calibration of the distributions of the drifted allele frequencies and measurements of LD across all 1000G reference populations and across all values of $F_{st}^{\star}$ pooled across regions is shown in Figure 3.7.

We begin by examining the calibration of the distribution of allele frequencies. Overall, the calibration of these distributions proves to be poor. An extreme rightward skew is present within each of these calibration plots across all values of $F_{st}^{\star}$ and across all 1000G reference populations. We speculate two reasons why this may be the case. Firstly, as before in the case of the OOA model, the 1000G reference populations and the UK Biobank GWAS population are not formed through a single population split. Rather, the UK Biobank and 1000G populations are both the result of a complex demographic history (Cook et al. 2020). This complex demographic history may increase or decrease drift between the population pairs in patterns that our model is unable to accommodate. The second possible reason

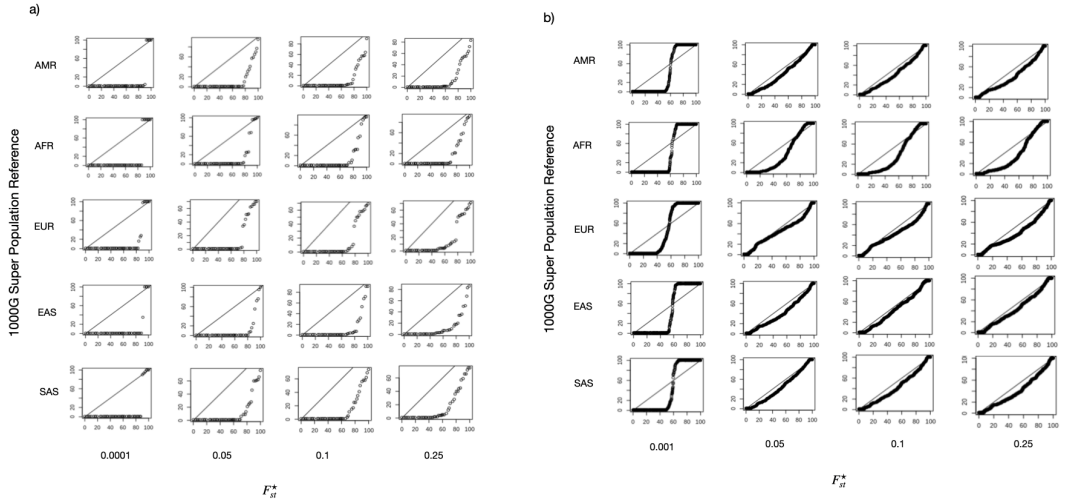

Figure 3.7: QQ plots showing the calibration of the distribution of a) the allele frequencies and b) LD measurements under our model across a series of regions from the UK Biobank GWAS using reference populations from the 1000G super populations for different values of $F_{st}^{\star}$. Each calibration plot corresponds to the results pooled from 10 regions from Chromosome 1 of the UK Biobank and a 1000G super population reference panel where weight matrices W were repeatedly sampled $P = 1000$ times to generate the distribution of the measurements. Each distribution was broken into 100 quantiles and for each allele frequency and LD measurement in the GWAS population we recorded the quantile in the distribution where the true value was located. This process was repeated across every SNP and SNP-SNP pair, generating a series of quantile counts. These counts were then pooled across all regions. The x-axis denotes the theoretical quantiles from a Uniform distribution and the y-axis denotes the distribution of the observed quantiles of the counts. The black line displays the relationship where the observed and theoretical quantiles are perfectly aligned indicating perfect calibration. Each row of subplots corresponds to the 1000G reference population set as R. Each column of subplots denotes the $F_{st}^{\star}$ value used to generate $\mathbb{W}$.

for the poor calibration of the distributions relates to the ascertainment biases of the different SNPs and samples that are included within the UK Biobank data-set (Bycroft et al. 2018) . These ascertainment biases are particularly important when setting the MAF threshold for the drifted allele frequencies generated through the haplotype weighting scheme. Appropriately choosing this threshold is very challenging compared to using *msprime* simulated data and there is not a defined lower bound in the GWAS population that we can rely on. It was previously shown that the ascertainment of SNPs changes the estimates of levels of drift between populations, especially when considering rare variants (Bhatia et al. 2013).

Because the threshold sets a lower bound on the allele frequencies we generate, we speculate that changing this threshold will therefore have a substantial impact on the calibration of our model and may do so in ways that are hard to predict due to the complicated demographic events in the history of the GWAS population.

Next, we examine the calibration of the distribution of LD measurements. Interestingly, with the exception of the lowest value of $F_{st}^{\star}$ (0.0001), we observe much stronger calibration of the distributions of LD measurements relative to the allele frequencies. This is similar to the results we saw previously using genetic data simulated under the OOA model (Figure 3.6). Interestingly, as we increase the value of $F_{st}^{\star}$, this does not affect the calibration of the distributions in a significant manner and our model appears similarly calibrated for values of $F_{st}^{\star}$ above 0.0001. Interestingly, in each of these cases, we observe a rightward skew of these distributions, similar to what we previously observed when assessing the calibration of our model under the demographic history of a population split (Figure 3.4 b).

3.3 Updating Weights

Under the Bayesian bootstrap with drift, we draw independent samples from gamma distributed weights. However, as demonstrated using the UK Biobank data (Figure 3.7), the resulting calibration of the distributions for the allele frequencies measurements were extremely poor. In these analyses, besides the level of drift between R and G, we did not use any other information about G such as the GWAS summary statistics. Therefore, the construction of an inference method that incorporates this data within the process of sampling W is an important future piece of work (See Chapter 5). A key building block to this approach is an alternative algorithm for sampling weights that lends itself to inference.

This process will involve two main changes to the haplotype weighting scheme. First, we will discretize the weight space rather than utilizing a continuous set of

gamma distributed weights. Secondly, we repeatedly sample W through marginal updates (ie. updating a single reference haplotype at a time), rather than drawing an entirely new set of weights across all reference haplotypes.

Firstly, we begin by discussing the need to discretize the weight state space. We note that the proportional weights have constrained support. The proportional weights are constructed such that $\sum_{i=1}^{M_{ref}} w_{ij}^{\star} = 1$ for all j and $w_{ij} \geq 0$ for all i and j leading to the weights being distributed along a simplex. Performing anything besides a trivial update on a simplex is an extremely challenging statistical problem, as the allowed ranges of the sampled quantities are interdependent (Betancourt 2012). Further, in Chapter 5, when we update W through a hidden Markov model (HMM), using continuous weights states rather than discrete weight states greatly complicates the implementation of our model.

Therefore, we discretize the set of continuous gamma distributed weights (Equation 3.4) into a series of equally spaced quantiles. The number of quantiles defined via K governs the size of the state space of weights. Note that the approach of choosing uniformly from evenly-spaced quantiles in the limit of an infinite number of quantiles is guaranteed to converge on the true distribution. Therefore, this construction provides a different way of generating the same distribution as in Equation 3.4 as $lim_{K \to \infty}$.

Secondly, using these discrete weight states, we discuss how we can perform marginal updates to weights in W to generate the samples in a manner that lends itself to downstream inference. We implement these marginal updates by updating a single *focal haplotype* while *freezing* the weights on the remaining $M_{ref} - 1$ haplotypes. When we undergo inference using this model in Chapter 5, these updates are performed using a dynamic programming algorithm that is very similar to the Li and Stephens model (Li and Stephens 2003). Algorithm 1 describes the steps to update the weight matrix W and Figure 3.8 illustrates this approach.

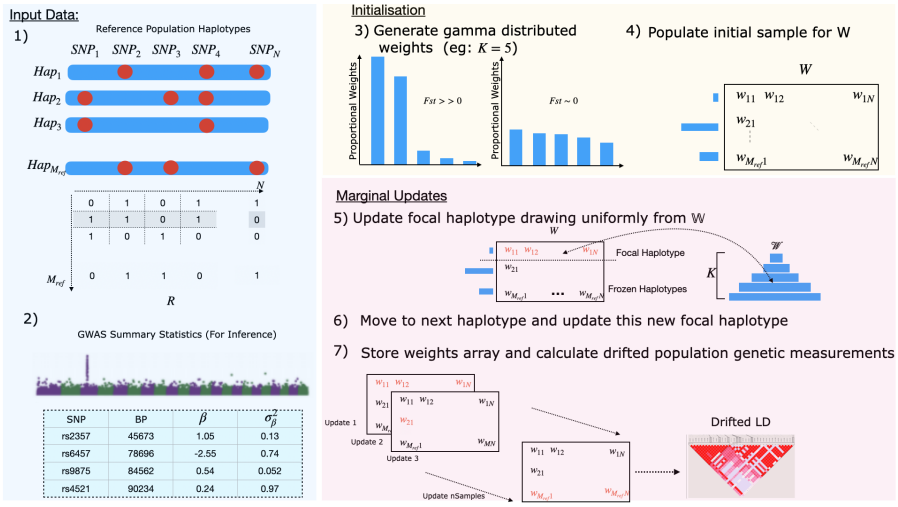

Figure 3.8: Marginally updating W. Steps 1) and 2) describe the reference population haplotypes and the (optional) GWAS summary statistics as the input data for our method. Step 3) displays the initialization of the weights, with $K = 5$ weight states. This is illustrated for both $\mathbb{W}$ set to a high value of $F_{st}^{\star}$ (left histogram) resulting in highly skewed weight states, as well as for a low value of $F_{st}^{\star}$ (right histogram) resulting in equally spaced weights. Step 4) shows the process of populating the initial weight matrix W^1. Lastly, steps 5-7 illustrates performing marginal updates to the weight matrix, through replacing the existing weight on the focal haplotype, with a potentially new weight sampled uniformly from $\mathbb{W}$. Using the newly updated weight matrix, we calculate the drifted population genetic measurements (LD and allele frequencies). Repeatedly calculating these measurement across the array of updated weight matrices leads us to generate a distribution of the population genetic measurements of interest.

Algorithm 1: Marginally Updating W

Result: Array of dimension: M_{ref} by N by Samples + 1
begin
  Sample a single weight from $\mathcal{U}(\mathbb{W})$ to populate W^1
  $l = 1$
  for $l \leq$ Samples **do**
    focalhap $= 1$
    for focalhap $\leq M_{ref}$ **do**
      Sample a single weight from $\mathcal{U}(\mathbb{W})$ to obtain $w^{'}$
      Replace $W^l_{focalhap,}$ with $w^{'}$
      focalhap = focalhap +1
      Store W^l in results array
      $l = l + 1$
    end
  end
end

We now demonstrate that the approach of marginally updating W using discrete

weights in $\mathbb{W}$ is equivalent to repeatedly updating W under the Bayesian bootstrap with drift. Specifically, we demonstrate that we obtain the theoretical distribution of the drifted allele frequencies under the Balding-Nichols model, conditional on the reference panel frequency and the level of F_{st}[6]. These simulations are useful to ascertain the resolution needed (ie. the value of K) for the discrete weight states to match the theoretical models for genetic drift.

To assess the accuracy of these distributions, we use a reference population R from the 1000G EUR population. We extract a small 0.1Mb region in Chromosome 22 with approximately 100 SNPs. We initialise the first sample for W by populating the weight matrix with a single weight drawn uniformly from $\mathbb{W}$. This initialization sets the allele frequencies at the start of the marginal updates equal to the reference population allele frequencies. Next, we perform the marginal updates to W (Algorithm 1) and at each sample we obtain a newly updated weight matrix. Using the updated weight matrix and R, we calculate the drifted allele frequencies across all SNPs for each sample. We perform 10 full updates through the reference population, corresponding to 8080 marginal updates. After we finish performing all successive updates to the weight matrix, we generate the distribution of the drifted allele frequencies for a single SNP in the population across all the samples. The resulting distribution of the drifted allele frequency is compared to the theoretical density using the Balding-Nichols model across a range of F_{st} values used to simulate the drifted allele frequencies and across and a range of discrete weight states K (Figure 3.9).

Overall, the empirical densities of the drifted allele frequencies we obtain from marginally updating W appropriately match the theoretical densities of drifted allele frequencies given by the Balding-Nichols model under a reference allele frequency of 19% and different levels of genetic drift. In particular, the empirical densities that we obtain from sampling W result in a mean of approximately 19% in each simulation,

[6]Because we are simulating directly under the generative model, we refer to F_{st} rather than $F^{\star}_{st}$.

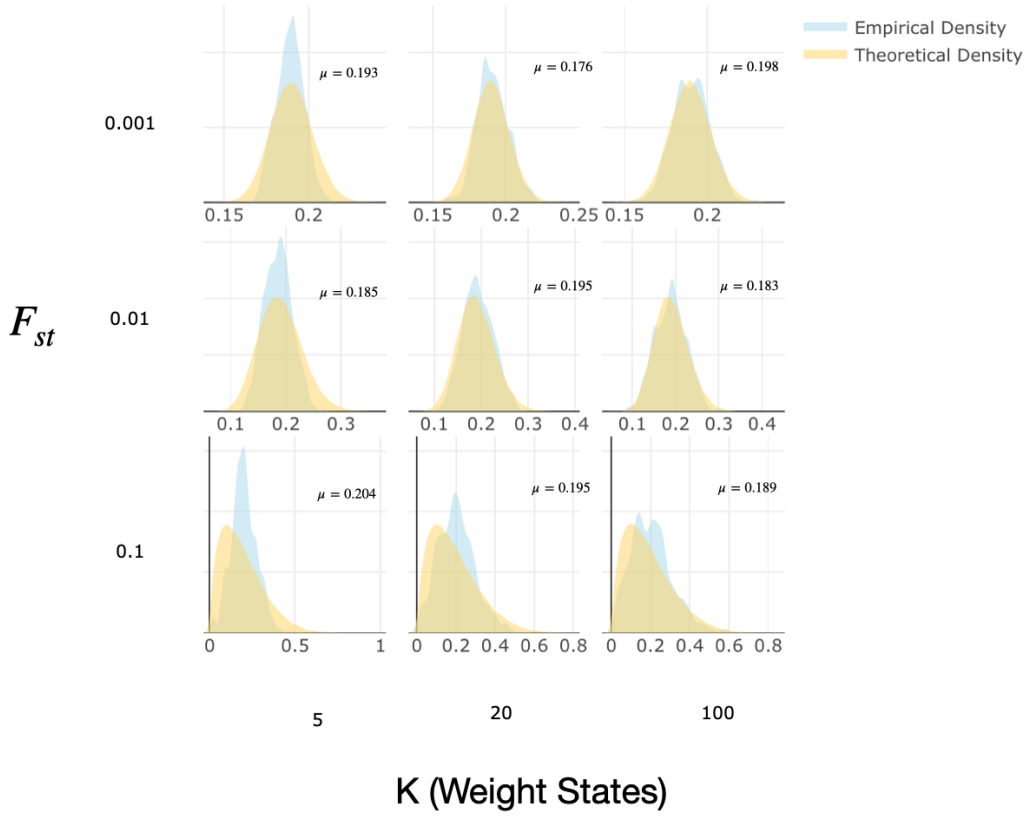

Figure 3.9: Theoretical density of drifted allele frequencies under Balding-Nichols genetic drift (yellow) compared to the empirical density obtained through marginally updating W (blue). Each row of subplots shows the F_{st} value set in the gamma distributed weights as well as within the Balding-Nichols theoretical distribution. Each column of subplots shows the number of weight states K used to generate the discrete gamma distributed weights. The same SNP at 19% reference allele frequency was selected in each simulation. The value μ above each density plot corresponds to the empirical mean of the drifted allele frequencies.

consistent with the theoretical expectation (Equation 1.12). We also observe that we achieve a fairly similar shape of the distribution as compared to the theoretical models, though note that the distribution of the allele frequencies obtained from the updates to W appear more normally distributed rather than beta distributed, especially under higher levels of F_{st}. Importantly, we observe that as we increase the number of discrete weight states K, the distribution converges onto the theoretical distribution. Interestingly, the manner in which the distribution converges, is

through generating an increased number of highly drifted allele frequencies that correspond to the tails of the Balding-Nichols distribution.

One reason for this result is due to the highly correlated sets of weights in W between each sample update. In other words, when we perform an update to W and generate a drifted allele frequency, it is challenging for our model at the next update to generate a drifted allele frequency that is significantly different than the previous allele frequency due to the manner in which we perform the marginal updates. Further, as we initialize the weights such that the first sample we draw will result in a drifted allele frequency equal to the reference allele frequency, it is computationally expensive to capture observations at the tails of the Balding-Nichols distribution. Overall, repeatedly marginally updating W in order to generate realistic distributions of drifted allele frequencies appears to be a challenging problem under high levels of F_{st}.

3.4 Conclusion

In this chapter, we described a method to generate the distribution of population genetic measurements in a genetically drifted population. These distributions have the potential to be highly useful for researchers in the field as they offer the ability to provide robustness to traditional reference panel point estimates. We noted that a significant limitation of this method relies on the availability of some pieces of individual level data to obtain measures of genetic drift between the reference and GWAS population. In the following chapter, we develop a method to infer the genetic drift from a limited subset of the GWAS summary statistics in a MLE framework.

Using simulated data, the model was well calibrated under the demographic history of a population split. However, the calibration of our model dramatically decreased when genetic data was simulated under the OOA model. A promising avenue for future investigation involves using simulated genetic data to explore

the calibration of our model under other demographic events such as bottlenecks, population expansions and migration. For these events it is unlikely that the parametrisation of the gamma distributed will be suitable. However, there are other relatively simple theoretical equations which describe distribution of allele frequencies in bottlenecks (Watterson 1984), migration (Barton and M. Slatkin 1986) and other demographic events which could be implemented within our model. Lastly, we applied our model to real data from the UK Biobank and the 1000G populations. Here, a wide array of ascertainment biases may have resulted in a poor calibration of the resulting distributions. In particular, we note that the allele frequency filter used to generate these distributions has a potentially large effect on the calibrations. Better understanding the sources and effects of these ascertainment biases in different GWAS and reference populations is another important direction of future research to improve our model.

In addition, we described how we can marginally update weights in W using discrete weights. We demonstrated that we can largely recover the theoretical distribution of allele frequencies under genetic drift using this method. However to obtain an accurate distribution of the drifted allele frequencies, W requires a large number of updates and $\mathbb{W}$ must contain a large enough number of weight states. These requirements are particularly true when needing to generate drifted SNPs with very rare MAF in the tails of the theoretical distribution.

We briefly suggest three potential improvements to our model. Firstly, we could generate the discrete weight state space $\mathbb{W}$ using Gaussian quadrature (Dillon and Langmore 2018) rather than breaking the gamma distribution into quantiles. The use of Gaussian quadrature is a more computationally expensive process, but will better approximate the environment of highly skewed weights under larger values of F_{st}. Secondly, we could model the effects of recombination. When we assessed the calibration of our model, we simulated genetic data with no recombination. We do so because recombination changes the genealogical history of the GWAS

samples across the genome. Therefore, the reference ancestor haplotypes will exchange *ancestral segments* to form the GWAS population. In Chapter 5 we describe how allowing the weights to evolve along the genome on a per-SNP basis mimics properties of recombination within the genealogical interpretation of our model. Thirdly, and most importantly, we could incorporate data and a likelihood into the marginal updates of W. This would lead us to a Gibbs sampling approach in order to carry out inference. Incorporating other sources of data, such as the GWAS summary statistics into the haplotype weighting scheme should improve the ability of our model to generate these distributions of the drifted population genetic measurements in a more accurate manner.

Overall, the framework of the haplotype weighting scheme, along with the ability to marginally update W, are powerful tools in our arsenal that will lead to methods that can infer properties of a GWAS population while only using existing reference population data.

4

Inferring Genetic Drift from GWAS Summary Statistics

Contents

In the previous chapter, we introduced a method which generated the distribution of the measurements of LD in a GWAS population that was genetically drifted from a reference population. This method required accurately setting the level of drift between the GWAS and reference population. In settings where the individual level GWAS data is provided, estimating drift is relatively straightforward. However, if the GWAS individual level data is not shared due to computational storage or genetic privacy concerns, researchers must instead rely on the GWAS summary

statistics to estimate drift. The use of the GWAS summary statistics for estimating drift is challenging for two reasons. Firstly, existing estimators for genetic drift require allele frequency data rather than the typical GWAS summary statistics (effect size and standard error) and which are rarely shared. Secondly, the GWAS summary statistics also contain noise due to sequencing errors, covariates and other confounding factors in the study. As a result, it is necessary to deconvolute the signals of noise from drift in the GWAS summary statistics. With these motivating reasons in mind, in this chapter we develop a method which jointly infers the levels of genetic drift relative to a reference population and the noise in a GWAS population using only the GWAS summary statistics.

We begin by modelling the effect of drift and noise on GWAS summary statistics using data generated from a simulated GWAS study along with empirical GWAS data from the UK Biobank. Next, by utilizing these models of drift and noise, we develop a maximum likelihood estimator (MLE) which infers the level of genetic drift between the reference and GWAS population. Importantly, the input for the MLE includes only the standard GWAS summary statistics and the allele frequencies obtained from a reference population. Further, we describe how the MLE can be extended to jointly infer the drift as well as the level of noise present in the GWAS population. We then assess the accuracy of the MLE using simulated coalescent data, first under the demographic history of a population split and second under populations formed through the Out of Africa (OOA) model. Lastly, we examine the performance of the method using real data from a UK Biobank GWAS population and the 1000G reference populations. We conclude that the method could prove valuable within the haplotype weighting scheme to ascertain the levels of genetic drift and to choose the most appropriate reference population to pair with the GWAS summary statistics.

4.1 Allele Frequencies and GWAS Variances

Summary statistics from a GWAS aim to contain pertinent information for researchers to perform additional downstream analyses and reproduce results from a study. However, the need to release data is balanced against both the responsibility to protect the genetic privacy of the individuals in the study and the computational constraints of sharing and storing large quantities of genetic data (Pasaniuc and Price 2017).

Due to conflicting institutional standards and an increasingly high bar necessary to protect the genetic privacy of the individuals within the study, a single universally agreed upon definition for a set of summary statistics does not exist. Work is underway to create a unified set of standards and protocols but progress has been slowed due to a variety of systematic and institutional factors (Lyon et al. 2021). Therefore, throughout this chapter, we assume that the summary statistics only contain the GWAS effect sizes and their standard errors across the variants in the study analyzed in a SNP by SNP manner (ie. not a multivariable model). Several large scale consortium studies release summary statistics in this limited manner. Examples of this limited release of summary statistics include the GIANT study investigating the genetic factors influencing height (Allen et al. 2010) and the DIAGRAM study investigating the genetic factors which cause Type 2 Diabetes (Morris et al. 2012).

Although the allele frequencies from a GWAS population may not be directly included in the summary statistics, they are implicitly encoded within the GWAS standard error of the effect size. To obtain the effect size and standard error of the effect size, GWAS studies typically rely on the additive model which assumes that there is a uniform, linear increase in risk for each copy of the disease associated allele in an individual (Bush and Moore 2012). In order to determine the underlying effect of the alleles, a logistic model is fitted to the GWAS data. In this model,

we assume that there is a binary case or control phenotype (1 or 0) and we use the genotypes x (0,1 or 2) on the samples in the GWAS as the predictors. We define the logistic model as:

$$\log(\frac{Pr(Y=1|X=x)}{Pr(Y=0|X=x)}) = \mu + \beta x. \tag{4.1}$$

The parameters that we estimate are μ, which represents the mean risk of genotype 0, and β, the effect (on the log odds ratio) of each copy of the disease causing allele on the mean phenotype. μ is often referred to as the logarithm of odds for genotype 0 and β is the log of odds ratio between genotype 1 and 0. Similarly 2β is the log odds ratio between genotypes 2 and 0.

Under this model, the relationship between the variance of the effect size estimate and the GWAS allele frequency for a single SNP j in the GWAS study without other covariates is:

$$Var(\hat{\beta}) = \frac{1}{Var(x)\frac{M_{GWAS}}{2}\phi(1-\phi)} \approx \frac{1}{\pi_j^G(1-\pi_j^G)M_{GWAS}\phi(1-\phi)}, \tag{4.2}$$

where ϕ denotes the proportion of cases in the GWAS, π_j^G denotes the allele frequency at SNP j and M_{GWAS} denotes the number of haplotypes in the GWAS study across both cases and controls (Vukcevic et al. 2011).

We assess the accuracy of Equation 4.2 using simulated GWAS data under the null model of no association. We construct a GWAS population of 10,000 haplotypes and assign randomly 50% of these haplotypes as cases and the remaining samples as controls. Each sample contains a single SNP with allele frequency drawn from $\mathcal{U}(0.02, 0.98)$. For each haplotype in the GWAS, we randomly assign a reference or alternate allele status to the SNP. We also pair haplotypes to form genotypes which result in 0, 1 or 2 copies of the alternate allele per genotype. We fitted the logistic model (Equation 4.1) to obtain the GWAS standard errors of the effect size which are squared to obtain the variances. This simulation is repeated

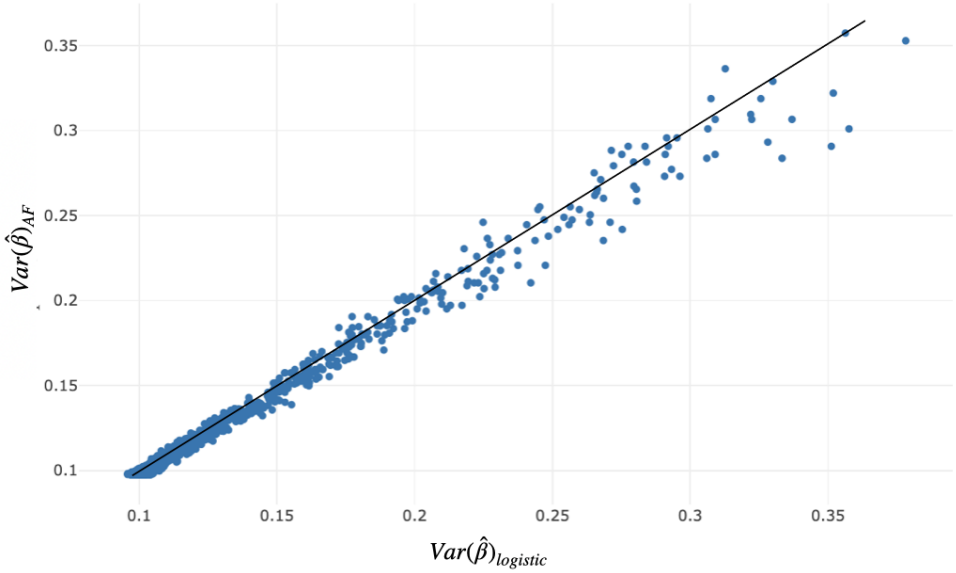

Figure 4.1: Relationship between the fitted and implied GWAS variances across 1000 replicate simulated GWAS experiments. Implied variances $Var(\hat{\beta})_{AF}$ are calculated using Equation 4.2. Fitted variances $Var(\hat{\beta})_{logistic}$ are obtained from the fit of the simulated GWAS data. Each point corresponds to a single GWAS simulation with a single SNP. The black line denotes the line y=x illustrating when the fitted and implied GWAS variances are equal.

across 1000 replicate GWAS populations, each time selecting a new allele frequency and generating a new set of genotypes across the GWAS samples. Figure 4.1 describes the relationship between the implied GWAS variances given by Equation 4.2 ($Var(\hat{\beta})_{AF}$) and the fitted variances from the logistic model ($Var(\hat{\beta})_{logistic}$) across replicate GWAS simulations.

The correlation in Figure 4.1 between the fitted GWAS variances and those implied by Equation 4.2 is extremely strong ($r^2 = 0.94$). However, as higher GWAS variances are generated, we observe that the correlation slightly decreases. These points correspond to SNPs with lower allele frequencies, thus implying that Equation 4.2 is not as accurate for lower frequency GWAS SNPs as opposed to common SNPs.

We further assessed the relationship between the GWAS allele frequencies and variances using real world GWAS data. We rely on the genetic and phenotype data provided by the UK Biobank (Bycroft et al. 2018). We select the phenotype primary

hypertension which has an ICD-10 code I10. Using the same quality control protocols as in the previous chapter (See `https://github.com/Nealelab/UK_Biobank_GWAS` for full details on quality control procedure), we filter imputed variants with an INFO Score < 0.8 and all variants with a MAF less than 0.1 percent. After quality control, 361,463 samples and roughly 13.7 million variants remain within the genotype-phenotype dataset. We generate GWAS summary statistics using BOLT-LMM (Loh et al. 2015). We classify 80,037 samples with a I10 code as cases and run BOLT-LMM with default settings.

We compare the GWAS variances fitted by BOLT-LMM[1] ($Var(\hat{\beta})_{BOLT}$) to the GWAS variances implied by Equation 4.2 using the allele frequencies from the post quality control UK Biobank GWAS population (Figure 4.2).

We demonstrate with the UK Biobank GWAS data that the GWAS variances fitted by BOLT consistently overestimate the implied variances from Equation 4.2 (Figure 4.2). Interestingly, this trend appears for the imputed variants but not for the genotyped variants.

This result implies that during the process of imputation, information is lost and generates missingness within the data. We test this hypothesis by examining the INFO-Score of the imputed data at each SNP. We compare the INFO-Score against the absolute value of the of the difference between the implied and fitted variances (Figure 4.3 a)) and show that for INFO-Scores less than 0.9, there is an increase in the inaccuracy of the implied variances as the INFO-Score decreases. This suggests that variants which are imputed with less confidence result in a more substantial differences in the variance relative to the variances implied by Equation 4.2. However, this pattern is also observed for SNPs with an INFO-Score greater than 0.9. This second trend is unexpected, and we speculate it may be caused by a combination of BOLT-LMM treating the imputed data in an abnormal fashion

[1]BOLT-LMM standard errors requires a transformation by dividing through $\phi(1-\phi)$ to convert from the quantitative scale to the binary scale when analyzing case-control traits as in the case of hypertension.

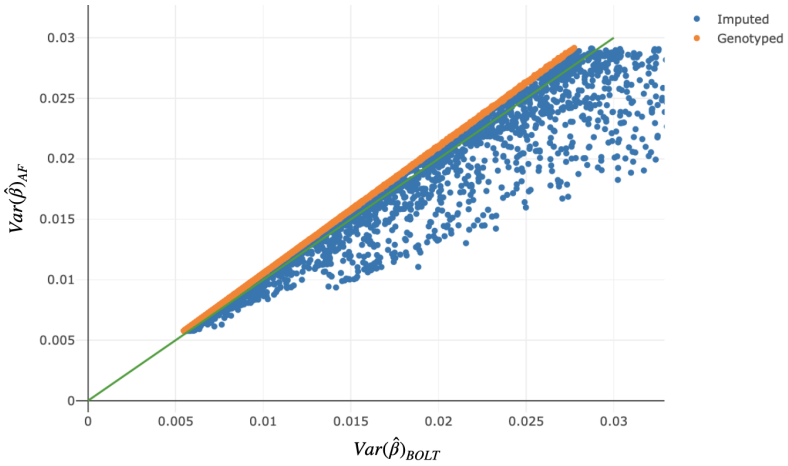

Figure 4.2: Relationship between the fitted and implied variances for the GWAS summary statistics on Chromosome 1 from a UK Biobank GWAS for hypertension. Implied variances $Var(\hat{\beta})_{AF}$ are calculated using Equation 4.2. Fitted variances $Var(\hat{\beta})_{BOLT}$ are obtained from BOLT-LMM. Each point corresponds to a variant in the GWAS summary statistics. Points are coloured as to whether they were genotyped (orange), or imputed (blue). The green line denotes the line y=x which illustrates when the fitted and implied GWAS variances are equal.

or an imputation quality control issue within the UK Biobank data. Examining this discrepancy in more detail (Figure 4.3 b) we observe that the implied variances using Equation 4.2 appear to reach a minimum variance at an INFO-Score of 0.9, whereas the variances fitted by BOLT do not reach a minimum and instead continue decreasing past an INFO-Score of 0.9. This leads to the absolute difference between the two variances increasing at this INFO-Score.

In comparison to the imputed data, the relationship between the implied GWAS variances and those fit by BOLT is much stronger for the genotyped data. Interestingly, we observe that the fitted variances from BOLT-LMM now underestimate their corresponding value using Equation 4.2. This underestimation could be a combination of a variety of factors including internal quality control performed by BOLT, which may remove samples at different SNPs based on internal metrics. Due to a much higher degree of correlation for the genotyped data as compared to the imputed data, within this chapter we will focus on applying the methods to estimate genetic drift using only genotyped GWAS data.

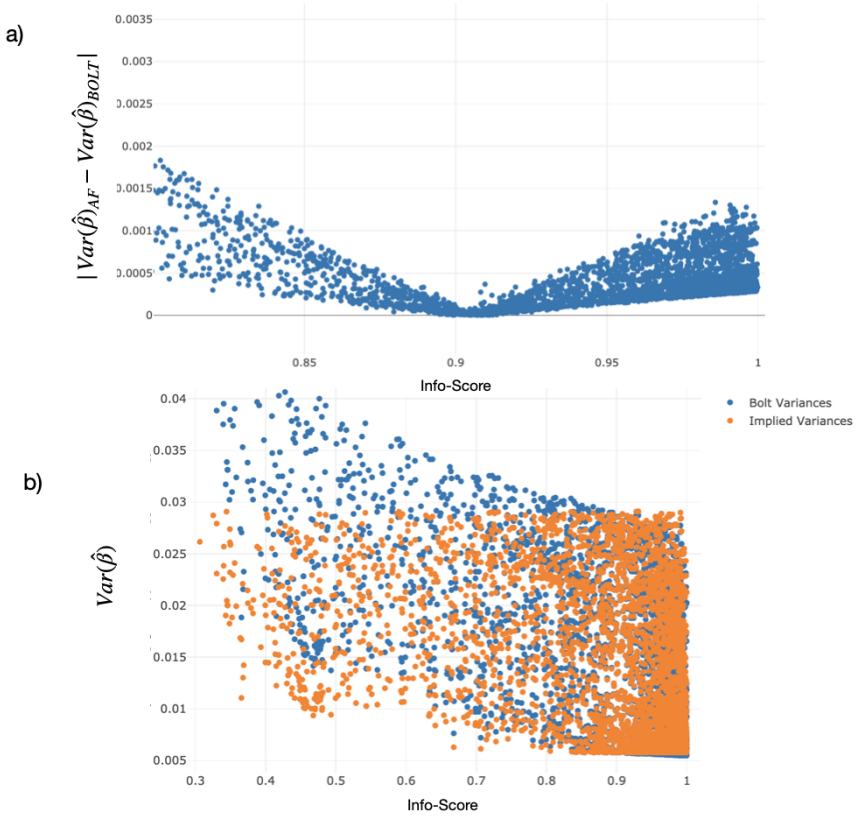

Figure 4.3: a)Relationship between the imputation INFO-Score and the absolute difference between the implied and fitted variances for a UK Biobank GWAS for primary hypertension. Each point corresponds to an imputed variant in the GWAS summary statistics on Chromosome 1. b) The relationship between the variance fitted by BOLT and the implied variances against the imputation INFO-Score. Note the different axes scales between a) and b).

4.1.1 Impact of Noise and Drift

In real-world settings, the relationship between the GWAS allele frequencies and the GWAS variances given by Equation 4.2 is not exact as sources of noise[2] are incorporated within the observed summary statistics in a GWAS study.

For convenience, we model noise as a random variable which acts as a multiplicative factor on top of the implied GWAS variances. Therefore, by examining the ratio of the fitted GWAS variances to the implied variances across each SNP, we can model the shape of the noise distribution.

As previously, in order to investigate the distribution of the ratios, we employ

[2]Although we refer to *noise*, we refer to the departure from the theoretical relationship captured in Equation 4.2 noting that this noise can have various causes.

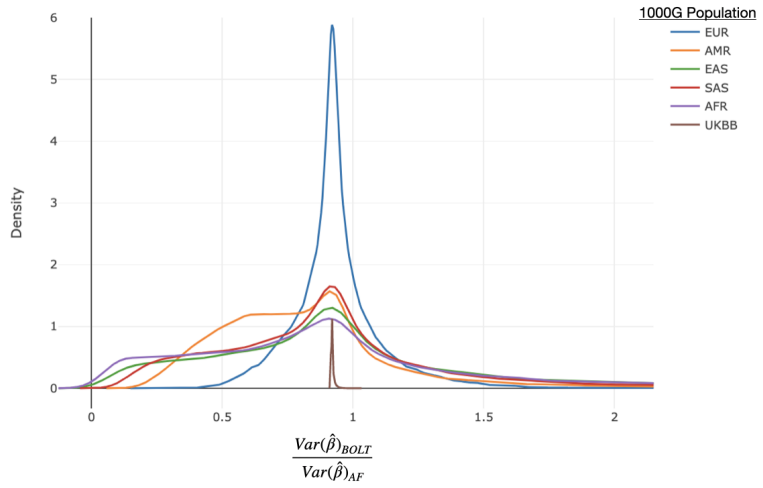

Figure 4.4: The distribution of the ratio of the fitted over the implied GWAS variances of genotyped variants on Chromosome 1 for a UK Biobank GWAS for primary hypertension. Each line corresponds to the choice of the population used to calculate the implied GWAS variances in Equation 4.2. The allele frequencies used in this transformation were calculated from the 1000G super-populations (EUR, AMR, EAS, SAS, AFR) or from the UK Biobank samples (UKBB).

data from the hypertension UK Biobank GWAS. Further, we examine the impact of substituting allele frequencies obtained from other reference populations rather than the GWAS population within Equation 4.2 to calculate the implied variances. The distribution of the ratios shifts depending on how closely the allele frequencies in the different populations match the in-sample UK Biobank allele frequencies.

We calculate the ratios using the allele frequencies obtained from the 1000G super-populations or from the UK Biobank GWAS samples. We filter out GWAS variances associated with variants that fall below 1% MAF in any of the 1000G super-populations. Figure 4.4 illustrates the distribution of the ratio of the fitted to the implied GWAS variances across different populations.

The mean of the distribution across each population is slightly below one. This result is in line with the results displayed in Figure 4.2 which suggests that we underestimate the fitted GWAS variances using Equation 4.2.

It is extremely challenging to disentangle the effect of noise from the effect

of using different population allele frequencies in Equation 4.2 on the shape of these distributions. In this analysis, we compare the same set of variants across populations and assume that the UK Biobank data-set is a highly quality controlled study (Bycroft et al. 2018). As a result, we expect the noise to be relatively low and homogeneous between populations. Therefore, we assume that differences between distributions are attributed solely to the drift in the population allele frequencies.

The distribution of ratios using the allele frequencies that are obtained from the UKBB possess a very narrow point mass[3] at a value slightly less than one. This suggests that the allele frequencies in the UK Biobank are well suited in approximating the GWAS variances using Equation 4.2. The EUR 1000G population distribution possesses the highest peak and the smallest variance amongst all other 1000G reference populations. Because the UK Biobank GWAS contains only samples with WBA, we expect to find a strong ancestral match to the samples within the 1000G EUR super-population. Consequently, we expect the allele frequencies between the EUR and UKBB population to match as well (Novembre et al. 2008). We observe that the concentration of ratios centered at one decreases across the AMR, SAS, EAS and AFR populations respectively. This ordering suggests that these populations contain allele frequencies that are increasingly drifted from the UK Biobank population, resulting in higher variances of these distributions. Lastly, we note that the 1000G AMR population distribution contains a unique leftward skew. This could be attributed to admixture across the AMR samples which may result in the allele frequencies matching the in-sample allele frequencies at some, but not all of the variants in the GWAS population.

[3]Due to the default binning behaviour of the density calculation, the peak for this distribution is underestimated for this population.

4.2 MLE of Drift and Noise

In the previous section, we described how substituting GWAS allele frequencies using a reference population that has drifted relative to the GWAS population alters the implied GWAS variances in Equation 4.2. In this section, we estimate the levels of drift given the GWAS variances and allele frequencies from a drifted reference population. In order to do so, we employ a maximum likelihood estimate (MLE). A MLE estimates the parameters of a probability distribution by maximizing a likelihood function such that the observed data is most probable under the statistical model.

4.2.1 Marginal Inference of $\hat{F_{st}}$

We utilize the Balding-Nichols model (Balding and Nichols 1995) to describe the distribution of allele frequencies in the GWAS population which has drifted relative to the reference population. Given the fitted GWAS variances $Var(\hat{\beta})$ and a matched set of allele frequencies from a reference population π^R, we derive the likelihood of the level of genetic drift given by F_{st} that separates these populations as:

$$\mathcal{L}(F_{st}|Var(\hat{\beta}), \pi^R). \tag{4.3}$$

Let X be the random variable which describes the GWAS allele frequency of a SNP which has drifted relative to a reference allele frequency. X is beta distributed and has a density function f_X parametrised by a shape and rate parameter $\alpha = \frac{1-F_{st}}{F_{st}}\pi^R$ and $\beta = \frac{1-F_{st}}{F_{st}}(1-\pi^R)$ where:

$$f_X(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{Beta(\alpha,\beta)}. \tag{4.4}$$

We perform a change of variables to obtain the probability density function for the GWAS variances in Equation 4.2, while ignoring the various constants for the purposes of readability as:

$$Y = \frac{1}{X(1-X)}. \tag{4.5}$$

The change of variables is performed for the range $0 < x < 1$ because the generation of negative GWAS variances is impossible. The function in Equation 4.5 is two-to-one in this range. Therefore, we perform the change of variables across the two one-to-one regions (X_1 and X_2) separately. When $0 < x < \frac{1}{2}$, the inverse function of Equation 4.5 is:

$$X_1 = \frac{1}{2} - \frac{(Y-4)^{\frac{1}{2}}}{2Y^{\frac{1}{2}}}. \tag{4.6}$$

Similarly, when $\frac{1}{2} < x < 1$ the inverse function of Equation 4.5 is:

$$X_2 = \frac{1}{2} + \frac{(Y-4)^{\frac{1}{2}}}{2Y^{\frac{1}{2}}}. \tag{4.7}$$

The transformed range of X_1 and X_2 is $y > 4$. Over the two-to-one portion of the transformation ($0 < x < 1$), the probability density function for Y is written as:

$$f_Y(y) = f_X(\frac{1}{2}+\frac{(y-4)^{\frac{1}{2}}}{2y^{\frac{1}{2}}})|(y^{\frac{3}{2}}(y+4)^{\frac{1}{2}})^{-1}|+f_X(\frac{1}{2}-\frac{(y-4)^{\frac{1}{2}}}{2y^{\frac{1}{2}}})|(y^{\frac{3}{2}}(y+4)^{\frac{1}{2}})^{-1}|, \tag{4.8}$$

where f_X is the beta probability density function in Equation 4.4. In summary, we can fully express the density function as:

$$f_Y(y;\alpha,\beta) = \frac{\frac{1}{2} + \frac{(y-4)^{\frac{1}{2}}}{2y^{\frac{1}{2}}}^{\alpha-1}(\frac{1}{2} + \frac{(y-4)^{\frac{1}{2}}}{2y^{\frac{1}{2}}})^{\beta-1}y^{\frac{3}{2}}(y+4)^{\frac{1}{2}})^{-1}}{Beta(\alpha,\beta)} - \frac{\frac{1}{2} - \frac{(y-4)^{\frac{1}{2}}}{2y^{\frac{1}{2}}}^{\alpha-1}(\frac{1}{2} - \frac{(y-4)^{\frac{1}{2}}}{2y^{\frac{1}{2}}})^{\beta-1}y^{\frac{3}{2}}(y+4)^{\frac{1}{2}})^{-1}}{Beta(\alpha,\beta)}. \tag{4.9}$$

The density function in Equation 4.9 describes the probability of observing the GWAS variance of a SNP under a level of drift relative to a reference allele frequency. Figure 4.5 a) illustrates an example of the density function. Using the

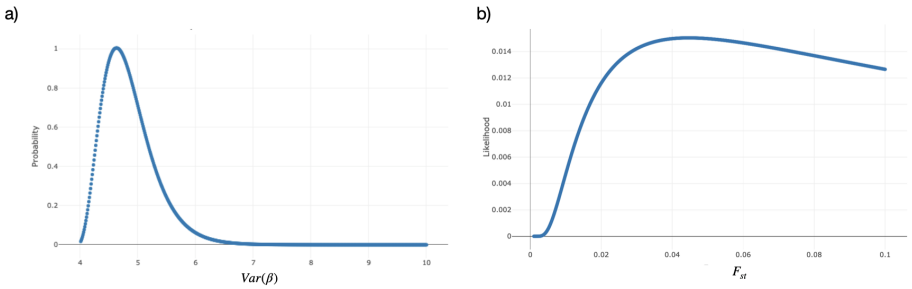

Figure 4.5: a) Probability of observing a GWAS variance given a F_{st} of 0.01 and π^R at 20%. b) Likelihood of F_{st} for an observed GWAS variance of 5.63 and π^R of 20%.

probability density function in Equation 4.9, we can obtain the likelihood for the level of drift F_{st} between the GWAS and reference population given the observed GWAS variance and reference allele frequency at a single SNP (Figure 4.5 b).

Under the assumption that the observed GWAS variances and reference allele frequencies across SNPs are independent, we obtain the log-likelihood across a set of N SNPs as the sum across the marginal log-likelihoods for each individual SNP as:

$$\ell(F_{st}|Var(\hat{\beta}), \pi^R) = \sum_{j=1}^{N} \log(f_Y(Var(\hat{\beta})_j|\pi_j^R, F_{st})). \tag{4.10}$$

.

In practice, these observations are not independent as a result of LD between SNPs. However, the effect of LD is typically ignored by estimators for genetic drift (Bhatia et al. 2013). Using standard MLE algorithms, we infer the level of genetic drift $\hat{F}_{st}$ to be that which maximises the function in Equation 4.10.

We assess the accuracy of the MLE by first using simulated GWAS data under a generative model. We begin by sampling 1000 reference allele frequencies π^R from $\mathcal{U}(0.02, 0.98)$. Next, we use the Balding-Nichols model to generate a set of drifted allele frequencies under a given level of F_{st}. Afterwards, we obtain the GWAS variances using Equation 4.2 while again ignoring all constants. We remove any SNP with a drifted minor allele frequency of less than 1% to mimic normal quality control processes in a GWAS analysis. We repeat this process 1000 times

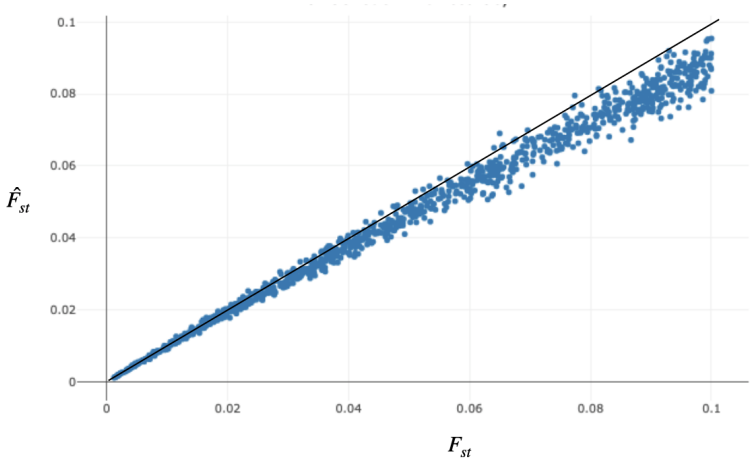

Figure 4.6: MLE accuracy across 1000 replicate simulations, each with 1000 SNPs generated under the Balding-Nichols model with a given level of F_{st}. The black line denotes the line y = x indicating a perfect estimate.

for different values of F_{st} drawn from $\mathcal{U}(0.001, 0.1)$ and for different sets of reference allele frequencies. The accuracy of the MLE is illustrated in Figure 4.6.

Overall, the MLE accurately estimates the value of F_{st} used to simulate the GWAS data. However, when the GWAS variances are simulated under very high levels of drift ($F_{st} > 0.1$), the accuracy of the MLE decreases resulting in the underestimation of the true value. We suspect that the reason for this underestimation is a consequence of the filtering step used in the data simulation. The removal of very rare and drifted SNPs leads to estimating lower levels of drift than expected.

4.2.2 Joint Inference of F_{st} and Noise

Previously, we described how various sources of noise may alter the observed GWAS variances. As a result, Equation 4.2 does not lead to an exact transformation between the GWAS allele frequencies and the observed GWAS variances. We now extend the MLE to jointly estimate both the F_{st} between the reference and GWAS population, as well as the level of noise present within the GWAS population.

We define Z as the random variable which describes the observed and noisy GWAS variances $Var(\hat{\beta}_{noise})$. As before, Y is the random variable which describes

the GWAS variances without noise. V is the random variable which describes the noise in the GWAS. A natural starting point for modelling the noise incorporated within the GWAS variances is through a gamma distribution (Fortune and Wallace 2019). We again assume that noise acts in a multiplicative fashion on top of the GWAS variances such that:

$$Z = YV. \tag{4.11}$$

Setting the shape and rate parameter in the gamma distribution to κ results in the gamma distribution possessing a mean of one (Equation 4.12). Therefore, the magnitude of noise in the GWAS is drawn from the following density:

$$f_V(v;\kappa) = \frac{\kappa^\kappa}{\Gamma(\kappa)} v^{\kappa-1} e^{-\kappa v}, \tag{4.12}$$

where $v > 0$ and $k > 0$.

We derive the density of Z by introducing a new random variable W defined as $W = V$. This substitution is a one-to-one transformation and we write V and W in terms of Z by setting $Y = \frac{Z}{W}$. The joint density of ZW is:

$$f_{ZW}(z,w) = \frac{f_{YV}(y,v)}{|\partial(z,w)/\partial(y,v)|_{y=\frac{z}{w},w=v}} \tag{4.13}$$

where the Jacobian is:

$$\frac{\partial(z,w)}{\partial(y,v)} = \begin{vmatrix} \partial z/\partial y & \partial z/\partial v \\ \partial w/\partial y & \partial w/\partial v \end{vmatrix} = \begin{vmatrix} v & y \\ 0 & 1 \end{vmatrix} = v. \tag{4.14}$$

Together the joint density is:

$$f_{ZW}(z,w) = \frac{f_Y(z/w) f_V(w)}{w}. \tag{4.15}$$

The range of $f_{ZW}(z,w)$ is $0 \leq z \leq \frac{Var(\hat{\beta})}{4}$ and $w \geq 0$. We obtain the marginal density for Z by integrating out the nuisance parameter w across the range $[0, \frac{Var(\hat{\beta})}{4}]$ as:

$$f_Z(z) = \int_0^{\frac{Var(\hat{\beta})}{4}} \frac{1}{w} f_Y(z/w) f_V(w) dw. \tag{4.16}$$

Finally, we substitute the probability density function in Equation 4.9 for f_Y and the gamma probability density function in Equation 4.12 for f_V to obtain the marginal density f_Z. Because the integral in Equation 4.16 possesses a non-elementary anti-derivative, we use standard numerical integration techniques to evaluate the integral in Equation 4.16 in the following sections.

Using the density f_Z, we evaluate the log-likelihood of the F_{st} and noise given a set of noisy GWAS variances and reference allele frequencies as the sum across the individual and independent likelihoods at each SNP as:

$$\ell(F_{st}, \kappa | Var(\hat{\beta}_{noise}), \pi^R) = \sum_{j=1}^{N} \log[f_Z(Var(\hat{\beta}_{noise})_j | \pi_j^R, F_{st}, \kappa)]. \tag{4.17}$$

We jointly estimate the drift parameter $\hat{F_{st}}$ and the gamma noise parameter $\hat{\kappa}$ by maximizing this joint likelihood surface.

We evaluate the accuracy of the MLE using simulated genetic data under the model of multiplicative gamma noise and genetic drift generated through the Balding-Nichols model. Given a value of F_{st} and a set of reference allele frequencies, a series of drifted allele frequencies and corresponding GWAS variances are generated as in the case of marginally estimating the F_{st}. In addition, we generate the noisy GWAS variances by sampling the gamma noise with a shape and rate parameter of κ and multiplying the noise on top of the GWAS variances. For each replicate simulation, we sample a value of F_{st} from $\mathcal{U}(0.001, 0.1)$ and additionally sample a noise value κ from $\mathcal{U}(0.01, 0.05)$. Again, we remove any SNPs in the GWAS population with a MAF less than 1%. The results of the joint inference of the drift and noise using the MLE across 1000 replicates, each with 1000 SNPs is depicted in Figure 4.7.

Overall, we achieve a reasonable degree of accuracy in estimating the F_{st} and the noise parameter κ using the MLE. Notably, the accuracy is weaker as compared

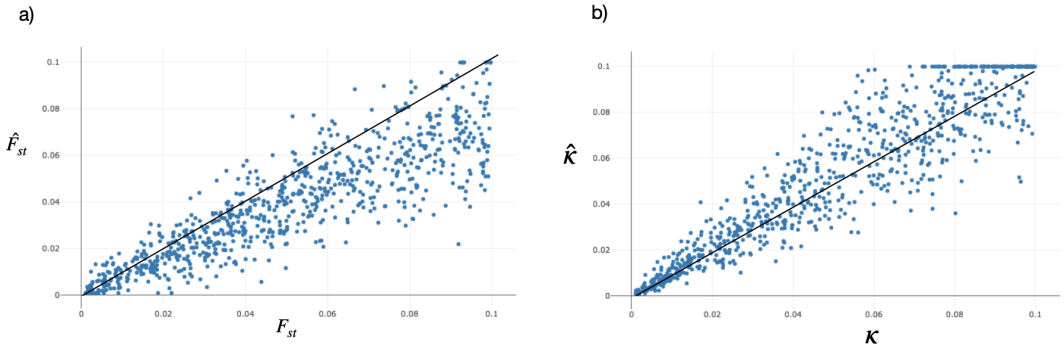

Figure 4.7: Accuracy of the joint MLE to estimate a)F_{st} and b) κ across 1000 replicate simulations. Each point corresponds to a replicate simulation with 1000 SNPs, which was simulated under a given level of noise κ and drift F_{st}. The black lines indicate the line y=x which illustrates when the estimated and true parameters are the same.

to the marginal inference of F_{st} (Figure 4.6). However, we expect the accuracy of the MLE to improve as the number of observations increase. The numerical integration to obtain the joint density f_Z is particularly slow, leading to a trade-off between increased speed or increased accuracy. As shown previously when marginally inferring the F_{st}, the joint inference typically performs less accurately when genetic data is simulated under high levels of F_{st} and results in underestimating the true level of genetic drift.

4.3 Application to Coalescent Data

4.3.1 Population Split

We now employ the joint MLE to infer the genetic drift and noise between a reference and a GWAS population simulated under the coalescent. To generate drift, we use *msprime* under the demographic history of a population split (Figure 3.3). In *msprime*, we set the effective population size of the ancestral population to 10,000 and the effective population size of the two descendant populations to 5,000. We form the GWAS and reference populations by sampling 1000 haplotypes from the respective descendant populations. Further, we set a human level mutation and recombination rate (1e-8) for a 1Mb sequence which generates approximately 1000

SNPs in a region. Using the allele frequencies obtained from the GWAS population, we generate the GWAS variances using Equation 4.2[4]. The noisy GWAS variances are generated by sampling a gamma distributed noise value for each SNP under a value of κ (Equation 4.12) and multiplying the noise by the GWAS variances obtained from Equation 4.2. We repeat this process 1000 times across replicate *msprime* population pairs. The estimates for both the noise and genetic drift are illustrated in Figure 4.8 across different population split divergence times and across different values of κ. We compare the values of F_{st} to the estimated values from *msprime* which employs a modified version of the Wakeley estimate and the ratio of coalescent times in the two populations (Eldon and Wakeley 2009). These estimates represent the individual-level data estimates for the F_{st}, and reflect just one method that researchers can use to directly estimate F_{st} from a pair of populations given the individual-level data. We note that a multitude of other F_{st} estimators are available (See Bhatia et al. 2013) including theoretical estimators for a population split (Nielsen et al. 1998). However, in this chapter we rely on the *msprime* estimator for consistency to apply and compare our method to the estimates for populations (eg. OOA and UKBB) which do not fit under these theoretical models.

Overall, as the divergence times within the population split increase in the past, the estimated values of drift increase as expected. However, we do not obtain the same values as estimated from *msprime*. As *msprime* uses a F_{st} calculation that is not directly related to the Balding-Nichols model of F_{st}, it is not surprising that we achieve different results. Because we aim to use this value of F_{st} to set the gamma distributed weights in the haplotype weighting scheme, estimating the exact same value of F_{st} as compared an estimator that uses individual level data is not of great concern. Rather, the ability of the estimate to discern between higher and lower levels of drift will prove more important downstream for applications in GWAS analyses in settings such as choosing the most appropriate reference population to pair with

[4]Again, we ignore the constants ϕ and M_{GWAS}.

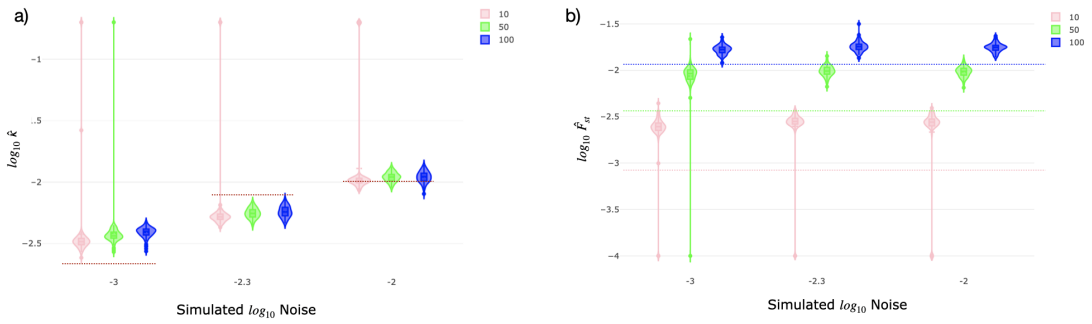

Figure 4.8: Density of the estimates using the joint MLE across 1000 replicates for a) noise and b) F_{st} using simulated coalescent data under the demographic history of a population split and using GWAS variances directly obtained from Equation 4.2. The colours represent the different divergence times within the population split and the x-axis describes the $\log_{10}$ noise κ that we layered on top of the GWAS variances. The y-axis denotes the $\log_{10}$ estimate obtained by the MLE. Each point corresponds to a replicate simulation with approximately 1000 SNPs. The red lines in a) denote the true simulated $\log_{10}$ noise in each set of simulations. The pink, green and blue lines in b) represent the mean value of $\log_{10}$ F_{st} estimated by *msprime* for each of the divergence times.

the GWAS summary statistics. Further, the estimates we obtain remain accurate for different values of noise that we incorporated on top of the GWAS variances, thereby implying that the MLE is robust to the level of noise within the GWAS.

Additionally, the MLE is reasonably accurate for inferring the noise within the GWAS population. As we increase the levels of noise that we simulate on top of the GWAS variances, we obtain increasingly higher estimates of noise. Again, for the purposes of this chapter, the accuracy of the MLE to infer the exact level of noise is not as significant as inferring the relative amount of noise between different settings as the noise term does not relate to a specific noise process such as the sequencing quality (ie. it is phenomenological rather than process driven).

In both cases of inferring the F_{st} and the noise, the MLE results in outlying estimates in some of the replicate regions. We suspect that these outliers are a result of the MLE algorithm hitting a boundary that we set for the drift and noise parameters leading to an incorrect estimation.

In order to generate more realistic noise, we randomly and evenly assign case or control status across the GWAS haplotypes. Next, we fit a logistic model to

the data (Equation 4.1) and obtain the GWAS standard error for each SNP which we square to obtain the GWAS variances. We employ the MLE as before, this time however accounting for ϕ and M_{GWAS} within Equation 4.2. The resulting estimates for F_{st} using Equation 4.17 are given in Figure 4.9.

Providing the MLE with the noisy GWAS variances obtained through fitting a logistic model as compared to directly generating the variances using Equation 4.2 results in very similar estimates. Again, as the time of the population split increases, we obtain higher levels of estimated F_{st} from the MLE. However, using the noisy GWAS variances from the logistic model as compared to directly generating the variances using Equation 4.2 results in a higher variation of the estimate for F_{st} across regions indicating that the noise from the model may lead to a higher degree of uncertainty within the method. Again, we note that the estimates are substantially different than the ones provided by *msprime* (ie. using the individual level data), although we see the same relative increase in F_{st} across the divergence times as previously in Figure 4.8.

4.3.2 Out of Africa

Next, we assess the accuracy of the MLE to obtain estimates of drift between pairs of reference and GWAS populations simulated under the Out of Africa (OOA) model (Figure 3.5). We use *msprime* to simulate this demographic history and generate all pairs of CHB, CEU and YRI populations as GWAS and reference populations[5]. We set both the recombination rate and mutation rate to 1e-8 and sample 1000 haplotypes from the present-day OOA populations. We generate GWAS variances by simulating summary statistics under the null model and fitting the data using a logistic model as in the previous section. The resulting estimates for the F_{st} are shown in Figure 4.10.

[5]Because the GWAS and reference label is defined without loss of generality, we include only the results of three of the pairings.

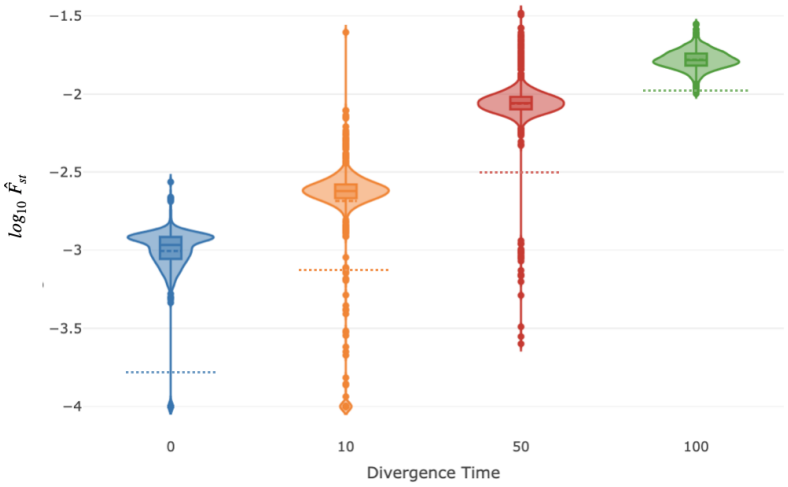

Figure 4.9: Density of the estimates of F_{st} under different values of divergence times using simulated coalescent data under the demographic history of a population split and GWAS variances obtained from a logistic model. The x-axis represents the different divergence times within the demographic event of a population split. The y-axis denotes the $\log_{10}$ estimates for F_{st} across the replicate population pairs. Each point corresponds to a replicate simulation with approximately 1000 SNPs. The dotted lines indicate the mean $\log_{10}$ estimate of F_{st} by *msprime*.

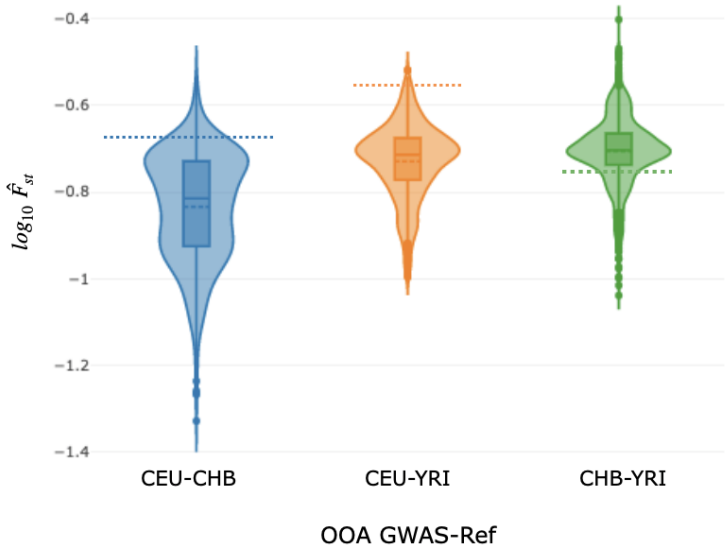

Figure 4.10: Density of the estimates from the MLE to infer F_{st} under different OOA population pairings across 1000 replicate regions using GWAS variances obtained from a logistic model. The colours represent the different population pairings. The y-axis denotes the log_{10} F_{st} estimate. Each replicate simulation contains approximately 1000 SNPs. The dotted lines correspond to mean $\log_{10}$ F_{st} estimate from *msprime*.

Overall, across each of the three OOA population pairings, we obtain higher estimates of genetic drift relative to the estimates we obtained earlier under a population split (Figure 4.9). Additionally, we observe there exists fairly wide variation in the estimates between replicate regions. In particular, the CEU-CHB population pairing displays the highest level of variation and also the lowest mean estimated F_{st}. We expect the levels of drift between these populations to be relatively high as has been previously documented using real world samples from these populations (Bhatia et al. 2013). Interestingly, the individual-level data estimates for F_{st} obtained from the *msprime* simulations are similar to estimates we obtain using the MLE, and more accurate than previously in the case of simulating data underneath a population split. However, we note that *msprime* considers the CEU-YRI population to contain higher levels of genetic drift than the CHB-YRI estimate whereas the MLE suggests that these estimates are roughly equal.

4.4 Application to UK Biobank

Finally, we assess the performance of the MLE to infer genetic drift and noise between a GWAS population in the UK Biobank and the 1000G super-populations as the reference populations.

As before, we employ the GWAS summary statistic for hypertension described earlier in this chapter. We obtain the summary statistics for Chromosome 1 and split the statistics into 1000 equally sized regions. Within each region we extract the same set of SNPs from each of the 1000G super-populations. Next, we jointly remove any SNPs with less than 1 % MAF in the UK Biobank GWAS population or in any of the 1000G reference populations. Therefore, we compare the same set of SNPs across each population pairing. Next, we run the MLE for each region using the observed GWAS variances obtained from BOLT-LMM and the allele frequencies from each reference population. These results are shown in Figure 4.11.

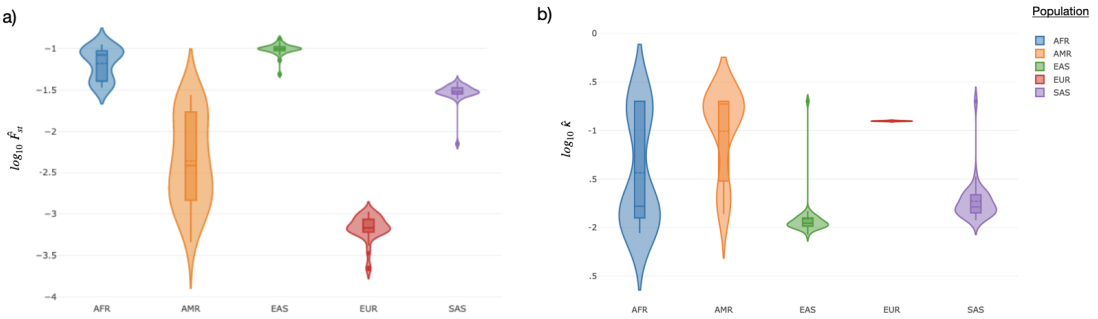

Figure 4.11: Estimating drift and noise using UK Biobank GWAS data and reference populations from the 1000G super-populations with the MLE. a) Density of estimates of F_{st} and b) κ from the joint MLE using summary statistics generated from a UK Biobank GWAS for hypertension across 1000 regions. The x-axis corresponds to using different 1000G super-populations to obtain the reference allele frequencies. The y-axis denotes the log_{10} estimate for F_{st} or κ across the regions.

Overall, the use of the 1000G EUR population as the reference panel results in estimating the lowest level of genetic drift relative to the GWAS summary statistics. As previously described, we expect the lowest level of drift between these two populations since the UK Biobank primarily consists of European samples. The AMR and SAS populations are estimated to contain the next two lowest levels of drift relative to the GWAS population. Interestingly, a wide variation of estimated F_{st} values exist across the different regions for the AMR population. We speculate that admixed samples which are present within the AMR super-population may result in certain regions containing higher levels of European ancestry leading to lower estimates of drift. Lastly, both the EAS and AFR populations show the highest estimates of drift relative to the GWAS population which is consistent with past research on the high levels of genetic differentiation separating these pairs of populations (Elhaik 2012).

Interestingly, the density of the noise estimates from the MLE contains several interesting features. We observe that using each of the 1000G super-populations with the exception of the EUR population results in a wide variation of estimates for the noise. We speculate two possibilities on the cause for this variation. Firstly, sources of noise (eg. sequencing errors or quality control issues) may be different

across the genome. Secondly, in some regions of the genome, genetic drift between a reference and GWAS population is incorrectly estimated as noise. Further, we observe that the mean estimates for noise are different across each of the 1000G populations. As we noted previously, since the 1000G and UK Biobank are both highly quality controlled data-sets, observing differential errors across the genome would be unexpected. As a result, we suggest that the more likely explanation for these patterns can again be explained by the MLE incorrectly classifying genetic drift as noise. Lastly, we observe an extremely small variation of estimated noise obtained using the 1000G EUR population. This result suggests that in cases when the reference and GWAS population exhibit very low genetic drift (eg. in the case of a UK Biobank and European population) it becomes harder for the MLE to differentiate signals of noise and drift.

4.5 Conclusion

We began this chapter by establishing the relationship between the GWAS variances and the GWAS allele frequencies. Using this relationship and the Balding-Nichols model of genetic drift, we developed a method to infer the underlying level of genetic drift relative to a reference population and noise within a GWAS population using only the GWAS summary statistics. The method presents a few limitations which we discuss below.

Firstly, when we established the relationship between the GWAS allele frequencies and the GWAS variances in Equation 4.2, we derived this equation under the setting of a binary GWAS phenotype. An interesting direction for future investigation would be to extend the results in this chapter to also consider quantitative phenotypes. The application of Equation 4.2 in the case of a quantitative phenotype is predicated on the assumptions that the effects of the variants on the phenotype variance are small and that the phenotype is processed by quantile normalisation or scaling of

the residuals after regressing out the covariate effects (Vukcevic et al. 2011). The evaluation of these assumptions using real GWAS data would greatly improve the utility of the method.

A second limitation of the method relates to the handling of missingness in the GWAS data. As illustrated in Figure 4.3, Equation 4.2 does not apply to imputed data as a result of missingness that arises during the process of imputation. Consequently, we restricted ourselves to using only the genotyped data for the methods in this chapter. This restriction assumes that we have access to information on which variants were genotyped and which were imputed in the GWAS. However, if this were not the case, we could restrict ourselves to only using variants which most commonly exist on the most widely available genotyping chips. Several online sources have been developed to curate a list of these variants (See `https://www.well.ox.ac.uk/~wrayner/strand/` for examples). Overall, the most important direction of future work would involve deriving the relationship between the levels of missingness and the expected transformation of the GWAS variances to the GWAS allele frequencies. Such work would allow us to more effectively incorporate the imputed data within the methods. One possible starting point would be to quantify the effect of the INFO-Score for the imputed variants within Equation 4.2.

A third limitation for the method is the challenge in interpreting estimates of noise and F_{st} within a real-world context. In the method, the value of F_{st} is estimated in the case of the Balding-Nichols model of genetic drift. However, the Balding-Nichols model considers a background reference allele frequency and models the distribution of the drifted allele frequencies in divided sub-populations. In this chapter, we considered the level of genetic drift between two present-day populations. Therefore, the Balding-Nichols model is not a perfectly accurate representation of the drift between these populations. Further, translating the estimated level of noise from the model is also challenging in a real-world context. The estimated value of κ is only a rough indicator for the amount of noise that we estimate within

the GWAS. Thus, it would be extremely useful to take into account the various sources of noises that may arise in a GWAS (sequencing, covariates, etc.) and characterise the distribution of the noise from these specific sources.

Fourthly, in this chapter, we assumed that we knew in advance of the various constants in Equation 4.2 such as M_{GWAS} and ϕ from the GWAS study. This is a reasonable assumption, as these constants are often times found in the accompanying manuscript. However, an interesting direction of future research would be to estimate these constants from the GWAS study using solely the GWAS summary statistics.

Despite these limitations, the method proves to be extremely advantageous for two reasons. Firstly, the estimates of genetic drift were found to be relatively accurate. By using simulated coalescent data under the demographic history of a population split while using either the GWAS variances implied by Equation 4.2 (Figure 4.8) or GWAS variances obtained from a logistic model fitted to the data (Figure 4.9) we estimated higher levels of F_{st} as the divergence time increased. In the case of the Out of Africa model, each population pairing possessed a relatively high value of F_{st}. Using the method on data from a UK Biobank GWAS, we correctly chose the European 1000G reference population as possessing the lowest level of drift relative to the UK Biobank GWAS population. Together, these results suggest that in real-world settings, the method provides relevant information for researchers when needing to select a suitable reference population given the GWAS summary statistics.

Secondly and most importantly, the MLE would display substantial utility within the haplotype-weighting scheme developed in the previous chapter. For instance, we could develop a two stage pipeline. The first stage would involve estimating the F_{st} between a reference population and the GWAS summary statistics using the MLE. The estimation could be applied across a series of reference populations to ascertain the most suitable population (ie. the population with the lowest drift relative to the GWAS population) for summary statistic analyses. The second stage would involve

performing the haplotype weighting scheme with the selected reference population and using the estimated level of F_{st} to construct the gamma distributed weights.

Overall, the results from this chapter highlight that on the surface, summary statistics do not appear to contain much of the pertinent information necessary to perform complex trait analyses. However, a thorough investigation into the properties of the GWAS summary statistics can still uncover critical information about the GWAS population that is useful for downstream summary statistic analyses.

5

Inferring the Distribution of Linkage Disequilibrium from GWAS Summary Statistics

Contents

In this chapter, we develop a method which infers the distribution of measurements of LD from GWAS summary statistics using the haplotype weighting scheme. The intent of this method is to account for uncertainty in the distribution of LD within the GWAS population for common statistical genetic applications such as summary statistic imputation and fine-mapping. The work in this chapter builds

on the work developed in Chapter 3 by incorporating information about likely sets of reference haplotype weights. Specifically, whereas in Chapter 3 we generated a series of LD matrices under a model of drift independently about knowledge regarding the GWAS population, here we consider how to use information available from the GWAS summary statistics.

We begin by deriving the likelihood function for generating drifted allele frequencies through the haplotype weighting scheme given the GWAS summary statistics. By incorporating this likelihood within the marginal weight updates developed in Chapter 3, we update the weights to perform inference. This process leads to our method: *InferLD*.

Using simulated genetic data, we demonstrate how our method accurately infers the in-sample GWAS LD measurements and allele frequencies. By employing real-world data with the GWAS summary statistics from the UK Biobank and reference populations from the 1000G Project, our method outcompetes traditional unweighted reference populations.

Next, we apply *InferLD* to the GWAS summary statistics from a simulated African-American admixed GWAS population and demonstrate how our method is useful for downstream statistical genetic applications such as fine-mapping. Further, we describe that by inferring the distribution of LD measurements, our method possesses a significant advantage over traditional reference populations due to its ability to improve the calibration of fine-mapping.

Lastly, we discuss an extension to *InferLD* where the weights change along the region under a hidden Markov model (HMM) framework and the advantages and disadvantages of this extension.

5.1 Incorporating Information from GWAS Summary Statistics

In Chapter 3, we previously described a method to marginally update the weights[1] for a focal reference haplotype across a row in W (See Table 5.1 for all notation and definitions used in this chapter). Under this framework, we sampled a weight from a set $\mathbb{W}$ comprised of K possible weights with uniform probabilities (the potential values representing the quantiles of the gamma distribution). Through this process, we generated a distribution of population genetic measurements such as the allele frequencies and LD measurements in a genetically drifted population relative to the reference population. In this chapter, we instead sample weights with probabilities defined by a likelihood function using the GWAS variances of the effect size $\sigma^2_{\hat{\beta}}$ as the observed data (Equation 5.1). We perform Gibbs sampling on W by repeatedly marginally updating the focal reference using this likelihood while freezing the remaining weights in W. In this process, we infer measurements of the GWAS population such as the allele frequencies or the LD. This likelihood is formally given as:

$$\mathcal{L}(W_{i,} = w_k^{'} | \sigma^2_{\hat{\beta}}, R). \tag{5.1}$$

We define the i^{th} haplotype in W as the focal haplotype. The focal haplotype is updated to one of K possible weight states from its current weight. Across all N SNPs and K weight states we obtain a matrix $\sigma^{2'}_{\hat{\beta}}$. Each element in $\sigma^{2'}_{\hat{\beta}}$ describes the value of the *proposed* GWAS variance at each SNP j if we update the focal haplotype to the k^{th} weight in $\mathbb{W}$.

In order to generate the proposed GWAS variances, we first generate the proposed drifted allele frequency for a given SNP j supposing that the weight on the i^{th} focal haplotype is updated to $w_k^{'}$ calculated as:

[1] Where the weights remain constant across the region for a given focal haplotype.

Table 5.1: Notation and associated definitions used in Chapter 5. When possible, notation from previous chapters were re-used.

Symbol	Definition
W	Weight matrix with dimensions (M_{ref},N)
W^1	First weight matrix initialized within the Gibbs sampling
$w^{'}$	Proposed weights for focal reference haplotype update
W^G	True weight matrix of GWAS population
K	Number of weight states
$\hat{K}$	Number of weight states set in *InferLD*
k	Index across weight states
$\mathbb{W}$	Gamma drift prior set of weight states
R	Reference haplotype panel with dimensions (M_{ref},N)
N	Number of SNPs in region
M_{ref}	Number of reference haplotypes
M_{GWAS}	Number of GWAS haplotypes
ϕ	GWAS case fraction
$\sigma^2_{\hat{\beta}}$	Observed GWAS variances
$\sigma^{2'}_{\hat{\beta}}$	Proposed GWAS variances
$\pi^{'}$	Proposed allele frequencies given focal haplotype update
π^G	Allele frequencies of GWAS population
Ω	Ratio of observed to proposed GWAS variances
θ	Likelihood of observing a GWAS variance
F_{st}	Genetic drift
$\hat{F_{st}}$	Genetic drift parameter set in *InferLD*
κ	Gamma likelihood shape and rate (noise) parameter
$\tilde{\kappa}$	Adjusted gamma likelihood shape and rate (noise) parameter
$\hat{\kappa}$	Noise parameter set in *InferLD*
$k^{'}$	Indexes of inferred hidden weight states within our HMM
i	Focal haplotype index
$j, j^{'}$	SNP index
$h, h^{'}$	Haplotype index
r_{inf}	Inferred linkage disequilibrium
r_{adj}	Adjusted linkage disequilibrium using shrinkage factor
s	Shrinkage factor
A	State transition matrix, with dimension K by K
B	Observation probability matrix with dimension K by N
ν	Initial state distribution
ρ	Recombination rate within HMM transition matrix
$x_j, x_{j'}$	SNP position in genome
α	forward-pass matrix of dimensions K by N
m,n	HMM weight states indexes

$$\pi^{'}_{kj} = \frac{\sum_{l'}^{M_{ref}} \left(\left(\sum_{l=1}^{M_{ref}} w_{lj} - w_{ij} + w^{'}_{k} \right) R_{l'j} \right)}{\sum_{l=1}^{M_{ref}} w_{lj} - w_{ij} + w^{'}_{k}}. \tag{5.2}$$

These proposed drifted allele frequencies are then transformed into $\sigma^{2'}_{\hat{\beta}}$ for each SNP j and weight state k using Equation 4.2 as:

$$\sigma^{2'}_{\hat{\beta}_{kj}} = \frac{1}{\pi^{'}_{kj}(1-\pi^{'}_{kj}) M_{GWAS} \phi(1-\phi)}. \tag{5.3}$$

Next, we derive a likelihood function for $\sigma^{2'}_{\hat{\beta}}$ given $\sigma^{2}_{\hat{\beta}}$, R and W. We assign a higher probability to weights which generate proposed GWAS variances that match closer to the observed GWAS variances. The process of matching these values implicitly matches the underlying GWAS allele frequencies encoded by the GWAS variances (See Chapter 4 for details). Moreover, the process of matching the allele frequencies between the weighted reference population and the GWAS population also allows us to match the LD measurements (See Chapter 3 for details).

In Chapter 4, we previously investigated the ratio of the fitted GWAS variances to the GWAS variances implied by Equation 4.2. The ratio of the observed to the implied GWAS variances was assumed to be gamma distributed with a variance proportional to the level of noise (Equation 4.3) parametrised with a shape and rate parameter κ. These ratios are utilized in the likelihood function in order to characterise the strength of matches. We define this ratio for a SNP j as:

$$\Omega_{kj} = \frac{\sigma^{2}_{\hat{\beta}_j}}{\sigma^{2'}_{\hat{\beta}_{kj}}} \tag{5.4}$$

where k is the index of the updated weight state for the focal haplotype.

We briefly investigate the relationship between the ratio of the observed GWAS variances to the implied GWAS variances and the allele frequencies in a simulated GWAS (Figure 5.1). Because these ratios are calculated using the GWAS allele frequencies, we expect this ratio to remain approximately at one across all SNPs. However, as seen in Chapter 4, the variance of these ratios become larger as the

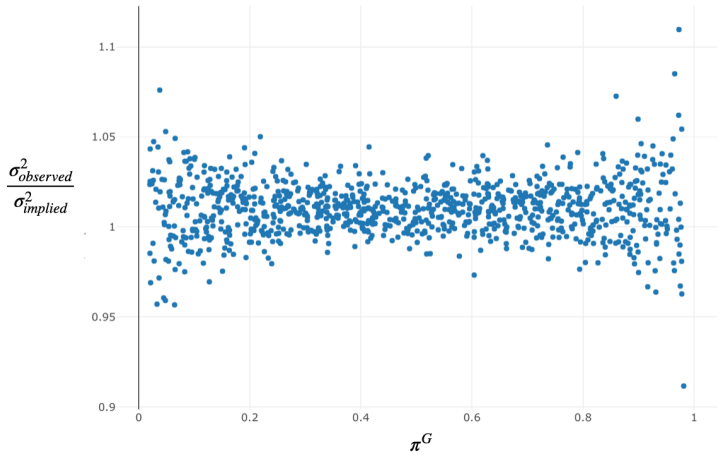

Figure 5.1: Ratio of the observed GWAS variances over the implied GWAS variances given by Equation 4.2 as compared to the GWAS allele frequency π^G. Each point corresponds to a single SNP from a simulated GWAS with 1000 SNPs and 10,000 cases and controls (See Figure 4.1 for details of this simulation).

MAF of the GWAS SNP decreases. This result suggests that the likelihood function should account for the increased noise within rare SNPs.

To empirically fit the relationship between the levels of noise and the GWAS allele frequency of a SNP, we split the simulated GWAS summary statistics into nine bins each spanning 10% allele frequency. Within each bin, we calculate the implied gamma shape (and rate) parameter κ using the ratio of the mean allele frequency over the variance of the allele frequency across all SNPs. We plot the gamma shape parameter against $\pi^G(1-\pi^G)$ where π^G is the GWAS allele frequency (Figure 5.2). The fit that we obtain is extremely strong with an r^2 correlation of 0.92. This leads us to adjust the shape (and rate) parameter in Equation 4.12 to:

$$\tilde{\kappa} = 4\kappa\pi^{'}_{kj}(1-\pi^{'}_{kj}), \tag{5.5}$$

where the factor of 4 exists such that the maximum value of $\tilde{\kappa}$ is κ when $\pi^{'} = 0.5$.

Thus, the probability that the weight $w^{'}_{k}$ is chosen is proportional to the likelihood function for the weight state k. Moreover, the likelihood function is in turn proportional to the density of the ratio of observed and implied GWAS variances.

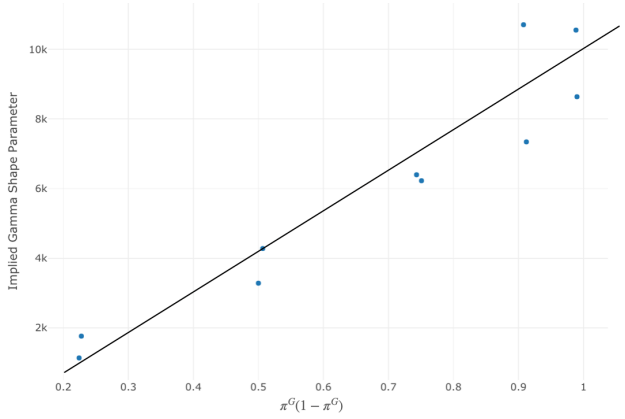

Figure 5.2: Relationship between the implied gamma shape parameter and mean value of $\pi^G(1-\pi^G)$ splitting 1000 SNPs into bins of 10% allele frequency each. The black line illustrates the fitted regression line with a r^2 correlation of 0.92.

For a given SNP j and weight state k the probability θ_{kj} is written as:

$$\theta_{kj} = \mathcal{L}(W_{ij} = w_k^{'} | \sigma^2_{\hat{\beta}_j}, R, \tilde{\kappa}) = \frac{\tilde{\kappa}^{\tilde{\kappa}}}{\Gamma(\tilde{\kappa})} \Omega_{kj}^{\tilde{\kappa}-1} e^{-\tilde{\kappa}\Omega_{kj}}. \tag{5.6}$$

We obtain the likelihood across all N SNPs in the region by calculating the multiplicative product of these likelihoods (assuming independent observations across SNPs) as:

$$\theta_k = \prod_{j=1}^{j=N} \theta_{kj}. \tag{5.7}$$

In order to update the focal haplotype in W, we sample a weight state[2] k from $\mathbb{W}$ proportional to its probability θ_k. Next, we proceed to marginally update the next focal haplotype in the same manner using the newly updated weight matrix. Updating W in this framework forms the basis of inference within *InferLD*. Moreover, as described in Chapter 3, by repeatedly sampling weights for W we can generate a distribution on the LD and the allele frequencies within the GWAS population.

[2]In the case of constant weights, this weight state is a vector that contains the same weight repeated N times.

5.2 Inference under a Generative Model

We now illustrate the implementation of *InferLD* using data simulated under a generative model. We begin by using the EUR 1000G population and extract a region on Chromosome 1 between positions 1065323 and 1534149 which contain 83 SNPs with a MAF greater than 1%. To generate the true drifted population G, we set $K = 100$ weight states in $\mathbb{W}$ under a given level of genetic drift. As in Chapter 3, we uniformly sample M_{ref} weights from $\mathbb{W}$ in an identical and independent manner with replacement to populate the true drifted weight matrix W^G.

Next, using Equation 3.1, we generate the set of GWAS allele frequencies π^G with W^G and R. These allele frequencies are converted into the implied GWAS variances using Equation 4.2. We then transform these implied variances into the noisy and observed GWAS variances $\sigma^2_{\hat{\beta}}$ by multiplying each of the implied GWAS variances by a noise value sampled from the gamma distribution (Equation 4.12) under a value of $\tilde{\kappa}$.

We set the parameters within *InferLD* to the same values we use to simulate the GWAS data. We perform updates to W once through the entire reference population equivalent to 808 marginal updates. When we apply our method to inferring more realistic GWAS data (ie. no longer under the generative model), we will perform multiple updates through W. Here, we perform a single update to illustrate the convergence of both the log-likelihoods and the correlation of the allele frequencies and LD measurements. We initialize the first weight matrix used in inference W^1 by uniformly sampling a weight in $\mathbb{W}$ and populating the entirety of W^1.

Figure 5.3 shows the result of inference across simulated data in four settings: a)high drift and low noise, b)high drift and high noise, c)low drift and low noise and d)low drift and high noise.

Overall, *InferLD* reaches a r^2 correlation above 0.9 between the inferred and the GWAS measurements in each setting. In the setting of low drift and high

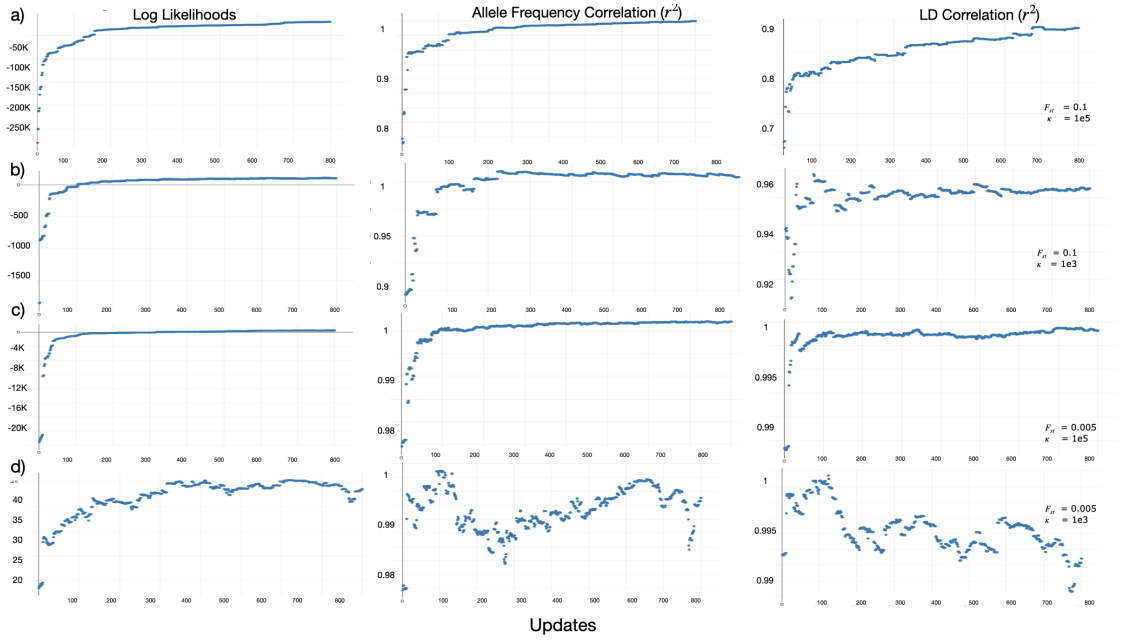

Figure 5.3: Inference accuracy using *InferLD* on simulated data under a generative model. GWAS summary statistics were simulated using a reference population from the 1000G EUR population under different levels of drift and noise. The corresponding parameter settings (drift and noise) used for both the generative data and the inference parameters are displayed to the right of the graphs in a)-d). The first column of graphs illustrates the log-likelihood across individual updates. The second column illustrates the r^2 correlation of the inferred allele frequencies to the GWAS allele frequencies across updates. Lastly, the third column illustrates the r^2 correlation of the inferred to the GWAS LD measurements (r) across updates. Note that each x-axis is on the same scale across all parameter combinations but y-axis changes scale between combinations.

noise (Figure 5.3 d), the likelihood curve displays a greater volatility. This is a consequence of the lower gamma noise parameter[3] thus making the likelihood surface less peaked around weight vectors that match the GWAS allele frequency. These results suggest that in order to obtain the highest degree of accuracy for inference, we should set $\hat{\kappa}$ as high as possible (ie. assume very little as noise in the GWAS).

Lastly, we observe that in the case of high drift (Figure 5.3 a and b), the accuracy of the allele frequency and the LD inference is smallest relative to the other parameter settings. This suggests that high levels of drift will be the most challenging for inference.

[3]Note that a higher κ shape parameter implies a lower level of noise in the construction of the likelihoods due to the properties of the gamma distribution.

5.3 Applications to GWAS Data

5.3.1 UK-Biobank and 1000G

We next evaluate *InferLD* using real GWAS summary statistics generated from a UK Biobank GWAS for hypertension as previously described in Chapter 4. We segment the GWAS summary statistics across Chromosome 1 into 2Mb chunks. We remove any region that contains less than 50 SNPs leaving 220 regions on which we perform inference using *InferLD*. The three 1000G super-populations used as reference are the following: the haplotypes from the European super-population (EUR), haplotypes from the African super-population (AFR) and haplotypes from both the European and African super-populations (EUR+AFR). We extract the same regions in the reference populations and jointly remove SNPs with less than 1% MAF. We run *InferLD* using parameters $\hat{\kappa} = 1e4$ and $\hat{K} = 1000$. To assess the impact of varying the level of $\hat{F_{st}}$, we run *InferLD* using three values of $\hat{F_{st}}$: 0.1, 0.01 and 0.001. For each region, we run the inference for 10 complete updates through the respective reference population. We extract the last 10 percent of samples from the inference and average the weight matrices element-wise across the samples. We use the the averaged weight matrix along with R to obtain the inferred allele frequencies and LD measurements (Equations 3.1, 3.3).

We compare the allele frequency correlation we obtain using *InferLD* to the allele frequency correlation using the unweighted reference populations (EUR, AFR or EUR+AFR) (Figure 5.4, Table 5.2). We calculate both an overall allele frequency correlation across all SNPs pooled across the 220 regions and additionally split the SNPs into different allele frequency groups. One group considers SNPs $\geq$ 40% MAF (labelled as group A2) and the other group considers SNPs $<$ 40% MAF (labelled as group A1). Similarly, for the LD inference, we calculate an overall correlation across all SNP-SNP pairs. We also split the SNP-SNP pairs into either a group

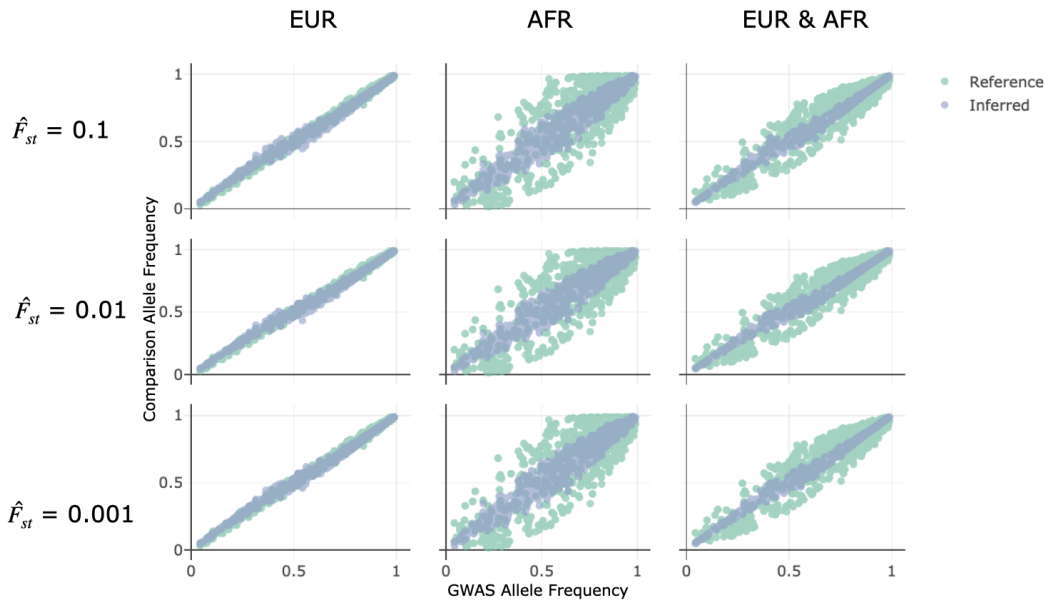

Figure 5.4: Accuracy of allele frequency inference with *InferLD* using the three reference populations (EUR, AFR or EUR+AFR) and UK Biobank hypertension GWAS summary statistics pooled across all regions. Columns correspond to the different populations used during inference while rows correspond to the value of $\hat{F}_{st}$. The x-axis denotes the true GWAS allele frequencies and the y-axis denotes either inferred allele frequency (blue) or the reference population allele frequency (green). The r^2 correlation values for these plots are given in Table 5.2

with highly linked SNPs (labelled as High) when $|r| \geq 0.3$ or into a group with relatively unlinked SNPs (labelled as Low) when $|r| < 0.3$.

The allele frequency inference performs best using the EUR population, followed by the EUR+AFR population and lastly by the AFR population. Interestingly, we observe that the inference performs better for SNPs in group A1 as compared to SNP in group A2. One explanation for the lower correlation of group A2 is a result of the transformation of the GWAS allele frequencies to the GWAS variances (Equation 4.2). This equation maps two allele frequencies (π^G and 1-π^G) to a single GWAS variance. Therefore, it is challenging to infer the GWAS allele frequency when the frequency approaches 50% as *InferLD* can not easily distinguish between these two allele frequencies. Promisingly, we also achieve similar accuracy between the EUR and the EUR+AFR population. This suggests that the addition of the African samples does not greatly reduce the accuracy of the allele frequency inference. This

is true although the African samples do not match the ancestral composition of the GWAS population. Lastly, in the case of an AFR reference population, we infer a significantly higher correlation when compared to the unweighted AFR population. The value of $\hat{F_{st}}$ set in *InferLD* was not found to have a meaningful impact on the allele frequency inference.

Similarly, we observe that the inference of the LD measurements is most accurate when *InferLD* uses the EUR reference population, followed by the EUR+AFR population and lastly the AFR population (Figure 5.5, Table 5.3). These results suggest that *InferLD* performs best when provided with an ancestrally matched reference population to the GWAS population. However, across all reference populations, the inference performs significantly worse for pairs of SNPs that are weakly linked with each other. These pairs of SNPs are inferred to contain much higher levels of LD than is actually present in the GWAS population. The inflation of the measurements of LD is a result of the inability of *InferLD* to account for recombination in the region. We explore an extension to *InferLD* to overcome this limitation at the end of the chapter. Interestingly, by setting lower values of $\hat{F_{st}}$, we reduce the extent of this inflation.

Therefore, in order to improve LD inference, a degree of *shrinkage* is applied, penalising and reducing the LD measurements for SNPs at an exponential rate (Wen and M. Stephens 2010). Applying shrinkage is a common technique used to improve correlations from a reference population (Benner et al. 2017). Given the inferred LD measurements r_{inf} we calculate r_{adj} by adjusting all non-diagonal elements in the LD matrix ($j \neq j^{'}$) in r_{inf} as:

$$r_{adjjj'} = exp(-|x_j - x_{j'}|s_{jj'})r_{inf_{jj'}}. \tag{5.8}$$

where $|x_j - x_{j'}|$ denotes the absolute *distance* between the SNPs and s is some shrinkage factor. In our analyses, for convenience we set s to a constant value.

Table 5.2: r^2 Allele frequency correlations to the true GWAS allele frequencies across different reference populations (EUR, AFR, EUR+AFR) $\hat{F_{st}}$ column corresponds to the drift parameter set during inference. Unweighted reference corresponds to using the reference population without performing inference. SNP frequency corresponds to either "All" denoting the correlation using all the SNPs across the pooled 220 regions, "A2" for SNPs with greater than 40% MAF in the GWAS population and "A1" for SNPs with less 40% MAF in the GWAS population.

Population	$\hat{F_{st}}$	SNP Frequency	r^2 Correlation
EUR	Unweighted Reference	All	0.994
		A1	0.995
		A2	0.840
	0.1	All	0.992
		A1	0.992
		A2	0.656
	0.01	All	0.993
		A1	0.993
		A2	0.672
	0.001	All	0.994
		A1	0.995
		A2	0.760
AFR	Unweighted Reference	All	0.721
		A1	0.771
		A2	0.106
	0.1	All	0.964
		A1	0.978
		A2	0.381
	0.01	All	0.965
		A1	0.976
		A2	0.383
	0.001	All	0.965
		A1	0.977
		A2	0.385
EUR+AFR	Unweighted Reference	All	0.893
		A1	0.917
		A2	0.245
	0.1	All	0.992
		A1	0.978
		A2	0.621
	0.01	All	0.993
		A1	0.992
		A2	0.632
	0.001	All	0.993
		A1	0.994
		A2	0.662

Table 5.3: r^2 LD correlations to the true GWAS LD measurements (r), across different reference populations (EUR, AFR, EUR+AFR). $\hat{F_{st}}$ column corresponds to the drift parameter set during inference. Unweighted reference refers to using the reference population without performing inference. Linkage column corresponds to either "All" denoting the correlation using all SNPs across the pooled 220 regions, "High" for pairs of SNPs with GWAS $|r|$ greater to or equal to 0.3, and "Low" for pairs of SNPs with GWAS $|r|$ less than 0.3.

Population	$\hat{F_{st}}$	Linkage	r^2 Correlation
EUR	Unweighted Reference	All	0.875
		High	0.977
		Low	0.673
	0.1	All	0.614
		High	0.971
		Low	0.329
	0.01	All	0.640
		High	0.958
		Low	0.355
	0.001	All	0.731
		High	0.971
		Low	0.452
AFR	Unweighted Reference	All	0.671
		High	0.761
		Low	0.096
	0.1	All	0.323
		High	0.845
		Low	0.090
	0.01	All	0.352
		High	0.846
		Low	0.091
	0.001	All	0.455
		High	0.867
		Low	0.092
EUR+AFR	Unweighted Reference	All	0.709
		High	0.871
		Low	0.371
	0.1	All	0.631
		High	0.958
		Low	0.338
	0.01	All	0.645
		High	0.959
		Low	0.335
	0.001	All	0.737
		High	0.965
		Low	0.466

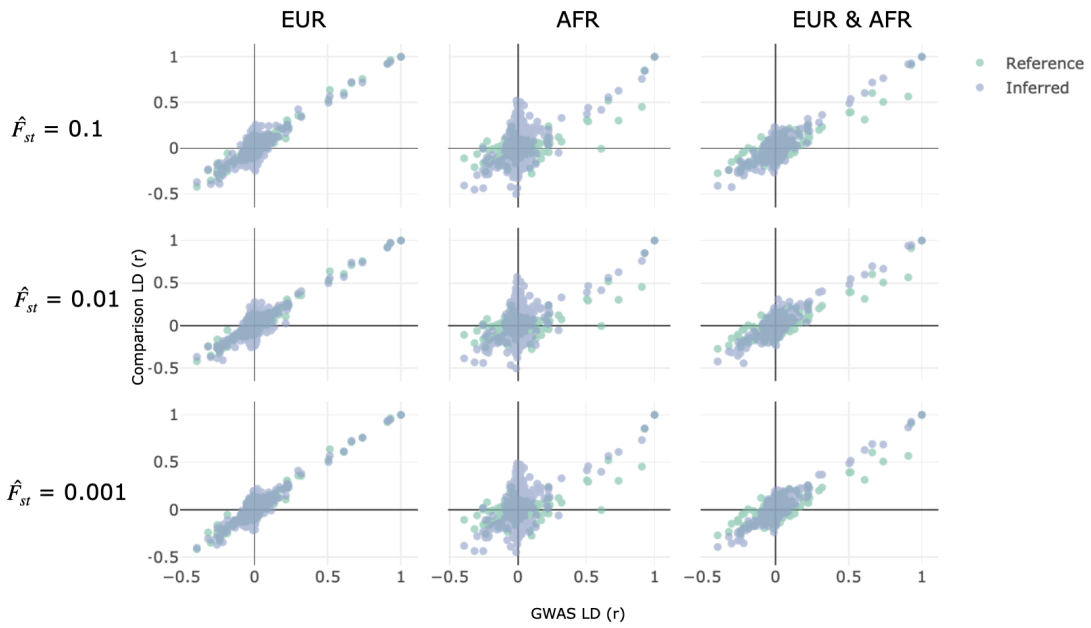

Figure 5.5: Accuracy of LD inference with *InferLD* using the three reference populations (EUR, AFR or EUR+AFR) and UK Biobank hypertension GWAS summary statistics pooled across all regions. Columns correspond to the different populations used during inference, while rows correspond to the value of $\hat{F}_{st}$. The x-axis denotes the true GWAS LD measurements and the y-axis denotes either the inferred LD measurement (blue) or the reference population LD measurement (green).

To assess the correct value of s used in our model, we perform the same simulations as in Figure 5.5 but now run LD inference with using the adjusted LD given by Equation 5.8. The results in Figure 5.6 show the resulting r^2 correlation to the true GWAS LD measurements across all pairs of variants using the three reference populations in *InferLD* and across a range of shrinkage factors s.

We observe that as the shrinkage factor increases, the value of the correlation in the EUR and EUR+AFR population remains roughly consistent at approximately 0.9. Interestingly, the correlation for our inferred LD LD measurements using the AFR reference population increases until it reaches approximately 0.75. Therefore, these results suggest that the shrinkage factor should be set to 0.05 to obtain the highest accuracy across all three reference populations. Note that rather than using this empirical approach, we could alternatively set s using a cross-validation approach (Zhu and M. Stephens 2017).

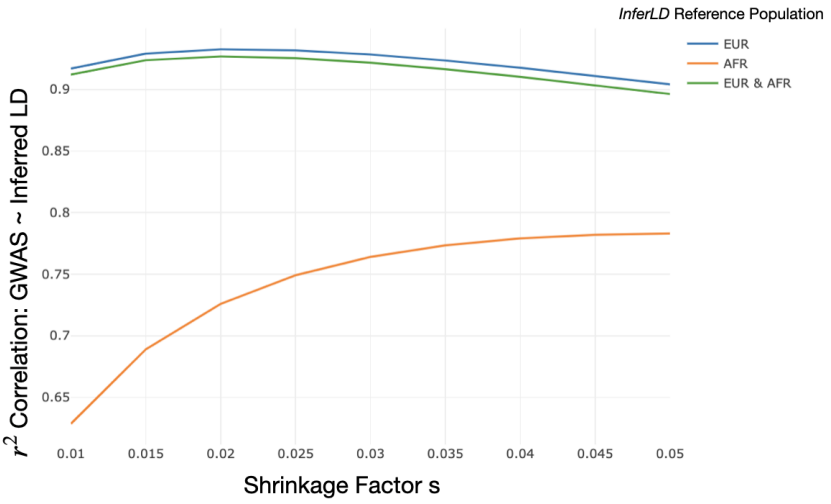

Figure 5.6: Assessing the correlation of *InferLD* under different shrinkage factors and across reference populations. Each simulation was performed as given in Figure 5.5 to generate the non-adjusted LD measurements and then shrunk using Equation 5.8 to account for the inflation in the LD inference.

We re-run inference with $\hat{F_{st}} = 0.001$. After applying shrinkage, the adjusted inferred LD measurements now out-perform the respective reference population correlations for all three reference populations across all SNPs (Figure 5.7).

5.3.2 Two-Way Admixture

We now assess performance of *InferLD* on a GWAS population composed of admixed individuals. As a result of the lack of genetic data collected for admixed populations, there is also a corresponding lack of reference populations available matched to these admixed populations (Atkinson et al. 2021). Under these circumstances, researchers typically rely on using reference populations formed of different marginal ancestries and attempt to match the correct ancestral composition of the admixed samples within the reference population. This setting demonstrates a highly suitable application for *InferLD* which can generate a weighted reference population that effectively will match the ancestral composition of the GWAS population.

In order to investigate the performance of *InferLD* in this setting, we simulate

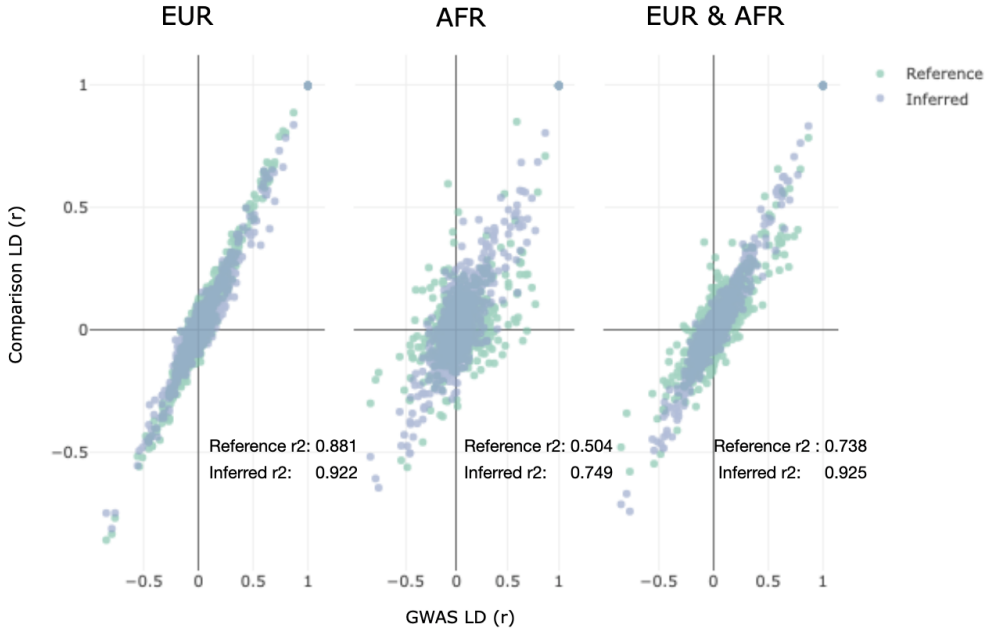

Figure 5.7: Accuracy of LD inference with *InferLD* using the three reference populations (EUR, AFR or EUR+AFR) and UK Biobank hypertension GWAS summary statistics pooled across all regions under a shrinkage factor $s = 0.05$. $\hat{F}_{st}$ was set to 0.001. The x-axis denotes the true GWAS LD measurements and the y-axis denotes either the inferred LD measurement (blue) or the reference population LD measurement (green).

an African-American admixed population. We roughly follow the same simulation framework as previously described in Chapter 2. Using *msprime*, we simulate a simplified version of the demographic model for American admixture using EUR, AFR and EAS populations (Browning et al. 2018). We generate admixture between the AFR and EUR populations 12 generations ago with 75/25 (AFR/EUR) ancestral proportions to form the admixed African-American population. We form each population with 5000 haplotypes. In each simulation, we generate a 1Mb region under a human-level mutation and recombination rate of 1e-8. Across all populations (EUR, EAS, AFR, and African-American), we jointly filter and remove SNPs that are below 1% MAF. Using the simulated African-American population, we use HAPGEN2 to generate a GWAS population with 10,000 cases and 10,000 controls. The GWAS population is simulated under a causal model of association with three causal SNPs chosen at random. Each of the three SNPs possesses heterozygous and homozygous log odds ratios for disease risk of 1.5 and 2.25

respectively. As we previously described in Chapter 2, we use these very high odds ratios to improve the accuracy of downstream fine-mapping and better compare the accuracy of using FINEMAP with different LD panels. Next, we use SNPTEST to generate association statistics using the haplotypes generated from HAPGEN. We provide the GWAS association statistics to FINEMAP. FINEMAP is run using the LD obtained from the four *msprime* populations (EUR, EAS, AFR reference and the African-American GWAS). Furthermore, we consider three other reference populations that reflect real-world scenarios. First, we employ a reference population (1000G Proportions) with the same proportions of EUR and AFR haplotypes as within the 1000G Project (1006 EUR and 1332 AFR haplotypes). We employ a reference population (Ref Admixed) which contains AFR and EUR haplotypes at a 75/25 ratio (1500 AFR and 500 EUR haplotypes) to match the admixture fraction of the GWAS population. Lastly, we employ the LD generated from *InferLD*.

We provide *InferLD* with the 1000G Proportions reference population. For inference, we perform 10 full updates[4] setting $\hat{F}_{st} = 0.001$, $\hat{\kappa} = 1e4$ and $\hat{K} = 1000$. We discard the first 90% of the updates as burn-in and calculate the element-wise average weight matrix across the last 10% of samples and calculate the LD measurements using Equation 3.3.

To assess the accuracy of FINEMAP, we utilise the two previously described metrics (See Chapter 2 for details): the weighted PIP and the number of true causal SNPs included within the credible set. The resulting FINEMAP metrics with the LD panels obtained from the different populations across 1000 replicate regions is described in Figure 5.8.

Overall, with respect to the weighted PIPs, FINEMAP performs most accurately when using the LD panel obtained from the GWAS population. Each of the marginal reference populations (EUR, AFR and EAS) performed the weakest as compared to the other populations. The *InferLD* population achieve an accuracy that is equal to

[4]Marginally updating each reference haplotype 10 times.

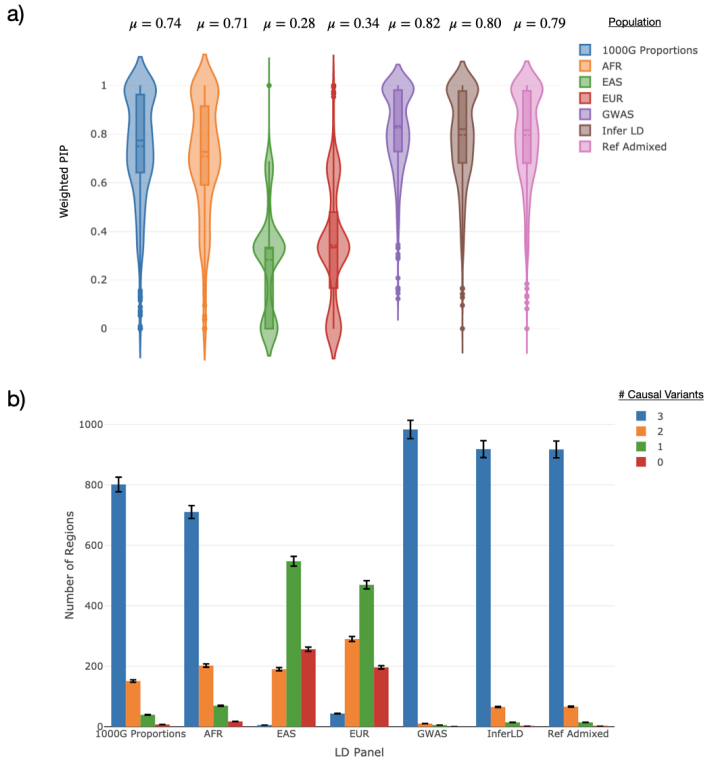

Figure 5.8: Accuracy of fine-mapping simulated admixed GWAS summary statistics using the LD panels obtained from different populations across 1000 replicate regions with FINEMAP. a) Distribution of the weighted PIPs. The value μ above the distribution indicates the mean across the replicates for a given population. FINEMAP was run with flag **–n-causal-snps 3**. b) The number of regions in which the 95 percent credible set marginally contained 0, 1, 2 or 3 causal SNPs across 1000 replicate regions. Error bars give $\pm$ 1.96 times the standard error of the mean.

the Ref Admixed population and is significantly higher (0.80 versus 0.74) than using a 1000G Proportion population. This result suggests that *InferLD* is appropriately weighting the different populations (EUR and AFR) within the 1000G Proportions population to achieve more accurate GWAS LD measurements. As previously illustrated in Chapter 2, we note that a wide variation exists across replicate regions in the weighted PIPs for each of the different populations. Furthermore, the number of regions in which three causal SNPs was found in the credible set was also higher (850 versus 795) in the *InferLD* population rather than the 1000G Proportions population further highlighting the strength of the method.

5.4 Leveraging Distributions of Linkage Disequilibrium

In the previous section, we evaluated fine-mapping accuracy by using LD inferred from a GWAS population with *InferLD*. In this setting, we provided FINEMAP with a single LD panel calculated using the element-wise average of W across a series of samples[5] outputted from *InferLD*. However, by providing FINEMAP with the mean LD from *InferLD*, we take away the ability of *InferLD* to generate a distribution of LD measurements. As we recall in Chapter 2, the point estimates of LD measurements exhibit significant variation between populations leading to different fine-mapping results. Consequently, the employment of the distribution of LD measurements within fine-mapping aims to provide robustness against this variation. However, FINEMAP and other state of the art fine-mapping software are unable to directly incorporate more than a single LD panel at one time. To circumvent this obstacle, we repeatedly sample LD panels from *InferLD* and provide each individual panel for independent FINEMAP runs using the same set of GWAS summary statistics. The averaging of FINEMAP results across separate runs ensures robustness against variation in LD measurements. The ability of *InferLD* to provide this robustness is evaluated by assessing the calibration of FINEMAP using the simulated admixed GWAS data described above.

Specifically, we sample 100 weight matrices from *InferLD* using the same parameters as before. For each sample, we calculate the LD measurements (Equation 3.3) and then run FINEMAP to obtain the resulting PIP for each SNP. Next, we average the PIPs across all 100 samples and then bucket SNPs into nine equally sized bins of 10% PIP each. Within each bin, two metrics are calculated: the mean of the PIPs and the fraction of causal SNPs contained within each bin (ie. proportion of true positives). The proportion of true positives is equal to the mean

[5]Post burn-in.

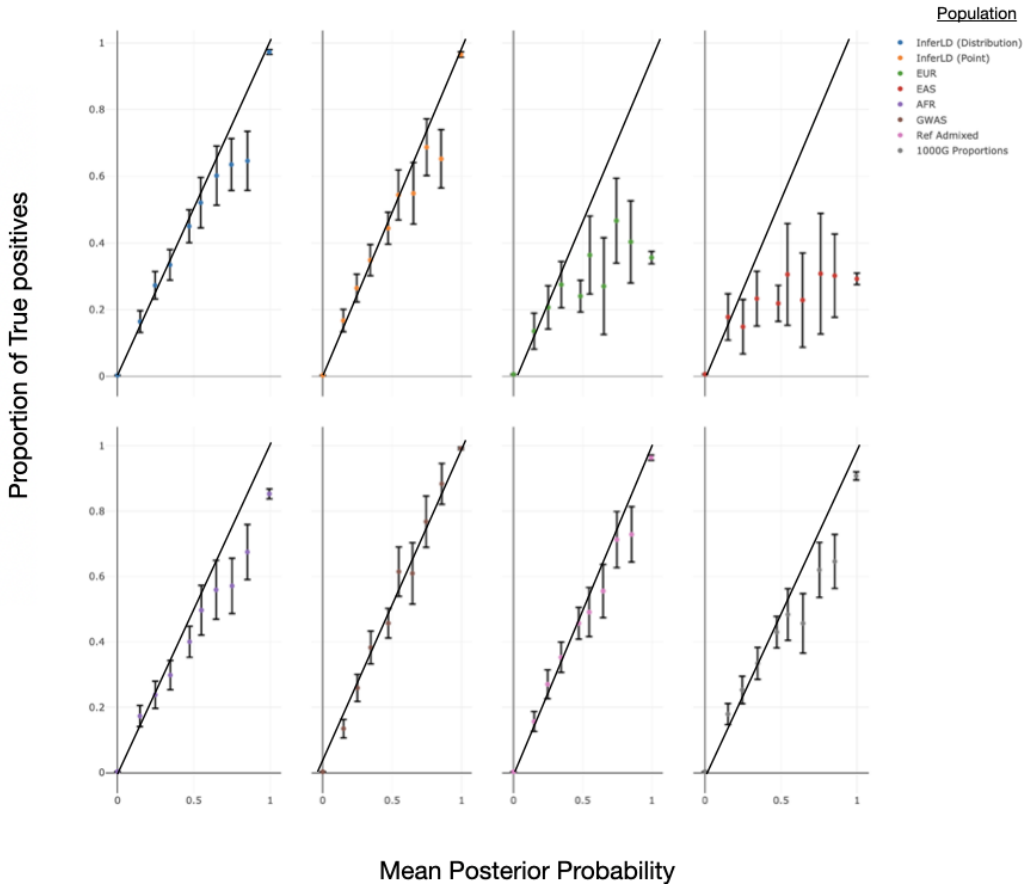

Figure 5.9: Calibration of fine-mapping using LD obtained from different populations. We note that the *InferLD* (Point) represents using the LD obtained from an average of samples from *InferLD* as in Figure 5.8 whereas *InferLD* (Distribution) uses the PIPs averaged across 100 FINEMAP runs with independent LD panels generated from *InferLD*. The black line is the line y=x which denotes perfect calibration. Error bars give ± 1.96 times the standard error of the mean.

PIP in each bin in a well calibrated credible set. We calculate these proportions and mean PIPs pooled across each of the 1000 replicate regions.

We compare the calibration of FINEMAP using LD panels obtained from the different populations investigated in the previous section (Figure 5.8). The two calibration metrics are calculated as previously described[6]. The calibration of FINEMAP in these different settings is shown in Figure 5.9.

The LD obtained directly from the GWAS population results in the best calibration of the FINEMAP PIPs. Using LD obtained from the Ref Admixed and

[6]There is no need to average PIPs across samples as FINEMAP is only run once per region.

InferLD (Distribution and Point) populations each appear roughly equally calibrated. The EUR and EAS populations appear the least calibrated. Importantly, LD obtained from *InferLD* outperforms the calibration of both the 1000G Proportions and the AFR population. Although we observe some differences in the resulting calibration between using *InferLD* with and without leveraging the distributions of LD measurements, it is unclear if leveraging the distributions in *InferLD* meaningfully improves calibration in this setting.

Interestingly, across the EUR, AFR, EAS and 1000G Proportions populations, the calibration of the PIPs appears to be anti-conservatively biased. In other words, the degree of confidence returned by FINEMAP is higher than expected given the data. The tendency of credible sets to produce PIPS that possess anti-conservative bias was recently documented in the case of simulated fine-mapping data with a single causal SNP (Hutchinson et al. 2020). Promisingly, using the LD obtained from *InferLD* removes this anti-conservative bias.

5.5 Recombining Weights

5.5.1 Motivation

Earlier, in order to correct for the inflation of LD measurements obtained from *InferLD*, we applied a shrinkage factor to the inferred measurements (Equation 5.8). To better understand why this inflation arises, we recall the genealogical interpretation of the haplotype weighting scheme within the forwards in time coalescent without recombination model (Figure 3.2). Importantly, as the name of this model suggests, recombination (ie. the exchange of genetic segments) between ancestor haplotypes is not modelled. In real-world settings, recombination has the effect of breaking down the associations between loci (ie. reducing measurements of LD between SNPs) (Pritchard and Przeworski 2001). Moreover, in the context of the haplotype weighting scheme, the weights represent the number of copies of the

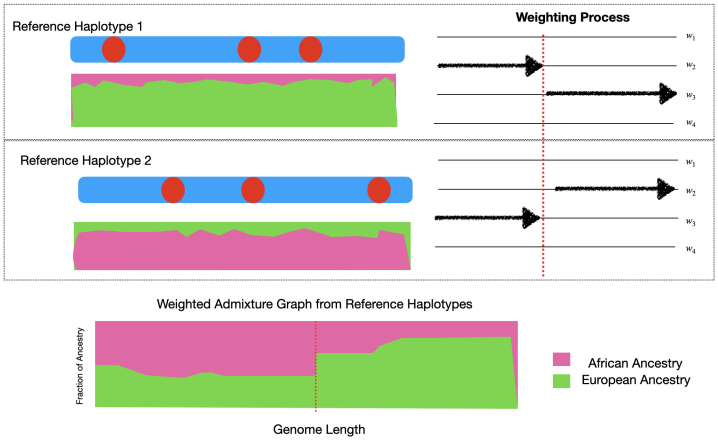

Figure 5.10: Example of using weights which recombine at a breakpoint halfway along the region for two reference haplotypes. The change in weights increases the proportion of the European reference haplotype in the second half of the region. Weights are ordered in size from lowest (w_1), to highest (w_4) with the black arrows denoting how the weights change along the region.

entire reference haplotype in the GWAS population and therefore we do not model the number of copies of the *genetic segments* of the reference haplotypes.

We now extend *InferLD* such that the reference haplotype weights vary along the region in a probabilistic fashion rather than remaining constant. By varying the weights along the region, we generate copies of the genetic segments from the reference haplotypes. When the copies of the genetic segments are stitched together (See Section 1.5.3 for details), we effectively mimic the process of recombination. This extension has the tangential benefit of improving inference in cases where GWAS samples possess admixed ancestries which vary regionally. Figure 5.10 describes how we emphasize the contributions of different ancestries along the region in the GWAS population by changing the weights along the reference haplotypes. Importantly, when the weights remain constant along the region, we are unable to model these changes in admixture[7].

[7]This is only true if the reference population does not contain any haplotype which displays the changes in admixture that are seen in the GWAS population.

5.5.2 Hidden Markov Model

Modelling how weights change across the region in the Gibbs sampling framework is naturally cast as a hidden Markov model (HMM). An HMM considers a sequence of observations to be the product of an unobserved Markov process, in which the sequence of underlying hidden states determines the probability of observing the data (See Durbin et al. 1998 for a complete description of this model). HMMs have been widely used in the field of statistical genetics including within genotype imputation (B. L. Browning and S. R. Browning 2016), haplotype phasing (S. R. Browning and B. L. Browning 2011) and multiple sequence alignment (Mount 2009).

The observed GWAS variances $\sigma^2_{\hat{\beta}}$ represent the sequence of observations $\{\sigma^2_{\hat{\beta}1}, \sigma^2_{\hat{\beta}2}, ..., \sigma^2_{\hat{\beta}N}\}$ across each SNP. The hidden states $\{w_1, w_2, ..., w_K\}$ represent the K discrete gamma distributed weights within $\mathbb{W}$. When we perform updates to the focal reference haplotype we will update the current weight to one of these weights at each SNP. Using standard HMM algorithms, we infer the sequence of hidden states from the observations (Figure 5.10). In this process, we infer a vector $k^{'}$ of length N. Using this vector, the focal reference haplotype at SNP j is updated to the weight state given by $k^{'}_{j}$.

The HMM is defined as $\lambda = (A, B, \nu)$. The matrix A denotes the state transition probabilities. Each element in A represents the probability of transitioning from weight state n to m between two adjacent SNPs j and $j + 1$ as:

$$a_{mn} = P(\text{state } w_n \text{ at } j + 1 | \text{state } w_m \text{ at } j). \tag{5.9}$$

To improve the computational complexity of the HMM algorithms, A has the property that transitions from the current weight state m to any of the other $K - 1$ weight states possess equal probability. Therefore, the diagonal elements in A are set to $1 - \rho$ and the non-diagonal elements are set to $\frac{\rho}{K-1}$, where ρ is the recombination

rate[8]. The non-diagonal elements in A represent the low probability[9] of a weight state transitioning to any other state. The diagonal elements in A represent the high probability of remaining in the current weight state. We can now write A fully as:

$$A = \begin{bmatrix} 1-\rho & \frac{\rho}{K-1} & \cdots & \frac{\rho}{K-1} \\ \frac{\rho}{K-1} & 1-\rho & & \vdots \\ \vdots & & & \\ \frac{\rho}{K-1} & \cdot & \cdot & 1-\rho \end{bmatrix}. \tag{5.10}$$

The observation matrix B contains elements which correspond to the probability of observing $\sigma^2_{\hat{\beta}j}$ given that we update the focal haplotype i to weight state w_n at SNP j while freezing the remaining weights in the Gibbs sampling framework. These elements are written as:

$$b_n(\sigma^2_{\hat{\beta}j}) = P(\sigma^2_{\hat{\beta}j}|\text{state } w_n \text{ at } j) = \mathcal{L}(\sigma^2_{\hat{\beta}j}|W_{ij} = w_n) \tag{5.11}$$

where the likelihood on the right side of this equation is the same likelihood function that we previously derived in Equation 5.6.

The initial state vector ν represents the probabilities of updating the focal haplotype to any of the K weight states at the first SNP in the sequence. For convenience, we assign the states in ν to be uniformly distributed.

Given the HMM $\lambda = (A, B, \nu)$ and the sequence of observations $\sigma^2_{\hat{\beta}}$, we now calculate $P(\sigma^2_{\hat{\beta}}|\lambda)$ which represents the likelihood of the observed sequence $\sigma^2_{\hat{\beta}}$ given the model. In order to obtain $P(\sigma^2_{\hat{\beta}}|\lambda)$, we use the forward-pass algorithm. For $j = 1, 2, ..., N$ and $n = 1, 2, ..., K$, we define:

$$\alpha_{nj} = P(\sigma^2_{\hat{\beta}1}, \sigma^2_{\hat{\beta}2}, ..., \sigma^2_{\hat{\beta}j}, W_{ij} = w_n|\lambda) \tag{5.12}$$

[8]This is the recombination rate within the context of the HMM, but discuss later the relationship between this rate and the human level recombination rate (1e-8).

[9]We allow for relatively few weight changes along the region, effectively limiting the number of recombination events in the model.

where $W_{ij} = w_n$ corresponds to the focal haplotype update at SNP j to weight state n.

These likelihoods are stored in a matrix α. An element α_{nj} denotes the probability of the partial observation sequence up to SNP j where the underlying Markov process is in state w_n at SNP j. In traditional HMM fashion, we calculate α_{nj} recursively as the following:

1. Let $\alpha_{n1} = \nu_n b_n(\sigma^2_{\hat{\beta}1})$ for $n = 1, 2, ..., K$
2. For $j = 2, 3, ..., N$ and $n = 1, 2, ..., K$ compute: $\alpha_{nj} = [\sum_{l=1}^{K} \alpha_{l,j-1} a_{ln}] b_n(\sigma^2_{\hat{\beta}j})$
3. From the recursive definition of the forward-pass algorithm (Equation 5.12), we obtain $P(\sigma^2_{\hat{\beta}}|\lambda) = \sum_{n=1}^{K} \alpha_{nN}$

Using the forward matrix, we are now able to efficiently sample the weight states from the posterior distribution for the weights on the focal reference haplotype conditioned upon the frozen haplotypes (ie. standard Gibbs sampling). In order to do so, we sample the sequence of weight states beginning at the end of the sequence at SNP $j = N$ and ending at the first SNP $j = 1$. In this process, we infer a vector $k^{'}$ of length N. We perform the algorithm as follows:

1. Sample the weight state for the N^{th} SNP given the final forward-pass column of K probabilities (ie. $\alpha_{,N}$). This gives us $k^{'}_{N}$
2. Calculate the probabilities across all K weight states at the $(N-1)^{st}$ SNP conditional on the weight state chosen at SNP N. In order to calculate these probabilities, we first obtain the transition probabilities from the weight state $k^{'}_{N}$ to any of the other possible K states (ie. $a_{j=k^{'}_{N},}$). We multiply each of the K transition probabilities by their respective likelihoods given by the $(N-1)^{st}$ column in the forward-pass matrix (ie. $\alpha_{,N-1}$)

3. Sample a weight state for the $(N-1)^{st}$ SNP. This gives us k'_{N-1}. Next, we move one SNP backwards in the sequence, towards the start of the sequence

4. Repeat Steps 2 and 3 until we reach the first SNP in the sequence.

Using this process, we obtain the vector k' which denotes the path through the weight state space across all N variants. The existing vector of weights at the focal reference haplotype is replaced with weight states given by the indexes k'. This enables us to perform the marginal updates on the focal reference haplotype. We repeat this process and update the next focal haplotype in the same manner (See Chapter 3, Algorithm 1).

5.5.3 Linkage Disequilibrium Inference

Because the weights in W are no longer constant along the region, we are not able to rely on Equation (3.3) to calculate measurements of LD. Therefore, a method is required to infer the underlying LD described by a weight matrix W and a reference population R where the haplotype weights vary across the region given the HMM.

As we previously described, the weights which vary across the region represent the number of copies of the genetic segments of the reference haplotypes. In this section, we derive a method to stitch together these genetic segments to form the GWAS haplotypes. The stitching together of GWAS haplotypes that are generated from W and R is analogous to tracing a path through W. The number of paths that pass through each element in W matches the corresponding weight at each element.

We begin by defining two constraints on the paths that are permitted in our model. Firstly, paths can only traverse left to right across W. Secondly, paths must travel through each SNP (ie. column) in W. Implicitly, we rely on the underlying Markovian model through which we generate W (ie. a defined j to $j+1$ process) when we construct these paths. It is a useful abstraction to consider paths as conduits through which weights can *flow* to haplotypes at adjacent SNPs.

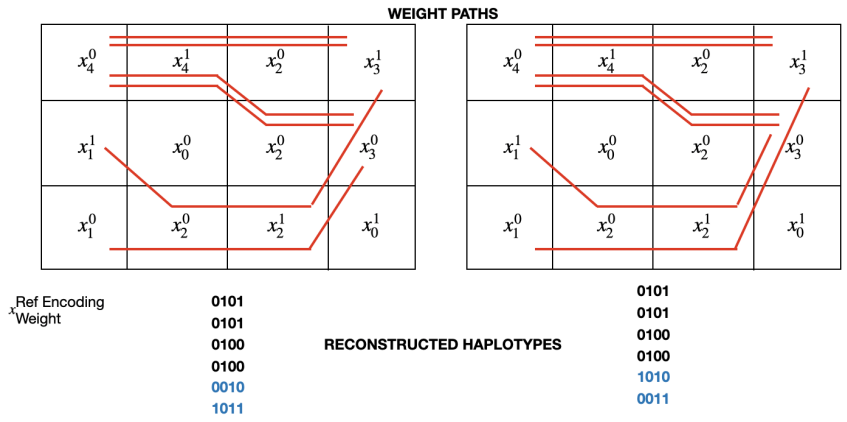

Figure 5.11: Illustrating two valid solutions of paths through W and R. The reconstructed haplotypes below demonstrate how the different paths will change the resulting inferred haplotypes (differences denoted in blue).

In other words, we aim to construct a directed graph where a node represents an element within W and the paths represent the edges. At each SNP j, each of the M_{ref} haplotypes transfer a proportion of their weights to each of the M_{ref} haplotypes at the next SNP $j+1$.

In most cases, the tracing of paths through W is a non-singular problem where multiple paths exist (Figure 5.11). To account for the non-singular nature of this inference, we integrate across all possible paths in W and obtain the *expected* set of paths through W. Because we often deal with normalised weights within the inference process rather than the absolute value of weights, these paths represent the expected flow of the proportion of weights from SNP j to SNP $j+1$ across each segment-segment connection.

The magnitude of this flow is stored in an array Q which has dimensions $(M_{ref}, M_{ref}, N-1)$. Along the depth of this array, we extract a matrix $Q^{j \to j+1}$ which describes the proportion of weight that flows from SNP j to the adjacent SNP $j+1$ between all M_{ref} haplotypes. The non-diagonal elements of the matrix $Q^{j \to j+1}$ represent weights exiting a haplotype h at SNP j and entering a different haplotype h' at SNP $j+1$. The diagonal elements represent the weight flowing *horizontally*, entering and exiting the same haplotype h. In cases when the weights

remain constant along the region, the array Q is filled with only identity matrices across each of the $N-1$ indexes. Importantly, when inferring LD, we aim to limit recombination (ie. the weights flowing non-horizontally between segment-segment connections) when possible.

We populate $Q^{j\rightarrow j+1}$ separately considering the diagonal and the non-diagonal elements. The formula for these two components is derived in a similar manner to the Markov global balance equations (Connors and Kumar 1989).

The diagonal elements in $Q^{j\rightarrow j+1}$ are calculated using Equation 5.13. Using the minimum value between the weights at haplotype h and h' between SNPs j and $j+1$ enforces minimum recombination by limiting non-horizontal flow of the weights.

$$Q_{hh}^{j\rightarrow j+1} = \frac{min(w_{hj}, w_{h,j+1})}{w_{hj}} \tag{5.13}$$

The non-diagonal elements in Q describe the flow from and to any pair of non-identical haplotypes h and h' at adjacent SNPs. The non-diagonal elements in $Q^{j\rightarrow j+1}$ are broken into two parts. First, we calculate the flow exiting haplotype h at SNP j (Equation 5.14). Secondly, we calculate the flow entering haplotype h' at SNP $j+1$ (Equation 5.15). These equations are given as:

$$\Delta_h^j = max(0, w_{hj} - w_{h,j+1}) \tag{5.14}$$

and

$$\Delta_{h'}^{j+1} = max(0, w_{h',j+1} - w_{h'j}). \tag{5.15}$$

To normalize the magnitude of the flow of the weights, we calculate the total flow entering and exiting each of the M_{ref} haplotypes as:

$$\Delta_{j+1} = \Sigma_{h=1}^{M_{ref}} \Delta_h^{j+1}. \tag{5.16}$$

Therefore, Equations 5.14, 5.15 and 5.16 combined allow us to calculate the non-diagonal component of $Q^{j \to j+1}$ as:

$$Q_{hh'}^{j \to j+1} = \frac{\Delta_h^j \Delta_h^{j+1}}{\Delta_{j+1} w_{h,}}. \tag{5.17}$$

Within our model, the covariance between a pair of SNPs represents the proportion of weights which flows from haplotypes h encoded 1 at SNP j to haplotypes h' also encoded 1 at SNP j'. We begin by deriving the covariance in cases where SNPs j and $j^{'}$ are adjacent. In other words, we consider SNPs j and $j+1$.

To obtain this covariance, we obtain the initial weight at all M_{ref} haplotypes at SNP j and then calculate the flow of these weights to the SNP $j+1$ given as $\text{diag}(w_{,j}^{\star}) \circ Q^{j \to j+1}$. However, as we are interested in only the segment-segment connections that contribute to the magnitude of the covariance (ie. haplotypes which are both encoded 1), we multiply this matrix-vector product by $(R_{,j} \otimes R_{,j+1})$. The resulting matrix M has dimensions (M_{ref} by M_{ref}) where each element corresponds to the marginal covariance between SNPs j and $j+1$ for each segment-segment connection. In summation, we write M as:

$$M = [\text{diag}(w_{,j}^{\star}) \circ Q^{j \to j+1}] \circ (R_{,j} \otimes R_{,j+1}), \tag{5.18}$$

where $\otimes$ represents the multiplicative outer product and $\circ$ represents the Hadamard product. Therefore, the total covariance between the SNPs j and $j+1$ across all segment-segment connections is given as:

$$cov_{j,j+1} = \sum_{p=1}^{M_{ref}} \sum_{q=1}^{M_{ref}} M_{pq}. \tag{5.19}$$

The process of calculating the covariance between non-adjacent SNPs (ie. *long range* covariance) is slightly more complex. This is due to the fact that Q encodes only the flow of weights between adjacent SNPs. Fortunately, since the weights are generated through the underlying Markov process, we are able to produce the *long*

range flow by calculating the cumulative matrix product across indexes along the depth of Q[10]. For instance, the long range flow between SNPs j and $j+2$ involves the matrix multiplication $Q^{j\to j+1}Q^{j+1\to j+2}$ to obtain $Q^{j\to j+2}$.

Therefore, in order to calculate the covariance between two non-adjacent SNPs, we first calculate the matrix C which describes the long range flow as:

$$C^{j\to j'} = \prod_{l=j}^{j'-l} Q^{l\to l+1}, \tag{5.20}$$

where $\prod$ denotes matrix multiplication. Therefore, we obtain a new expression for M as:

$$M = [\text{diag}(w^{\star}_{,j}) \circ C^{j\to j'}] \circ (R_{,j} \otimes R_{,j'}). \tag{5.21}$$

In summation, the covariance and the allele frequencies[11] under R and W are sufficient to calculate measurements of LD (Equation 1.1, 1.2)

5.5.4 Drawbacks of Recombining Weights

The advantages of extending *InferLD* to allow for weights that vary along the region come with a few drawbacks. Firstly, this extension leads to the number of columns in W scaling with the length of the region. Therefore, storing W in memory is computational taxing as the size of the region increases. Secondly, performing updates to W under the HMM also presents a significant computational challenge. For instance, performing the forward-pass algorithm under a best-case scenario[12] has run-time complexity $O(KN)$. Performing the backwards sampling also proves computationally challenging and requires repeated sampling of the probabilities in α to generate the weight state sequence.

[10]The long run behaviour of the transitions under a Markov chain can be calculated by cumulatively multiplying in succession each individual transition matrix (Pollett and Stewart 1994).

[11]Calculated as before when the weights remained constant using Equation 3.1.

[12]Recombination happening with equal probability to any other haplotype.

However, the most significant drawback of this extension is the loss of long range LD - a phenomenon we label as *leakage*. Leakage is a consequence of the memory-less Markov process which underlies the generation of Q. In this process, if long range LD exists in the GWAS population, there is no feasible manner to convey this information in the process of inferring Q. Observe that when we form Q, we implicitly average across the sets of possible paths (Figure 5.11). Therefore, we would expect that because there is only a maximum value of r of 1 that can be achieved from our model, this averaging will regress the LD we infer towards a value less than 1.

Briefly, we illustrate using the approaching of the recombining weights on real GWAS data from the UK Biobank and the 1000G AFR reference population and illustrate the appearance of this phenomenon. We choose to use the AFR reference population to pick a population where under the previous inference method when the weights remained constant we obtained a significant improvement of the correlation to the GWAS LD measurements over the unweighted reference population. We use the same hypertension GWAS summary statistics as described earlier in Chapter 4. We extract a single 1Mb region in these summary statistics with approximately 500 SNPs and extract the matching region within the EUR reference population. We filter and jointly remove any SNPs below 1% MAF in either population. Next, we perform inference using *InferLD* setting $\hat{F_{st}} = 0.001$, $\hat{\kappa} = 1e4$ and $\hat{K} = 1000$. The HMM model parameters are described above, and within our HMM transition matrix we set a value of ρ that corresponds to our recombination rate to 0.001. As mentioned before, we discuss within the conclusion a more robust method to set ρ. We discard the first 90% of the updates as burn-in and calculate the element-wise average weight matrix across the last 10% of samples and calculate the LD measurements using the methods derived in Section 5.5.3 (Equation 5.21).

We compare the LD heatmap of the r measures obtained from our method compared to LD obtained from the reference population and the true GWAS LD measurements. We further compare the correlation of the values to the true

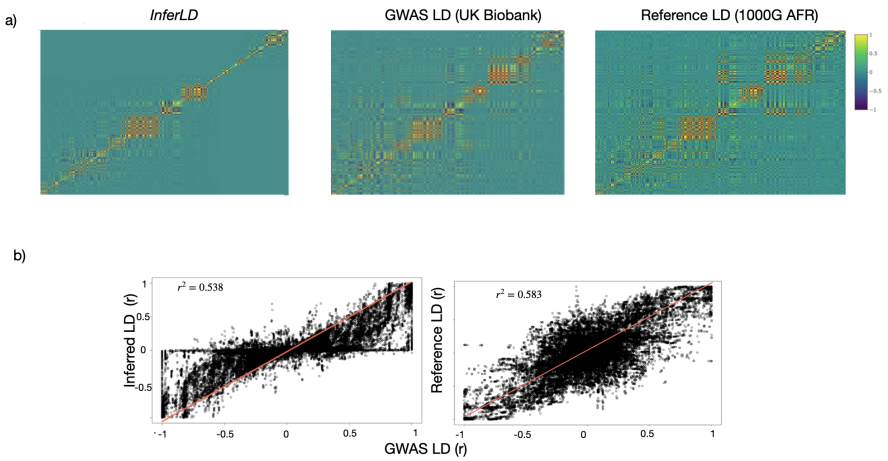

Figure 5.12: LD inference with *InferLD* using recombining weights. a) Heatmaps of the three LD matrices: *InferLD*, GWAS (UK Biobank) and reference (1000G AFR). Colour bar on the right indicates the LD measure r. b) Scatter-plot of the LD measures between the inferred LD and the GWAS LD (left) and the reference LD and the GWAS LD (right). Measurements of LD are given as r. Correlations are denoted in the upper left corner of the plots. The red line indicates perfect correlation y = x.

GWAS LD measurements in both the reference and inferred setting. These results are shown in Figure 5.12.

Overall, the main feature we observe in Figure 5.12 a) is the loss of long range LD when inferring the LD measurements using *InferLD*. In particular, we observe that the magnitude of this loss increases with distance between the SNPs, and at a far greater rate than would be expected by a simple LD decay by distance phenomenon (Pritchard and Przeworski 2001). Further, in Figure 5.12 b), we observe that the loss of the LD using *InferLD* results in the shrinkage of our inferred measurements such that we observe many SNP-SNP measurements to be zero when in the GWAS population they are non-zero. These features result in a lower of correlation that we obtain with *InferLD* relative to the reference population (r^2 0.538 vs 0.583). Importantly, note the extremely reduced correlation of using this extension for *InferLD* as compared to earlier when we forced the weights to remain constant along the region (Figure 5.7). As explained previously, these features are the result of LD leakage within our LD inference process. For this reason, we decided to not pursue the framework with recombining weights any further.

5.6 Conclusion

In this chapter, we developed a method named *InferLD* which infers the population genetic measurements from a GWAS population using the GWAS summary statistics and a reference population. Inference is performed by repeatedly updating the weights on the reference haplotypes with a likelihood function using Gibbs sampling.

By using GWAS summary statistics from both a generative model and the UK Biobank, our method was able to accurately infer the in-sample GWAS LD measurements and allele frequencies. We demonstrated that *InferLD* was able to improve fine-mapping accuracy using simulated admixed GWAS data and outcompete traditional reference populations. Finally, we showed that *InferLD* has the potential to leverage the distribution of LD measurements in order to improve the calibration of fine-mapping.

InferLD is similar to other methods which adjust reference populations to better capture measurements within the GWAS population given the GWAS summary statistics. In Chapter 1, we discussed the details of two of these methods, *DISTMIX* (D. Lee et al. 2015) and *Adapt-Mix* (Park et al. 2015). *InferLD* has several advantages over these methods. Firstly, we weight individual haplotypes in the reference population as compared to the sub-populations. Secondly, these two methods are tailored to produce only the imputed summary statistics, limiting their use. Instead, within *InferLD*, we directly infer the measurements of LD in the GWAS population, making our method much more flexible for downstream use in statistical genetic applications. Thirdly, these methods output a single point estimate of the LD measurements. In contrast, *InferLD* has the unique ability to generate the distribution of LD measurements which could prove valuable in downstream summary statistics analyses. One important direction for future work entails directly comparing *InferLD* to these methods in different simulated settings.

We observed that *InferLD* outcompetes the fine-mapping accuracy of different reference populations. However, these fine-mapping simulations lack a high degree of realistic genetic variation. In particular, the simulated GWAS admixed population is formed using only the reference populations (AFR and EUR). In real world settings, admixture involves an amalgamation of a plethora of different ancestries (Hellenthal et al. 2014). Furthermore, it is unrealistic to directly use the reference populations which form the simulated GWAS samples in the fine-mapping analyses. In commonly used reference populations including the 1000G Project, several complex layers of population structure are present (Henn et al. 2010). Therefore, using mixtures of the 1000G EUR and 1000G AFR sub-populations to model African-American admixture GWAS population will therefore likely result in lower accuracy in downstream analyses (Medina-Gomez et al. 2015). Consequently, an important future piece of work involves the evaluation of *InferLD* on more realistic simulated admixed GWAS data (Skotte et al. 2019).

Another direction for future investigation involves improving the inference of *InferLD*. More appropriate techniques could be developed in order to infer the LD measurements from W and R when the weights in W change along the region. One idea that is relatively easy to implement involves providing *InferLD* with a realistic human recombination map (Myers et al. 2005) for use in the HMM transition matrix. Utilizing this map in inference leads to weights changing state only in situations when the data is very strong. However, as we continually decrease the the opportunity of the weights to recombine we end up approaching the setting when the weights remain constant along the region. Therefore, a trade-off must be assessed in terms of the computational effort to run the HMM, the number of weight recombination events we infer (and if they provide much if any benefit) and the consequences of these events as they pertain to increased LD leakage and decreasing the accuracy of our inference. Another approach to improve the inference of LD from W is to develop an alternate algorithm that aims to minimize the

amount of recombination in the region. For instance, a greedy approach could be developed to infer a set of paths that matches the weights and minimizes the total number of recombination events.

Unfortunately, the ability of *InferLD* to obtain the distribution of LD measurements rather than just a single point estimate was shown to have limited effectiveness in fine-mapping applications. Note that in this chapter, we considered only the fine-mapping applications for our method. In Chapter 2, we described a pipeline to perform the joint summary statistic imputation and downstream fine-mapping. We suspect that performing joint analyses such as these would face the same challenges we have outlined here. Another limitation within the analyses in this chapter was the manner in which we implemented FINEMAP to incorporate distributions of LD measurements. Specifically, since FINEMAP was unable to directly incorporate the distributions of LD, we ran FINEMAP independently and averaged the results. We note that using the averaged FINEMAP results within *InferLD* (ie. as a distribution) produced well calibrated PIPs at roughly the same level as the point estimates for *InferLD*. Therefore, we suggest that a more nuanced approach and developing fine-mapping software that directly incorporates these distributions may prove more useful to leverage the information contained within the distributions of LD. In the next and final chapter, we discuss this approach in more detail.

InferLD is a worthwhile attempt to utilize existing reference population data in order to better analyse GWAS summary statistics. Most importantly, we hope that *InferLD* motivates researchers to consider modelling the distribution of measurements of LD rather than routinely relying on point estimates in order to provide additional robustness in their methods.

6 Conclusions

Contents

In this thesis, we aimed to better understand the impact of genetic drift on the distributions of LD and incorporate these distributions into statistical genetic applications. To finish, we first discuss the overall results from this work and the surprises and challenges we encountered along the way. Finally, we highlight outstanding problems that have arisen. Importantly, these problems are addressed within the context of how to best collect and analyze the next generation of genetic data.

6.1 Discussion

In Chapter 2, we described the distribution of LD using simulated coalescent data and the extent to which stochastic forces such as recombination and genetic drift play

a role in shaping these distributions. Although the development of this framework to generate this distribution through conditioning on pairs of SNPs was simple, it culminated in an understanding of the shape and characteristics of this distribution. We discovered the existence of a surprisingly high variation in LD measurements between different sets of samples drawn from a panmictic population. Genetic drift between populations further increased this variation. By using a coalescent simulator such as *msprime*, the development of this framework proves extremely flexible for assessing other demographic events besides the population split in our analyses. Further, in the context of statistical genetic applications, point estimates of LD fail to capture underlying variation and therefore leads to variation in downstream accuracy within these applications. These results specifically highlight the perils of using point estimates of LD in statistical genetic applications.

In Chapter 3, the haplotype weighting scheme was developed as a stepping stone for inferring properties of a GWAS population. By repeatedly sampling weights on the reference haplotypes in this scheme, we generated well calibrated distributions of both the drifted allele frequencies and the LD measurements within a population under a demographic history of a population split. Unfortunately, this scheme proved to be a poor fit for modelling genetic data simulated under the OOA model and proved to be even a poorer fit when applied to the UK Biobank. An important future research goal will be to understand the role of demographic events in the history of real world populations and the impact of ascertainment biases in our model. We concluded this chapter by developing an equivalent method to perform this scheme through discretizing the weight state space and performing marginal updates to our weight matrices. This framework motivated us to consider using the GWAS summary statistic data to infer properties of the GWAS population.

However, on the surface, summary statistics did not contain information required to perform inference. This led us to analyse two important quantities contained within these statistics including the drift relative to a reference population and the

level of noise in the GWAS population. We characterized the distribution of these two quantities and constructed a MLE enabling us to estimate these two quantities. One challenge we faced was the difficulty in incorporating these estimates within downstream inference. The drift and noise terms estimated in the MLE were not informative in setting the levels of drift and noise in *InferLD*.

Lastly, in Chapter 5, we aimed to account for uncertainty in the distribution of LD within the GWAS population for statistical genetic applications such as summary statistic imputation and fine-mapping. We used the haplotype weighting scheme along with the GWAS summary statistics to develop a Gibbs sampling algorithm that performed likelihood based updates to the weights. By inferring measurements of LD from a GWAS population using only the summary statistics, *InferLD* was shown to outcompete the accuracy of unweighted reference populations. Further, our method has the ability to account for the uncertainty within the distribution of LD within the GWAS population. However, an outstanding challenge remains to better leverage these distributions in fine-mapping applications. We observed that the use of simplistic ways of implementing this distribution into existing tools did not improve the calibration of downstream results.

6.2 Weighted Tree Sequences

In Chapter 3, we modelled genetic drift through our haplotype weighting scheme by considering an ancestral population which splits into a descendant GWAS and reference population. According to the genealogical interpretation of our model, the reference population samples were used to approximate the distribution of the ancestors to the GWAS population. This distribution was constructed by placing weights on the reference haplotypes to reflect this genealogical interpretation. In order to even more closely reflect the genealogical relationship of the GWAS and reference population, we can form this distribution through weighting the ancestral

nodes within the genealogical tree of the reference haplotypes. A tree is fundamental to biology and both encodes and summarises the outcomes of evolutionary processes (Kelleher, Y. Wong, et al. 2019). Currently, a large number of methods exist to infer trees from genetic data (Wooding 2004). Trees have previously been inferred from reference population data including the 1000G Project and the Simons Genome Diversity Project (Mallick et al. 2016).

We envisage that the weight on a node within the tree results in an increased or decreased number of descendant samples that fall below this node proportional to the weight. Tracing the weights on the nodes on the tree downwards to the descendants leads to generating copies of the reference haplotypes which represents the drifted GWAS population.

The placement of weights on the genealogical tree rather than on the descendant reference haplotypes requires a new inference method entirely different from the Gibbs sampling framework used in this thesis. In order to assess if a set of weights on the nodes within the genealogical tree is well fitted to the GWAS data, we may be able to reuse the likelihood function we derived in Chapter 5. One possible approach for inferring the set of weights on the nodes involves using a similar inference algorithm within *tsdate*. This algorithm uses a prior for the ancestral nodes in the tree sequence and a time grid to infer the age of nodes using a belief propagation approach based on an HMM where the age of nodes are hidden states (Wohns et al. 2021). Using a similar approach as *tsdate*, we would consider the discrete weights as the hidden states.

This process could be extended to allow for recombination. This will lead to the consideration of a *tree-sequence* rather than a single marginal tree (Kelleher, Y. Wong, et al. 2019). In this process, the weights vary across the different trees of the tree-sequence similar to how weights vary under the HMM framework presented in Chapter 5.

6.3 Development of Tools Leveraging Distributions of Linkage Disequilibrium

In this thesis, we described the variation in measurements of LD across populations. Consequently, we modelled and characterized the distribution of LD measurements to capture this variation in a statistical framework. However, existing statistical genetic tools do not have the ability to directly incorporate the distributions of LD. Here, we consider how tools employed within fine-mapping and rare variant association studies could be modified to more powerfully leverage the distribution of LD to improve robustness and calibration.

Firstly, we consider how existing fine-mapping approaches could be modified to more directly leverage these distributions. The most promising direction for this approach relies on the work of improving the coverage of credible sets in Bayesian genetic fine-mapping (Hutchinson et al. 2020). In this work, the authors presented a method to re-estimate the coverage of credible sets by repeatedly simulating GWAS summary statistics based on the SNP correlation structure under a given causal model. This method re-estimates the coverage of credible sets using rapid simulations based on the observed or estimated SNP correlation structure to find the *adjusted coverage estimate*. This work is extended to find the *adjusted credible sets*, which contains the smallest set of variants such that their adjusted coverage estimate meets the target coverage. However, in this method, the GWAS summary statistics are simulated using a multivariate normal approximation with a LD panel from a reference population. A worthwhile extension involves using the framework developed in this method but instead simulating the summary statistics using a correlation matrix obtained from *InferLD* given the reference population and the GWAS summary statistics.

Secondly, we consider how to leverage the distribution of measurements of LD within rare variant association studies. Rare variant studies are of great biological

interest to researchers. For instance, rare variants uncovered in these studies may provide insight into traits under negative selection (Gibson 2012). Due to the under-powered nature of such studies, information is aggregated across genes rather than searching for associations at every variant. This aggregation is performed through over-dispersion or burden tests (Auer and Lettre 2015). Within these rare variant analyses, the appropriateness of matching LD reference panels to the GWAS population is even more critical than in a traditional GWAS analysis with common variants (S. Lee et al. 2013). As above, summary statistic methods for rare variant association studies typically employ a multivariate normal approximation as the null distribution for the observed z-scores. Therefore, the application of *InferLD* to infer the in-sample LD more accurately than a reference population would prove useful.

6.4 Data: More Diverse and More Widely Shared

In Chapter 5, we illustrated that *InferLD* was well suited for the application of fine-mapping admixed GWAS data. However, as discussed in Chapter 1, there currently exists a relatively scarce availability of diverse and other admixed populations for evaluating our method (Martin et al. 2019).

Fortunately, there is hope on the horizon as researchers become increasingly aware of the necessity to analyze diverse populations in GWAS experiments (Wojcik et al. 2019). One recently curated repository of diverse genetic data is the Pan UK Biobank (`https://pan.ukbb.broadinstitute.org/`). This repository includes GWAS summary statistics analyzed from all samples within the UK Biobank. Importantly, more than 20,000 individuals with primarily non-European ancestries are included in these database. Even more importantly, summary statistics from this project were freely released to the research community. This database would prove a suitable benchmark for *InferLD* and other statistical genetic tools that aim to more accurately analyze summary statistics from diverse populations.

The genetic community is built upon a robust culture of data sharing (Byrd et al. 2020). The availability of publicly shared data has allowed for new and improved approaches that better analyze large collections of genetic data. Further, this availability offers opportunities for secondary analysis by researchers who were not involved in the original data collection. However, the increased ability to deanonymize public summary statistics and uniquely identify individuals within genetic studies is an ever-present risk for both the research community and the study participants (Piskol et al. 2013). Although this risk is partially mitigated through controlled access to the data through sharing platforms such as the database of Genotypes and Phenotypes (dbGaP) (`https://www.ncbi.nlm.nih.gov/gap/`), a substantial risk still exists (Deelen et al. 2015,Brouard et al. 2019,Piskol et al. 2013). The efficient and effective sharing of genetic data, including the logistical challenges involved, is an important consideration for researchers who are mindful of these risks.

We now consider how the work in this thesis allows for more efficient and effective sharing of LD matrices between researchers. Using only the GWAS summary statistics and a reference population, we demonstrated that *InferLD* obtains the distribution of the LD measurements from a GWAS population. These distributions (ie. series of LD samples) possess three important qualities that lead to efficient and effective data sharing. Firstly, these distributions provide more reliable information than a single point estimate of LD from a population. Secondly, the sharing of LD obtained from *InferLD* can be carried out in an extremely efficient manner. In the non-HMM version of *InferLD*, LD inferred from our method can be shared using the vector of weights across the reference haplotypes and the associated reference panel used during inference. Thirdly, the distribution of LD obtained by using a reference population avoids the privacy concerns involved with sharing GWAS LD.

In conclusion, our hope is that the work in this thesis does not only advance the understanding of the biological basis of complex traits and diseases, but also contributes to increasing data-sharing efforts in the field of statistical genetics.

Through sharing and collaboration, researchers are able to continually push the bounds of scientific research by improving the reproducibility and robustness of their discoveries.

References

Adulcikas, John et al. (Feb. 15, 2019). "Targeting the Zinc Transporter ZIP7 in the Treatment of Insulin Resistance and Type 2 Diabetes". In: *Nutrients* 11.2. pmid: 30781350.

Allen et al. (Oct. 14, 2010). "Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height". In: *Nature* 467.7317, pp. 832–838. pmid: 20881960.

Anderson-Trocmé, Luke et al. (Jan. 1, 2020). "Legacy Data Confound Genomics Studies". In: *Molecular Biology and Evolution* 37.1, pp. 2–10. URL: https://doi.org/10.1093/molbev/msz201 (visited on 02/16/2021).

Ardlie, Kristin G., Leonid Kruglyak, and Mark Seielstad (Apr. 2002). "Patterns of Linkage Disequilibrium in the Human Genome". In: *Nature Reviews Genetics* 3.4 (4), pp. 299–309. URL: https://www.nature.com/articles/nrg777 (visited on 05/12/2021).

Asimit, Jennifer L. et al. (Sept. 2016). "Trans-Ethnic Study Design Approaches for Fine-Mapping". In: *European Journal of Human Genetics* 24.9 (9), pp. 1330–1336. URL: https://www.nature.com/articles/ejhg20161 (visited on 05/13/2021).

Atkinson, Elizabeth G. et al. (Feb. 2021). "Tractor Uses Local Ancestry to Enable the Inclusion of Admixed Individuals in GWAS and to Boost Power". In: *Nature Genetics* 53.2 (2), pp. 195–204. URL: https://www.nature.com/articles/s41588-020-00766-y (visited on 06/20/2021).

Auer, Paul L. and Guillaume Lettre (Feb. 23, 2015). "Rare Variant Association Studies: Considerations, Challenges and Opportunities". In: *Genome Medicine* 7.1, p. 16. URL: https://doi.org/10.1186/s13073-015-0138-2 (visited on 06/21/2021).

Bai, Wei-Yang et al. (Sept. 25, 2020). "Genotype Imputation and Reference Panel: A Systematic Evaluation on Haplotype Size and Diversity". In: *Briefings in Bioinformatics* 21.5, pp. 1806–1817. URL: https://doi.org/10.1093/bib/bbz108 (visited on 05/13/2021).

Balding, David J. and Richard A. Nichols (June 1, 1995). "A Method for Quantifying Differentiation between Populations at Multi-Allelic Loci and Its Implications for Investigating Identity and Paternity". In: *Genetica* 96.1, pp. 3–12. URL: https://doi.org/10.1007/BF01441146 (visited on 05/13/2021).

Barton, N. H. and Sarah P. Otto (Apr. 2005). "Evolution of Recombination Due to Random Drift". In: *Genetics* 169.4, pp. 2353–2370. pmid: 15687279. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1449609/ (visited on 03/21/2021).

Barton, N. H. and M. Slatkin (June 1986). "A Quasi-Equilibrium Theory of the Distribution of Rare Alleles in a Subdivided Population". In: *Heredity* 56.3 (3), pp. 409–415. URL: https://www.nature.com/articles/hdy198663 (visited on 06/09/2021).

Benner, Christian et al. (May 15, 2016). "FINEMAP: Efficient Variable Selection Using Summary Data from Genome-Wide Association Studies". In: *Bioinformatics* 32.10, pp. 1493–1501. URL: https://doi.org/10.1093/bioinformatics/btw018 (visited on 05/13/2021).

Benner et al. (Oct. 5, 2017). "Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-Wide Association Studies". In: *American Journal of Human Genetics* 101.4, pp. 539–551. pmid: 28942963. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5630179/ (visited on 02/07/2021).

Berisa and Pickrell (Jan. 15, 2016). "Approximately Independent Linkage Disequilibrium Blocks in Human Populations". In: *Bioinformatics* 32.2, pp. 283–285. pmid: 26395773. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4731402/ (visited on 05/10/2021).

Betancourt, M. J. (2012). *Cruising The Simplex: Hamiltonian Monte Carlo and the Dirichlet Distribution*. arXiv: 1010.3436 [physics]. URL: http://arxiv.org/abs/1010.3436 (visited on 02/08/2021).

Bhatia, Gaurav et al. (Sept. 2013). "Estimating and Interpreting FST: The Impact of Rare Variants". In: *Genome Research* 23.9, pp. 1514–1521. pmid: 23861382. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3759727/ (visited on 05/26/2021).

Bodmer, Walter F. and Luigi L. Cavalli-Sforza (Aug. 1968). "A Migration Matrix Model for the Study of Random Genetic Drift". In: *Genetics* 59.4, pp. 565–592. pmid: 5708302. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1212023/ (visited on 06/08/2021).

Broekema, R. V., O. B. Bakker, and I. H. Jonkers (2021). "A Practical View of Fine-Mapping and Gene Prioritization in the Post-Genome-Wide Association Era". In: *Open Biology* 10.1 (), p. 190221. URL: https://royalsocietypublishing.org/doi/10.1098/rsob.190221 (visited on 05/26/2021).

Brouard, Jean-Simon et al. (June 21, 2019). "The GATK Joint Genotyping Workflow Is Appropriate for Calling Variants in RNA-Seq Experiments". In: *Journal of Animal Science and Biotechnology* 10.1, p. 44. URL: https://doi.org/10.1186/s40104-019-0359-0 (visited on 06/23/2021).

Browning, Brian L. and Sharon R. Browning (Jan. 7, 2016). "Genotype Imputation with Millions of Reference Samples". In: *The American Journal of Human Genetics* 98.1, pp. 116–126. URL: https://www.sciencedirect.com/science/article/pii/S0002929715004917 (visited on 06/18/2021).

Browning, Sharon R. and Brian L. Browning (Sept. 16, 2011). "Haplotype Phasing: Existing Methods and New Developments". In: *Nature Reviews. Genetics* 12.10, pp. 703–714. pmid: 21921926. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3217888/ (visited on 06/18/2021).

Browning et al. (May 24, 2018). "Ancestry-Specific Recent Effective Population Size in the Americas". In: *PLOS Genetics* 14.5, e1007385. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007385 (visited on 05/08/2021).

Bush, William S. and Jason H. Moore (Dec. 27, 2012). "Chapter 11: Genome-Wide Association Studies". In: *PLOS Computational Biology* 8.12, e1002822. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822 (visited on 04/05/2021).

Bycroft, Clare et al. (Oct. 2018). "The UK Biobank Resource with Deep Phenotyping and Genomic Data". In: *Nature* 562.7726 (7726), pp. 203–209. URL: https://www.nature.com/articles/s41586-018-0579-z (visited on 06/08/2021).

Byrd, James Brian et al. (Oct. 2020). "Responsible, Practical Genomic Data Sharing That Accelerates Research". In: *Nature Reviews Genetics* 21.10 (10), pp. 615–629. URL: https://www.nature.com/articles/s41576-020-0257-5 (visited on 06/23/2021).

Chan, Yvonne L., Christian N. K. Anderson, and Elizabeth A. Hadly (Apr. 21, 2006). "Bayesian Estimation of the Timing and Severity of a Population Bottleneck from Ancient DNA". In: *PLOS Genetics* 2.4, e59. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020059 (visited on 05/21/2021).

Charlesworth (Apr. 28, 2006). "Balancing Selection and Its Effects on Sequences in Nearby Genome Regions". In: *PLOS Genetics* 2.4, e64. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020064 (visited on 05/20/2021).

Charlesworth, Magnus Nordborg, and Deborah Charlesworth (Oct. 1997). "The Effects of Local Selection, Balanced Polymorphism and Background Selection on Equilibrium Patterns of Genetic Diversity in Subdivided Populations". In: *Genetics Research* 70.2, pp. 155–174. URL: https://www.cambridge.org/core/journals/genetics-research/article/effects-of-local-selection-balanced-polymorphism-and-background-selection-on-equilibrium-patterns-of-genetic-diversity-in-subdivided-populations/1F946DDC83F5277DC964D76329A6BB66 (visited on 05/25/2021).

Chen, Wenan et al. (Nov. 1, 2016). "Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics". In: *Genetics* 204.3, pp. 933–958. URL: https://doi.org/10.1534/genetics.116.188953 (visited on 05/10/2021).

Cockerham, C. Clark (1969). "Variance of Gene Frequencies". In: *Evolution* 23.1, pp. 72–84. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1969.tb03496.x (visited on 05/10/2021).

Connors, Daniel P. and P. R. Kumar (Nov. 1, 1989). "Simulated Annealing Type Markov Chains and Their Order Balance Equations". In: *SIAM Journal on Control and Optimization* 27.6, pp. 1440–1461. URL: https://epubs.siam.org/doi/abs/10.1137/0327074 (visited on 06/19/2021).

Cook, James P, Anubha Mahajan, and Andrew P Morris (Sept. 29, 2020). "Fine-Scale Population Structure in the UK Biobank: Implications for Genome-Wide Association Studies". In: *Human Molecular Genetics* 29.16, pp. 2803–2811. URL: https://doi.org/10.1093/hmg/ddaa157 (visited on 06/08/2021).

Coop, Graham and Molly Przeworski (Jan. 2007). "An Evolutionary View of Human Recombination". In: *Nature Reviews Genetics* 8.1 (1), pp. 23–34. URL: https://www.nature.com/articles/nrg1947 (visited on 05/24/2021).

Dangerfield, C. E. et al. (June 2012). "A Boundary Preserving Numerical Algorithm for the Wright-Fisher Model with Mutation". In: *BIT Numerical Mathematics* 52.2,

pp. 283–304. URL: http://link.springer.com/10.1007/s10543-011-0351-3 (visited on 05/26/2021).

Das, Sayantan et al. (Oct. 2016). "Next-Generation Genotype Imputation Service and Methods". In: *Nature Genetics* 48.10 (10), pp. 1284–1287. URL: https://www.nature.com/articles/ng.3656 (visited on 05/26/2021).

Deelen, Patrick et al. (Mar. 27, 2015). "Calling Genotypes from Public RNA-Sequencing Data Enables Identification of Genetic Variants That Affect Gene-Expression Levels". In: *Genome Medicine* 7.1, p. 30. URL: https://doi.org/10.1186/s13073-015-0152-4 (visited on 06/23/2021).

DeGiorgio, Michael, James H. Degnan, and Noah A. Rosenberg (Oct. 1, 2011). "Coalescence-Time Distributions in a Serial Founder Model of Human Evolutionary History". In: *Genetics* 189.2, pp. 579–593. pmid: 21775469. URL: https://www.genetics.org/content/189/2/579 (visited on 06/08/2021).

Dillon, Josh and Ian Langmore (Jan. 9, 2018). *Quadrature Compound: An Approximating Family of Distributions*. arXiv: 1801.03080 [stat]. URL: http://arxiv.org/abs/1801.03080 (visited on 02/20/2021).

Donnelly (Oct. 1986). "Partition Structures, Polya Urns, the Ewens Sampling Formula, and the Ages of Alleles". In: *Theoretical Population Biology* 30.2, pp. 271–288. pmid: 3787504.

Donnelly and Tavare (1995). "Coalescents and Genealogical Structure Under Neutrality". In: *Annual Review of Genetics* 29.1, pp. 401–421. pmid: 8825481. URL: https://doi.org/10.1146/annurev.ge.29.120195.002153 (visited on 02/08/2021).

Durbin, Richard et al. (Apr. 23, 1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. 332 pp. Google Books: HUUhAwAAQBAJ.

Eldon and Wakeley (Feb. 1, 2009). "Coalescence Times and F ST Under a Skewed Offspring Distribution Among Individuals in a Population". In: *Genetics* 181.2, pp. 615–629. URL: https://academic.oup.com/genetics/article/181/2/615/6062965 (visited on 06/24/2021).

Elhaik, Eran (Nov. 21, 2012). "Empirical Distributions of FST from Large-Scale Human Polymorphism Data". In: *PLOS ONE* 7.11, e49837. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049837 (visited on 04/04/2021).

Erlich, Yaniv and Arvind Narayanan (June 2014). "Routes for Breaching and Protecting Genetic Privacy". In: *Nature reviews. Genetics* 15.6, pp. 409–421. pmid: 24805122. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151119/ (visited on 05/24/2021).

Ethier, S. N. and Thomas Nagylaki (Feb. 1, 1989). "Diffusion Approximations of the Two-Locus Wright-Fisher Model". In: *Journal of Mathematical Biology* 27.1, pp. 17–28. URL: https://doi.org/10.1007/BF00276078 (visited on 05/26/2021).

Ewens, Warren J. (2004). *Mathematical Population Genetics 1: Theoretical Introduction*. 2nd ed. Interdisciplinary Applied Mathematics, Mathematical Population Genetics. New York: Springer-Verlag. URL: https://www.springer.com/gp/book/9780387201917 (visited on 05/23/2021).

Fadista, João et al. (Aug. 2016). "The (in)Famous GWAS P -Value Threshold Revisited and Updated for Low-Frequency Variants". In: *European Journal of Human Genetics*

24.8 (8), pp. 1202–1205. URL: https://www.nature.com/articles/ejhg2015269 (visited on 05/13/2021).

Faye, Laura L. et al. (Aug. 8, 2013). "Re-Ranking Sequencing Variants in the Post-GWAS Era for Accurate Causal Variant Identification". In: *PLOS Genetics* 9.8, e1003609. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003609 (visited on 05/13/2021).

Fisher, R. A. (1931). "The Evolution of Dominance". In: *Biological Reviews* 6.4, pp. 345–368. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-185X.1931.tb01030.x (visited on 03/02/2021).

Flannick, Jason et al. (Apr. 2014). "Loss-of-Function Mutations in SLC30A8 Protect against Type 2 Diabetes". In: *Nature Genetics* 46.4 (4), pp. 357–363. URL: https://www.nature.com/articles/ng.2915 (visited on 05/26/2021).

Fortune, Mary D and Chris Wallace (June 1, 2019). "simGWAS: A Fast Method for Simulation of Large Scale Case–Control GWAS Summary Statistics". In: *Bioinformatics* 35.11, pp. 1901–1906. URL: https://doi.org/10.1093/bioinformatics/bty898 (visited on 04/04/2021).

Foss, E. et al. (Mar. 1, 1993). "Chiasma Interference as a Function of Genetic Distance." In: *Genetics* 133.3, pp. 681–691. pmid: 8454209. URL: https://www.genetics.org/content/133/3/681 (visited on 05/20/2021).

Gentle, James E. (July 28, 2009). *Computational Statistics*. Springer Science & Business Media. 732 pp. Google Books: mQ5KAAAAQBAJ.

Gibson, Greg (Jan. 18, 2012). "Rare and Common Variants: Twenty Arguments". In: *Nature Reviews. Genetics* 13.2, pp. 135–145. pmid: 22251874.

Goldstein, David B. and Michael E. Weale (July 24, 2001). "Population Genomics: Linkage Disequilibrium Holds the Key". In: *Current Biology* 11.14, R576–R579. URL: https://www.sciencedirect.com/science/article/pii/S0960982201003487 (visited on 05/12/2021).

Good, Benjamin H. (Dec. 11, 2020). "Linkage Disequilibrium between Rare Mutations". In: *bioRxiv*, p. 2020.12.10.420042. URL: https://www.biorxiv.org/content/10.1101/2020.12.10.420042v1 (visited on 06/24/2021).

Griffiths (Apr. 1, 1981). "Neutral Two-Locus Multiple Allele Models with Recombination". In: *Theoretical Population Biology* 19.2, pp. 169–186. URL: https://www.sciencedirect.com/science/article/pii/0040580981900162 (visited on 05/26/2021).

Griffiths, Lessard Jenkins, and Sabin (Dec. 1, 2016). "A Coalescent Dual Process for a Wright–Fisher Diffusion with Recombination and Its Application to Haplotype Partitioning". In: *Theoretical Population Biology* 112, pp. 126–138. URL: https://www.sciencedirect.com/science/article/pii/S0040580916300491 (visited on 05/26/2021).

Griffiths and Tavaré (Aug. 1, 1994). "Ancestral Inference in Population Genetics". In: *Statistical Science* 9.

— (Jan. 1, 1998). "The Age of a Mutation in a General Coalescent". In: *Stochastic Models - STOCH MODELS* 14, pp. 273–295.

Guan, Yongtao and Matthew Stephens (Sept. 1, 2011). "Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems". In: *The Annals of Applied Statistics* 5.3. URL: https://projecteuclid.org/journals/annals-of-applied-

statistics/volume-5/issue-3/Bayesian-variable-selection-regression-for-genome-wide-association-studies-and/10.1214/11-AOAS455.full (visited on 05/10/2021).

Gutenkunst, Ryan N. et al. (Oct. 23, 2009). "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data". In: *PLOS Genetics* 5.10, e1000695. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000695 (visited on 02/09/2021).

Hans, Chris, Adrian Dobra, and Mike West (June 1, 2007). "Shotgun Stochastic Search for "Large p" Regression". In: *Journal of the American Statistical Association* 102.478, pp. 507–516. URL: https://amstat.tandfonline.com/doi/abs/10.1198/016214507000000121 (visited on 05/17/2021).

Hellenthal, Garrett et al. (Feb. 14, 2014). "A Genetic Atlas of Human Admixture History". In: *Science (New York, N.Y.)* 343.6172, pp. 747–751. pmid: 24531965.

Henn, Brenna M. et al. (Oct. 15, 2010). "Fine-Scale Population Structure and the Era of next-Generation Sequencing". In: *Human Molecular Genetics* 19.R2, R221–R226. URL: https://doi.org/10.1093/hmg/ddq403 (visited on 06/19/2021).

Hill, W. G. and Alan Robertson (June 1, 1968). "Linkage Disequilibrium in Finite Populations". In: *Theoretical and Applied Genetics* 38.6, pp. 226–231. URL: https://doi.org/10.1007/BF01245622 (visited on 05/11/2021).

Hill, W. G. and Weir (Feb. 1, 1988). "Variances and Covariances of Squared Linkage Disequilibria in Finite Populations". In: *Theoretical Population Biology* 33.1, pp. 54–78. URL: https://www.sciencedirect.com/science/article/pii/0040580988900044 (visited on 05/11/2021).

Holsinger, Kent E., and Bruce S. Weir (Sept. 2009). "Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting FST". In: *Nature reviews. Genetics* 10.9, pp. 639–650. pmid: 19687804. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4687486/ (visited on 02/08/2021).

Hoppe, Fred M. (June 1, 1987). "The Sampling Theory of Neutral Alleles and an Urn Model in Population Genetics". In: *Journal of Mathematical Biology* 25.2, pp. 123–159. URL: https://doi.org/10.1007/BF00276386 (visited on 05/26/2021).

Howie et al. (June 19, 2009). "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies". In: *PLOS Genetics* 5.6, e1000529. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000529 (visited on 05/23/2021).

Hudson (Apr. 1983). "Properties of a Neutral Allele Model with Intragenic Recombination". In: *Theoretical Population Biology* 23.2, pp. 183–201. pmid: 6612631.

— (Mar. 1985). "The Sampling Distribution of Linkage Disequilibrium under an Infinite Allele Model without Selection". In: *Genetics* 109.3, pp. 611–631. pmid: 3979817. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1216291/ (visited on 05/11/2021).

— (1995). "The Coalescent Process and Background Selection". In: *Philosophical Transactions: Biological Sciences* 349.1327, pp. 19–23. JSTOR: 56119.

Hurst, Laurence D. (Feb. 2009). "Genetics and the Understanding of Selection". In: *Nature Reviews Genetics* 10.2 (2), pp. 83–93. URL: https://www.nature.com/articles/nrg2506 (visited on 05/20/2021).

Hutchinson, Anna, Hope Watson, and Chris Wallace (Apr. 13, 2020). "Improving the Coverage of Credible Sets in Bayesian Genetic Fine-Mapping". In: *PLOS Computational Biology* 16.4, e1007829. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007829 (visited on 05/26/2021).

Jennings, H. S. (Mar. 1917). "The Numerical Results of Diverse Systems of Breeding, with Respect to Two Pairs of Characters, Linked or Independent, with Special Relation to the Effects of Linkage". In: *Genetics* 2.2, pp. 97–154. pmid: 17245880. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1193714/ (visited on 05/12/2021).

Jones, A. T., J. R. Ovenden, and Y.-G. Wang (Oct. 2016). "Improved Confidence Intervals for the Linkage Disequilibrium Method for Estimating Effective Population Size". In: *Heredity* 117.4 (4), pp. 217–223. URL: https://www.nature.com/articles/hdy201619 (visited on 05/25/2021).

Keinan, Alon et al. (Oct. 2007). "Measurement of the Human Allele Frequency Spectrum Demonstrates Greater Genetic Drift in East Asians than in Europeans". In: *Nature Genetics* 39.10 (10), pp. 1251–1255. URL: https://www.nature.com/articles/ng2116 (visited on 06/08/2021).

Kelleher, Jerome, Alison M. Etheridge, and Gilean McVean (May 4, 2016). "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes". In: *PLOS Computational Biology* 12.5, e1004842. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004842 (visited on 05/26/2021).

Kelleher, Jerome, Yan Wong, et al. (Sept. 2019). "Inferring Whole-Genome Histories in Large Population Datasets". In: *Nature Genetics* 51.9 (9), pp. 1330–1338. URL: https://www.nature.com/articles/s41588-019-0483-y (visited on 06/21/2021).

Kichaev, Gleb et al. (Oct. 30, 2014). "Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies". In: *PLOS Genetics* 10.10, e1004722. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004722 (visited on 05/10/2021).

Kimura and Crow (Apr. 1964). "The Number of Alleles That Can Be Maintained in a Finite Population". In: *Genetics* 49.4, pp. 725–738. pmid: 14156929. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210609/ (visited on 05/12/2021).

Kingman, J. F. C. (Sept. 1, 1982). "The Coalescent". In: *Stochastic Processes and their Applications* 13.3, pp. 235–248. URL: https://www.sciencedirect.com/science/article/pii/0304414982900114 (visited on 05/26/2021).

Kotz, Samuel and N. Balakrishnan (1997). "Advances in Urn Models during the Past Two Decades". In: *Advances in Combinatorial Methods and Applications to Probability and Statistics*. Ed. by N. Balakrishnan. Statistics for Industry and Technology. Boston, MA: Birkhäuser, pp. 203–257. URL: https://doi.org/10.1007/978-1-4612-4140-9_14 (visited on 05/12/2021).

Kruglyak, L. (June 1999). "Prospects for Whole-Genome Linkage Disequilibrium Mapping of Common Disease Genes". In: *Nature Genetics* 22.2, pp. 139–144. pmid: 10369254.

Laurie, Cathy C. et al. (Sept. 2010). "Quality Control and Quality Assurance in Genotypic Data for Genome-Wide Association Studies". In: *Genetic epidemiology* 34.6, pp. 591–602. pmid: 20718045. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3061487/ (visited on 05/23/2021).

Lee, Donghyung et al. (Oct. 1, 2015). "DISTMIX: Direct Imputation of Summary Statistics for Unmeasured SNPs from Mixed Ethnicity Cohorts". In: *Bioinformatics (Oxford, England)* 31.19, pp. 3099–3104. pmid: 26059716.

Lee, Seunggeun et al. (July 11, 2013). "General Framework for Meta-Analysis of Rare Variants in Sequencing Association Studies". In: *American Journal of Human Genetics* 93.1, pp. 42–53. pmid: 23768515. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3710762/ (visited on 05/15/2021).

Li, Yun et al. (2009). "Genotype Imputation". In: *Annual Review of Genomics and Human Genetics* 10.1, pp. 387–406. pmid: 19715440. URL: https://doi.org/10.1146/annurev.genom.9.081307.164242 (visited on 05/13/2021).

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

Li and Brendan Yu Keating (Oct. 31, 2014). "Trans-Ethnic Genome-Wide Association Studies: Advantages and Challenges of Mapping in Diverse Populations". In: *Genome Medicine* 6.10, p. 91. URL: https://doi.org/10.1186/s13073-014-0091-5 (visited on 05/26/2021).

Loh, Po-Ru et al. (Mar. 2015). "Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts". In: *Nature Genetics* 47.3 (3), pp. 284–290. URL: https://www.nature.com/articles/ng.3190 (visited on 04/04/2021).

Lynch, Michael et al. (Nov. 2016). "Genetic Drift, Selection and the Evolution of the Mutation Rate". In: *Nature Reviews Genetics* 17.11 (11), pp. 704–714. URL: https://www.nature.com/articles/nrg.2016.104 (visited on 05/26/2021).

Lyon, Matthew S. et al. (Jan. 13, 2021). "The Variant Call Format Provides Efficient and Robust Storage of GWAS Summary Statistics". In: *Genome Biology* 22.1, p. 32. URL: https://doi.org/10.1186/s13059-020-02248-0 (visited on 04/04/2021).

Maller, Julian B. et al. (Dec. 2012). "Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases". In: *Nature Genetics* 44.12 (12), pp. 1294–1301. URL: https://www.nature.com/articles/ng.2435 (visited on 05/26/2021).

Mallick, Swapan et al. (Oct. 2016). "The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations". In: *Nature* 538.7624 (7624), pp. 201–206. URL: https://www.nature.com/articles/nature18964 (visited on 06/21/2021).

Marchini, Jonathan et al. (July 2007). "A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes". In: *Nature Genetics* 39.7 (7), pp. 906–913. URL: https://www.nature.com/articles/ng2088 (visited on 05/11/2021).

Marchini and Howie (July 2010). "Genotype Imputation for Genome-Wide Association Studies". In: *Nature Reviews Genetics* 11.7 (7), pp. 499–511. URL: https://www.nature.com/articles/nrg2796 (visited on 05/26/2021).

Marees, Andries T. et al. (Feb. 27, 2018). "A Tutorial on Conducting Genome-wide Association Studies: Quality Control and Statistical Analysis". In: *International Journal of Methods in Psychiatric Research* 27.2. pmid: 29484742. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/ (visited on 06/09/2021).

Martin, Alicia R. et al. (Apr. 2019). "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities". In: *Nature Genetics* 51.4 (4), pp. 584–591. URL: https://www.nature.com/articles/s41588-019-0379-x (visited on 05/26/2021).

McVean (Oct. 2002). "A Genealogical Interpretation of Linkage Disequilibrium". In: *Genetics* 162.2, pp. 987–991. pmid: 12399406.

— (Mar. 1, 2007). "The Structure of Linkage Disequilibrium Around a Selective Sweep". In: *Genetics* 175.3, pp. 1395–1406. pmid: 17194788. URL: https://www.genetics.org/content/175/3/1395 (visited on 08/06/2020).

Medina-Gomez, Carolina et al. (2015). "Challenges in Conducting Genome-Wide Association Studies in Highly Admixed Multi-Ethnic Populations: The Generation R Study". In: *European Journal of Epidemiology* 30.4, pp. 317–330. pmid: 25762173. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4385148/ (visited on 06/19/2021).

Meirmans, Patrick G. and Philip W. Hedrick (2011). "Assessing Population Structure: FST and Related Measures". In: *Molecular Ecology Resources* 11.1, pp. 5–18. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2010.02927.x (visited on 05/22/2021).

Mercader, Josep M. and Jose C. Florez (2017). "The Genetic Basis of Type 2 Diabetes in Hispanics and Latin Americans: Challenges and Opportunities". In: *Frontiers in Public Health* 5. URL: https://www.frontiersin.org/articles/10.3389/fpubh.2017.00329/full (visited on 06/25/2021).

Morales, Joannella et al. (Feb. 15, 2018). "A Standardized Framework for Representation of Ancestry Data in Genomics Studies, with Application to the NHGRI-EBI GWAS Catalog". In: *Genome Biology* 19.1, p. 21. URL: https://doi.org/10.1186/s13059-018-1396-2 (visited on 05/10/2021).

Morris, Andrew P et al. (Sept. 2012). "Large-Scale Association Analysis Provides Insights into the Genetic Architecture and Pathophysiology of Type 2 Diabetes". In: *Nature Genetics* 44.9 (9), pp. 981–990. URL: https://www.nature.com/articles/ng.2383 (visited on 04/04/2021).

Mount, David W. (July 2009). "Using Hidden Markov Models to Align Multiple Sequences". In: *Cold Spring Harbor Protocols* 2009.7, pdb.top41. pmid: 20147223.

Myers, Simon et al. (Oct. 14, 2005). "A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome". In: *Science* 310.5746, pp. 321–324. pmid: 16224025. URL: https://science.sciencemag.org/content/310/5746/321 (visited on 06/19/2021).

Nei and Li (Sept. 1973). "Linkage Disequilibrium in Subdivided Populations". In: *Genetics* 75.1, pp. 213–219. pmid: 4762877.

Nicholson, George et al. (2002). "Assessing Population Differentiation and Isolation from Single-Nucleotide Polymorphism Data". In: *Journal of the Royal Statistical Society:*

Series B (Statistical Methodology) 64.4, pp. 695–715. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00357 (visited on 05/10/2021).

Nielsen, Rasmus et al. (June 1998). "MAXIMUM-LIKELIHOOD ESTIMATION OF POPULATION DIVERGENCE TIMES AND POPULATION PHYLOGENY IN MODELS WITHOUT MUTATION". In: *Evolution; International Journal of Organic Evolution* 52.3, pp. 669–677. pmid: 28565245.

Nordborg and Tavaré (Feb. 2002). "Linkage Disequilibrium: What History Has to Tell Us". In: *Trends in Genetics* 18.2, pp. 83–90. URL: https://linkinghub.elsevier.com/retrieve/pii/S016895250202557X (visited on 08/06/2020).

Novembre, John et al. (Nov. 2008). "Genes Mirror Geography within Europe". In: *Nature* 456.7218 (7218), pp. 98–101. URL: https://www.nature.com/articles/nature07331 (visited on 05/10/2021).

Ohta (Mar. 1, 1982). "Linkage Disequilibrium Due to Random Genetic Drift in Finite Subdivided Populations". In: *Proceedings of the National Academy of Sciences* 79.6, pp. 1940–1944. pmid: 16593171. URL: https://www.pnas.org/content/79/6/1940 (visited on 05/12/2021).

Ohta and Kimura (Sept. 1969). "Linkage Disequilibrium at Steady State Determined by Random Genetic Drift and Recurrent Mutation". In: *Genetics* 63.1, pp. 229–238. pmid: 5365295.

— (Aug. 1971). "Linkage Disequilibrium between Two Segregating Nucleotide Sites under the Steady Flux of Mutations in a Finite Population". In: *Genetics* 68.4, pp. 571–580. pmid: 5120656. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1212677/ (visited on 05/22/2021).

Park, Danny S. et al. (June 15, 2015). "Adapt-Mix: Learning Local Genetic Correlation Structure Improves Summary Statistics-Based Analyses". In: *Bioinformatics* 31.12, pp. i181–i189. URL: https://doi.org/10.1093/bioinformatics/btv230 (visited on 05/13/2021).

Pasaniuc, Bogdan and Alkes L. Price (Feb. 2017). "Dissecting the Genetics of Complex Traits Using Summary Association Statistics". In: *Nature Reviews Genetics* 18.2, pp. 117–127. URL: http://www.nature.com/articles/nrg.2016.142 (visited on 04/04/2021).

Pasaniuc, Bogdan, Noah Zaitlen, et al. (Oct. 15, 2014). "Fast and Accurate Imputation of Summary Statistics Enhances Evidence of Functional Enrichment". In: *Bioinformatics* 30.20, pp. 2906–2914. URL: https://doi.org/10.1093/bioinformatics/btu416 (visited on 05/26/2021).

Pickrell and Pritchard (2012). "Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data". In: *PLoS genetics* 8.11, e1002967. pmid: 23166502.

Piskol, Robert, Gokul Ramaswami, and Jin Billy Li (Oct. 3, 2013). "Reliable Identification of Genomic Variants from RNA-Seq Data". In: *The American Journal of Human Genetics* 93.4, pp. 641–651. URL: https://www.sciencedirect.com/science/article/pii/S0002929713003832 (visited on 06/23/2021).

Pollett, P. K. and D. E. Stewart (1994). "An Efficient Procedure for Computing Quasi-Stationary Distributions of Markov Chains with Sparse Transition Structure". In: *Advances in Applied Probability* 26.1, pp. 68–79. JSTOR: 1427580.

Pritchard and Przeworski (July 2001). "Linkage Disequilibrium in Humans: Models and Data". In: *American Journal of Human Genetics* 69.1, pp. 1–14. pmid: 11410837. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1226024/ (visited on 05/09/2021).

Pritchard, Matthew Stephens, and Peter Donnelly (June 1, 2000). "Inference of Population Structure Using Multilocus Genotype Data". In: *Genetics* 155.2, pp. 945–959. pmid: 10835412. URL: https://www.genetics.org/content/155/2/945 (visited on 02/07/2021).

Reich, D. E. et al. (May 10, 2001). "Linkage Disequilibrium in the Human Genome". In: *Nature* 411.6834, pp. 199–204. pmid: 11346797.

Rogers, Alan R (Aug. 1, 2014). "How Population Growth Affects Linkage Disequilibrium". In: *Genetics* 197.4, pp. 1329–1341. URL: https://doi.org/10.1534/genetics.114.166454 (visited on 05/15/2021).

Rubin, Donald B. (Jan. 1981). "The Bayesian Bootstrap". In: *The Annals of Statistics* 9.1, pp. 130–134. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-1/The-Bayesian-Bootstrap/10.1214/aos/1176345338.full (visited on 05/18/2021).

Rüeger, Sina, Aaron McDaid, and Zoltán Kutalik (May 21, 2018). "Evaluation and Application of Summary Statistic Imputation to Discover New Height-Associated Loci". In: *PLoS Genetics* 14.5. pmid: 29782485. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5983877/ (visited on 05/31/2021).

Sajantila, A et al. (Oct. 15, 1996). "Paternal and Maternal DNA Lineages Reveal a Bottleneck in the Founding of the Finnish Population." In: *Proceedings of the National Academy of Sciences of the United States of America* 93.21, pp. 12035–12039. pmid: 8876258. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC38178/ (visited on 02/08/2021).

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson (Aug. 2018). "From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping". In: *Nature Reviews Genetics* 19.8 (8), pp. 491–504. URL: https://www.nature.com/articles/s41576-018-0016-z (visited on 05/26/2021).

Schierup, Mikkel, Deborah Charlesworth, and Xavier Vekemans (Sept. 1, 2000). "Schierup, MH, Vekemans, X, Charlesworth, D. The Effect of Hitchhiking on Genes Linked to a Balanced Polymorphism in a Subdivided Population. Genet Res, 76: 63-73". In: *Genetical research* 76, pp. 63–73.

Schiffels, Stephan and Richard Durbin (Aug. 2014). "Inferring Human Population Size and Separation History from Multiple Genome Sequences". In: *Nature genetics* 46.8, pp. 919–925. pmid: 24952747. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4116295/ (visited on 05/20/2021).

Schmegner, Claudia et al. (Nov. 2005). "Genetic Variability in a Genomic Region with Long-Range Linkage Disequilibrium Reveals Traces of a Bottleneck in the History of the European Population". In: *Human Genetics* 118.2, pp. 276–286. pmid: 16184404.

Schwarze, Katharina et al. (Oct. 2018). "Are Whole-Exome and Whole-Genome Sequencing Approaches Cost-Effective? A Systematic Review of the Literature". In: *Genetics in Medicine* 20.10 (10), pp. 1122–1130. URL: https://www.nature.com/articles/gim2017247 (visited on 05/23/2021).

Scott, W. K. et al. (Mar. 2000). "Fine Mapping of the Chromosome 12 Late-Onset Alzheimer Disease Locus: Potential Genetic and Phenotypic Heterogeneity". In: *American Journal of Human Genetics* 66.3, pp. 922–932. pmid: 10712207.

Shi, Huwenbo et al. (June 4, 2020). "Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data". In: *American Journal of Human Genetics* 106.6, pp. 805–817. pmid: 32442408.

Skotte, Line et al. (2019). "Ancestry-Specific Association Mapping in Admixed Populations". In: *Genetic Epidemiology* 43.5, pp. 506–521. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22200 (visited on 06/19/2021).

Slatkin (May 15, 1987). "Gene Flow and the Geographic Structure of Natural Populations". In: *Science* 236.4803, pp. 787–792. pmid: 3576198. URL: https://science.sciencemag.org/content/236/4803/787 (visited on 05/22/2021).

— (May 1994). "Linkage Disequilibrium in Growing and Stable Populations". In: *Genetics* 137.1, pp. 331–336. pmid: 8056320.

— (June 2008). "Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future". In: *Nature Reviews Genetics* 9.6 (6), pp. 477–485. URL: https://www.nature.com/articles/nrg2361 (visited on 05/26/2021).

Smith, J. M. and J. Haigh (Feb. 1974). "The Hitch-Hiking Effect of a Favourable Gene". In: *Genetical Research* 23.1, pp. 23–35. pmid: 4407212.

Spain, Sarah L. and Jeffrey C. Barrett (Oct. 15, 2015). "Strategies for Fine-Mapping Complex Traits". In: *Human Molecular Genetics* 24.R1, R111–R119. URL: https://doi.org/10.1093/hmg/ddv260 (visited on 05/17/2021).

Stephan, Wolfgang, Yun S. Song, and Charles H. Langley (Apr. 1, 2006). "The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci". In: *Genetics* 172.4, pp. 2647–2663. pmid: 16452153. URL: https://www.genetics.org/content/172/4/2647 (visited on 05/25/2021).

Su, Zhan, Jonathan Marchini, and Peter Donnelly (Aug. 15, 2011). "HAPGEN2: Simulation of Multiple Disease SNPs". In: *Bioinformatics* 27.16, pp. 2304–2305. URL: https://doi.org/10.1093/bioinformatics/btr341 (visited on 05/17/2021).

Tajima, F. (Oct. 1993). "Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis". In: *Genetics* 135.2, pp. 599–607. pmid: 8244016. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205659/ (visited on 05/12/2021).

Takahata and Nei (July 1984). "FST and GST Statistics in the Finite Island Model". In: *Genetics* 107.3, pp. 501–504. pmid: 17246223. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1202339/ (visited on 05/26/2021).

Tam, Vivian et al. (Aug. 2019). "Benefits and Limitations of Genome-Wide Association Studies". In: *Nature Reviews Genetics* 20.8, pp. 467–484. URL: http://www.nature.com/articles/s41576-019-0127-1 (visited on 05/26/2021).

Tataru, Paula, Thomas Bataillon, and Asger Hobolth (Nov. 1, 2015). "Inference Under a Wright-Fisher Model Using an Accurate Beta Approximation". In: *Genetics* 201.3,

pp. 1133–1141. pmid: 26311474. URL: https://www.genetics.org/content/201/3/1133 (visited on 05/26/2021).

Thornton, Kevin and Peter Andolfatto (Mar. 2006). "Approximate Bayesian Inference Reveals Evidence for a Recent, Severe Bottleneck in a Netherlands Population of Drosophila Melanogaster". In: *Genetics* 172.3, pp. 1607–1619. pmid: 16299396.

Tran, Tat Dat, Julian Hofrichter, and Jürgen Jost (2013). "An Introduction to the Mathematical Structure of the Wright–Fisher Model of Population Genetics". In: *Theory in Biosciences* 132.2, pp. 73–82. pmid: 23239077. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4269093/ (visited on 05/26/2021).

Visscher, Peter M. et al. (July 6, 2017). "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *American Journal of Human Genetics* 101.1, pp. 5–22. pmid: 28686856. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501872/ (visited on 05/26/2021).

Vukcevic, Damjan et al. (2011). "Disease Model Distortion in Association Studies". In: *Genetic Epidemiology* 35.4, pp. 278–290. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20576 (visited on 06/11/2021).

Wakeley (Feb. 1, 2009). "Coalescent Theory: An Introduction". In: 58.

Wallace, Chris (Apr. 20, 2020). "Eliciting Priors and Relaxing the Single Causal Variant Assumption in Colocalisation Analyses". In: *PLOS Genetics* 16.4, e1008720. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008720 (visited on 05/26/2021).

Wallace, Chris et al. (June 24, 2015). "Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping". In: *PLOS Genetics* 11.6, e1005272. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005272 (visited on 05/25/2021).

Wang, Gao et al. (2020). "A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.5, pp. 1273–1300. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12388 (visited on 05/26/2021).

Watterson, G. A. (Apr. 1, 1975). "On the Number of Segregating Sites in Genetical Models without Recombination". In: *Theoretical Population Biology* 7.2, pp. 256–276. URL: https://www.sciencedirect.com/science/article/pii/0040580975900209 (visited on 05/12/2021).

— (1976). "The Stationary Distribution of the Infinitely-Many Neutral Alleles Diffusion Model". In: *Journal of Applied Probability* 13.4, pp. 639–651. JSTOR: 3212519.

— (Dec. 1, 1984). "Allele Frequencies after a Bottleneck". In: *Theoretical Population Biology* 26.3, pp. 387–407. URL: https://www.sciencedirect.com/science/article/pii/004058098490042X (visited on 06/09/2021).

Weir (1979). "Inferences about Linkage Disequilibrium". In: *Biometrics* 35.1, pp. 235–254. JSTOR: 2529947.

Weir and C. Clark Cockerham (Nov. 1969). "Group Inbreeding with Two Linked Loci". In: *Genetics* 63.3, pp. 711–742. pmid: 5399257. URL:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1212379/ (visited on 05/11/2021).

Weir and C. Clark Cockerham (1984). "Estimating F-Statistics for the Analysis of Population Structure". In: *Evolution* 38.6, pp. 1358–1370. JSTOR: 2408641.

Wen, Xiaoquan and Matthew Stephens (Sept. 2010). "USING LINEAR PREDICTORS TO IMPUTE ALLELE FREQUENCIES FROM SUMMARY OR POOLED GENOTYPE DATA". In: *The annals of applied statistics* 4.3, pp. 1158–1182. pmid: 21479081. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3072818/ (visited on 05/08/2021).

Wiehe, Thomas et al. (Feb. 2000). "Distinguishing Recombination and Intragenic Gene Conversion by Linkage Disequilibrium Patterns". In: *Genetics Research* 75.1, pp. 61–73. URL: https://www.cambridge.org/core/journals/genetics-research/article/distinguishing-recombination-and-intragenic-gene-conversion-by-linkage-disequilibrium-patterns/CE7925219E4C0D02CADBE1EDF12FD5CF (visited on 05/11/2021).

Wilson, Melanie A. et al. (Sept. 2010). "Bayesian Model Search and Multilevel Inference for SNP Association Studies". In: *The Annals of Applied Statistics* 4.3, pp. 1342–1364. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-3/Bayesian-model-search-and-multilevel-inference-for-SNP-association-studies/10.1214/09-AOAS322.full (visited on 05/10/2021).

Wohns, Anthony Wilder et al. (Apr. 15, 2021). "A Unified Genealogy of Modern and Ancient Genomes". In: *bioRxiv*, p. 2021.02.16.431497. URL: https://www.biorxiv.org/content/10.1101/2021.02.16.431497v2 (visited on 06/21/2021).

Wojcik, Genevieve L. et al. (June 2019). "Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits". In: *Nature* 570.7762 (7762), pp. 514–518. URL: https://www.nature.com/articles/s41586-019-1310-4 (visited on 06/21/2021).

Wong (Dec. 15, 1998). "Generalized Dirichlet Distribution in Bayesian Analysis". In: *Applied Mathematics and Computation* 97.2, pp. 165–181. URL: https://www.sciencedirect.com/science/article/pii/S0096300397101400 (visited on 06/08/2021).

Wooding, Stephen (May 2004). "Inferring Phylogenies." In: *American Journal of Human Genetics* 74.5, p. 1074. pmid: null. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1181971/ (visited on 06/21/2021).

Wright (Mar. 1, 1931). "Evolution in Mendelian Populations". In: *Genetics* 16.2, pp. 97–159. pmid: 17246615. URL: https://www.genetics.org/content/16/2/97 (visited on 03/02/2021).

— (Dec. 1, 1945). "The Differential Equation of the Distribution of Gene Frequencies". In: *Proceedings of the National Academy of Sciences* 31.12, pp. 382–389. pmid: 16588707. URL: https://www.pnas.org/content/31/12/382 (visited on 05/26/2021).

— (Mar. 1951). "The Genetical Structure of Populations". In: *Annals of Eugenics* 15.4, pp. 323–354. pmid: 24540312.

Wu, Yue, Eleazar Eskin, and Sriram Sankararaman (Feb. 13, 2020). "A Unifying Framework for Imputing Summary Statistics in Genome-Wide Association Studies".

In: *Journal of Computational Biology* 27.3, pp. 418–428. URL: https://www.liebertpub.com/doi/full/10.1089/cmb.2019.0449 (visited on 05/17/2021).

Zhang, Weihua et al. (Dec. 28, 2004). "Impact of Population Structure, Effective Bottleneck Time, and Allele Frequency on Linkage Disequilibrium Maps". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.52, pp. 18075–18080. pmid: 15604137.

Zheng, Jie et al. (Jan. 15, 2017). "LD Hub: A Centralized Database and Web Interface to Perform LD Score Regression That Maximizes the Potential of Summary Level GWAS Data for SNP Heritability and Genetic Correlation Analysis". In: *Bioinformatics* 33.2, pp. 272–279. URL: https://doi.org/10.1093/bioinformatics/btw613 (visited on 05/26/2021).

Zhu, Xiang and Matthew Stephens (2017). "BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES". In: *The Annals of Applied Statistics* 11.3, pp. 1561–1592. pmid: 29399241.

— (Oct. 19, 2018). "Large-Scale Genome-Wide Enrichment Analyses Identify New Trait-Associated Genes and Pathways across 31 Human Phenotypes". In: *Nature Communications* 9.1 (1), p. 4361. URL: https://www.nature.com/articles/s41467-018-06805-x (visited on 05/23/2021).

Zuvich, Rebecca L. et al. (Dec. 2011). "Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality". In: *Genetic epidemiology* 35.8, pp. 887–898. pmid: 22125226. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3592376/ (visited on 05/10/2021).