A systematic comparison of methods for decoding the genetic architecture of phenotypic traits

Laura Hardy

Supervised by Eric Stone

Word count: 9394

Introduction: 2019

Results: 1929

Discussion: 5446

A thesis submitted in partial fulfilment of the degree of

Bachelor of Philosophy (Science) (Honours)

Research School of Biology

The Australian National University

October 2023

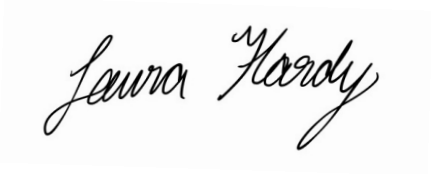
Abstract

A central pursuit of genetics is to infer the genetic basis of phenotypic variation. Fundamental to this challenge is the elucidation of genetic architecture, comprised of the number, location and relative effect of genes on a phenotypic trait. Many methods have been developed to resolve this covariation in genes and traits, and in this thesis we will systematically compare three: the standard variant-by-variant Genome-Wide Association Study (GWAS), a well-known variable selection approach (Least Absolute Shrinkage and Selection Operator, or LASSO), and a novel method we call the E-value. To achieve this, we propose a new framework for isolating a discrete set of factors influencing the performance of such methods so that their individual and synergistic impacts can be quantified. In this framework, we control population structure, population size, the number of variants, and the number and heritability of trait-associated variants. Then for each configuration of these parameters, over many replicates, we generate genotype and phenotype and quantify how well their simulated relationship is resolved by each method. Although GWAS remains a common choice for initial analyses, our results indicate that LASSO is far better suited to the challenge of resolving genetic architectures. In contrast, our new method — the E-value — performed unexpectedly poorly. The framework we contribute, and the inference it empowers, enhances our understanding of the methodologies used to resolve genotype-phenotype relationships. We hope this translates to improvements in future studies seeking to understand or leverage the genetic basis of important traits such as human disease.

Declaration

This thesis is an account of research undertaken between February 2023 and October 2023 at the Research School of Biology, the Australian National University, Canberra, Australia.

Except where otherwise indicated, the material presented in this thesis is my original work.



Laura Marie Hardy

October 2023

Acknowledgements

I can honestly say that Honours has been my favourite year of my degree, and that is primarily due to all of the wonderful people I've met and spent time with this year, while getting my brain crammed full of biology, statistics, mathematics and general good vibes.

First and foremost, thank you to my supervisor, Eric Stone. People had told me having a good supervisor was important but I had no real notion of this until this year, where I have absolutely hit the supervisor jackpot. Thank you for making so much time for me, for all of the fascinating on-topic (and off-topic) conversations, for injecting my brain with very cool quantitative biology, for many, many emails and for your unwavering support. It's been an honour to be your student.

Thank you to Elle Saber, without whom my graphs would be ugly and my R skills generally much worse, and Andy Bachler, without whom I might still be navigating the treacherous labyrinth of the NCI supercomputer. Shout out also to the wider ANU Biological Data Sciences Institute and associated parties for all of the excellent cross-desk banter, forums, free food and general solidarity.

I would like to acknowledge my housemates, Chrissie, Saskia and in particular Tammy who has also been on the Honours grind this year. Thanks for bearing with me, cooking yummy dinners, listening to me waffle on about my project, all the late-night kitchen chats and carpool karaoke to uni every day.

And lastly, thank you to my examiners, Gavin Huttley, Dan Noble and Damien Farine for all your wonderful comments, encouragement and feedback throughout this year. I hope you enjoy reading this thesis!

Table of contents

Abbreviations and terms

Notation	Description
Genotype	The genetic material of an individual, usually referring to the specific alleles present at a given locus or set of loci.
Phenotype	The observable characteristics or traits of an organism.
Allele	A version of a gene.
LD	Linkage disequilibrium: the non-random association of alleles at two or more loci. This thesis uses r^2 to quantify the degree of LD between a pair of variants.
Diallelic	Pertaining to a genetic locus that has only two different alleles.
MAF f	Minor allele frequency: the frequency at which the less common allele occurs in a particular population.
Genetic architecture	The number, location and relative effect of genes on a phenotypic trait.
GWAS	Genome-wide association study. Though elsewhere used quite generally, we use "GWAS" to describe one specific methodology investigated here.
Association study	A methodology for resolving genetic architecture which assesses genetic variants on the basis of p-value of their association with a phenotype.
LASSO	Least Absolute Shrinkage and Selection Operator. A variable selection method.
Variable selection	The process of selecting a subset of relevant features or variables for use in model construction.
E-value	Both the name of a novel methodology for conducting an association test between genotype and phenotype, and the measure of association itself, defined as the expected empirical conditional p-value of a genetic variant.
SNP	Single nucleotide polymorphism.
Variant	Genetic variant: could refer to a SNP, copy number variation or sequence variation.

Causal variant	A genetic variant that contributes to phenotype.
Haplotype	A combination of alleles at adjacent locations (loci) on a chromosome that are inherited together.
Additive	Pertaining to the combined effect of individual alleles, such that the effect of two alleles together is the sum of their individual effects.
Dominant/recessive	Describes the relationship between two alleles of a gene, where the dominant allele masks the effect of the recessive allele in heterozygotes.
Epistasis	The interaction between genes where the effect of one gene is modified or masked by the presence of another gene. In a model, epistasis manifests as a deviation from additivity.
n	Number of individuals in a population.
p	Number of genetic variants.
m	Number of causal variants.
Heritability h^2	The proportion of phenotypic variance that is explainable by genetic variance.
Population structure	The presence of multiple genetically distinct subpopulations that differ in their mean phenotypic values.
d	A parameter used to generate population structure in the simulation framework.
Causal rank	The rank of the causal variant with the highest significance (e.g. lowest p-value) out of all causal variants.
Empirical false positive rate	If the causal rank is k, then the empirical false positive rate is $k - 1$, representing the number of false positives, divided by the total number of variants p. We use this as our measure of performance.
Simpson's paradox	A phenomenon in which a trend between two variables in a population appears, disappears or reverses when the population is split into subgroups
L1 norm	The sum of the absolute values of the components in a vector.
L2 norm	The square root of the sum of the squared components of a vector; commonly known as the Euclidean norm.

List of figures and tables

Figures

Tables

Chapter 1

Introduction

1.1 Overview

A fundamental goal of genetics is to understand the relationship between genotype and phenotype; what versions of genes (i.e. alleles) an individual has and the characteristics (e.g. eye colour) held by that individual. Phenotypes that are heritable can be explained at least in part by genotype, with the unexplained variability attributed to the "environment", broadly defined (Falconer & Mackay 1996; Lynch & Walsh 1998). In this way, we have a function F mapping genotype X (signal) and environment ϵ (noise) to phenotype Y.

$$Y = F(X) + \epsilon \tag{1.1}$$

This function is a mathematical encapsulation of the concept of genetic architecture of traits: the number, location and relative effect of genes on a phenotypic trait (Vinkhuyzen 2013). There are numerous methods to infer the nature of this function, many of which categorise as Genome-Wide Association Studies (GWAS) (Uffelmann et al. 2021), machine learning (Nicholls et al. 2020) or variable selection approaches (Che et al. 2020; He & Lin 2021). Understanding the genetic architecture of a trait allows us to better predict risk (Korte & Farlow 2013; Uffelmann et al. 2021), heritability (Vinkhuyzen 2013), genetic correlations (Uffelmann et al. 2021) and the underlying biology of a phenotype (Nicholls et al. 2020; Uffelmann et al. 2021).

In this chapter, we outline the principles underlying simple association studies which assess genetic variants on the basis of p-value of their association with a phenotype, which we will narrowly refer to as GWAS. We identify the limitations of GWAS and explain the biological and statistical foundation of parameters affecting its ability to resolve genetic architecture, highlighting the gap in current understanding of the combinatorial impact of these parameters. We also introduce variable selection approaches as an alternative to

standard GWAS. The chapter concludes by setting out the aims of the thesis, which purposes to provide a comprehensive understanding of methodological behaviour under a range of conditions and hence a guide for method choice in future studies.

1.2 Genome-wide association studies

Among approaches for interrogating the relationship between genotype and phenotype, the most simplistic is likely still the most popular (Dickson et al. 2010; Uffelmann et al. 2021). This approach proceeds by comparing single nucleotide polymorphisms (SNPs) or alternatively, variations in copy number or sequences, in a group of individuals with differing trait values (Korte & Farlow 2013; Uffelmann et al. 2021). Such genomic differences are collectively called variants, and often act as marker loci for regions of the genome containing one or more genes in linkage with the polymorphic marker loci (Korte & Farlow 2013; Nicholls et al. 2020; Russ et al. 2022; Mackay, Stone & Ayroles 2009). The output of GWAS is a list of p-values, effect sizes and their directions generated from association tests of each of the variants under consideration, reported in blocks of correlated variants that show significant association with the trait (Nicholls et al. 2020; Uffelmann et al. 2021). GWAS allows for the exploration of many complex, polygenic traits, and can uncover novel genetic loci linked to specific traits across diverse populations (Boyle, Li, & Pritchard 2017; Falconer & Mackay 1996; Holland et al. 2020; Korte & Farlow 2013; Lynch & Walsh 1998; O'Connor et al. 2019; Watanabe 2019; Zeng et al. 2018).

Association studies have a broad range of applications. Uncovering trait architecture is a motivation in and of itself, but GWAS can also be used to determine the biological cause of heritable phenotypes (Korte & Farlow 2013), estimate the heritability of a trait (Vinkhuyzen 2013), calculate genetic correlations (Uffelmann et al. 2021), make clinical risk predictions for physical and mental disease (Korte & Farlow 2013), inform drug development programs (Nicholls et al. 2020) and also as control variables in epidemiological studies (Ufflemann et al. 2021). To do so, the trait-associated variants reported by the study must further be differentiated into putative causal and non-causal variants, with the former subject to additional computational analysis and/or laboratory validation (Korte & Farlow 2013; Nicholls et al. 2020). False negatives — causal variants missed in the GWAS screen — remain unexplained biology, whereas false positives may dilute understanding and inflate cost in follow up study (Ioannidis et al. 2001). Mitigating these concerns motivates a systematic

quantification of how association methods such as GWAS perform across a range of applications and conditions.

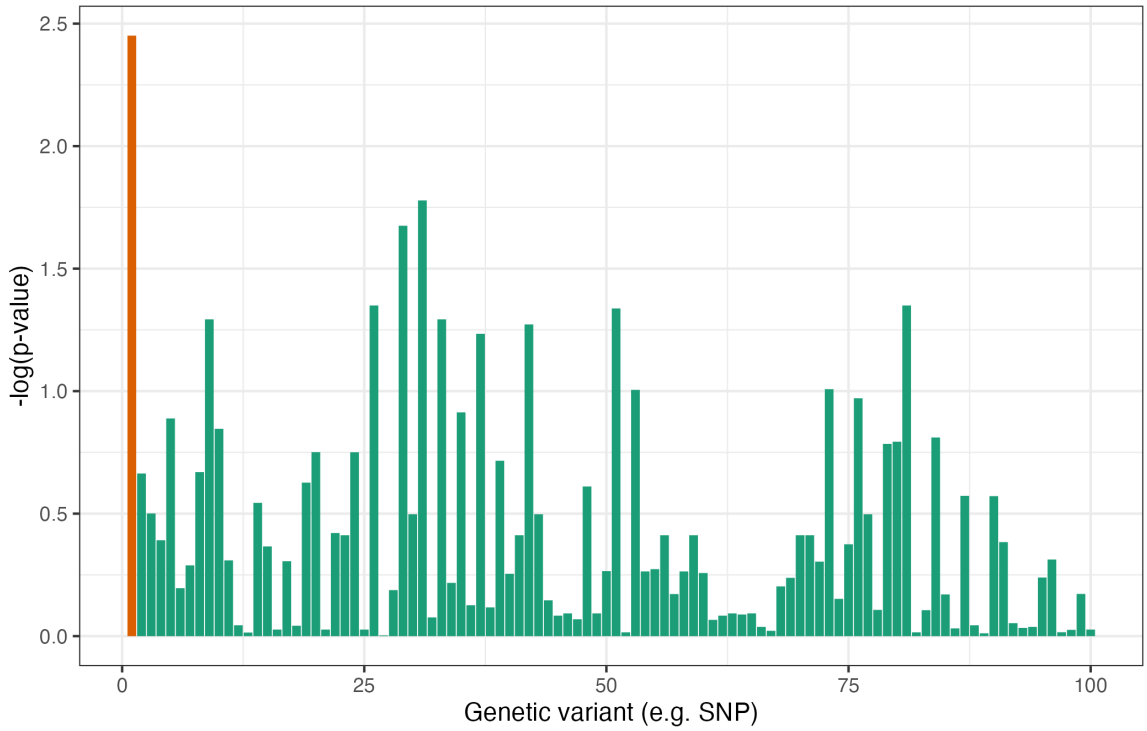


Figure 1.1: Illustration of a GWAS. The x-axis indexes 100 SNPs from a sample of phenotyped individuals. Each SNP is individually tested for association with the phenotype, yielding 100 p-values. Here, as is customary, the y-axis shows the results as -log10(p-value). Typically, either a significance threshold is used to cull the set of SNPs for further bulk study, or the SNPs are ranked by significance and investigated more sequentially. In the figure, the first SNP (in red) by index also ranks first in significance. The data shown here was simulated as part of this study as will be explained in the Methods.

1.3 Limitations of GWAS

Indeed, although popular, GWAS is prone both to false positives and false negatives. By design, the approach seeks a signal of the difference in mean phenotype value between variant types, implying that association tests will detect any randomly chosen partition with a large mean difference, casual variant or not (Strope et al 2015; Uffelmann et al. 2021). False positives will occur when, by chance, a non-causal variant shows a stronger correlation with the phenotype than does a true causal variant (Dickson et al. 2010; Korte & Farlow 2013). As GWAS studies have scaled up in use, more associated variants than are practical to investigate are reported by inconsistent and sometimes un-replicated studies (Dickson et al. 2010; Ioannidis et al. 2001; Nicholls et al. 2020). Targeting the study of specific traits by corresponding methods evidenced to have a low false negative rate under the trait architecture should allow more stringent thresholds of p-values to correct for multiple testing and false discovery (Nicholls et al. 2020).

A key phenomenon that limits the efficacy of GWAS is linkage disequilibrium (LD), the phenomenon of nonrandom association between alleles at two or more loci (Skelly, Magwene & Stone 2016; Nicholls et al. 2020). Physically proximal loci are often in LD, as recombination is slow to break up the association (Lynch & Walsh 1998; Slatkin 2008); however, physically distant or even independently segregating loci can also have this property (Skelly, Magwene & Stone 2016). Having a more structured population, i.e. having different subpopulations with distinct allele frequencies, corresponds to increased levels of local LD (Cardon & Palmer 2003; Ewens 2004; Skelly, Magwene & Stone 2016). LD can assist in finding casual variants as it broadens the scope of reported variants by the GWAS (Dickson et al. 2010; Nicholls et al. 2020; Uffelmann et al. 2021). Conversely, it can hinder by providing a tightly-knit block of variants which cannot be differentiated between in terms of causality (Kruglyak 1999; Skelly, Magwene & Stone 2016; Sutter et al. 2004). There are a range of other factors influencing the rate of false discovery and false negatives, some of which we highlight below.

1.4 Variable selection approaches

Variable selection (or feature selection) is the process of selecting a subset of predictive variables (features) for use in the construction of a model (Chowdhury & Tanvir 2020). There are many different variable selection approaches, from those which filter variables according to their correlation, variance or informativeness (Sánchez-Maroño, Alonso-Betanzos & Tombilla-Sanromán 2007), to forward or backward selection methods which respectively add or remove variables according to model performance (Chowdhury & Tanvir 2020; Ratner 2010), to methods which penalise using the L1 or L2 norms of coefficients (Tibshirani 1996; Waldmann et al. 2013), and more. Variable selection is a fundamental aspect of many modelling tasks, providing interpretable and high-performing models at a reduced computational cost (Chowdhury & Tanvir 2020; Ratner 2010). These methods have successfully been applied in the context of genotype-phenotype mapping before (Puthiyedth et al. 2021; Srivstava & Chen 2010; Waldmann et al. 2013), proving particularly adept at dealing with correlation structure between predictors (Hebiri & Lederer 2012; Nouira & Azencott 2021). In this thesis we will evaluate the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996) as indicative of wider variable selection approaches in an association study.

1.5 Factors influencing association studies: methods and previous work

The increasing ease with which genomic data can be obtained has made GWAS a common component of trait dissection. The list of essential ingredients is minimal: for n individuals, we require an $n \times 1$ vector Y of trait measurements and an $n \times p$ matrix X of genetic data on p segregating sites (usually SNPs). Even so, there is rich complexity in how genetic signal manifests when X is used in a statistical model to explain the variation in the trait Y. There is substantial appreciation of this complexity in the literature (Dickson et al. 2010; Falconer & Mackay 1996; Holland et al. 2020; Korte & Farlow 2013; Lynch & Walsh 1998; Mackay, Stone & Ayroles 2013; O'Connor et al. 2019; Russ et al. 2022; Skelly, Magwene & Stone 2016; Uffelmann et al. 2021; Vinkhuyzen et al. 2013; Watanabe 2019; Zeng et al. 2018); however, systematic studies are lacking and needed to properly elucidate the impact of various factors in isolation and combination.

Although the individual effect of some factors influencing GWAS results are well understood, how they act synergistically to influence the results of an association study has not been holistically quantified. Small sample size (i.e., n) can decrease the power of the statistical test to distinguish phenotypic signal from environmental or sampling noise (Casella & Berger 2002; Lynch & Walsh 1998), obscuring causal variants (Bodmer & Bonilla 2008; de Koning & Haley 2005; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Manolio et al. 2009). Increasing the number of variants p to be investigated will increase the expected number of non-causal variants whose p-values are above the threshold by chance (Casella & Berger 2002; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Uffelmann et al. 2021), as p-values are uniformly distributed under the null hypothesis (Casella & Berger 2002). This phenomenon is known as the multiple testing problem. Allele frequency can also affect outcomes; the proportion of phenotypic variance that can be attributed to a SNP (heritability) depends on allele frequency (Falconer & Mackay 1996; Vinkhuyzen et al. 2013; Visscher et al. 2006; Yang et al. 2010), and rare variants may create synthetic associations with common (causal) variants in LD, resulting in false positives (Dickson et al. 2010; Platt, Vilhjalmsson & Nordborg 2010). Furthermore, GWAS struggles to detect variants from a collection of many loci with small effects on a trait, and is by comparison far more effective at identifying a few common variants with large effects (Dickson et al. 2010; Holland et al. 2020; Liu, Li, & Pritchard 2019; Mackay, Stone & Ayroles 2009; Strope et al. 2015; Watanabe 2019).

However, not all of the features whose impact we might wish to quantify are immediately obvious.

The proximal goal of this work is to identify a discrete (albeit partial) list of factors that may influence the performance of small association studies, and then vary those factors in a systematic way to evaluate how performance is affected. To do so comprehensively requires the development of methodology and a code base for running numerical simulations. An important part of this is to decide which aspects of genetic complexity are within scope, how they can be discretised, and what constitutes their dynamic range within which variable performance can be observed. Equally important is to control unwanted sources of variation that threaten to obscure the signal. With that in mind, we developed the framework described below and in subsequent sections. To our knowledge, there is no equivalent framework for generating genotype-phenotype data from a set of predetermined parameters and conducting an association study in the literature at present.

1.6 Aims and scope

In this thesis, we present a framework for simulating idealised and structured populations to allow us to assess three methods used to perform genotype-phenotype association studies. The overarching aim was quantifying the performance of association studies on simulated populations by individually and combinatorially manipulating a discrete set of factors that impact their performance. We evaluate two existing association studies (methods); GWAS, described above, and the LASSO, as representative of variable selection approaches, as well as one novel method, dubbed the E-value. The parameters these are tested under are population size, number of genetic variants, population structure, number of causal variants, and the heritability of each causal variant. We also propose a new scalar measure to quantify population structure and apply this to simulated, structured populations as a correlate of performance of association studies.

We then present the results from the simulations, comparing the chosen methods while illustrating how the above factors combine to impact performance, sometimes counter to intuition. In doing so, we demonstrate the effectiveness of the simulation framework and provide insights into the conditions under which particular methods are unsuitable. Finally, we explore the implications of these results for method choice in future trait architecture studies, and the extension of the simulation framework to the study of other complex traits.

Methods

This thesis evaluates the ability of three methods: GWAS, the E-value and LASSO, to resolve genetic architecture from simulated datasets. In this chapter, we present a framework for generating genotype and phenotype data governed by a discrete set of factors: population size, number of genetic variants, population structure, number of causal variants, and the heritability of each causal variant. We begin by introducing the simplest case as a demonstrative example, then describe a systematic approach to individually and combinatorially manipulate these factors and quantify their impact on method performance over thousands of repetitions. Noting that population structure is uniquely non-numeric among the factors, we also introduce a new scalar quantification of population structure and demonstrate its sensibility.

2.1 Simplest test case

Consider a haploid population of $n = 20$ individuals whose polymorphisms are diallelic, with $p = 100$ variants encoded in a binary (0/1) manner. Collect these variants in an $n \times p$ matrix X where the allele frequencies are drawn from the Ewens sampling distribution (Ewens 1972) and the $0/1$ states are assigned to the n individuals at random (as in Skelly, Magwene & Stone 2016). Then, simulate a $n \times 1$ quantitative phenotype vector Y controlled by a single variant corresponding to the first column of the matrix, X_1.

$$Y = f(X_1) + \epsilon \tag{2.1}$$

We specify $f(X_1) = \beta X_1$ for some coefficient β that represents the mean phenotypic difference between individuals with the "1" allele and those with the "0". The error term ϵ follows a normal distribution with mean zero and variance σ^2. Thus, Y is the sum of a deterministic component $f(X_1)$ and a stochastic component $\epsilon \sim N(0, \sigma^2)$ that encodes genetic signal and environmental noise, respectively. Later, we will derive β as a function of

heritability, MAF of X_1 and σ^2, but for now we will fix $\beta = 2$, and $\sigma = 1$, under the assumption that the within-class variance and standard deviation is 1.

We simulate all individual phenotype values with a normal distribution and add the mean difference $\beta = 2$ to individuals with state (allele) 1 in variant 1. Restricting our methods to GWAS for now, each variant is tested against the phenotype with a t-test to obtain a vector of p-values, and we record where the causal variant falls in a ranking of significance by ordering this vector. We also record the minor allele frequency (MAF) of the causal variant.

The results of one such simulation were already displayed in Figure 1.1. Recall that the first SNP, X_1, had the highest bar and hence the smallest p-value, ranking first in significance among the SNPs. As this was the causal SNP used to generate the phenotype, GWAS correctly ranked the true causal variant as most significant. Consider, however, the meaning had X_1 be ranked k for some $k > 1$; in a sequential follow-up study, there would be $k - 1$ false discoveries ahead of the causal variant.

To summarise, this process generated a haploid, idealised population, in which there was one causal variant with a mean difference between phenotypes of 2. Now that we have established a procedure, we adjust the input parameters and repeat the process, testing each method in turn.

2.2 Ploidy and genetic architecture

All simulations are based on a haploid, diallelic population where, as in Section 2.1, alleles are encoded as 0 or 1. There are two alleles and two genotypes, and the numerical encoding is straightforward:

$$Y = \sum_{i=1}^{m} \beta_i X_i + \epsilon \tag{2.2}$$

Here m is the number of causal variants. We restrict our simulation of genetic architecture to additive functions (as seen above) mapping between X and Y, and increase the complexity of the architectures by varying the number of causal variants and their effect sizes (described in Section 2.6). Further discussion of the rationale behind this restriction and the extension of this framework to other genetic mappings and polyploidy can be found in Chapter 4.

2.3 Dimensionality

We were motivated more by incidental GWAS than by large, designed studies. In particular, we wished to focus on values of n that were both small (relative to say human disease GWAS) and realistic (as reflected in sufficient applications). From a basis of small n, we chose values of p that were sufficient to demonstrate a range of performance. This is similar to many applications in statistical learning where the key performance indicator is the ratio between n and p; varying this ratio for values of n guided our range of p. Hence, we tested the following combinations of values:

$$(n, p) = (20,100),\ (100,100),\ (20,1000),\ (100,1000) \tag{2.3}$$

2.4 Allele frequencies

It is well known that the power to detect an association depends on allele frequency, with more common variants being easier to detect (Dickson et al. 2010; Korte & Farlow 2013; Vinkhuyzen et al. 2013). We provide a mathematical explanation for this below to motivate the choices that follow.

Let X_1 represent a causal variant from a genotype whose matrix is $X_{n\times p}$. Suppose X_1 has k ones, and $(n-k)$ zeroes. Then the MAF, which we will denote f, in the population for variant X_1, is given by:

$$f = \mathbb{E}X_1 = \frac{1}{n}\sum_{j=1}^{n} X_{j1} = \frac{k}{n} \tag{2.4}$$

Here, $\mathbb{E}$ denotes expected value. Let Y be given by the following additive, haploid model, where m is the number of causal variants:

$$Y = \sum_{i=1}^{m} \beta_i X_i + \epsilon \tag{2.5}$$

The total variance in phenotype Y under this model can be divided into variance explained and unexplained by genetics:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}\left(\sum_{j=1}^{m} \beta_i X_i\right) + \mathrm{Var}(\epsilon) \\ &= \sum_{j=1}^{m} \beta_i^2 \mathrm{Var}(X_i) + \mathrm{Var}(\epsilon) \end{aligned} \tag{2.6}$$

where the second equality follows under the assumption that the m variants are pairwise independent and do not covary.

The higher the proportion of total variance that can be explained by a particular genetic variant X_1, the higher the power to detect an association between X_1 and phenotype Y. Calculating this variance, then, we see that it depends solely on allele frequency as stated.

$$\begin{aligned} \mathrm{Var}(X_1) &= E(X_1 - E(X_1))^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(X_{j1} - f\right)^2 \\ &= \frac{1}{n}\left(k(1-f)^2 + (n-k)f^2\right) \\ &= \frac{k}{n}(1-f)^2 + \left(1 - \frac{k}{n}\right)f^2 \\ &= f(1-f)^2 + (1-f)f^2 \\ &= f(1-f)(1-f+f) \\ &= f(1-f) \end{aligned} \tag{2.7}$$

In other words, a variant causing a mean difference β contributes $\beta^2 f(1-f)$ to the genetic variance, making common variants easier to "see" by statistical methods.

We sought to control the effect of allele frequency while acknowledging its variability; therefore, we chose a common distribution of allele frequencies for all of our simulations. Given that, the obvious choice was the Ewens-Watterson distribution (Ewens 1972; Watterson 1974).

The Ewens-Watterson distribution is a probability distribution on the allele frequencies in a population, depicted for a population of size $n = 20$ in Figure 2.1. It describes a

population at mutation-drift equilibrium in the absence of selection, and requires that $n << p$, both of which are desired conditions for our simulation. For each variant (column) in the matrix, we first sample the allele frequency (number of ones) from the Ewens distribution and then assign the 0/1 states to individuals by a process described below.

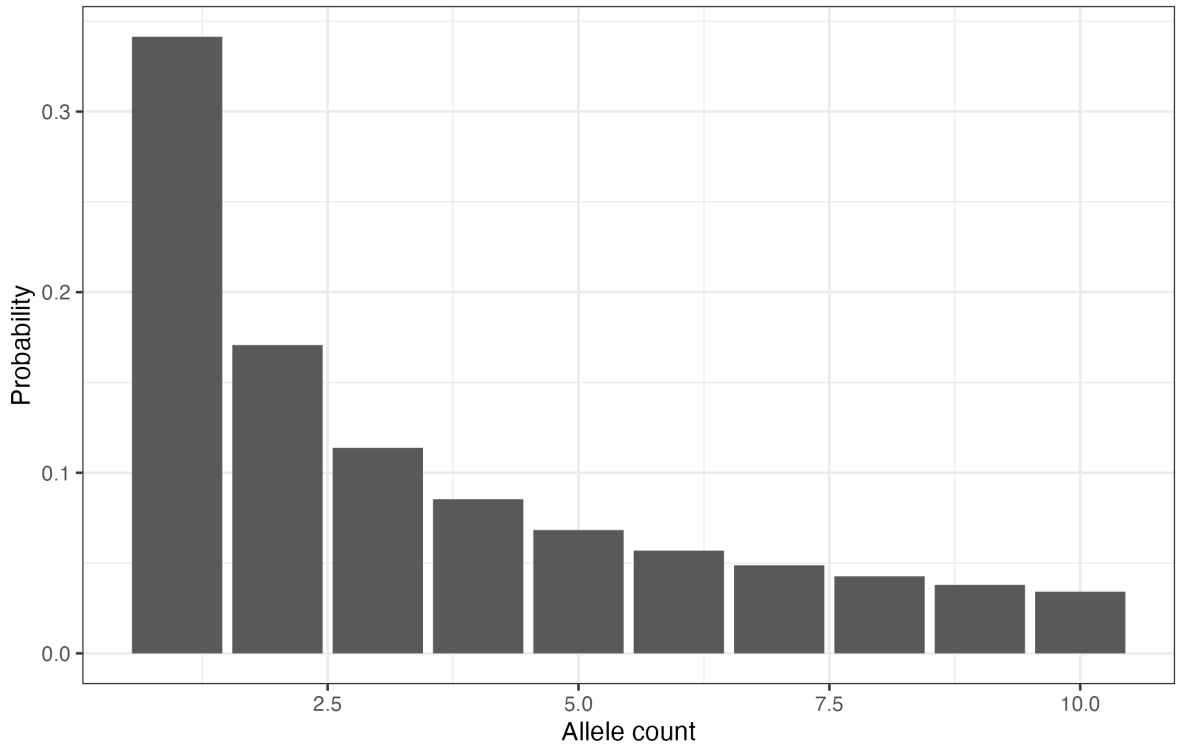


Figure 2.1: The Ewens-Watterson distribution. The probability of a particular number of minor alleles occurring in a population of size 20 of a diallelic genetic variant (e.g. SNP) under the Ewens-Watterson distribution.

2.5 Population structure

Population structure refers to the organisation of genetic variation in a population, i.e. having separate subpopulations with distinct allele frequencies (Cardon & Palmer 2003). We predict that adding population structure will obscure the genetic architecture by increasing the density of correlated variants; in other words, that resolution of genetic signal will be worsened with increasing variant collinearity (Korte & Farlow 2013; Kruglyak 1999; Skelly, Magwene & Stone 2016; Sutter et al. 2004; Uffelmann et al. 2020). However, we want to disentangle the contribution of variant collinearity, i.e. global LD, from demography and perturbations to the allele frequency spectrum (e.g. an increase of rare variants in an expanding population). Our solution was to simulate population substructure within an overall population whose allele frequencies follow the Ewens distribution. Taking a combinatorial approach as opposed to using a population genetic model also permits independent and systematic control of input factors and prevents the introduction of local LD.

We began by fixing n and p to one of the specified pairs of values and representing genotype as a matrix X. For each genetic variant (column) in the matrix, we sampled the minor allele count k from the Ewens distribution, so that we had k ones and $n-k$ zeroes to allocate to the column. Next, we split the column into five subpopulations of either 4 or 20 individuals, corresponding to $n = 20$ or $n = 100$, respectively. This subdivision facilitates our mechanism of introducing structure, which allocates ones and zeroes to each subpopulation depending on both what is already in that subpopulation, and how many ones and zeroes remain overall (denote this N_1 and N_0).

The concept resembles other "urn models" used in combinatorics and population genetics (Johnson, Norman & Kotz 1977). Within each subpopulation, the allele assigned to the first individual is selected using a binomial function with one trial, one observation and probability of receiving a one given by:

$$\text{prob} = \frac{N_1}{N_1 + N_0} \tag{2.8}$$

For the very first individual in the population, chosen at random for each variant, this probability is simply $\frac{k}{n}$. To assign alleles to subsequent individuals in the subpopulation, let n_0, n_1 respectively denote the number of zeroes and ones in the subpopulation so far and let $0 \leq d < 1$ be a fixed value that modulates the strength of dependency. Provided $N_0, N_1 \neq 0$, define the probability of receiving a one as:

$$\text{prob} = (1-d)\frac{N_1}{N_1 + N_0} + d\frac{n_1}{n_1 + n_0} \tag{2.9}$$

If at any point $N_1 = 0$, then henceforth prob $= 0$, whereas if $N_0 = 0$, then henceforth prob $= 1$.

Thus, d is a parameter for generating structure which represents how much the probability of being assigned a one depends on what is already in the subpopulation. As d approaches 1, the distribution of alleles across subpopulations is affected more strongly by the alleles already assigned to each subpopulation. In our simulations, we fix d prior to creating each X and iterate through increments of the interval [0,1) to generate populations with

different amounts of structure. Figure 2.2 below visualises the effects of increased structure on a simulated population. The increments were chosen based off the relationship between our parameter for generating structure and our measurement of structure to be discussed below. We tested the following values for d:

Table 2.1: Values of d used for simulation.

From	To	By increments of	Total number of values
0	0.45	0.05	10
0.475	0.825	0.025	15
0.85	0.99	0.01	15

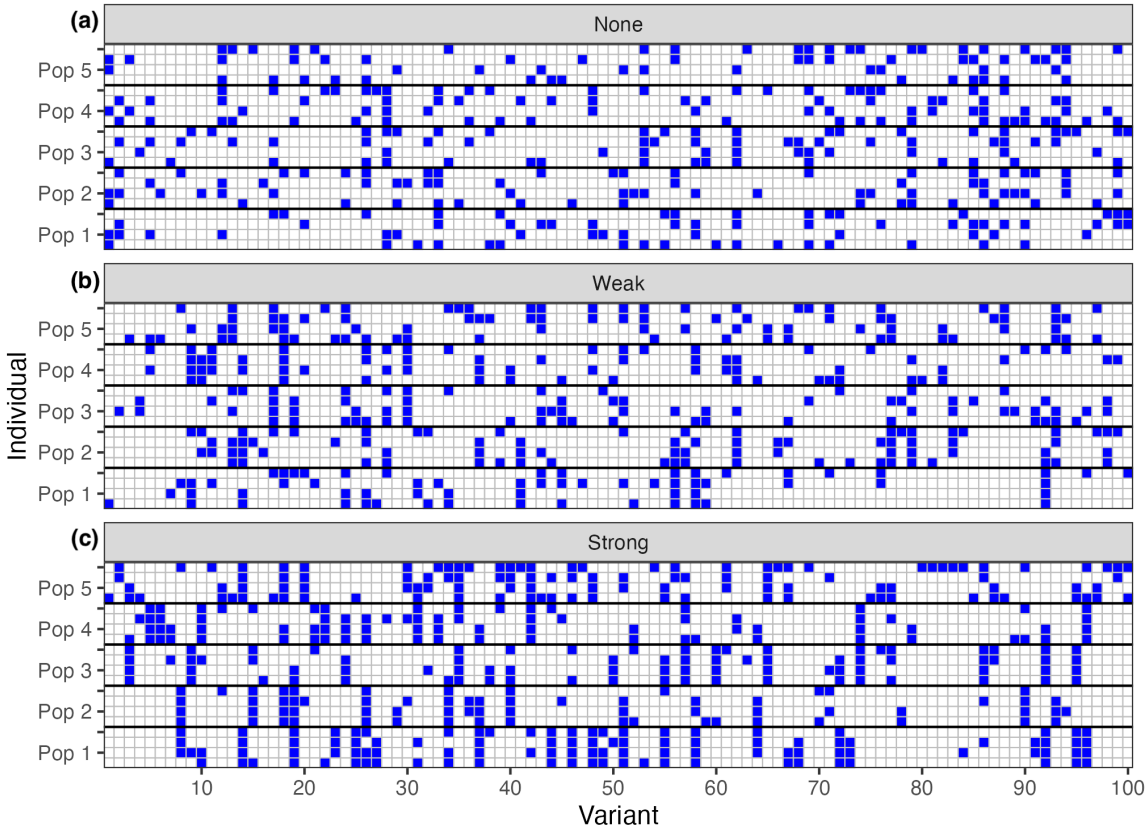


Figure 2.2: Increasing structure results in concentration of minor alleles into the same subpopulations. These grids indicate the positioning of minor alleles (1s, coloured blue) and major alleles (0s, coloured white) in the genotype matrix X for 3 distinct populations of 20 individuals and 100 variants. The populations were simulated according to the methods outlined above under (a) no structure (d = 0) (b) weak structure (d = 0.4) and (c) strong structure (d = 0.8), split into 5 subpopulations as shown.

2.5.1 Quantifying Population Structure

Our study motivated a new matrix-based statistic to quantify population structure. We sought a statistic that was both fit for purpose here and generally applicable with a rigorous underpinning. We quantify structure as the average pairwise r_{ij}^2, where r_{ij} is the correlation

between variant i and variant j. In other words, for a given population $X_{n\times p}$, we measure structure by:

$$\text{average LD} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{p^2} \tag{2.10}$$

The population structure we have introduced shapes our $n \times p$ matrix X so that the columns are correlated (i.e. LD, but due to structure rather than linkage) and thus so their pairwise correlations to the phenotype vector Y are dependent. Figure 2.3 represents the correlation matrix for the structured populations whose genotype matrices are represented by Figure 2.2, demonstrating the role of structure in increasing correlation.

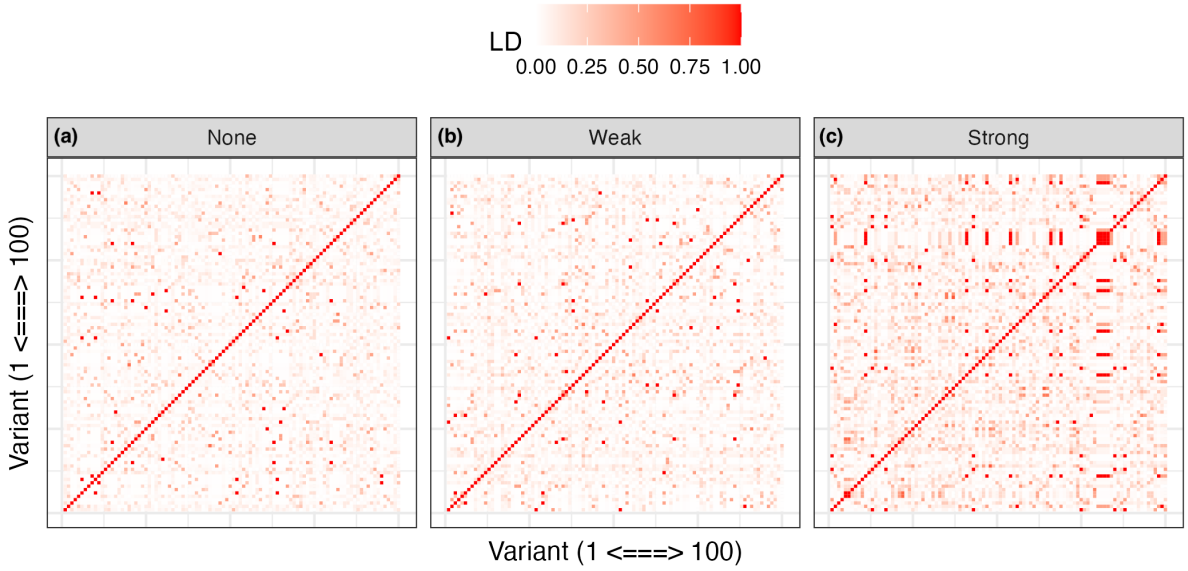


Figure 2.3: Heat map of correlation between genetic variants in genotype matrix. The correlation between the columns of the three X matrices represented in Figure 2.2 is depicted here using a heat map. As in Figure 2.2, the three populations are of 20 individuals and 100 variants with (a) no structure (d = 0) (b) weak structure (d = 0.4) and (c) strong structure (d = 0.8).

Recall that LD is the phenomenon of nonrandom association between alleles at two or more loci, or in other words, when we see more correlation than we would expect between two unlinked genetic variants. By first standardising the columns of X, we may represent population structure in terms of covariance between individuals with the $n \times n$ matrix XX^T, while representing pairwise LD between variants with the $p \times p$ matrix X^TX. These two matrices have distinct eigenvectors but common eigenvalues, and those eigenvalues form a connection between population structure and association mapping.

This approach is based on the singular value decomposition $X = U\Sigma V^T$, where U is an $n \times n$ orthogonal matrix, V is an $p \times p$ orthogonal matrix and $\Sigma = [S \quad 0]$ is an $n \times p$ matrix consisting of S, a diagonal matrix of singular values (not necessarily nonzero), and appropriately sized 0 matrices (Anton & Rorres 2014). Let $D = \Sigma^2 = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}$. The singular values are the square roots of the eigenvalues of the symmetric matrix:

$$X^T X = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^2 V^T = VDV^T = R_{p\times p} \tag{2.11}$$

which is the matrix of pairwise correlations between variants whose squared entries are depicted in Figure 2.3.Now consider another symmetric matrix with the same set of positive eigenvalues:

$$XX^T = (U\Sigma V^T)(U\Sigma V^T)^T = US^2U^T = G_{n\times n} \tag{2.12}$$

This matrix, sometimes called a genetic relationship matrix, describes the covariance between individuals and encodes population structure. In fact, the eigendecomposition of such matrices has been used extensively to identify population structure, as well as to control for it in association studies (Capelli et al. 2006; Lovell et al. 2005; Menozzi, Piazza & Cavalli-Sforza 1978; Patterson, Price & Reich 2006; Price et al. 2006; Stoneking et al. 1997).

Thus, global LD and population structure are bound together by a common set of positive eigenvalues which we will call $\lambda_i = D_{ii} = (S_{ii})^2$. LD, in turn, can obscure association signals. If, for example, X_i is casual and correlated with X_j, then the stronger their correlation (as quantified by $r(X_i, X_j)^2$), the more likely it is that $r(X_j, Y)^2$ exceeds $r(X_i, Y)^2$ by chance, thus pushing X_i below X_j in the significance ranking. Taken genome-wide, the average pairwise r^2 among columns of X is indicative of the efficacy of association studies in the sense that it helps quantify the contribution of structure (through the aforementioned eigenvalues) to obscuring the ranking of associations and hence the resolution of causal variants. We will show how this is done.

Since $r(X_i, X_i) = 1$ for all i, the trace of R, which is the sum of its diagonal entries, is:

$$\mathrm{Tr}(R) = p \tag{2.13}$$

As λ_i are the eigenvalues of $X^T X$, it follows that

$$\sum_{i=1} \lambda_i = p \tag{2.14}$$

Note also that by the eigendecomposition of R that

$$\begin{aligned} R^2 &= (VDV^T)(VDV^T) \\ &= VD^2V^T \end{aligned} \tag{2.15}$$

is also an eigendecomposition.

Hence,

$$\begin{aligned} \mathrm{Tr}(R^2) &= \sum_{i=1} \lambda_i^2 \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} r_{ij}^2 \end{aligned} \tag{2.16}$$

In other words, the sum of the squared eigenvalues is equal to the sum of all p^2 pairwise squared correlations between the p variants, and dividing by p^2 yields our "Average LD" measure of population structure.

This can be recognised as a ratio of norms, a type of measure used in linear algebra and signal processing to describe the "effective dimension" of a subspace (as in Yiming et al. 2021). Specifically, recalling that the Schatten p-norm of a matrix A is

$$\|A\|_{\mathrm{p}} = \left(\mathrm{Tr}(A^{\mathrm{p}})\right)^{\frac{1}{\mathrm{p}}} \tag{2.17}$$

we have:

$$
\begin{aligned}
\frac{\|R\|_2^2}{\|R\|_1^2} &= \left(\frac{\|R\|_2}{\|R\|_1}\right)^2 \\
&= \frac{\mathrm{Tr}(R^2)}{\mathrm{Tr}(R)^2} \\
&= \frac{\sum \lambda_i^2}{(\sum \lambda_i)^2} \\
&= \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} r_{ij}^2}{p^2} \\
&= \text{average LD}
\end{aligned}
\tag{2.18}
$$

Thus, we have constructed a principled, scalar measure of population structure that has an intrinsic connection to LD and hence confusion in GWAS. It remains to connect this measure to our algorithmic parameter d for generating structure.

Figure 2.4 depicts this relationship for each of our study dimensions. The plot provides the intuition behind our choice of increments for d; for lower values of d the populations are clustered more tightly at low values of average LD, while for higher input values we produce more of a range of higher measurements of structure.

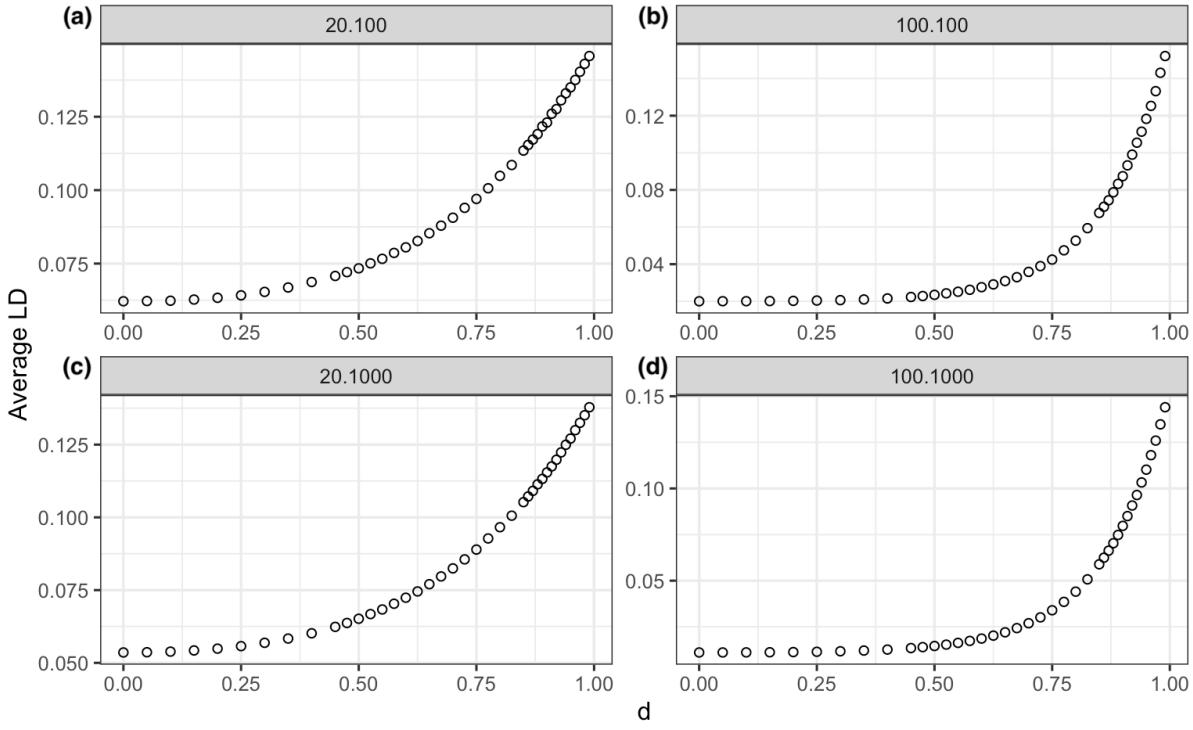


Figure 2.4: The relationship between average LD and the input parameter for generating structure, d. Each point indicates the average value of average LD calculated for the 1000 populations of specified dimensions generated at each value of d. The dimensions of the simulated populations were (n, p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000).

2.6 Generating a phenotype

2.6.1 Effect sizes and heritability

To ensure meaningful comparisons, we sought to avoid confounding the factors that govern each portion of simulations. This forced trade-offs due to the interplay between population structure, allele frequency and statistical power. It is well known, for example, that rare variants (e.g. those with small MAF f, often thresholded at 0.05) require larger effect sizes for detection in association studies (Casella & Berger 2002; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Watanabe et al. 2019); this was also noted in Section 2.4, where we demonstrated that variant causing a mean difference of β contributes $\beta^2 f(1-f)$ to the genetic variance, implying that rare variants explain less of the total phenotypic variance for a given mean difference. Moreover, for more complex architectures, one must specify both the overall genetic variance explained (via the function that maps genotype to phenotype) and how each causal variant contributes. We chose overall genetic variance explained as our experimental factor for manipulation, calculating effect sizes accordingly. Specifically, we grouped simulations by heritability (h^2); that is, the proportion of variance in the trait explained by genetics (Vinkhuyzen et al. 2013). For a single causal variant, the heritability is a function of both the mean difference and the allele frequency (Falconer & Mackay 1996). This is worked out below:

Given a genotype matrix $X_{n\times p}$, arbitrarily choose the first column X_1 to be the causal variant, and let the MAF of X_1 be f. From the calculations in Section 2.4, we know that

$$\text{Var}(X_1) = f(1-f) \tag{2.19}$$

And if $Y = \beta X_1 + \epsilon$, for $\epsilon \sim N(0,\, \sigma^2)$, we have

$$\begin{aligned} h^2 &= \frac{\text{variance explained by genetics}}{\text{total variance}} \\ &= \frac{Var(\beta X_1)}{Var(Y)} \\ &= \frac{\beta^2 f(1-f)}{\beta^2 f(1-f) + \sigma^2} \end{aligned} \tag{2.20}$$

Rearranging for β we get:

$$\beta = \sigma \sqrt{\frac{h^2}{(1-h^2)f(1-f)}} \tag{2.21}$$

For a single causal variant, we use this formula to determine the expected mean difference between allelic states in the Y vector. Because β scales with σ, without loss of generality we set $\sigma = 1$ for all simulations. Heritability values in trait studies can vary widely depending on the trait being studied, the population under investigation, and the methods used for estimation. For complex traits, typically less than 10% of the phenotypic variation is explained by SNPs reported by GWAS (Russ et al. 2022; Vinkhuyzen et al. 2013), with some exceptions reaching up to 50% (Haddad et al. 2006; Jostins et al. 2012). We considered the ability of our methods to identify causal variants with heritability ranging from 0.1 to 0.5 by increments of 0.1.

2.6.2 Number of causal variants

Previous association studies have found that most traits are influenced by many causal variants with small effect (Dickson et al. 2010; Holland et al. 2020; Liu, Li, & Pritchard 2019; Mackay, Stone & Ayroles 2009; Strope et al. 2015; Watanabe 2019). We chose therefore to test a range of numbers of causal variants within our small n, p simulations and partition the heritability amongst them for small effect sizes.

Here we faced another set of trade-offs. Our framework and code base allows for arbitrary mappings from genotype to phenotype, but with increasing complexity comes increased difficulty in drawing apples-to-apples comparisons. We decided to restrict attention here to models in which the *m* causal variants contribute equally and additively to the trait as

$$Y = \sum_{i=1}^{m} \beta_i X_i + \epsilon \tag{2.22}$$

Thus, using the same fixed values for h^2 we replace f with f_i for each X_i and replace h^2 with $\frac{h^2}{m}$, calculating the coefficient value for each X_i as:

$$\beta_i = \sigma \sqrt{\frac{\frac{h^2}{m}}{\left(1 - \frac{h^2}{m}\right) f_i \, (1 - f_i)}} \tag{2.23}$$

We tested the following numbers of causal variants:

$$m = 1,\ 2,\ 5,\ 10 \tag{2.24}$$

In each simulation, the first m columns of X were taken to be the causal variants without loss of generality, since the columns of X are generated independently of one another and according to the same input parameters.

2.7 Methodologies

After generating each (X, Y) pair, we apply three association study methods in succession: standard GWAS, a new "GWAS-like" approach we call the E-value, and the feature selection approach LASSO. The output of each is a list of genetic variants ordered according to their relative significance under the particular method.

2.7.1 GWAS

By far the most familiar ranking of variable importance is the p-value. Under GWAS, each column vector of X is tested against Y, resulting in a p-value for the corresponding variable. The induced ranking is by statistical significance, with p-values ordered from smallest to largest.

In our study, the GWAS p-value can be equivalently computed via linear regression, a t-test or a correlation test. Having initially used both

the 'lm' (for regression) and 't.test' functions in R, computational efficiency led us to later compute correlations and significance value directly.

2.7.2 The E-value

Consider a single variable of interest in an otherwise null experiment. Let P_i denote its p-value, and recall that a p-value quantifies the probability of observing a result as or more extreme under the null hypothesis (Casella & Berger 2002). By definition, the expected proportion of variables with a p-value smaller than (or equal to) P_i is exactly P_i. By extension, if there are a total of p variables, then the expected number of variables with a p-value smaller than (or equal to) P_i is the product $p \times P_i$. In other words, $p \times P_i$ is the expected time to first success, i.e. when the variable of interest is encountered in the ranking. This all directly follows from fact that p-values are uniformly distributed on [0,1] under the null hypothesis (Casella & Berger 2002).

It turns out that not all null experiments are created equal. While each variable may correspond to a p-value distributed as uniform $U[0,1]$, correlation structure among the variables may induce a joint dependence structure among the p-values. In other words, the more correlated a pair of variables, the more tightly coupled their p-values (as in Figure 2.5). Thus, correlation structure can obscure the source of signal in an experiment. The p-value does not capture this phenomenon; under GWAS, we test each variant X_i against Y independently of the others. This motivates us to introduce a statistic which does, and which hopefully improves performance as a result.

Let P_j be a random variable denoting the p-value for variant j under the null hypothesis. Then each P_j is marginally distributed as Uniform(0,1), but they need not be independent of one another. Given the realised p-value $P_j = p_j$, define the E-value E_j as:

$$E_j = \mathbb{E}\left[\frac{1}{p-1}\sum_{k\neq j}\left(\mathbf{1}_{\{P_k<P_j|P_j=p_j\}} + \frac{1}{2}\mathbf{1}_{\{P_k=P_j|P_j=p_j\}}\right)\right] \tag{2.25}$$

Here $\mathbf{1}$ denotes an indicator function and the expectation is taken over the joint null distribution of p-values. In other words, E_j denotes the average proportion of spurious predictors with a p-value smaller than that of variant j, given that the p-value of variant j is as observed. The second term of the summand accounts for completely redundant predictors, i.e. variants whose column entries in X are identical, by treating their relative rank to each other as arbitrary. While the p-value reflects how strongly a signal deviates from the null case, the

E-value also accounts for ambiguity in that signal's source. In this context we may consider the E-value as an attempt to calibrate each association signal by its corresponding "neighbourhood of correlation": the correlation between an individual variant with the rest of the genome, creating a new (and hypothetically better) ranking of the variants.

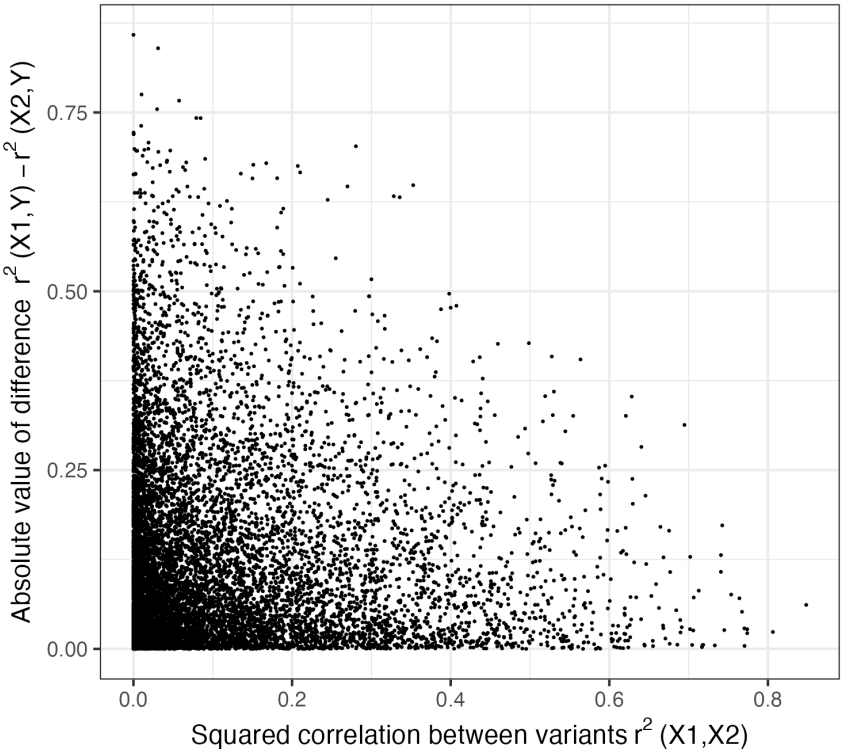


Figure 2.5: The more correlated a pair of variables, the more tightly coupled their p-values. Plot of 10000 simulations in which there are two variants and a phenotype vector whose correlations are randomly sampled from the $N(0,1)$ distribution. The squared correlation between the two variants $r^2_{X_1,X_2}$ is compared with the absolute value of the difference between the squared correlations $r^2_{X_1,Y}$ and $r^2_{X_2,Y}$ of each variant and phenotype (which are used to calculate their p-values) respectively. Note that here it is not assumed that either variant is causal.

To illustrate the difference between an E-value and a p-value calculated for the same variant, we will consider two simple limiting cases. Suppose we have a trait vector $Y_{n\times 1}$ determined by a single causal variant X_1, and that there is just one other (non-causal) variant X_2 under consideration. The LD between the two variants can be quantified as the squared correlation $r^2_{X_1X_2}$. In the absence of any linkage disequilibrium, $r^2_{X_1X_2} = 0$ and

$$\begin{aligned} E_1 &= \mathbb{E}\left[\frac{1}{2-1}\left(\mathbf{1}_{\{P_2<P_1|P_1=p_1\}} + \frac{1}{2}\mathbf{1}_{\{P_2=P_1|P_1=p_1\}}\right)\right] \\ &= \mathbb{P}(P_2 < P_1 \,|\, P_1 = p_1) + \frac{1}{2}\mathbb{P}(P_2 = P_1 \,|\, P_1 = p_1) \\ &= p_1 \end{aligned} \tag{2.26}$$

As there is no correlation between X_1 and X_2, and P_2 is uniformly distributed under the null hypothesis, this simply equates to the p-value of X_1. An equivalent derivation holds for X_2.

At the other extreme, the variants are identical, with $r^2_{X_1X_2} = 1$ and

$$\begin{aligned} E_1 &= \mathbb{E}\left[\frac{1}{2-1}\left(\mathbf{1}_{\{P_2<P_1|P_1=p_1\}} + \frac{1}{2}\mathbf{1}_{\{P_2=P_1|P_1=p_1\}}\right)\right] \\ &= \mathbb{P}(P_2 < P_1 \,|\, P_1 = p_1) + \frac{1}{2}\mathbb{P}(P_2 = P_1 \,|\, P_1 = p_1) \\ &= 0.5 \end{aligned} \tag{2.27}$$

Similarly for E_2. In other words, if the two variants are perfectly correlated, then they are indistinguishable in the ranking. Extending this to larger simulations with many non-causal variants that are not in correlation with X_1 and X_2, they will continue to place next to each other in the ranking, according to their E-value (no longer 0.5). Figure 2.6 below illustrates how E-value adjusts the p-values of variants according to their correlation with a causal variant and the strength of association between that causal variant and the phenotype across a scope of r^2 values.

Calculating the E-value in general posed a challenge, particularly as its computational complexity is of order p^2 (as compared to $O(p)$ for GWAS). To mitigate this, we again calculated the the relevant statistics and their significance levels directly.

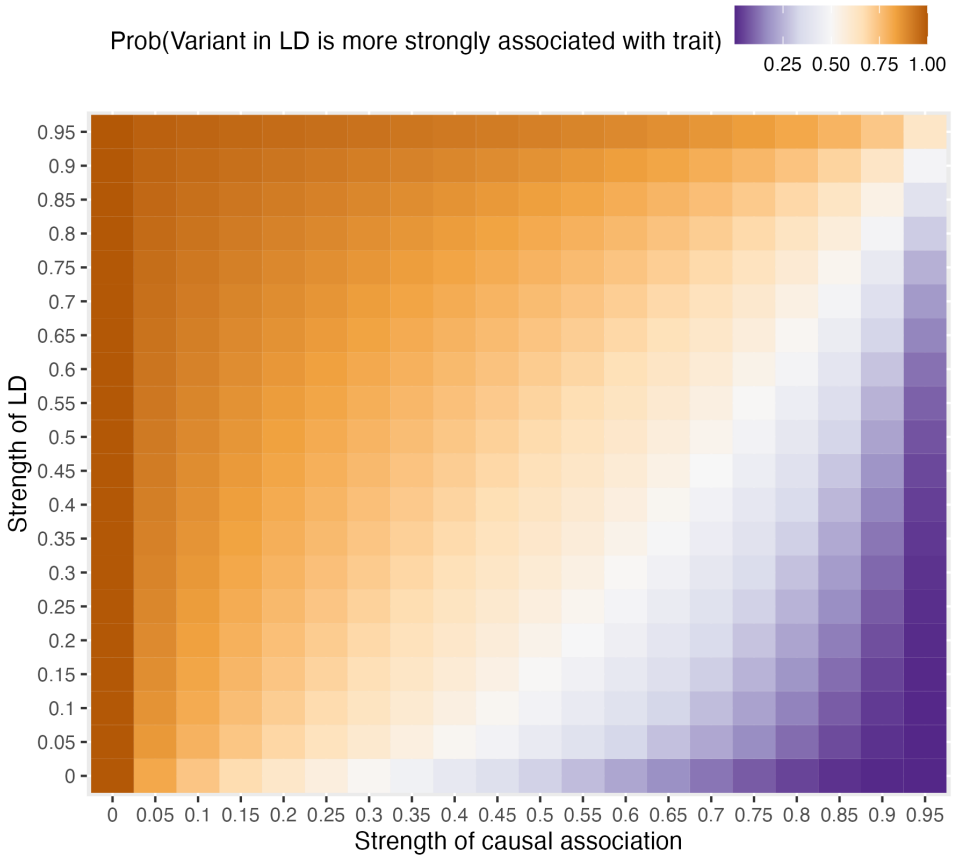


Figure 2.6: When LD with a non-causal variant and strength of causal association are large, then the strength of non-causal association is also large. E-value adjusts for this by computing the probability that the non-causal variant is more strongly associated with the trait than the causal variant. Heat map, for an indicative sample size, of the probability that a non-causal variant X_2 in LD with a causal variant X_1 is more strongly associated with a trait Y than the causal variant. The strength of the LD and causal association, measured by $r^2_{X_1,X_2}$ and $r^2_{X_1,Y}$ respectively, are as specified on the y and x-axes. When these are the only two variants, the probability is the E-value.

2.7.3 The LASSO

LASSO, or Least Absolute Shrinkage and Selection Operator, is a variable selection approach for linear regression and other statistical models. By introducing an additional summand to an Ordinary Lease Squares (OLS) expression, it is able to derive a unique solution for the estimated coefficients $\hat{\beta}_i$ in cases where the number of predictors X_i far outweighs the number of observations (i.e. $p >> n$) (Tibshirani 1996). The LASSO aims to minimise the following expression:

$$\sum_{i=1}^{p} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_i X_i) \right)^2 + \lambda \sum_{i=1}^{p} |\hat{\beta}_i| \tag{2.28}$$

where the first term is the familiar "residual sum of squares" and the second is a penalty on the magnitude of the regression coefficients. The nature of this penalty shrinks the $\hat{\beta}_i$ estimates, forcing most of them to zero (Tibshirani 1996). As in GWAS, each X_i is one of p variants, and each coefficient $\hat{\beta}_i$ represents the estimated effect of that variant on the trait. The size of the coefficient indicates the predicted relative importance of the variable, so these are used as our ranking mechanism: we order the coefficients by descending size of their absolute value and report the causal rank.

In OLS with many predictors, their interpretation is difficult and accuracy is often low, and when $p \geq n$ OLS fails entirely (Weisberg 1985). By contrast, LASSO avoids both pitfalls, simultaneously minimising variance (first terms) and restricting the number of predictor variables to interpret (second term) (Tibshirani 1996). As such, it produces interpretable models that are stable under small changes to the data (Tibshirani 1996). When predictors are correlated, as genetic variants often are, LASSO tends to pick one of them as representative of the group and shrink the others to zero (Nouira & Azencott 2021). Thus, as we increase LD through the injection of population structure, it is conceivable that the LASSO will more often shrink causal variants to select a correlated, non-causal variant. We used the 'glmnet' R package to implement the LASSO on our simulated dataset (https://cran.r-project.org/web/packages/glmnet/glmnet.pdf). The software uses 10-fold cross validation to select a model out of a sequence of models which is used to plot the relationship between the parameter λ and the mean squared error. As λ increases, mean square error increases and eventually variables are dropped. The authors of the LASSO recommend using the λ that is within one standard error of the minimum because it typically leads to a more parsimonious model with fewer included variables and a minimal loss in mean squared error. (Friedman et al. 2010). With the chosen value for λ it is a simple matter to get a function that will generate the coefficients for the model. As the coefficients of many of the X_i are shrunk to zero by LASSO, there is potential for "ties" in significance ranking between variants. However, in our study the success rate of LASSO in finding the causal variant is so high that this is a non-issue: causal variants are never reported to have no significance (Figure 3.5).

2.8 Performance metric

In biology, and more broadly, one is often faced with a large set of variables (matrix X, a.k.a features or covariates), within which only a few are expected to have relevance to a

response (vector Y). This has the structure of a standard prediction problem in which Y is modelled as a function of X and a common approach, irrespective of methodology, is to quantify the importance of each variable and use that as a ranking of potential relevance. One biological application of this ranking is sequential experimental validation, which is the case here. Sequential validation gives a linear cost to computational false positives, motivating our interest in "time to first success". Suppose each variable associates either with a null distribution (negative) or an alternative distribution (positive); the time to first success corresponds to the first positive case encountered when proceeding down the variable importance ranking. In this study, the negative case refers to a non-causal variant, and the positive to a causal variant.

The output of our GWAS, E-value and LASSO algorithms is a *causal rank:* the position (rank) of the causal variant within the ordering of p-values or coefficients calculated by the algorithm for all the variants. If there are multiple causal variants, we take the smallest rank. For a causal rank of k, where $k > 1$, there are $k - 1$ non-causal variants that are computational false positives. In light of the above, we call this the *empirical false discovery rate* and use it to create a metric of method performance. As we are comparing between different numbers of total causal variants, we will scale the value $k - 1$ by $1/p$, and it is scaled empirical false discovery rate that will measure performance in all of our simulations.

2.9 Overall experimental design

The parameters that we have described in this section can be separated into those that govern the simulation of X, those that affect the generation of Y and the methodologies for resolving the function mapping between X and Y. We wanted to conduct sufficient simulations under each combination of parameters to capture a breadth of population genetic variation without making the procedure computationally prohibitive. As such, we structured the study as follows (Figure 2.7):

The genotype matrix is a function of its dimensions (n, p) [4 levels] and algorithmic degree of dependency d [40 levels]. For each of these $40 \times 4 = 160$ specifications, we simulated 1000 independent X matrices as previously described. To quantify the relationship at each (n, p) level between d and our principled measure of structure (average LD, see Figure 2.4), we computed the mean value across the 1000 replicates.

Additionally, for each these $40 \times 4 \times 1000 = 160000$ genotype matrices, we considered varying degrees of heritability [5 levels] and numbers of causal variants [4 levels]; in each of these $5 \times 4 = 20$ cases, we generated a single phenotype vector Y. In total, we thus created $160000 \times 20 = 3200000$ distinct (X, Y) pairs. For each, we applied GWAS, E-value and the LASSO, reporting the rank and MAF of the first identified causal variant. Let $\Theta = (n, p, d, h^2, \#\text{causal})$ denote the vector of parameters governing the simulations. Then, as described above, each unique Θ corresponds to 1000 repetitions; however, each (Θ, X) pair generates a single phenotype vector Y.

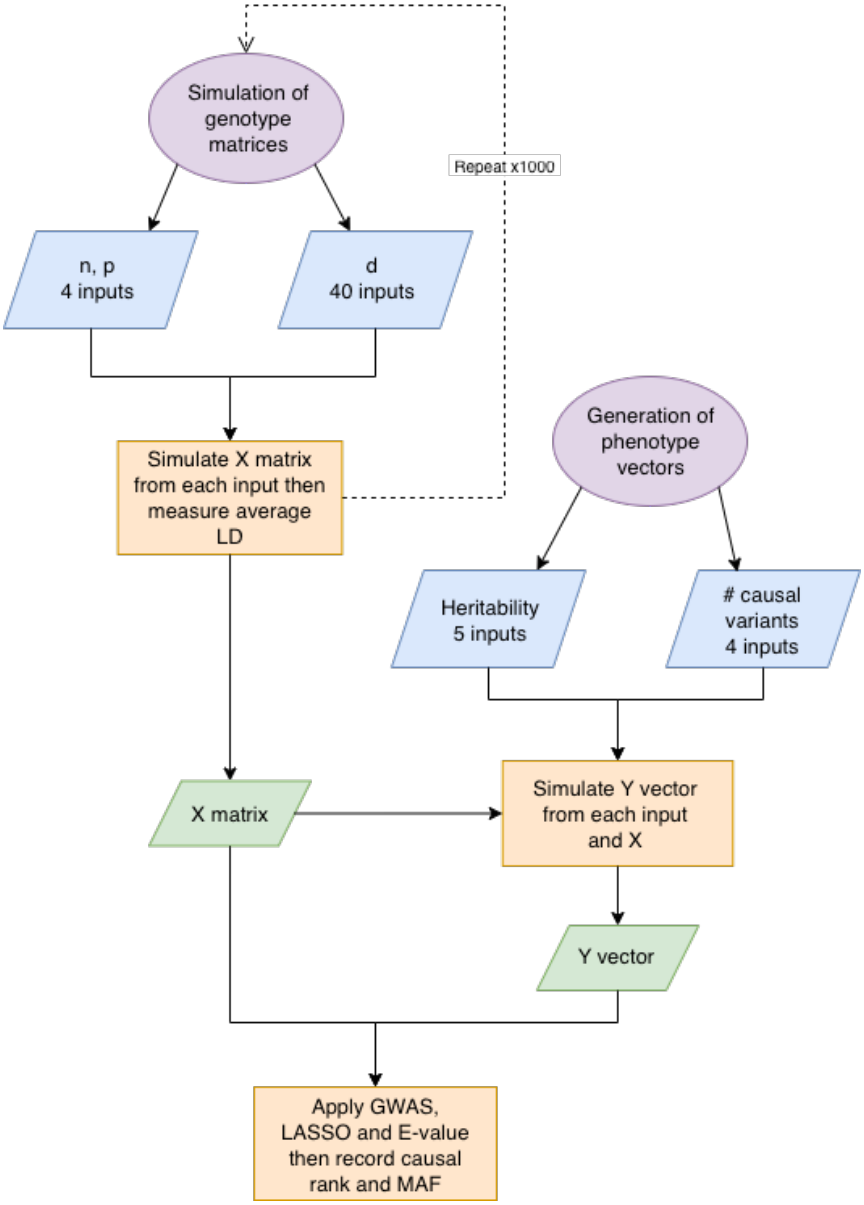


Figure 2.7: Experimental design workflow. Flow chart of the process of simulating genotype matrices, generating phenotype vectors and applying the association study methods to them as described in Chapter 2 of this thesis.

2.10 Algorithmic implementation

The methods described in this chapter and analyses of results were conducted using R version 4.1 (www.r-project.org). The libraries used were tidyverse (tidyverse.org) and glmnet (https://cran.r-project.org/web/packages/glmnet/glmnet.pdf).

The simulation of the genotype and phenotype datasets and application of GWAS, the E-value and the LASSO were performed on the National Computational Infrastructure (NCI) supercomputer Gadi. Gadi is a 4,962 node supercomputer comprising Intel Sapphire Rapids, Cascade Lake, Skylake and Broadwell CPUs and NVIDIA V100 and DGX A100 GPUs which uses the Rocky Linux 8 operating system. All other analysis was performed using RStudio on a MacBook Pro with a 2 GHz Dual-Core Intel Core i5 processor on which the operating system was macOS Monterey Version 12.6.

The script used to generate the data and apply the methodologies to it can be found in the GitHub repository honoursthesis (https://github.com/laura-hardy/honoursthesis).

Chapter 3

Results

This chapter reports on the application and subsequent evaluation of association study methodologies presented in the Methods (Chapter 2). Having constructed a framework for the simulation of genotype-phenotype data under prescribed conditions, we applied the methodologies to the generated data, aiming to assess their ability to recover the function mapping between genotype and phenotype. We present the results of this performance assessment, first under the perturbation of one parameter at a time, then as a product of a combination of conditions.

We find that the simulation framework and proposed measures of structure and method performance are fit for purpose. Applying the methodologies to simulated populations, we attain results for individual factor manipulations which are aligned with previous predictions, and observe the combinatorial impact of these factors on performance.

Throughout this chapter, "Performance" will refer to our performance metric as defined in Section 2.8, the scaled empirical false positive rate. As such, "good" performance will correspond to a low scaled empirical false positive rate, and "bad" performance to a high one.

3.1 Simulated populations

Our implementation on NCI of the design in Section 2.9 required some elementary parallelisation. We divided the task into four separate scripts according to (n, p) and then split them into many parallel jobs. This, and the corresponding run time and compute usage, is summarised in Table 3.1.

Table 3.1: Compute time and service unit utilisation on Gadi for sets of simulations.

(n, p) of simulations	Number of (X, Y) pairs simulated	Number of parallel jobs on Gadi	Approximate time per job (minutes)	Approximate number of service units per job (SU)	Total compute (kSU)	Total linear compute time (days)
20, 100	800000	400	5	0.8	0.32	1.375
20, 1000	800000	400	120	22	8.8	33
100, 100	800000	400	10	2	0.8	2.75
100, 1000	800000	800	90	16	12.8	50

3.2 Testing individual factors

In this section, we investigate how performance varies with individual factors by averaging (marginalising) across the remainder of factors for specified (n, p) and method. We quantify performance as the empirical false discovery rate, scaled by $\frac{1}{p}$.

3.2.1 As heritability increases, performance of GWAS and the E-value improves correspondingly, while the LASSO demonstrates consistent high performance

As expected, both GWAS and the E-value perform better for higher heritability values, although the gains begin to plateau at $h^2 = 0.3$ when $n = 100$, and $h^2 = 0.4$ when $n = 20$ (Figure 3.1). The LASSO's performance stays relatively constant for each heritability (Figure 3.1) since average performance, while a good benchmark for comparison with the other methods, lacks sufficient resolution to demonstrate the impact of heritability on the LASSO. As we will see throughout this section, the LASSO demonstrates consistently high performance across all scenarios (Figure 3.2, 3.3, 3.4), so we will be unable to comment extensively on it until Section 3.3, when we study cross-sections of more input parameters.

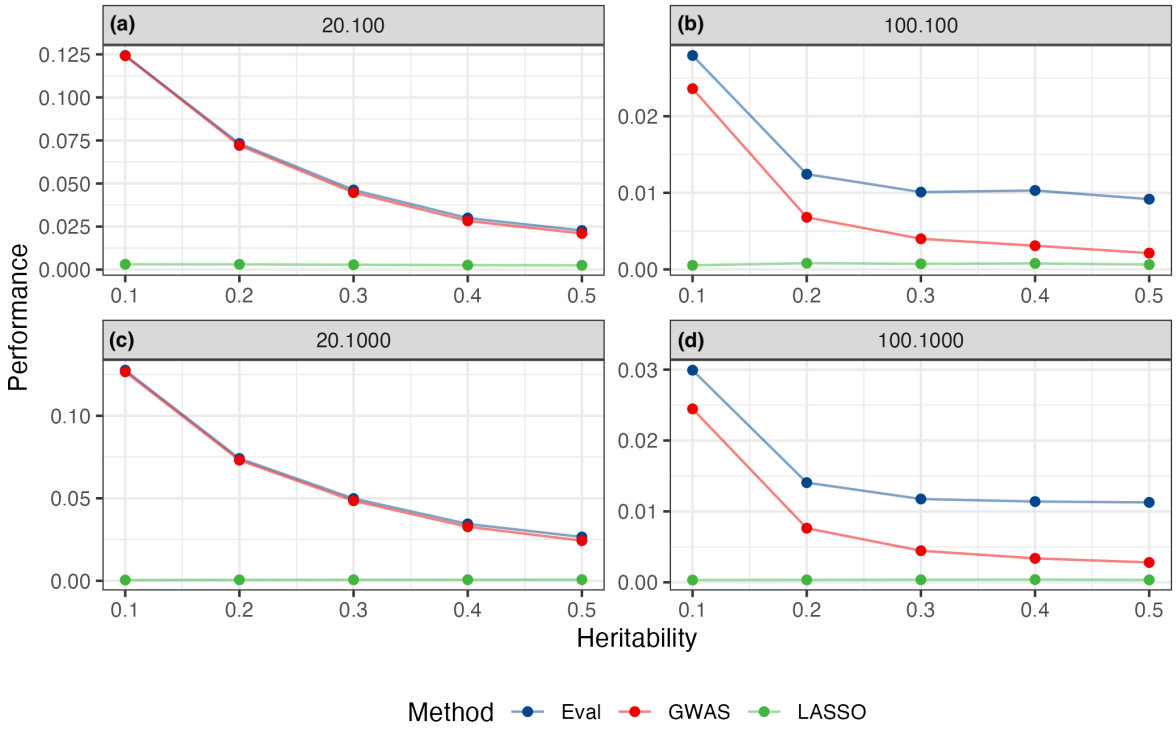


Figure 3.1: Performance improves as heritability increases. Each point represents average performance for one method over 200000 simulations generated under the indicated heritability value**.** The dimensions of each population are (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000). Solid lines on the graph connect the plotted average performance values, coloured by method. The error bars in this graph are sufficiently small to be omitted.

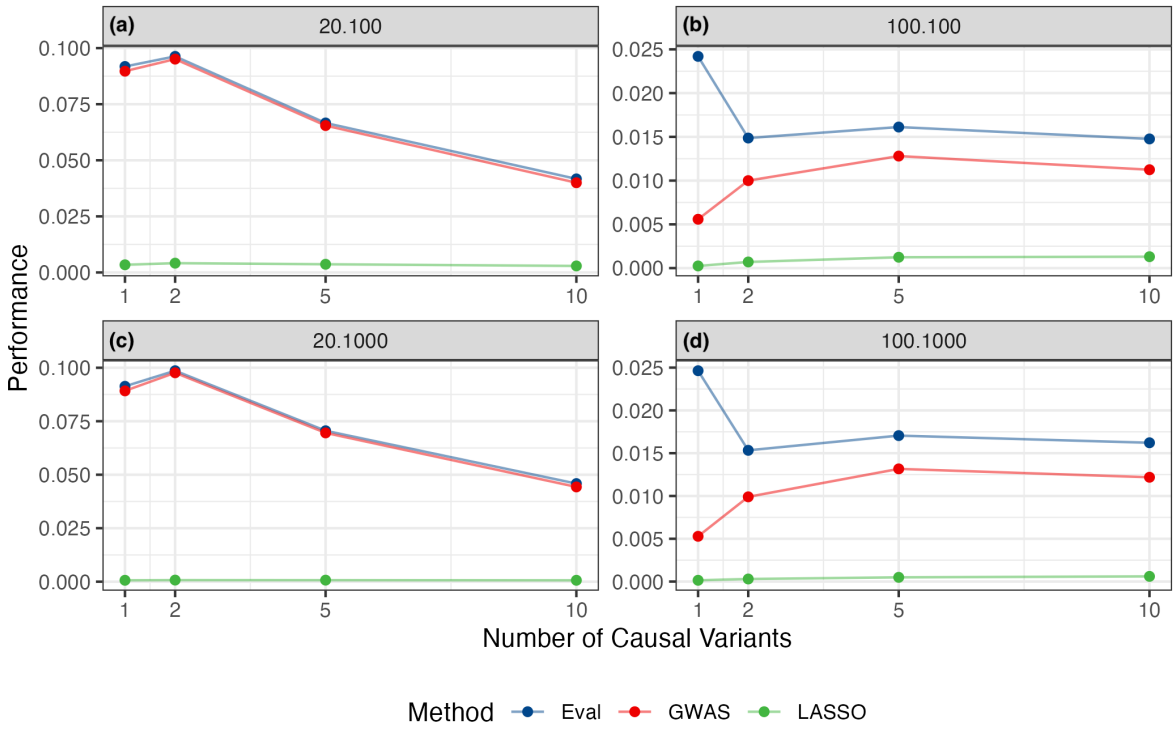


Figure 3.2: Average performance of methods against the number of causal variants used to generate phenotype. Each point is an average of performance across 250000 simulations (50000 for each heritability value) with phenotypes generated by the specified number of causal variants. The dimensions of each population are (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000). Solid lines on the graph connect the plotted average performance values, coloured by method. The error bars in this graph are sufficiently small to be omitted.

3.2.2 In most circumstances, the "time" to find the first causal variant is reduced when there are many causal variants used to generate the phenotype

Figure 3.2 exposes that for $n = 20$, both GWAS and the E-value perform worst when there are 2 causal variants, but see improved performance for further increases to the number of causal variants. The LASSO retains stable, high performance levels. Contrastingly, when $n = 100$, the E-value and GWAS diverge, with the E-value performing worst with 1 causal variant and then remaining relatively constant, while GWAS performing best with one causal variant and worsens as the number increases. The LASSO, although still superior, worsens slightly as number of causal variants increases in the $n = 100$ case. Despite this, the worst performance value for $n = 100$ for all methods is still better than the best performance value for $n = 20$. All three methods demonstrate minimal change to performance when increasing the number of causal variants from 5 to 10.

Recall that total heritability is partitioned amongst the causal variants, so each depicted average performance in Figure 3.2 encompasses a range of simulations with differing effect sizes. These plots are less informative than the corresponding ones in Section 3.3 that put performance at each value for number of causal variants in the context of the heritability assigned to that simulation.

3.2.3 Performance of GWAS and the E-value is worse on more highly structured populations when n is 100, and fluctuates as structure changes when n is 20, while the LASSO is minimally affected

We notice that average performance across intervals of average LD gets worse for GWAS and the E-value in particular when $n = 100$ (Figure 3.3). Although we do not see the same story when $n = 20$, later, when we consider how structure and MAF of the causal variant interact, what we observe is quite different. The LASSO also depicts relatively minor perturbation as structure varies, but still performs strongly across the breadth of population structures (Figure 3.3).

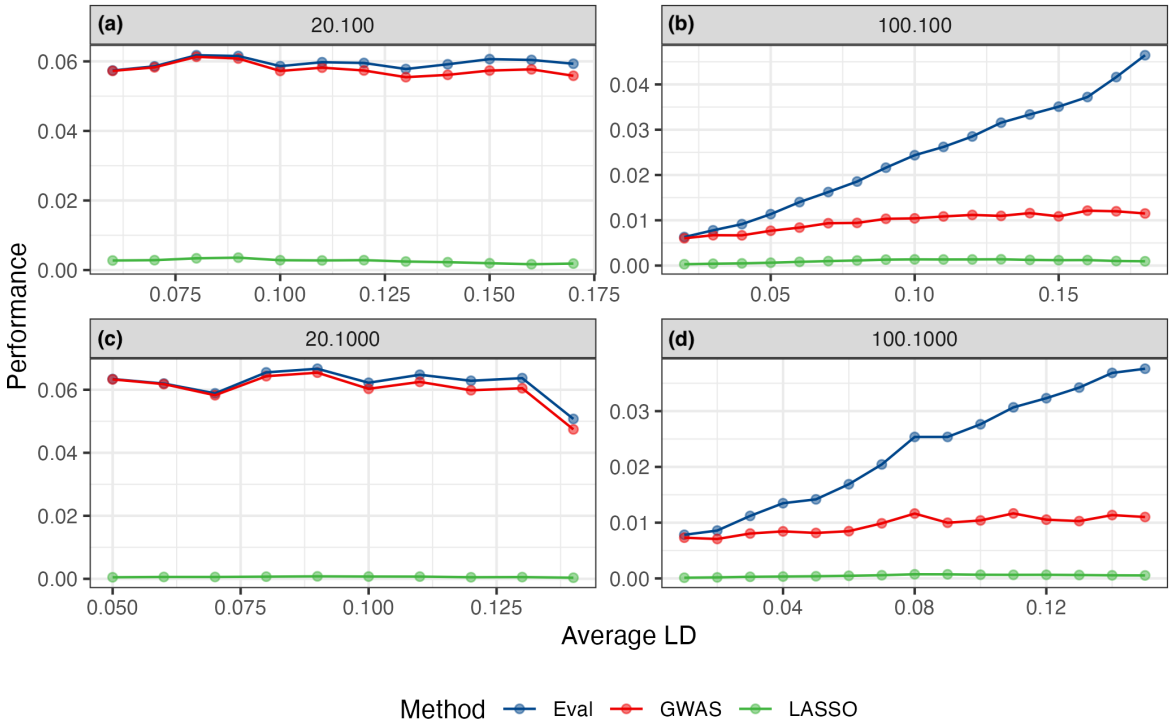


Figure 3.3: Population structure impacts average performance. Each point represents average performance of one method across populations whose average LD falls within an interval of width 0.01. Intervals of average LD with less than 2000 simulations contained within them were omitted. The dimensions of each population are (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000). Solid lines on the graph connect the plotted average performance values, coloured by method. The error bars in this graph are sufficiently small to be omitted.

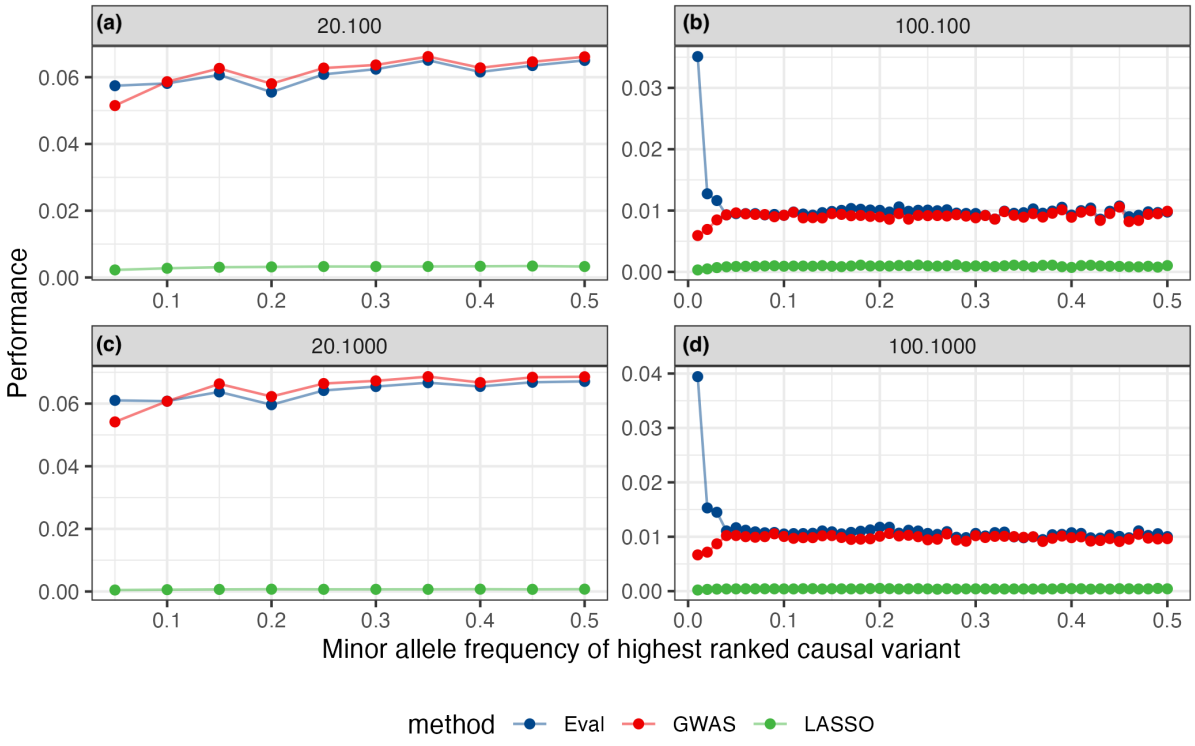


Figure 3.4: Rare variants receive better performance scores, on average, when they are the first causal variants to be identified. Performance, which is a function of the number of false positives ranked higher (by significance, e.g. p-value) than the highest ranked causal variant, is averaged here over all simulations in which the highest ranked causal variant had the MAF shown. The dimensions of each population are (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000). Solid lines on the graph connect the plotted average performance values, coloured by method. The error bars in this graph are sufficiently small to be omitted.

3.2.4 All methods give rare causal variants a better average rank than common variants when they are the first causal variant to be identified, except for the E-value when n=100

All three methods demonstrate fairly consistent performance on average across common alleles. When the first causal variant to be identified is a rare one (MAF ≤ 0.05), it is on average given a slightly better scaled empirical false positive rate (better performance) by GWAS and the LASSO, relative to common causal alleles. The same is true for the E-value when $n = 20$, but when $n = 100$, average performance worsens for rare alleles, particularly for singletons (Figure 3.4). Note that Figure 3.4 does not discriminate between how many causal variants there were or what effect size they had, and both of these factors evidently influence performance (Figure 3.2 and Figure 3.1).

3.2.5 The LASSO significantly outperforms GWAS and the E-value, while GWAS exhibits similar patterns to the E-value when n=20 and outperforms it when n=100

The LASSO performs better than both GWAS and the E-value, both on average overall (Figure 3.5) and averaged by heritability (Figure 3.1), number of causal variants (Figure 3.2), average LD (Figure 3.3) and MAF of highest ranked causal variant (Figure 3.4). When $n = 20$, we observe nearly identical performance averages between GWAS and the E-value for heritability (Figure 3.1), number of causal variants (Figure 3.2) and average LD (Figure 3.3). Contrastingly, when $n = 100$, GWAS consistently outperforms the E-value (Figures 3.1, 3.2, 3.3, 3.5), with the gap widening for highly structured populations (Figure 3.3). Importantly, performance on specific simulations is consistent between GWAS and the E-value; for populations and corresponding phenotypes where the E-value performs poorly, GWAS also performs poorly, and vice versa (Figure 3.5).

3.2.6 Increasing n improves performance of GWAS and the E-value, while p has no impact on average performance of any method

The average performance of GWAS and the E-value showed a marked improvement as n was increased from 20 to 100 (Figure 3.5). The effect of n on average performance of the LASSO was negligible at the current resolution (Figure 3.5). We observed that p has no impact on performance averaged across heritability (Figure 3.1), number of causal variants (Figure 3.2), average LD (Figure 3.3) or MAF of highest ranked causal variant (Figure 3.4).

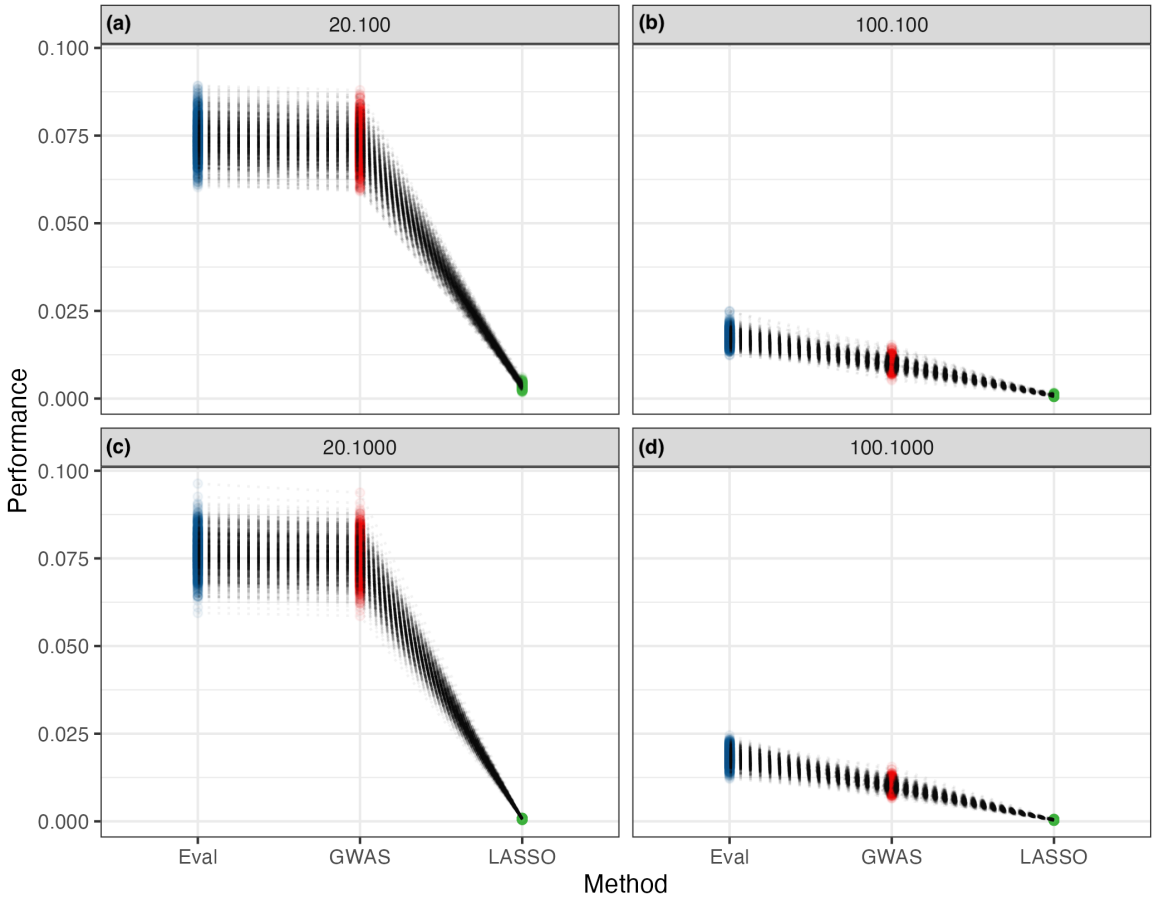


Figure 3.5: Average performance over sets of specific simulations is consistent between GWAS and the E-value, while the LASSO outperforms them both. There are 1000 points plotted for each method, each of which represents an average of performance of the given method across 40 distinct populations and 20 different phenotypes per population. These 800 population-phenotype pairs are generated under all 800 unique combinations of possible input values for structure, heritability and number of causal variants. The dotted lines between methods indicate correspondence between averages taken over the same set of populations. The dimensions of each population are (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000). The points are densely stacked on top of one another, and are coloured blue (E-value), red (GWAS) or green (LASSO) by method.

3.3 Testing combinations of factors

In this section we explore method performance (scaled empirical false discovery rate) with cross-sections of multiple parameters, i.e. how does performance vary due to the interaction of multiple factors.

3.3.1 The higher the heritability distributed amongst a fixed number of causal variants, the better GWAS and the E-value perform

Figure 3.6 dissects the performance for each method (grouped columns) at each study size (grouped rows). Within each grouping, results are further aggregated by heritability (column) and number of causal variants (row). Note that along rows of simulations whose

phenotypes were generated by the same number of causal variants, performance of GWAS and the E-value increases correspondingly with heritability for all n, p.

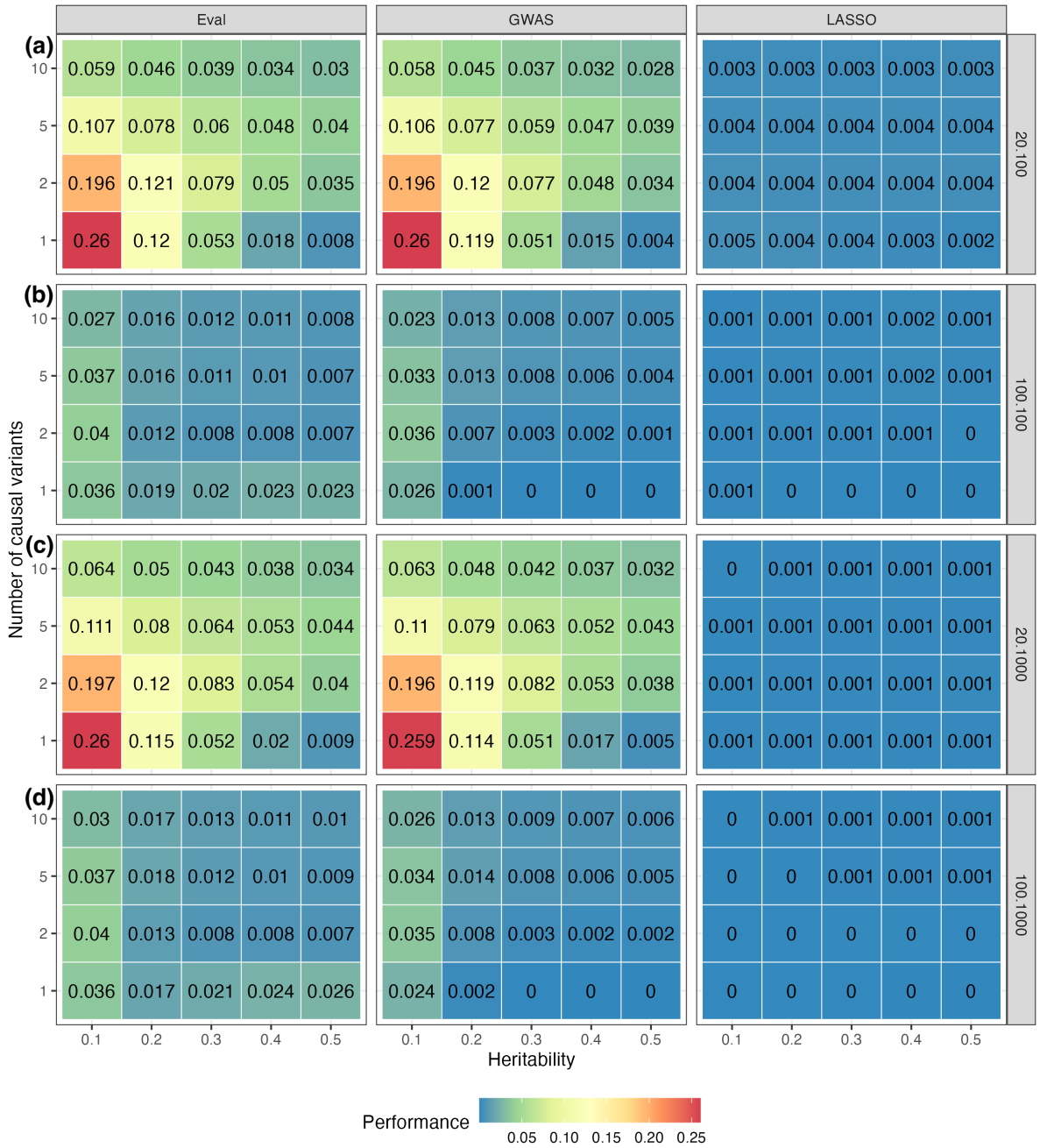


Figure 3.6: Heat map of the performance of each method for varying heritabilities and number of causal variants. The average performance of the E-value (left), GWAS (centre) and the LASSO (right) over the specified heritability and number of causal variants. Each colour-coded block is labelled with the average performance of the 40000 simulations aligned with the input parameters used to generate it. The dimensions of the simulated populations are (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000).

3.3.2 For low heritabilities, GWAS and the E-value performance tracks with increased number of causal variants, but this does not hold as heritability increases

As the number of causal variants increases along an axis of fixed heritability (left to right within groups), performance does not display a consistent pattern for either method (Figure 3.6). For $n = 20$, when $h^2 = 0.1$ performance of both methods improves as the number of causal variants increases, but for higher heritabilities the worst scores are attained when there are 2 or 5 causal variants. Contrastingly, for $n = 100$ GWAS performs extremely well for 1 causal variant, while the E-value does best for 2 and performs worst on average at 1.

Further cross-sectioning the performance of the E-value at $n = 100$ by structure, we observe that for one causal variant, the performance in highly structured populations is far worse than for other numbers of causal variants (Figure 3.7).

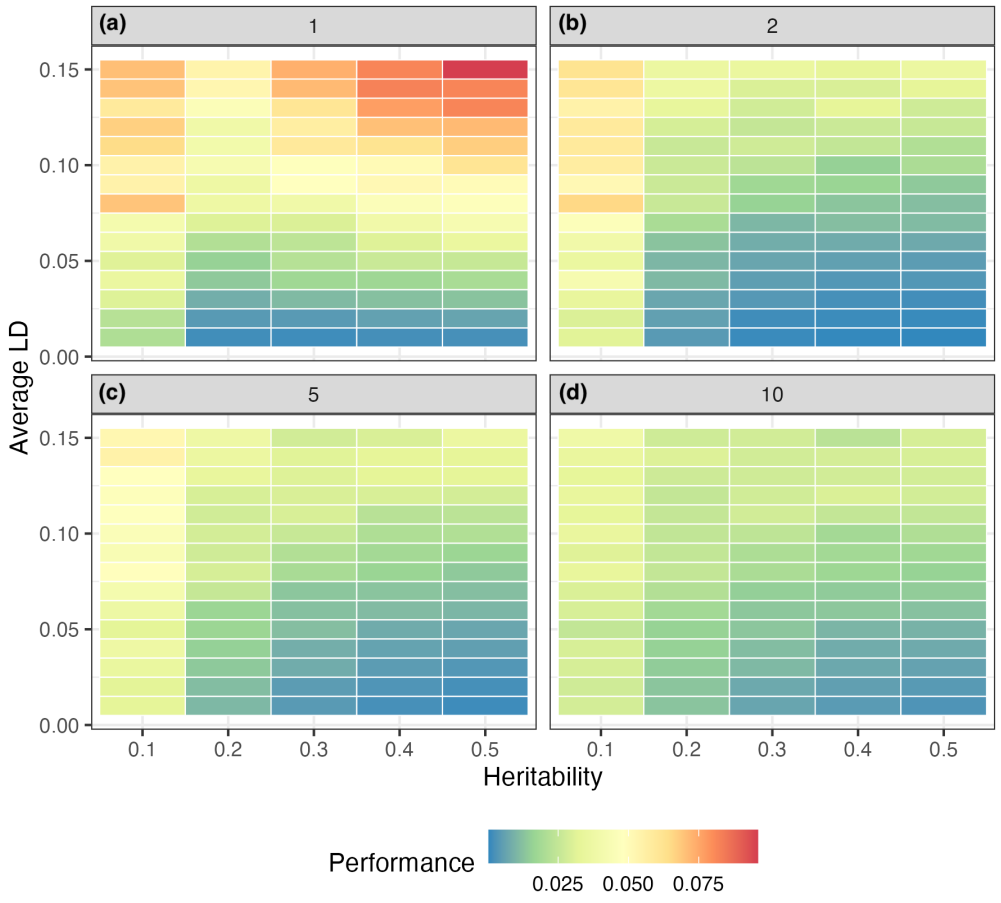


Figure 3.7: The E-value performs poorly on highly structured populations when there is only one causal variant, particularly for high heritabilities. Heat map of average performance across populations with the specified heritability, average LD, and number of causal variants (a) 1, (b) 2, (c) 5 or (d) 10. The result is shown for simulations with $n = 100$ individuals and $p = 1000$ variants, but is indicative of performance at $n = 100$, $p = 100$ (Figure A1). We fixed a cutoff of average LD ≤ 0.15 so that all averages were taken over at least 100 simulations.

3.3.3 The LASSO performs very well for all heritabilities and numbers of causal variants, but when n=20 and p=100 it exhibits similar trends to GWAS

The average performance of the LASSO for any combination of heritability and number of causal variants is never worse than 0.005 (or 0.002 if excluding $n = 20, p = 100$) (Figure 3.6). For $n = 20$, $p = 100$, its performance also increases with heritability for fixed number of causal variants, and varies with the number of causal variants for fixed heritability, in an analogous manner to GWAS.

3.3.4 Under increasing structure, all three methods are worse at identifying common variants, but GWAS and the LASSO perform much better than the E-value when the first variant they find is rare

The average performance of GWAS, the E-value and the LASSO is worse for populations with high levels of average LD and when the first causal variant to be identified is common (Figure 3.8). However, when the MAF of the causal variant which determines the performance score is low (< 0.05 when $n = 20$ and < 0.02 when $n = 100$), performance is comparable to an unstructured population (Figure 3.8). The only exception to this is the E-value, which performs particularly badly when MAF $= 0.01$ (i.e. when the causal variant is a singleton) for $n = 100$ (Figure 3.9).

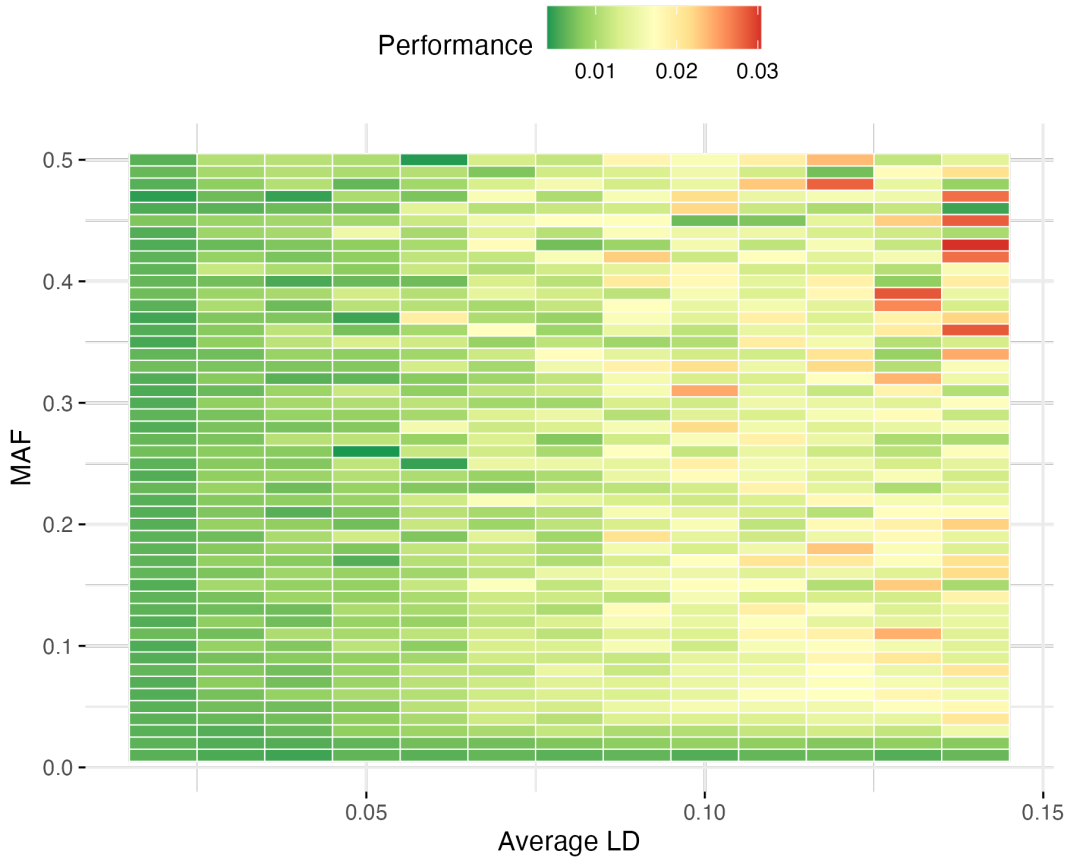


Figure 3.8: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for GWAS on a population of 100 individuals and 100 causal variants, but the trend in results is comparable for GWAS and the LASSO for all dimensions, and for the E-value when $n = 20$ (see figures A2-A10).

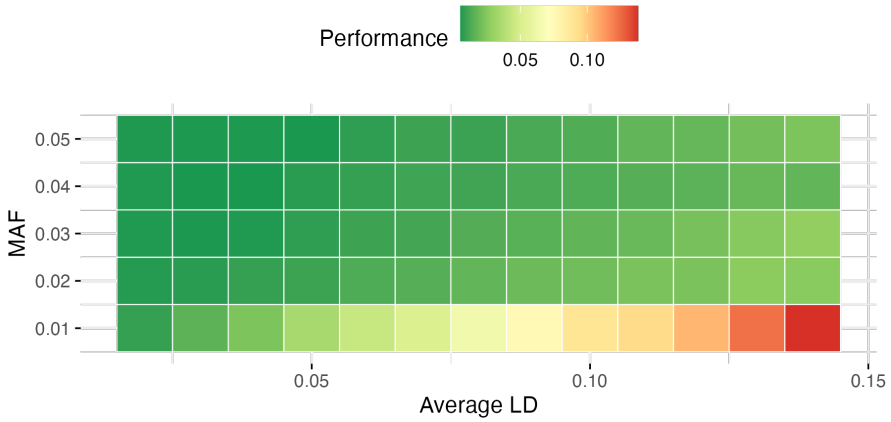


Figure 3.9: Increasing structure decreases the performance of the E-value, particularly for very rare variants. Heat map of average performance of E-value over the simulations with the specified average LD and MAF of highest ranked causal variant. MAF values are truncated at 0.05 for higher resolution of the performance at singletons. This result is depicted for a population of 100 individuals and 100 causal variants, but the results are comparable for $n = 100$, $p = 1000$ (see figure A11).

3.3.5 Increasing structure perturbs the allele frequency distribution of the causal variant that contributes to the performance score away from the Ewens sampling distribution

Recall that we sampled overall allele frequencies for each population from the Ewens distribution and used these to simulate genetic variants. By fixing heritability, we disentangled a variant's allele frequency from the statistical power to detect it, suggesting that — absent other factors — the distribution of minor allele frequencies among the most highly ranked causal variant should mirror the initial Ewens distribution (depicted in Figure 2.1). This was not always the case. When $n = 20$, all methods favour rare variants in highly structured populations (Figure 3.10) and when $n = 100$, GWAS and the LASSO both identify high proportions of singletons. Contrastingly, the E-value performs badly on singletons when $n = 100$ (Figure 3.9), preferring doubletons in highly structured populations (Figure 3.11).

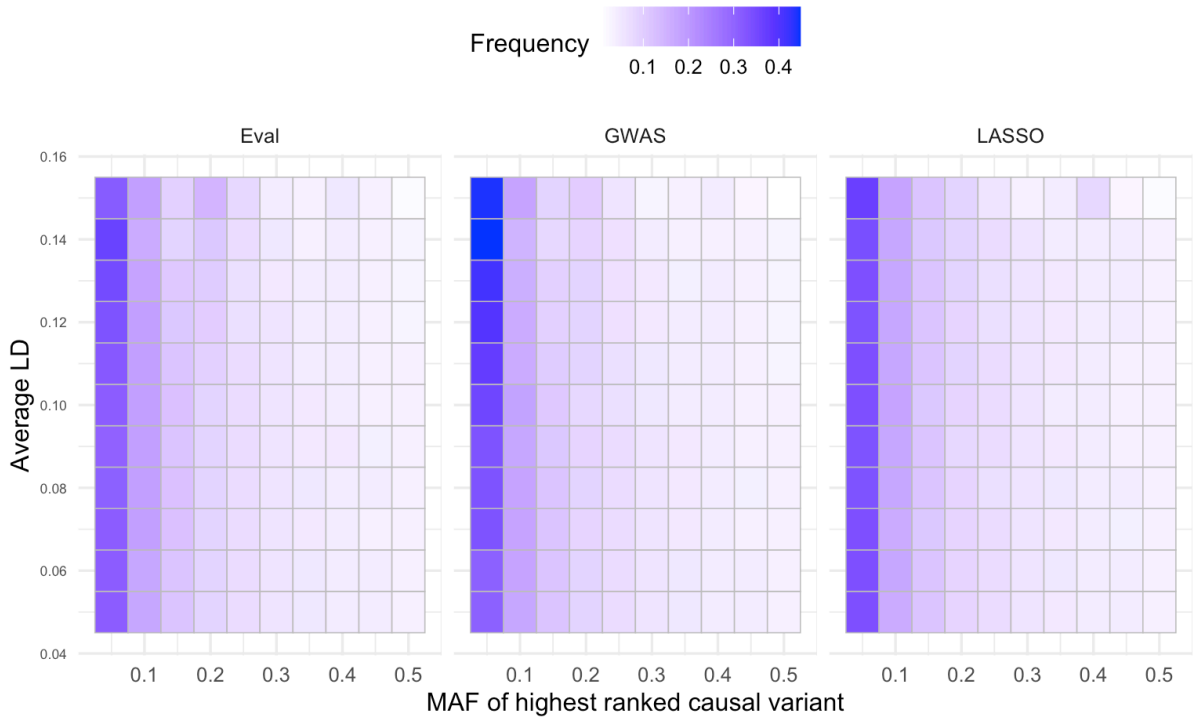


Figure 3.10: Increasing structure in very small populations makes rare variants more visible relative to common variants. Heat map depicting the frequency distribution of the causal variant first identified by the specified method in populations with increasing amounts of structure. These simulations were of populations with $n = 20$ individuals and $p = 1000$ variants, but the results are comparable for $n = 20, p = 100$ (Figure A12).

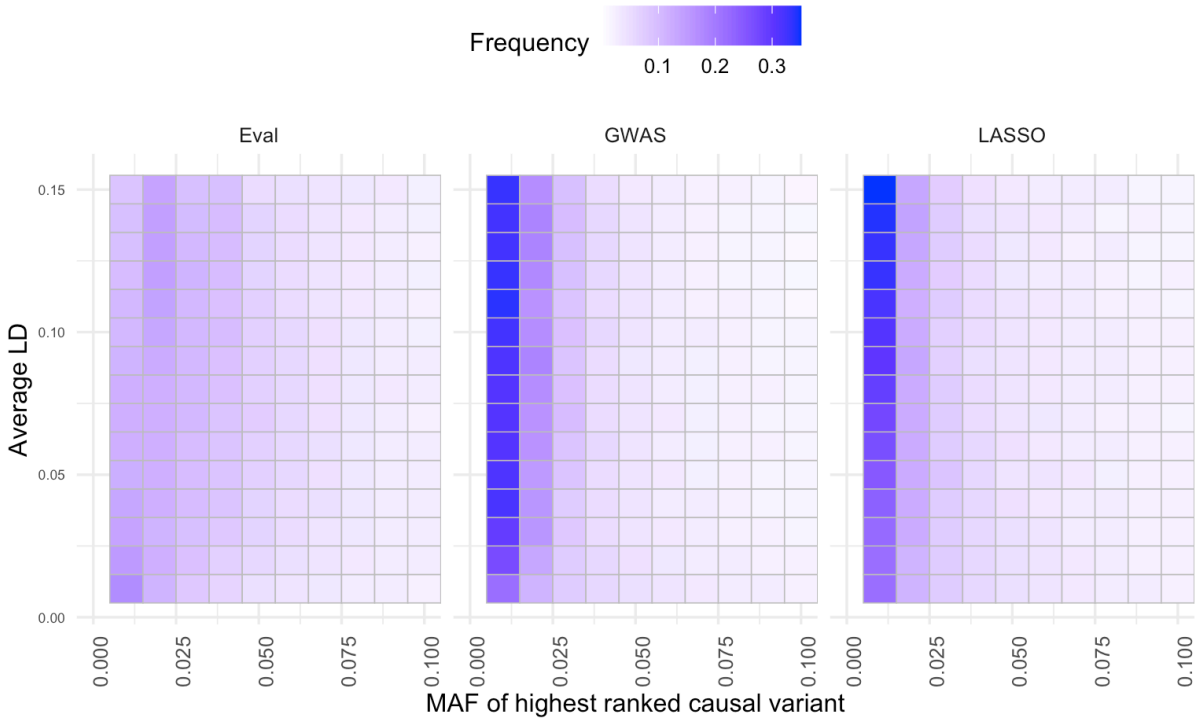


Figure 3.11: Increasing structure in small populations makes rare variants more visible to GWAS and the LASSO relative to common variants, while the E-value prefers doubletons to singletons. Heat map depicting the frequency of the first variant to be identified by the specified method in populations with increasing amounts of structure. This plot is truncated at MAF = 0.1 for higher resolution of the frequency of rare variants. The simulations were of populations with $n = 100$ individuals and $p = 1000$ variants, but the results are comparable for $n = 100$, $p = 100$ (Figure A13).

When we take a further cross-section of these results against the number of causal variants used to simulate phenotype, it becomes evident that in the $n = 20$ case, GWAS and the E-value show increased propensity to identifying singletons over common variants as the number of causal variants increases (Figure 3.12). Alternatively, when $n = 100$, the more causal variants there are to choose from, the more GWAS and the LASSO rank singletons highly, and the more E-value selects doubletons (Figure 3.13). We did not observe any analogous notable difference in the distributions when slicing by heritability (data not shown).

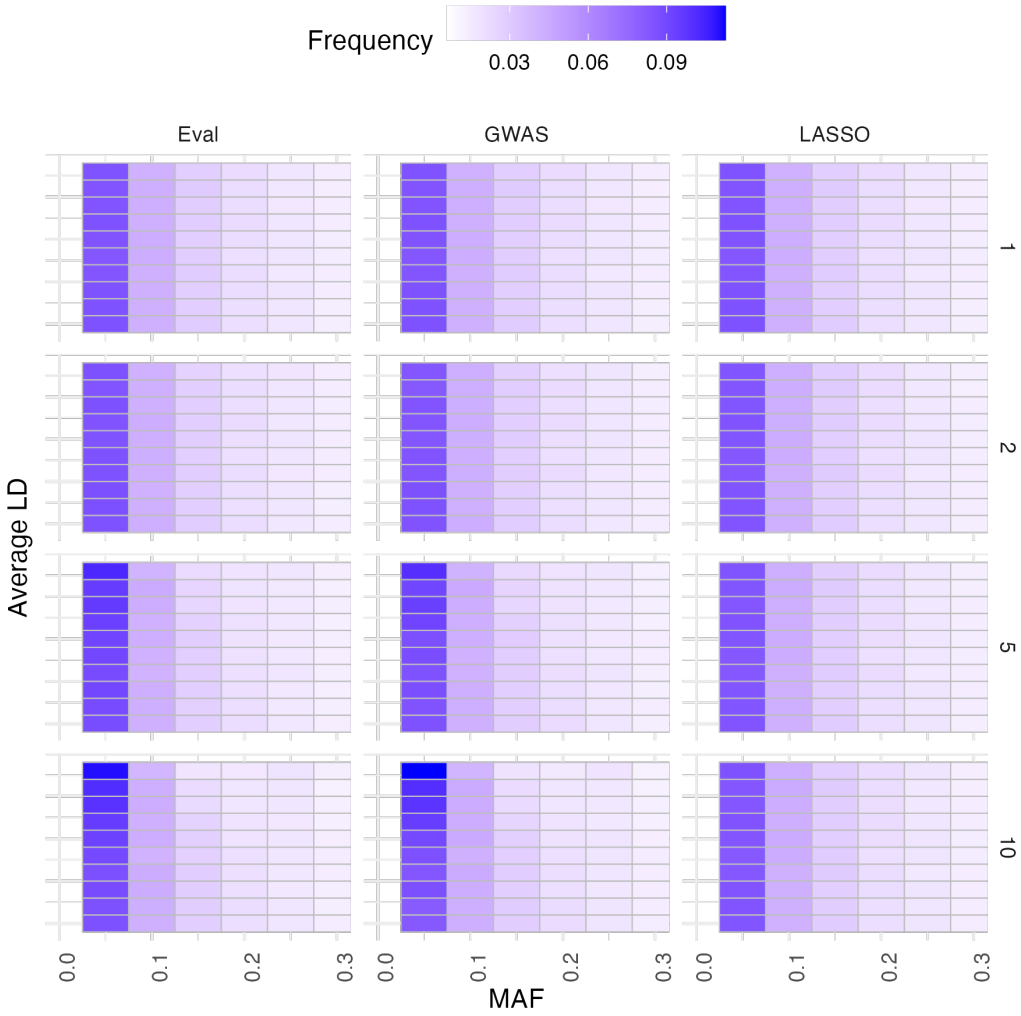


Figure 3.12: In very small populations, higher numbers of causal variants skew the MAF distribution of identified variants further toward the rare end for GWAS and the E-value. Heat map depicting the frequency of the first variant to be identified by the specified method (shown on the top axis) in populations with increasing structure (shown on the left axis, increasing bottom to top for each panel), simulated with 1, 2, 5 or 10 causal variants (shown on the right axis). This plot is truncated at MAF = 0.3 for higher resolution of the frequency of rare variants. The simulations were of populations with $n = 20$ individuals and $p = 1000$ variants, but the results are comparable for $n = 20$, $p = 100$ (Figure A14).

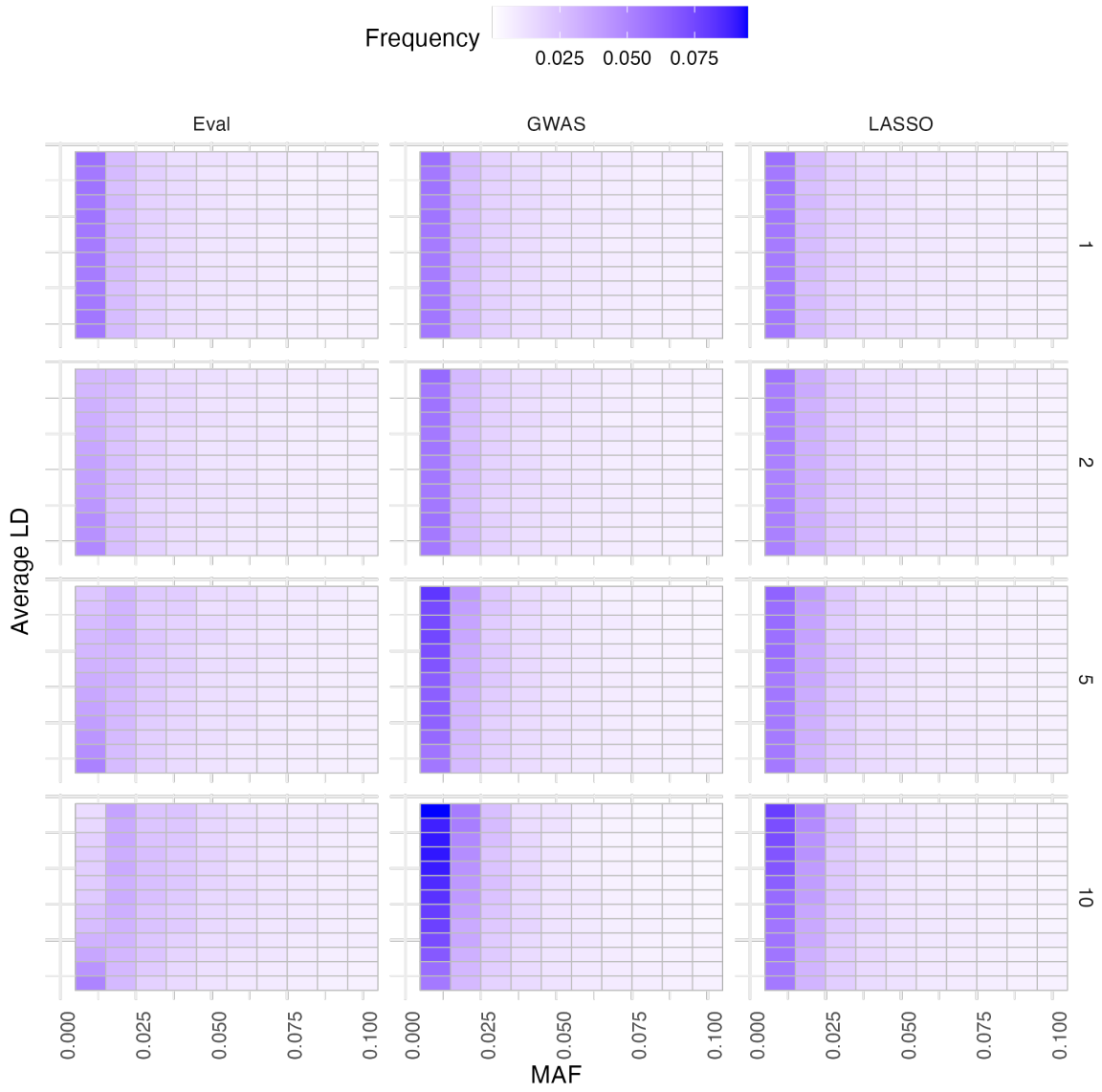


Figure 3.13: In small populations, higher numbers of causal variants skew the MAF distribution of identified variants further toward singletons for GWAS and the LASSO, while the E-value skews towards doubletons. Heat map depicting the frequency of the first variant to be identified by the specified method (shown on the top axis) in populations with increasing structure (shown on the left axis, increasing bottom to top for each panel), simulated with 1, 2, 5 or 10 causal variants (shown on the right axis). This plot is truncated at MAF = 0.1 for higher resolution of the frequency of rare variants. The simulations were of populations with $n = 100$ individuals and $p = 100$ variants, but the results are comparable for $n = 100$, $p = 1000$ (Figure A15).

3.3.6 Although all three methods display overall correlation in performance, for a fixed number of causal variants, the only correlation is between GWAS and the E-value.

When $n = 20$, the correlation between average performance of GWAS and the E-value is close to 1, as are the correlations split by number of causal variants (as seen in Figure 3.14). When $n = 100$, the overall correlation is negative, but the correlations split by number of causal variants are positive, however still significant (Figure 3.14). This is not the case for the two other pairwise comparisons of methods. When $n = 20$, a correlation of approximately 0.3

($p = 100$) or 0.1 ($p = 1000$) is reported between the LASSO and the other two methods, however when performance averages are divided by number of causal variant, the correlation is close to zero (Figure 3.15, 3.16). Similarly, when $n = 100$ GWAS and the LASSO are positively correlated (with magnitude approximately 0.6), and the E-value and the LASSO negatively correlated (with magnitude approximately 0.4), the correlations for each number of causal variants are negligible (Figure 3.15, 3.16).

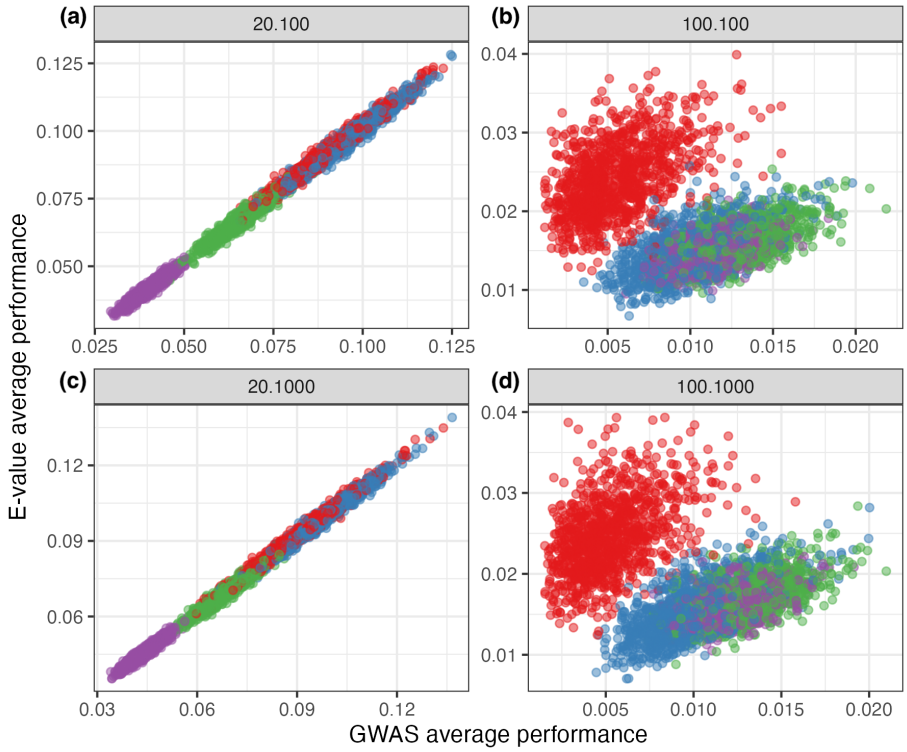


Figure 3.14: Average performance of GWAS against the E-value over sets of the same simulations. Each plot depicts simulations of the specified dimensions (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000), and contains 4000 points. Every point represents an average of method performance across 40 distinct populations (each with differing level of structure) and 5 different phenotypes (heritabilities) per population, for a fixed number of causal variants used to generate phenotype. They are aligned with GWAS performance on the x-axis and E-value performance on the y-axis and coloured according to number of causal variants: red (1), blue (2), green (5) and purple (10).

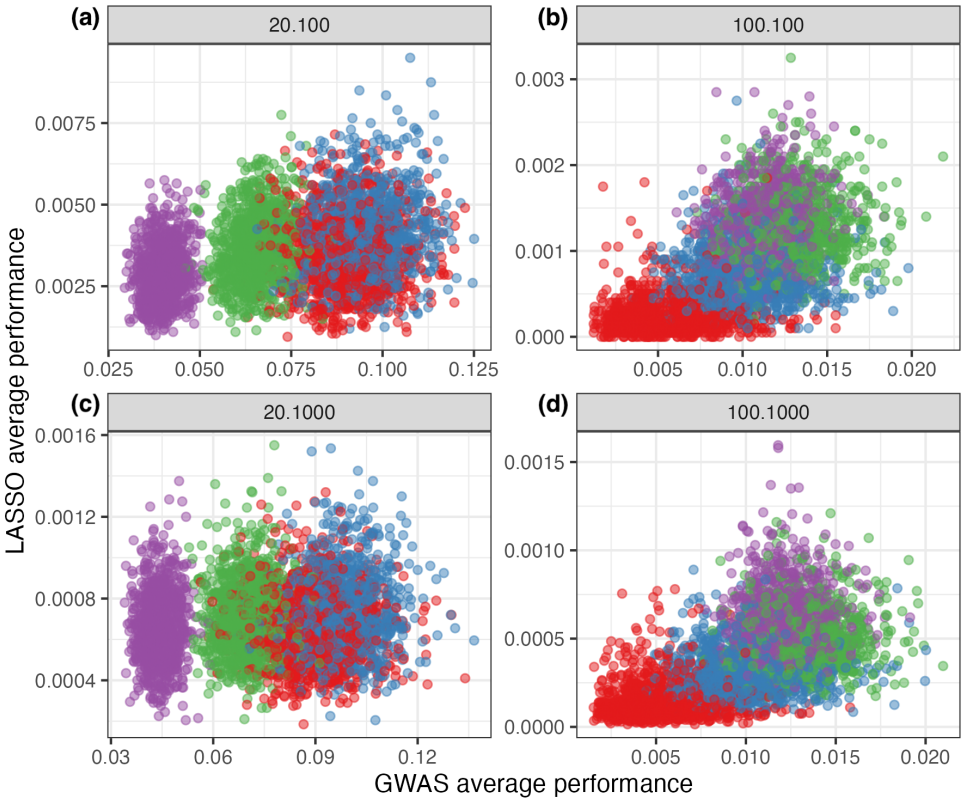


Figure 3.15: Average performance of GWAS against the LASSO over sets of the same simulations. Each plot depicts simulations of the specified dimensions (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000), and contains 4000 points. Every point represents an average of method performance across 40 distinct populations (each with differing level of structure) and 5 different phenotypes (heritabilities) per population, for a fixed number of causal variants used to generate phenotype. They are aligned with GWAS performance on the x-axis and LASSO performance on the y-axis and coloured according to number of causal variants: red (1), blue (2), green (5) and purple (10).

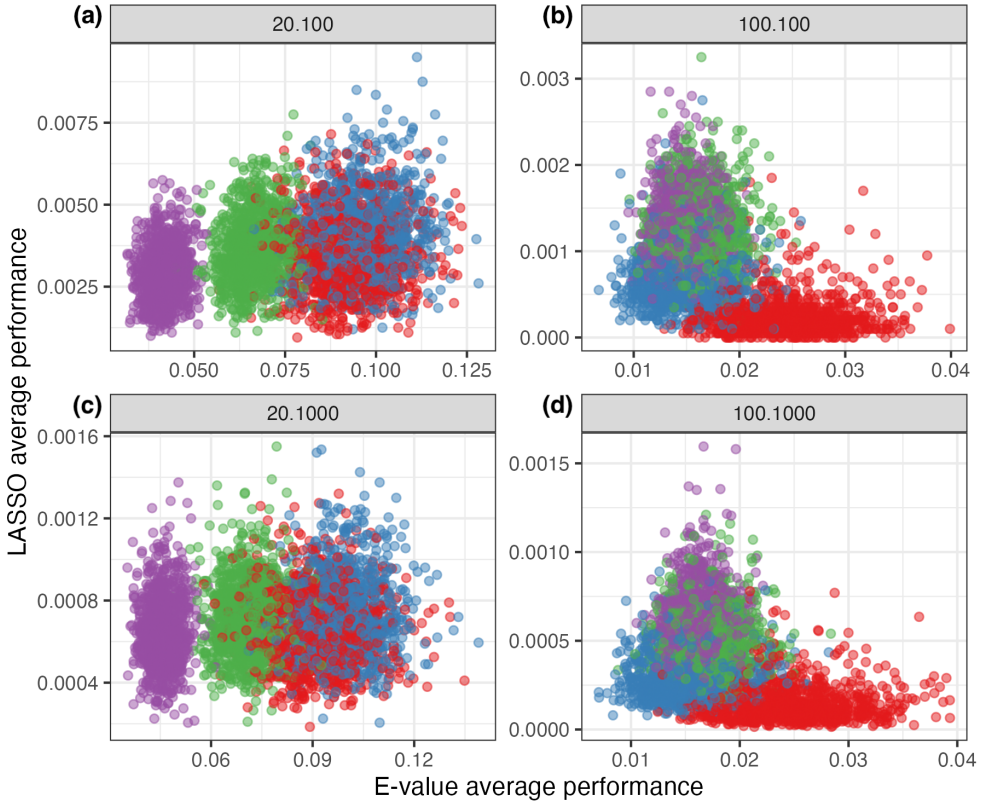


Figure 3.16: Average performance of the E-value against the LASSO over sets of the same simulations. Each plot depicts simulations of the specified dimensions (n,p) = (a) (20,100) (b) (100,100) (c) (20,1000) (d) (100,1000), and contains 4000 points. Every point represents an average of method performance across 40 distinct populations (each with differing level of structure) and 5 different phenotypes (heritabilities) per population, for a fixed number of causal variants used to generate phenotype. They are aligned with E-value performance on the x-axis and LASSO performance on the y-axis and coloured according to number of causal variants: red (1), blue (2), green (5) and purple (10).

Chapter 4

Discussion

Association studies are a critical tool for the exploration and characterisation of the genetic architecture of phenotypic traits. An understanding of the precise relationship between a genotype and the phenotype under scrutiny is not only inherently valuable knowledge but may also be applied to estimate the heritability of a trait (Vinkhuyzen et al. 2013), calculate genetic correlations (Uffleman et al., 2021), make disease risk predictions (Korte & Farlow, 2013), inform drug development programs (Nicholls et al., 2020) and also as control variables in epidemiological studies (Ufflemann et al., 2021). In this thesis we have quantified the performance of three association study methods: GWAS, the LASSO and the E-value. We have seen that overall, LASSO performs extraordinarily well and the novel E-value relatively poorly, using GWAS as a benchmark. We were also able to compare patterns of method performance against fine increments of factors that are known to influence association studies, both individually and in combination, via our simulated dataset. The outcomes of this comparison included insights into the detrimental effect of population structure and consequent interactions between structure and other confounders on association studies. In this chapter we will decipher the biological underpinnings of these results as well as propose some possible improvements or extensions to the study design that could further our interrogation of genotype-phenotype relationships.

4.1 The simulation framework

A key component of this thesis was the development of the framework for simulating genotype and phenotype data. This framework was designed to iterate through an appropriate range of parameters and take their combinations to produce a broad scope of testing scenarios. We will scrutinise some of these design choices here.

4.1.1 Breadth and depth

The populations we simulated in this study were of very small ($n = 20$) and small ($n = 20$) number of individuals, with the number of variants p ten or one hundred times larger than n. The two variables we selected from these categories have provided a good indication of method behaviour under each circumstance (see Section 4.2.1). In contrast, we input reasonably fine increments of our other experimental factors. However, it is well established that transitioning from small n to large n will improve the power of statistical studies (Casella & Berger 2002) including association studies (de Koning & Haley 2005; Jansen et al. 2019; Korte & Farlow 2013; Lee et al. 2018; Mackay, Stone & Ayroles 2009; Uffelmann et al. 2021), so this was a dimension we consciously omitted from this study. Similarly, the impact of p being significantly larger than n is already understood (Casella & Berger 2002; Mackay, Stone & Ayroles 2009; Uffelmann et al. 2021).

There is also a computational tradeoff to consider in deciding both the upper bounds of n and p and the precision with which their impact is studied. Achieving accurate performance averages (with many repetitions) at larger (n, p) while not exceeding our compute resources (Table 3.1) would require restricting our study of population structure or heritability. Sacrificing some granularity in n and p permitted us to be more incremental in assessing these less well studied factors.

4.1.2 Linkage disequilibrium and haplotypes

In most association studies, the goal is localisation. Under the premise that nearby SNPs are in strong LD, the idea is less to find a causal SNP than to identify a region of interest within which a causal SNP resides (Korte & Farlow 2013; Russ et al. 2022; Mackay, Stone & Ayroles 2009). This brings us to the notion of haplotypes: sets of genetic determinants located on a single chromosome that tend to be inherited together. Regions of the genome which show little evidence for recombination and contain only small numbers of distinct haplotypes are defined as haplotype blocks (Altshuler, Donnelly & The International HapMap Consortium 2005). In human GWAS, for example, rather than using every SNP, haplotype blocks are represented by a "tagging SNP" (Altshuler, Donnelly & The International HapMap Consortium 2005; Hästbacka et al. 1992; Keres et al. 1989). If association is detected between the tagging SNP and a trait, then the block may be interrogated in a focused, follow-up study (Uffelmann et al. 2021).

In our simulations, we avoid local LD due to haplotype structure, as if we only had tagging SNPs. While we consider LD an important feature to vary (as it is SNP collinearity), we wanted to avoid the conflation between localisation and identification, and we wanted to exclude variability in haplotype block length. Thus, we used population substructure as a dial to vary global LD (as described in Section 2.5) while limiting local LD to what arises stochastically.

4.1.3 Ploidy and genetic architecture

After careful consideration, we decided to base all of our simulations on a haploid population. In doing so, we consciously sacrificed some realistic complexity for enhanced transparency and interpretability. Because we model LD without the complications of recombination and haplotype blocks, the impact of ploidy manifests in the complexity of mapping alleles to genotypes. In our haploid implementation, we consider diallelic variants represented by the entries 0 or 1 in the matrix X. There are two alleles and two genotypes, and the numerical encoding is straightforward:

$$Y = \sum_{i=1}^{m} \beta_i X_i + \epsilon \tag{4.1}$$

Here m is the number of causal variants, and $\epsilon \sim N(0,1)$. The analogous diploid encoding, by contrast, begins with two alleles (say A and a) and three genotypes (AA, Aa and aa). The effect of encoding is to specify the mapping from genotype to phenotype. For example, to encode AA=Aa=aA=1 and aa=0 means that model is dominant/recessive because the heterozygote shares a phenotype with one of the homozygote classes. Alternatively, an additive encoding of AA=2, Aa=1, aa=0 indicates that the (mean) phenotype changes linearly with the dosage of an allele (where change means to increase or decrease depending on which allele is chosen as the reference 0).

There is little to be gained from including this added complexity. For our purposes, a dominant/recessive model behaves like a haploid model with modified allele frequencies. That is to say, the phenotype is dichotomised into two genotypic classes: the combined AA/Aa class and the aa class. Moreover, the complexity of an additive model is mimicked by a more complex haploid genetic architecture. In other words, if we have two causal variants, each

encoded as 0/1, then they form the possible "haplotypes" 00/01/10/11. If the effects of both variants are equal and independent, as considered in our simulations, then this is equivalent to an additive diploid model:

$$Y = (\beta_1 X_1 + \beta_2 X_2) + \epsilon \tag{4.2}$$

We did not consider epistasis, in which the genetic component of the trait is determined by interactions between causal variants (e.g. adding a term such as $\beta X_1 X_2$), but the framework naturally extends to encompass this phenomenon (see Section 4.5).

4.1.4 The stochastic component of phenotype

Heritable phenotypes are determined by a combination of the genotype and environment of individuals (Falconer & Mackay 1996; Lynch & Walsh 1998). Above, we discuss mechanisms of simulating the genetic component of phenotype, while the stochastic component is contributed by the standard normal. In our experiments, each individual's phenotype is generated by a weighted sum of the entries (0/1) of their causal variants, plus a noise term sampled from $N(0, \sigma^2)$. We assume both that σ is the same across all subpopulations, and further that $\sigma = 1$, noting that these assumptions affect our ability to resolve the genetic architecture. However, our study design mitigates the impact of the latter assumption, in that our effect sizes scale with σ (see Section 2.6.1). The possibility of having different variance for each subpopulation will be discussed in Section 4.5.

4.2 Quantifying the performance of association studies

For each unique combination of dimension (n, p), structure d, heritability h^2 and number of causal variants, we simulated 1000 distinct (X, Y) pairs which our three methods were applied to. We quantified the performance of these methods using the scaled empirical false positive rate: the number of variants ranked above the highest-ranked causal variant in an ordering by significance (p-value, E-value or model coefficient), divided by p. Our goal was to understand both the individual impact of the factors on method performance, as well as how they act in concert to obscure or bolster genetic signal. We will explore the results presented in Chapter 3, providing a biological interpretation and noting any limitations of the study design.

4.2.1 How does dimensionality affect performance?

The number of genetic variants, p, does not affect our measure of performance. Across all input parameters and cross-sections of those parameters, the difference between $p = 100$ and $p = 1000$ was negligible (Figure 3.5, 3.6). Hence, as the *scaled* empirical false positive rate is equal at $p = 100$ and $p = 1000$, the number of false positives at $p = 1000$ is ten times the number at $p = 100$. This aligns with our expectations and established theory; any non-causal variant is governed by the null distribution, so for example in GWAS, the non-causal p-values are Uniform(0,1) distributed (Casella & Berger 2002). There is then a certain probability that any non-causal variant will score above a fixed threshold, such as the Bonferroni threshold (as in Strope et al. 2015) or simply the p-value of the highest ranked causal variant (as in this thesis). It stands to reason that the more non-casual variants there are, the higher the expected number that will pass above the threshold. This "multiple testing problem" is prolific within association studies (Dickson et al. 2010; Korte & Farlow 2013; Nicholls et al. 2020; Strope et al. 2015), as they conduct vast numbers of simultaneous tests for very large p. Correction measures such as stringent p-value thresholds can combat the increasing false positive rate (Nicholls et al. 2020; Strope et al. 2015; Uffleman et al. 2021), but not without decreasing the power to detect true positives with small effect sizes (Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Watanabe et al. 2019). These are important considerations in any real-world association study, as from a practical perspective the cost of interrogating reported variants in a follow-up study is linear in p.

This highlights a potential flaw in our performance metric, in that it does not differentiate between causal variants that rank first (empirical false positive rate of 0) with a p-value of 10^{-3} and variants that rank first with a lower p-value. Indeed, the expected rank of a causal variant with p-value=10^{-3} is 1. Although the metric is appropriate for assessing our small-scale simulations, it lacks the ability to determine whether a variant would meet a particular significance threshold. Adapting the performance metric to accomodate this need would make the experimental framework more suitable for large-scale simulations or application to real datasets.

Unlike p, the number of individuals n in the population has a pronounced impact on performance. The ability of GWAS and the E-value to identify casual variants decreases significantly for simulations where $n = 20$ compared to $n = 100$ (Figure 3.5), and even the LASSO admits a slight worsening of performance in the former case which is visible at higher resolutions (Figure 3.5). Given that a key recommendation to improve the accuracy of

association studies is to increase the sample size (de Koning & Haley 2005; Jansen et al. 2019; Korte & Farlow 2013; Lee et al. 2018; Mackay, Stone & Ayroles 2009; Uffelmann et al. 2021), this is hardly surprising.

But the detrimental effect of small sample size on performance exceeds that of a typical statistical measure. The very small $n = 20$ case we studied came with additional complications, including the unexpected interaction between MAF and structure (see Section 3.3.5). We will illustrate the combinatoric rationale for this added difficulty by example. Suppose we are generating a population, as in the Methods, with $n = 20$ individuals and $p = 1000$ variants, sampling alleles from the Ewens distribution. The expected number of singletons (variants with just one 1, i.e. MAF = 0.05) is

$$\frac{1}{1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{10}} \times p \approx 0.341 \times 1000 = 341 \tag{4.3}$$

The number of unique possible singletons is 20, so on average we expect $341 \div 20 \approx 17$ identical copies of each singleton in our simulated population. It is impossible for any statistical method to distinguish between these identical singletons, known as genetically indistinguishable SNPs (Lawrence et al. 2005; Skelly, Magwene & Stone, 2016). With only one causal variant, the probability of selecting one of these singletons out of all possible variants is $34.1\,\%$, and this will increase for multiple causal variants. This has a significant impact on our ability to correctly identify causal variants, beyond the usual consequences of small sample size.

4.2.2 How does genetic architecture affect performance?

We interrogated the influence of genetic architecture on performance by varying the strength and division of genetic signal within an additive, haploid population. Initially, we tested the ability of association study methods to identify just one causal variant endowed with all of the signal. We then progressed to varying the amount of signal that was divided up between an increasing portion of the total available genetic variants. The latter circumstance is more indicative of a realistic trait architecture study, which increasingly report large numbers of causal variants with small effect (Korte & Farlow 2013; Holland et al. 2020;

Nicholls et al. 2020; Vinkhuyzen et al. 2013). The two crucial parameters at play here were the number of causal variants, and the heritability partitioned amongst them.

For a fixed number of causal variants, increasing heritability improves performance (Figure 3.6). This follows logically from the definition of heritability; the proportion of variation in a trait that is explained by genetics (Vinkhuyzen et al. 2013). By adding more heritability to each causal variant, we enlarge their effect sizes (β is proportional to h^2) and hence heighten the power of association studies to detect them (Casella & Berger 2002; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Watanabe et al. 2019).

The story is not so straightforward for changes to the number of causal variants at a fixed heritability. At very low heritabilities, GWAS and the E-value perform very poorly, but their performance improves as the number of causal variants increases (Figure 3.6). This suggests that they both have insufficient power to detect the variants with small effect sizes, but improve because the average success in a random selection increases when there are more causal variants. This is illustrated by the following.

Suppose we have a population with n individuals, and p variants, m of which are causal. However, assume the heritability of the causal variants is so low as to be negligible, so that they behave as non-causal variants in an association study and their p-values or equivalent are sampled from the Uniform(0,1) distribution. Then the probability of attaining a particular rank R is $\frac{1}{p}$.

Let $m = 2$, so that the causal variants are X_1, X_2, and let $Z = \min\{X_1, X_2\}$. We will derive the expected causal rank, $\mathbb{E}Z$. The cumulative distribution function of Z is

$$\begin{aligned} F(z) &= \mathbb{P}(Z \leq z) \\ &= 1 - \mathbb{P}(\min\{X_1, X_2\} > z) \\ &= 1 - \mathbb{P}(X_1 > z)\mathbb{P}(X_2 > z) \\ &= 1 - (1 - z)^2 \end{aligned} \tag{4.4}$$

where the 3rd equality follows from the fact that X_1, X_2 are independent and the 4th because they are sampled from the $U(0,1)$ distribution. The probability density function is

$$\begin{aligned} f(z) &= \frac{d}{dz}\left[1 - (1 - z)^2\right] \\ &= 2(1 - z) \end{aligned} \tag{4.5}$$

Hence, the expected causal rank is

$$\begin{aligned}\mathbb{E}Z &= \int_0^1 2(1-z)z \; dz \\ &= 2\left[\frac{z^2}{2} - \frac{z^3}{3}\right]_0^1 \\ &= 2\left[\frac{1}{2} - \frac{1}{3}\right] \\ &= \frac{1}{3}\end{aligned} \tag{4.6}$$

Extending this to the general case then, we have that if $Z = \min\{X_1, X_2, \ldots, X_m\}$ then the expected causal rank of m variants is:

$$\mathbb{E}Z = \frac{1}{m+1} \tag{4.7}$$

Therefore, as we increase the number of causal variants in a low-heritability population, our average performance will improve.

At high heritabilities, increasing the number of causal variants produces the opposite result (Figure 3.6). Having fewer causal variants is favourable when there is a large amount of genetic signal available, so that individual variants have a stronger effect size and are more visible to an association test (Casella & Berger 2002; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Watanabe et al. 2019). The sole exception occurs when $n = 100$, where the E-value performs worst on one causal variant (Figure 3.6). This is a consequence of the E-value's poor performance on highly structured populations of 100 individuals simulated with one causal variant (Figure 3.7), which we discuss in Section 4.3.

For mid-range heritabilities, GWAS and the E-value perform best when there are 2-5 causal variants (Figure 3.6). This is likely a product of the interplay between the benefits of partitioning heritability less finely and having a higher chance of uncovering the correct variant because there are more of them. The ambiguity here is an artefact of the performance metric, which takes only the first causal variant to be discovered into account when measuring

success. The construction of our simulation process was intended to extricate individual factor effects from the complexity found in real population-trait relationships (Dickson et al. 2010; Falconer & Mackay 1996; Holland et al. 2020; Korte & Farlow 2013; Lynch & Walsh 1998; Mackay, Stone & Ayroles 2013; O'Connor et al. 2019; Russ et al. 2022; Skelly, Magwene & Stone 2016; Uffelmann et al. 2021; Vinkhuyzen et al. 2013; Watanabe 2019; Zeng et al. 2018). That the performance metric should generate such ambiguity here reflects a major flaw in its design and application. Although it captures the notion of reducing time to first success and is effective for most other factor comparisons, it is unfit for the purpose of comparing across different numbers of causal variants.

Finally, for all combinations of n, p, heritability and number of causal variants, the LASSO performs so well it is hard to discern any trends (Figure 3.6). The sole exception is when $n = 20, p = 100$, where despite still far outperforming the other methods, it follows a similar pattern of performance to GWAS (Figure 3.6). We will explore the rationale for the LASSO's high performance in Section 4.4.

Scrutiny of average performance on the same sets of repetitions for differing numbers of causal variants yields interesting results. All three pairwise comparisons between methods portray a phenomenon known as Simpsons paradox, in which a trend between two variables in a population appears, disappears or reverses when the population is split into subgroups (Simpson 1951). The LASSO and GWAS appear positively correlated (Figure 3.15), and the LASSO and E-value negatively so (Figure 3.16), but divided into subgroups by number of causal variants, both these correlations disappear. GWAS and the E-value are positively correlated overall, but at $n = 20$ this is true also for each number of causal variants, which is not the case at $n = 100$ where the correlation reverses (Figure 3.14). An achievement of the study design is that it allows us to observe such phenomena by peeling apart the complex interactions between multiple factors.

4.3 Predicting performance on a structured population

A key aim of this thesis was to study the effect of population structure on association studies, and in particular to disentangle its effect from other confounding factors such as MAF, effect size and genetic architecture. It is well established that population stratification can negatively impact the performance of a GWAS, and as a result, techniques have been developed to combat its effect (Kang et al. 2008; Price et al. 2006; Yu et al. 2006; Zhang et al.

2010; Zhao et al. 2007). What is lacking, however, is an explicit quantification of its action under a broad scope of possible confounding circumstances, which this thesis addresses.

A persistent challenge we faced was managing the interplay between demography, the allele frequency spectrum and population structure. We combatted this by sampling overall allele frequencies from the Ewens distribution (Ewens 1972), then splitting the population into 5 subpopulations and assigning the alleles to each subpopulation according to both its present composition and how many alleles remained to be assigned (as in Figure 2.2, for example). The greater the probability of allocating a minor allele to an individual depended on the subpopulation they were a part of, the higher the overall structure in the population (Figure 2.4). This design allowed us to vary the extent of variant collinearity (as measured by average pairwise r^2) without influencing the overall allele frequencies and therefore statistical power. It would be interesting to properly consider the mathematics of this algorithm as a formal urn model (Johnson, Norman & Kotz 1977); however, that was beyond the scope of this thesis.

4.3.1 How does population structure affect association studies?

Our results indicated that this design was able to inject a reasonable range of structure into the population, successfully quantified by our proposed measure into levels of average LD spanning from 0.05 to 0.15 and higher (Figure 2.4). We were then able to compare performance between levels of structure, noting that on average, increasing structure diminished the ability to identify causal variants of the E-value substantially and GWAS less so (Figure 3.3). The behaviour of GWAS aligned with our expectations and the literature (Cardon & Palmer 2003; Kruglyak 1999; Skelly, Magwene & Stone 2016; Sutter et al. 2004), but E-value did not perform as predicted. As has been a trend throughout this study, it is difficult to comment on the LASSO's performance using averages over just one parameter due to its consistent overall success. LASSO performed very well over structured populations on average (Figure 3.3), but did demonstrate some variation in performance when subdivided by MAF (Figure 3.8), discussed below. Our next step, which yielded some interesting outcomes, was to study the methods under the interaction between population structure and other aspects of the genetic architecture.

In sharp contrast to the overall trend, performance appears to improve with increased structure when we restrict to cases where the MAF of the first causal variant to be identified is low (Figure 3.8). A notable exception to this is the E-value when $p = 100$, which

demonstrates poor performance for high structure on singletons (MAF = 0.01) but relatively improved performance on doubletons (MAF = 0.02) (Figure 3.9). Note that the “first causal variant to be identified” is the variant whose rank is higher than any other causal variants, and is therefore used to define performance in terms of the scaled empirical false positive rate. MAF is not an input parameter, but rather is an output, as it is measured after the methods have ranked the variants. This is important to remember when it comes to interpreting the results of our manipulation of structure faceted by MAF.

4.3.2 Examining the interplay of structure, minor allele frequency and effect size

In a highly structured population, we are more likely to observe “clumping" of 0s and 1s into the same subpopulations, as seen in Figure 2.2c. At higher allele frequencies, the result is a variant whose matrix column consists of blocks which make it virtually indistinguishable from other common variants. As average LD increases, so too does the number of genetically indistinguishable variants (Skelly, Magwene & Stone 2016). To illustrate this, consider a completely unstructured population of $n = 100$ individuals, where there are 100 unique singleton patterns, but $\binom{100}{2} = 4950$ unique doubletons, with the number of unique variant patterns increasing correspondingly to MAF. Instead, in the limiting case of highly structured populations, all minor alleles reside within the same population, so for example if MAF = 0.2 and $n = 100$ with 5 subpopulations, then there will only be 5 unique variant patterns. Increasing structure therefore can drastically increase the number of perfectly or almost perfectly correlated variants. Such variants will be ranked into a group by association studies, so that other, rarer, variants will either be ranked entirely above or below such a group. If ranked above, average performance at small MAF will increase and if below, it will not contribute to average performance at small MAF. A common causal variant could fall anywhere in the hierarchy of the group if they are all truly indistinguishable, increasing the likelihood of false positives (and thus decreasing the performance score). This is detrimental both to average performance and to performance specifically at common variants. In contrast, population structure “ignores" singletons.

This phenomenon is reflected in our study of the distribution of MAF in the highest-ranked causal variants. As we deliberately drew our overall allele frequencies from the same distribution each time, any deviation from this distribution in the causal variants was an artefact of the methods. We observed that as structure increased, the distribution became

biased towards rare variants, particularly towards singletons (Figure 3.10, 3.11), except for the E-value which preferred doubletons to singletons in simulations with 100 individuals (Figure 3.11). When further cross-sectioning on the number of causal variants and heritability, we saw these behaviours exacerbated for increases to the former and no change in the latter case (Figure 3.12, 3.13), supporting the above explanations.

Because we standardised the effect size, which affects visibility under an association study (Casella & Berger 2002; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Watanabe et al. 2019), for each causal variant using a fixed heritability and its own MAF, we can remove effect size as a possible explanatory variable here. We can explain the pattern of performance at different MAF exhibited by GWAS and the LASSO by this. However, this is not the case in real association studies, as is discussed below.

Of the three methods, LASSO is most adept at dealing with these highly structured conditions, and E-value the least. The behaviour of the E-value when $n = 100$ juxtaposes both the overall trend and that of GWAS and the LASSO under more specific circumstances (Figure 3.9). It also performs far worse on average than even GWAS at high structures. The E-value ranks variants using their expected empirical conditional p-values: the average proportion of spurious predictors with a p-value smaller than that of the variant under study. By design, it adjusts for LD by shrinking the E-values of causal variants which are correlated with other, non-causal variants, smaller than their p-values. Thus, a possible explanation for its poor performance on structured populations is that it has overcompensated for the high levels of LD and excessively penalised all causal variants.

We observe that under a high level of structure and on a small population, it reports a higher amount of doubletons than singletons as the first-ranked causal variant (Figure 3.11, 3.13), despite there being half as many doubletons (not necessarily unique) in the population as a whole under the Ewens distribution. We previously noted that rare variants are favoured at high structures as they manifest as more genetically distinguishable patterns. This holds true, except for singletons, which the E-value treats distinctly differently. As singletons by nature either have correlation r^2 with other variants close to 0 or 1, the E-value penalty to their p-value is appreciably larger (see Section 2.7).

These observations are only possible because our study design allowed us to separate and compare the contributions of MAF, structure, heritability and the number of causal variants independently. It is essential to note that in real association studies, the same considerations do not apply. As we saw in Section 2.4, effect size, which is a main

determinant of statistical detection power, is dependent on MAF. For a constant effect size, it is far easier to detect common variants than rare ones (Casella & Berger 2002; Dickson et al. 2010; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Watanabe et al. 2019). What we have done essentially amounts to evening the playing field for variants of any allele frequency. There were two key elements of the simulation framework that permitted this: drawing the overall allele frequency from the Ewens distribution, and calibrating effect sizes by fixed heritability and MAF. As a result, we are able to predict that GWAS and the E-value will perform worse on structured populations, and the LASSO will be relatively unaffected, although this will depend on the MAF of causal variants.

4.4 Comparing method performance

Across all combinations of input parameters, LASSO dramatically surpasses GWAS and the E-value. This is observed both on average (Figure 3.5) and when restricted to a particular genetic architecture (Figure 3.6). The only circumstances in which the average performance of GWAS is comparable to the LASSO is when $n = 100$ and there are only 1-2 causal variants with large heritability ($h^2 \geq 0.3$) and hence large effect size (Figure 3.6). To justify this success, we must consider the design and motivation behind the LASSO.

As we explained in Section 2.3.7, the LASSO mitigates the effect of correlation between predictor variables by effectively choosing a representative from the correlated group and shrinking all other coefficients to 0 (Nouira & Azencott 2021). In a highly structured population, this manifests as correctly identifying the casual variant, and conferring no importance onto any other variants that are in LD with it. While GWAS and the E-value suffer progressively as structure increases, the LASSO is adept at resolving genetic architecture, except in conditions where the variants are genetically indistinguishable. Previous studies have also found success applying the LASSO to structured populations due to its ability to deal with correlation structure between variants (Hebiri & Lederer 2012; Nouira & Azencott, 2021). Thus, on average across all simulated populations (which are increasingly structured), it outperforms the other methods.

Under circumstances where the effect size is small or partitioned amongst multiple causal variants, many of these will fall below the threshold for identification by GWAS, which tests each variant individually against phenotype (Casella & Berger 2002; Korte & Farlow 2013; Mackay, Stone & Ayroles 2009; Watanabe et al. 2019). In contrast, the LASSO

optimises all coefficients in conjunction with one another by its nature as a variable selection approach (Tibshirani 1996). Variants with less genetic signal are therefore more readily identified by LASSO than GWAS. We have already discussed the flaws in the design of the E-value and its tendency to over-penalise causal variants. Of the three methods, LASSO demonstrates preeminent ability to adapt to any simple or complex genetic architecture. As such, we recommend it as a method for future trait architecture studies.

4.5 Future directions

A pivotal aspect of this thesis was the development of the novel framework for simulating genotype-phenotype data. This also reflects a considerable opportunity for future work in the extension of this framework to more complex architectures. All that is required is the adjustment of the mapping used to generate phenotype from genotype.

In this study, we solely addressed quantitative traits, but many association studies focus on binary traits, such as the presence/absence of a genetic disease (Dickson et al. 2010; Nicholls et al. 2020; Russ et al. 2022; Uffelmann et al. 2021). By fixing a threshold for Y values on either side of which we assign entries a binary 0/1, we could simply and efficiently alter the experimental procedure to focus on binary traits; alternatively, we could follow the logic of generalised linear models and generate Bernoulli 0/1 phenotypes on the basis of probability differences linked to allelic state via the logit function. However, some additional consideration would be needed to ensure the case/control frequency does not match allele frequency (Uffelmann et al. 2021).

As we discussed in Section 4.1.3, another avenue for extension is incorporating ploidy. Additive and dominant structures are already possible through the current framework with an adjusted MAF distribution. To consider epistasis, or the non-linear interaction between loci (Russ et al. 2022), we could use an analogous mathematical expression:

$$Y_k = \sum_{i,j} \beta_{ij} X_{ki} X_{kj} + \epsilon \tag{4.8}$$

Here Y_k refers to the phenotype entry for a particular individual in the population, and β_{ij} is the unique coefficient for effect size of the combined interaction between causal variants X_i and X_j. Further thought would be required to derive a formula for β_{ij} in a comparable

manner to the additive case. As the entries in the variants' column vectors are either 0 or 1, both minor alleles must be present in the individual for a nonzero interaction effect on phenotype, thus encoding epistasis.

At present, there is no heterogeneity in our mapping between genotype and phenotype. We could adjust this by setting a different variance σ^2 for the normal distribution used to assign a phenotype to individuals in each subpopulation. One issue with this is that it would be difficult to maintain a constant heritability whilst doing so, which is an important part of our experimental framework (as discussed in Section 4.3).

Finally, consider the tuning parameter used in the LASSO. Although we utilised the recommended λ that is within one standard error of the minimum (Friedman, Hastie & Tibshirani 2010), it is possible that even greater prediction accuracy might be attained by experimenting with this parameter (Hebiri & Lederer 2012). This could be considered as an additional input factor within our framework for testing LASSO specifically.

4.6 Conclusion

In this thesis, we have presented a novel framework for the simulation of genotype data and the corresponding generation of phenotype data according to a specified genetic architecture. By manipulating those specifications we were able to interrogate and subsequently report on the action of association study methods on a diverse range of possible population and trait types. We quantified the performance of GWAS, the LASSO and the E-value across primarily small populations and numbers of genetic variants, whilst varying input parameters such as the number of causal variants, their heritability and the correlation structure in the population. To address the latter, we proposed a new quantification of population structure in terms of average LD, which was fit for purpose.

The framework design was a central achievement of this thesis in that it permitted the disentanglement of contributions from different factors affecting method performance. Our examination of these contributions prompted insight into the behaviour of methods on structured and idealised populations with simple and complex genetic architecture. Overall, despite the ubiquity of GWAS, LASSO dominated its performance across all simulated scenarios, easily resolving causal variants in conditions where GWAS tends to fail. Contrary to our expectations, the E-value does not noticeably outperform GWAS under any circumstances. In future, we hope that the framework we have developed may augment the

resolution of genotype-phenotype relationships, empowering future association studies and providing a recommendation for method choice.

References

Altshuler, D., P. Donnelly, & The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437: 1299–1320.

Anton, H. & C. Rorres. 2013. Singular Value Decomposition. In: Anton, H. & C. Rorres (11th ed). Elementary linear algebra: applications version. Wiley, pp 514-520.

Bodmer, W. & C. Bonilla. 2008. Common and rare variants in multifactorial susceptibility common diseases. *Nature Genetics*. 40(6): 695–701.

Boyle, E. A., Y. I. Li & J. K. Pritchard. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 169: 1177–1186.

Capelli C, Redhead N, Romano V, Cali F, Lefranc G. Population structure in the Mediterranean basin: A Y chromosome perspective. *Annals of Human Genetics*. 2006;70:207–225.

Cardon, L. & L. Palmer. 2003. Population stratification and spurious allelic association. *Lancet* 361(9357): 598-604.

Casella, G. & R.L. Berger. 2002. Hypothesis testing. In: Casella G. & R.L Berger (2nd ed). Statistical Inference. Duxbury Press, pp. 373-413.

Che, K., X. Chen, M. Guo, C. Wang, & X. Liu. 2020. Genetic Variants Detection Based on Weighted Sparse Group Lasso. *Frontiers in Genetics* 11.

Chowdhury, M.Z.I. & T.C. Tanvir. 2020. Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*. 8(1): e000262.

de Koning, D. J. & C. S. Haley. 2005. Genetical genomics in humans and model organisms *Trends in Genetics*. 21: 377–381.

Dickson, S., K. Wang, I. Krantz, H. Hakonarson, & D. Goldstein. 2010. Rare variants creat synthetic genome-wide associations. *PLoS Biology* 8(1): 1-12.

Ewens, W. 1972. The sampling theory of selectively neutral alleles. *Theoretical Populatio Biology* 3: 87-112.

Ewens, W.J. 2004. Many Loci. In: Ewens, W.J. Mathematical population genetics: I. Theoretical Introduction. pp 241-274.

Falconer, D. S. & T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. Addison Wesley Longman, Harlow.

Friedman, J., T. Hastie & R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 33: 1–22.

Haddad, S., C. Chen, S. Santangelo, & J. Seddon. 2006. The genetics of age-related macula degeneration: a review of progress to date. *Survey of Ophthalmology*. 51: 316–63.

Hästbacka, J., de la Chapelle, A., Kaitila, I., et al. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics* 2: 204–211.

He, Q. & D. Lin. 2011. A variable selection method for genome-wide association studies. *Bioinformatics* 27(1): 1-8.

Hebiri, M. & J.C. Lederer. 2012. How Correlations Influence Lasso Prediction. *arXiv*. [Online] Available: https://arxiv.org/pdf/1204.1605.pdf.

Holland, D. et al. 2020. Beyond SNP heritability: polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLOS Genetics*. 16, e1008612.

Ioannidis, J.P., E.E. Ntzani, T.A. Trikalinos & D.G. Contopoulos-Ioannidis. 2001. Replication validity of genetic association studies. *Nature Genetics* 29: 306–09.

Jansen, P. R. et al. 2019. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature Genetics*. 51: 394–403.

Johnson, N.L. & S. Kotz. 1977. Urn Models and Their Application: An Approach to Modern Discrete Probability Theory. Wiley.

Jostins, L., S. Ripke, R. Weersma, R. Duerr, D. McGovern, et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491: 119–24.

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly & E. Eskin. 2008. Efficient control of population structure in model organism association mapping. *Genetics*. 178(3): 1709–1723.

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* 245(4922): 1073-1080.

Korte, A. & A. Farlow. 2013. The advantages and limitations of trait analysis with GWAS: review. *Plant methods* 9(29): 1-9.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*. 22: 139–144.

Lahoucine, S., A. Molina, A. Cánovas, & J. Casellas. 2019. Screening for epistatic selection signatures: A simulation study. *Scientific Reports* 9(1026): 1-5.

Lawrence, R., D.M. Evans, A.P. Morris, X. Ke, S. Hunt et al. 2005. Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. *Genome Research* 15: 1503–1510.

Lee, J. J. et al. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*. 50: 1112–1121.

Liu, X., Y. I. Li & J. K. Pritchard. 2019. Trans effects on gene expression can drive omnigenic inheritance. *Cell*. 177: 1022–1034.e6.

Lovell, A., C. Moreau, V. Yotova, F. Xiao & S. Bourgeois. 2005. Ethiopia: Between Sub-Saharan Africa and western Eurasia. *Annals of Human Genetics*. 69: 275–287.

Lynch, M. & B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, Massachusetts.

Mackay, T., E. Stone, & J. Ayroles. 2009. The genetics of quantitative traits: challenges and prospects. *Nature reviews* 10: 565-577.

Manolio, T.A., F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, et al. 2009. Finding the missing heritability of complex diseases. *Nature*. 461(7265): 747–753.

Menozzi, P., A. Piazza & L. Cavalli-Sforza. 1978. Synthetic Maps of Human Gene Frequencies in Europeans. *Science* 201, No. 4358: 786-792.

Nicholls, H., C. John, D. Watson, P. Munroe, M. Barnes, & C. Cabrera. 2020. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Frontiers in genetics* 11(350): 1-15.

Nouira, A. & C.A. Azencott. 2022. Multitask group Lasso for Genome Wide association Studies in diverse populations. *Pacific Symposium on Biocomputing*. 27: 163-174.

O'Connor, L. J. et al. 2019. Extreme polygenicity of complex traits is explained by negativ selection. *American Journal of Human Genetics*. 105: 456–476.

Patterson, N., A.L. Price & D. Reich. 2006. Population structure and eigenanalysis. *PLoS Genetics* 2(12): e190.

Platt, A., B.J. Vilhjalmsson & M. Nordborg. 2010. Conditions under which genome-wide association studies will be positively misleading. *Genetics*. 186(3): 1045–1052.

Price, A., N. Patterson, R. Plenge, M. Weinblatt, N. Shadick, & D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8): 904-909.

Puthiyedth, N., N. Zhang, Z. Wang & Y. Yan. 2021. Performance Comparison of LASSO Variants with Genome-Wide Association Studies (GWAS). In: 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Houston, TX, USA, pp. 1682-1684.

Ratner, B. 2010. Variable selection methods in regression: Ignorable problem, outing notab solution. *Journal of Targeting, Measurement and Analysis for Marketing*. 18: 65–75.

Russ D., J. Williams, V. Cardoso, L. Bravo-Merodio, S. Pendleton, F. Aziz, et al. 2022. Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models. *PLoS ONE* 17(2).

Sánchez-Maroño, N., A. Alonso-Betanzos & M. Tombilla-Sanromán. 2007. Filter Methods for Feature Selection – A Comparative Study. In: H. Yin, P. Tino, E. Corchado, W. Byrn & X. Yao (eds). Intelligent Data Engineering and Automated Learning - IDEAL 2007. *Lecture Notes in Computer Science* 4881. Springer, Berlin, Heidelberg.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 13: 238–241.

Skelly, D., P. Magwene, & E. Stone. 2016. Sporadic, global linkage disequilibrium betwee unlinked segregating sites. *Genetics* 202: 427-437.

Slatkin, M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*. 9: 477–485.

Srivastava, S. & L. Chen. 2010. Comparison between the SSVS and the LASSO for Genome Wide Association Studies. *Communication in Information and Systems*. 10(1): 39-52.

Stoneking, M., J.J. Fontius, S.L. Clifford, H. Soodyall & S.S. Arcot. 1997. Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Research*. 7: 1061–1071.

Strope, P., D. Skelly, S. Kozmin, G. Mahadevan, E. Stone, P. Magwene, et al. 2015. The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic an genotypic variation and emergence as an opportunistic pathogen. *Genome research* 25: 762-774.

Sutter, N. B., M. A. Eberle, H. G. Parker, B. J. Pullar, E. F. Kirkness et al. 2004. Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Research*. 14: 2388–2396.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Roya Statistical Society. Series B (Methodological)* 58(1): 267–88.

Uffelmann, E. & D. Posthuma. 2021. Emerging Methods and Resources for Biological Interrogation of Neuropsychiatric Polygenic Signal. *Biological Psychiatry*. 89(1): 41-53

Uffelmann, E., Q. Huang, N. Munung, J. de Vries, Y. Okada, A. Martin, et al. 2021. Genome- wide association studies. *Nature Reviews* 1(59): 1-21.

Vinkhuyzen, A., N. Wray, J. Yang, M. Goddard, & P. Visscher. 2013. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annual Review of Genetics* 47: 75-95.

Visscher, P.M., et al. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *Public Library of Science Genetics*. 2: e41.

Waldmann, P., G. Mészáros, B. Gredler, C. Fuerst & J. Sölkner. 2013. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*. 4.

Watanabe, K. et al. 2019. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*. 51: 1339–1348.

Watterson, G. 1974. The Sampling Theory of Selectively Neutral Alleles. *Advances in Applied Probability* 6(3): 463–88.

Weisberg, S. 1985. Applied Linear Regression (2nd ed). Wiley.

Xu Y., A. Narayan, H. Tran, C.G. Webster 2021. Analysis of the ratio of ℓ1 and ℓ2 norms in compressed sensing. *Applied and Computational Harmonic Analysis*. 55:486-511.

Yang, J., B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*. 42: 565–69.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*. 38(2): 203–208.

Zeng, J. et al. 2018. Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*. 50: 746–753.

Zhang, Z., E. Ersoz, C.Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, J.M. Ordovas, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*. 42(4): 355–360.

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, et al. 2007. An Arabidopsis example of association mapping in structured samples. *Public Library of Science Genetics*. 3(1): e4.

Appendices

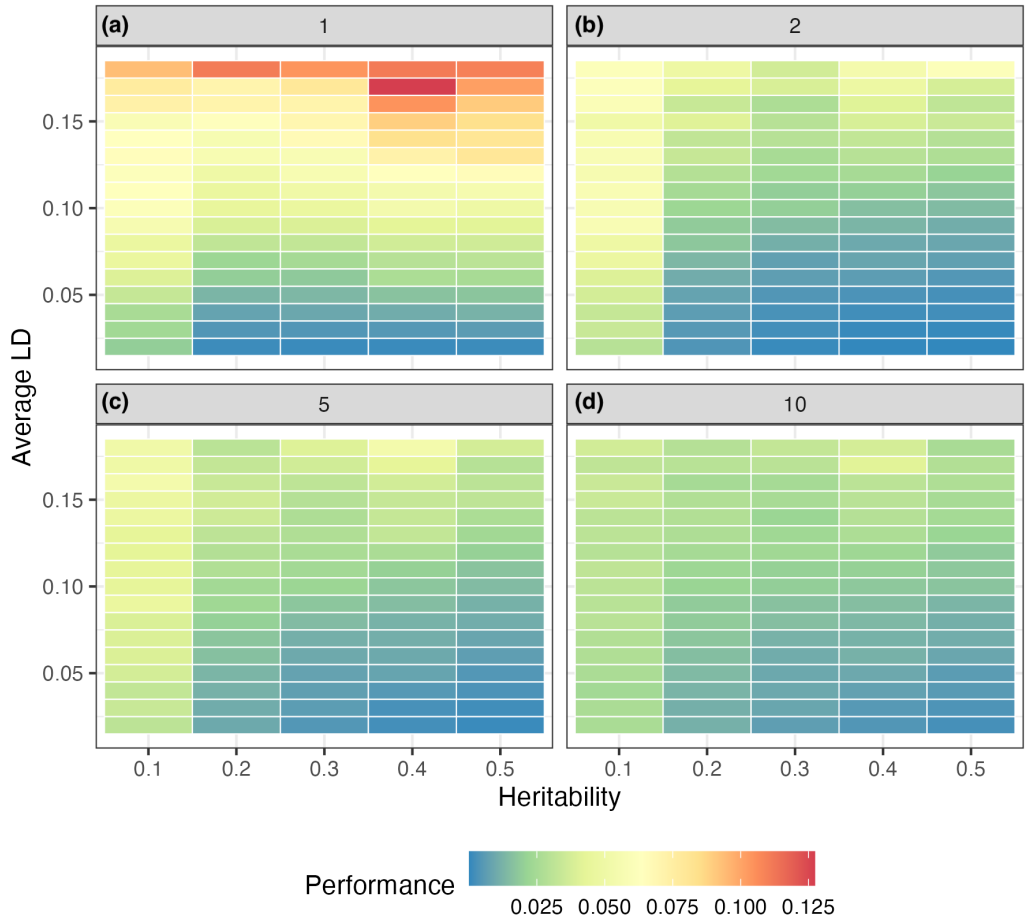


Figure A1: The E-value performs poorly on highly structured populations when there is only one causal variant, particularly for high heritabilities. Heat map of average performance across populations with the specified heritability, average LD, and number of causal variants (a) 1, (b) 2, (c) 5 or (d) 10. The result is shown for simulations with $n = 100$ individuals and $p = 100$ variants. We fixed a cutoff of average LD ≤ 0.15 so that all averages were taken over at least 100 simulations.

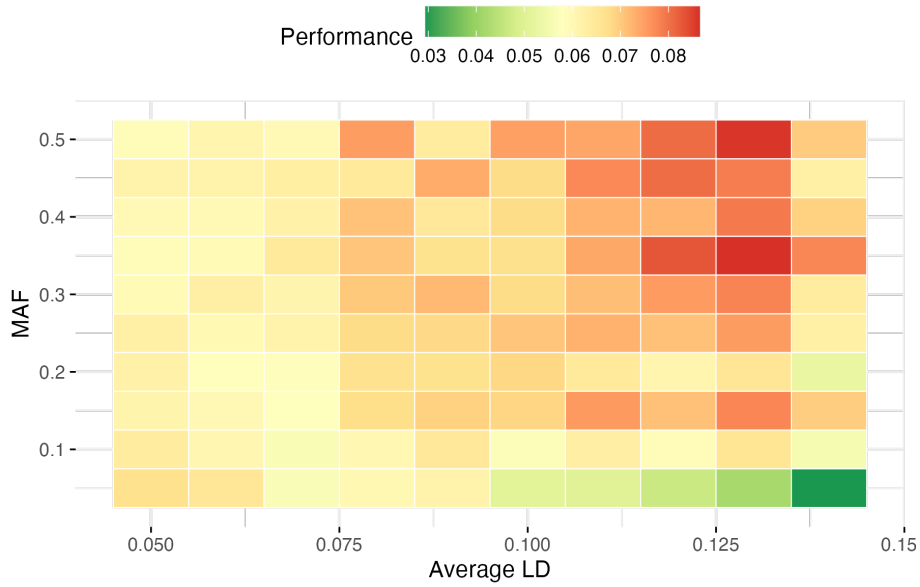


Figure A2: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for GWAS on a population of 20 individuals and 100 causal variants.

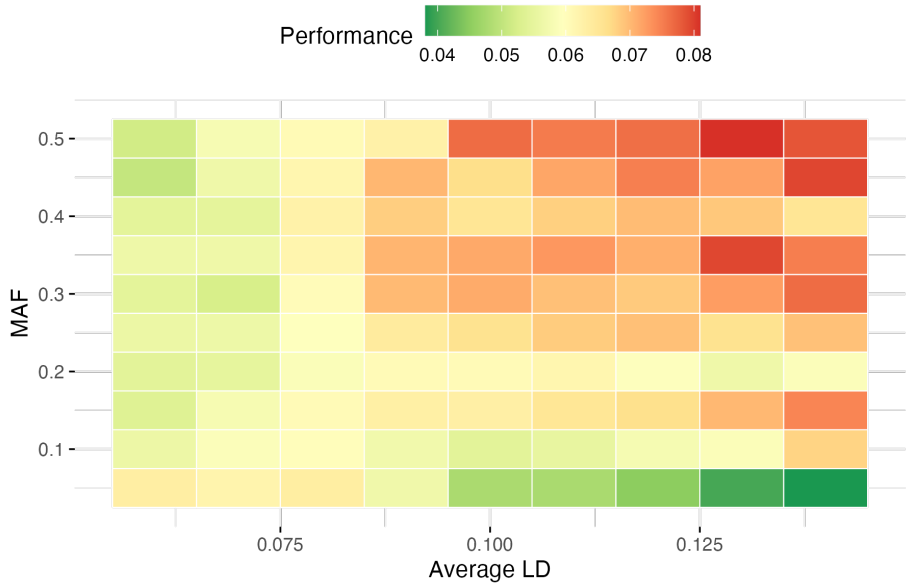


Figure A3: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for GWAS on a population of 20 individuals and 1000 causal variants.

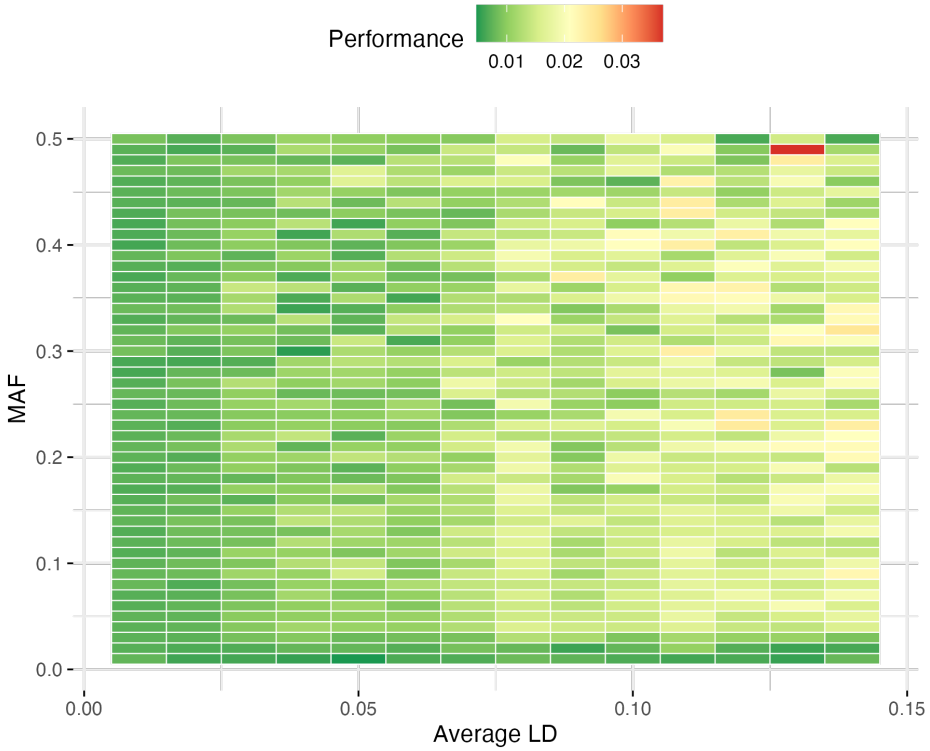


Figure A4: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for GWAS on a population of 100 individuals and 1000 causal variants.

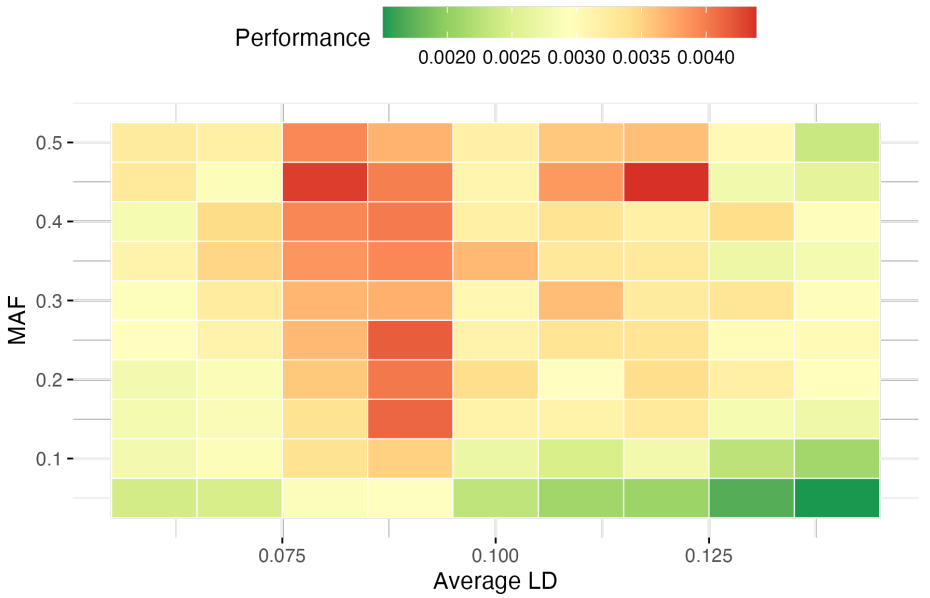


Figure A5: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for LASSO on a population of 20 individuals and 100 causal variants.

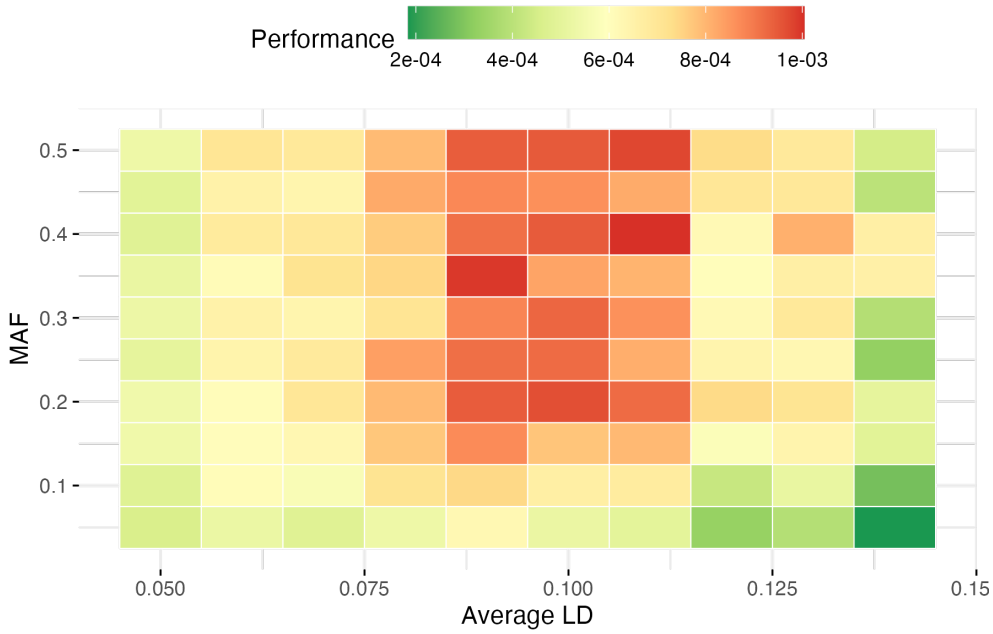


Figure A6: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for LASSO on a population of 20 individuals and 1000 causal variants.

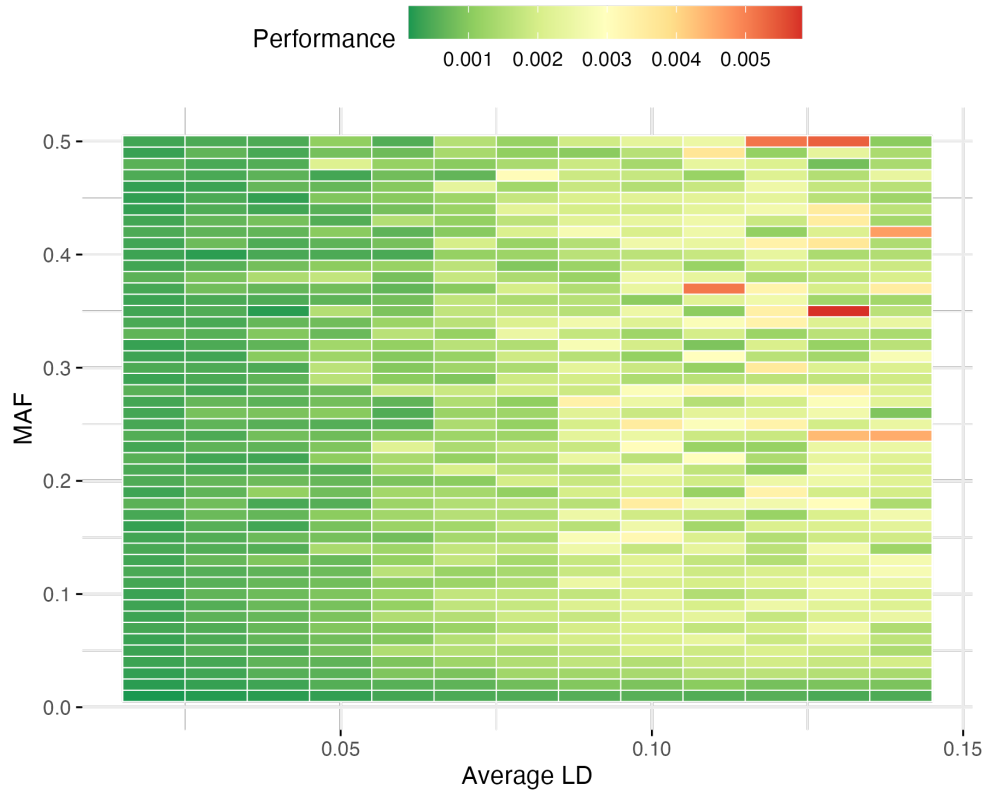


Figure A7: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for LASSO on a population of 100 individuals and 100 causal variants.

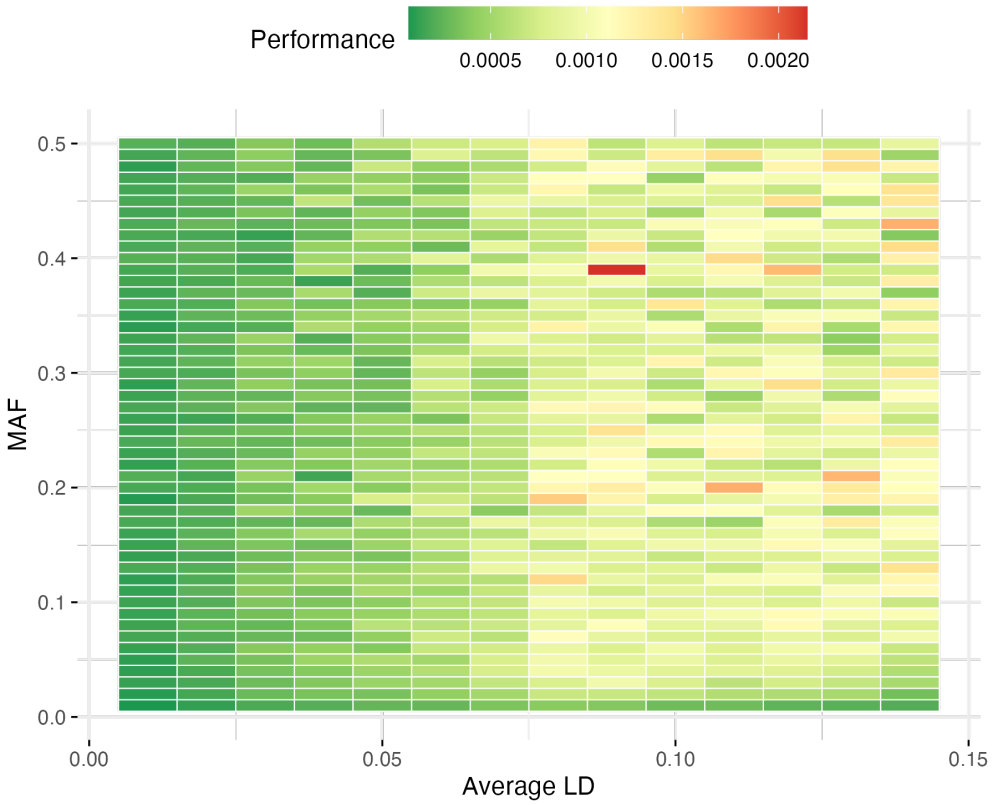


Figure A8: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for LASSO on a population of 100 individuals and 1000 causal variants.

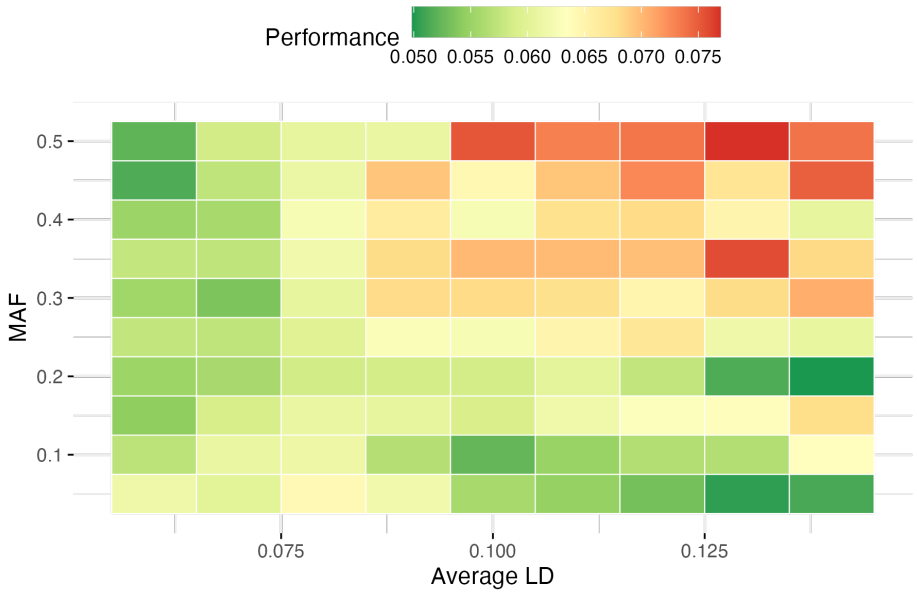


Figure A9: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for E-value on a population of 20 individuals and 100 causal variants.

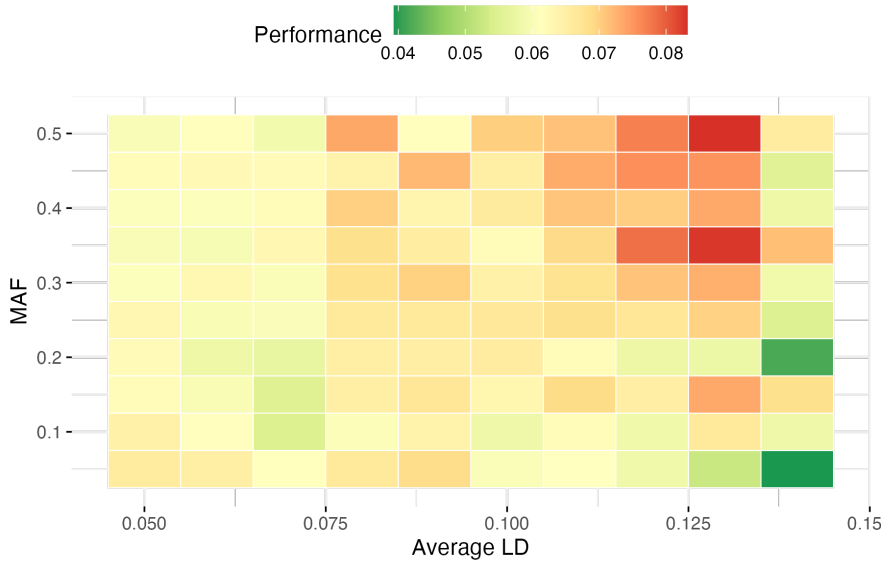


Figure A10: Increasing structure make all methods worse at identifying common variants and GWAS and the LASSO better at identifying rare ones. Heat map of average performance over the simulations with the specified average LD and MAF of highest ranked causal variant. This result is depicted for E-value on a population of 20 individuals and 1000 causal variants.

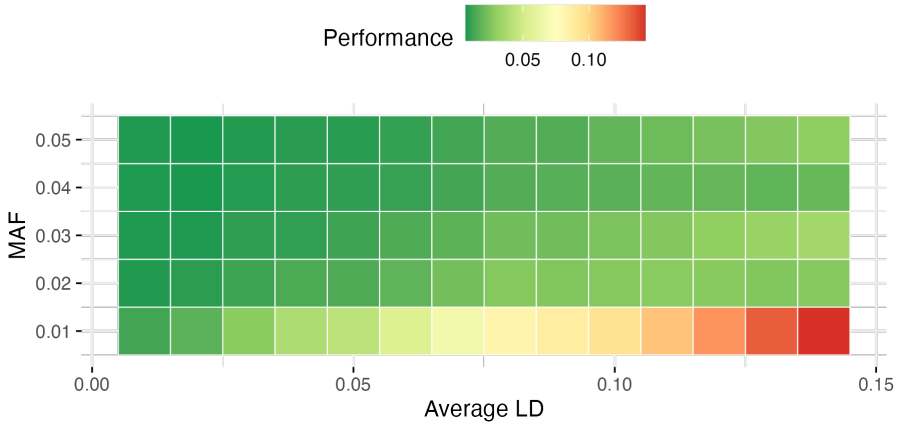


Figure A11: Increasing structure decreases the performance of the E-value, particularly for very rare variants. Heat map of average performance of E-value over the simulations with the specified average LD and MAF of highest ranked causal variant. MAF values are truncated at 0.05 for higher resolution of the performance at singletons. This result is depicted for a population of $n = 100$ individuals and $p = 1000$ causal variants.

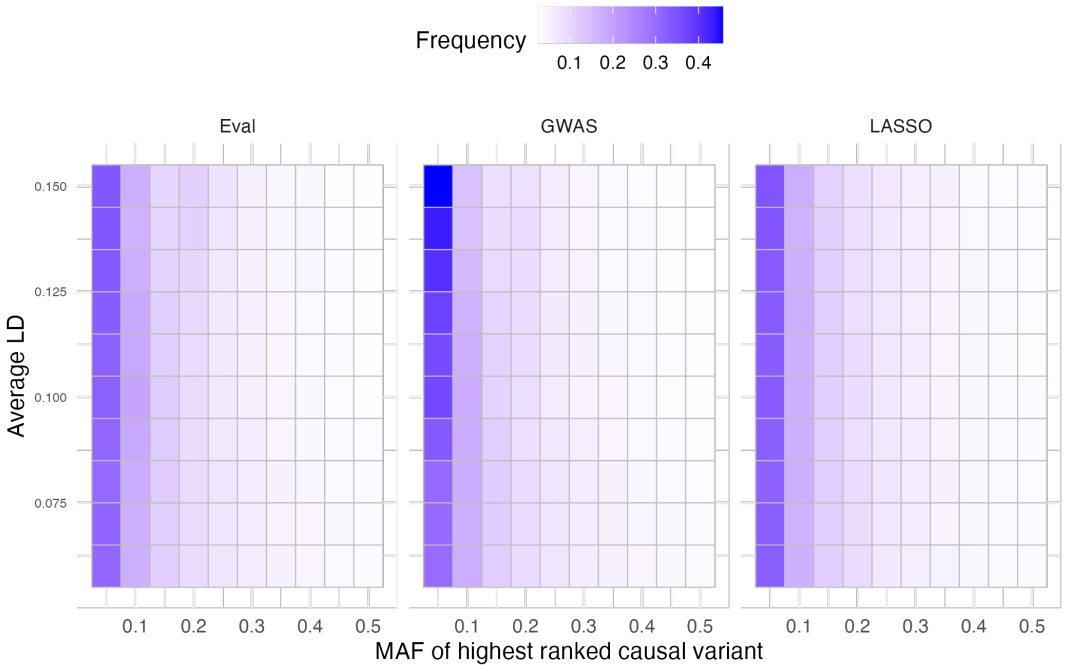


Figure A12: Increasing structure in very small populations makes rare variants more visible relative to common variants. Heat map depicting the frequency distribution of the causal variant first identified by the specified method in populations with increasing amounts of structure. These simulations were of populations with $n = 20$ individuals and $p = 100$ variants.

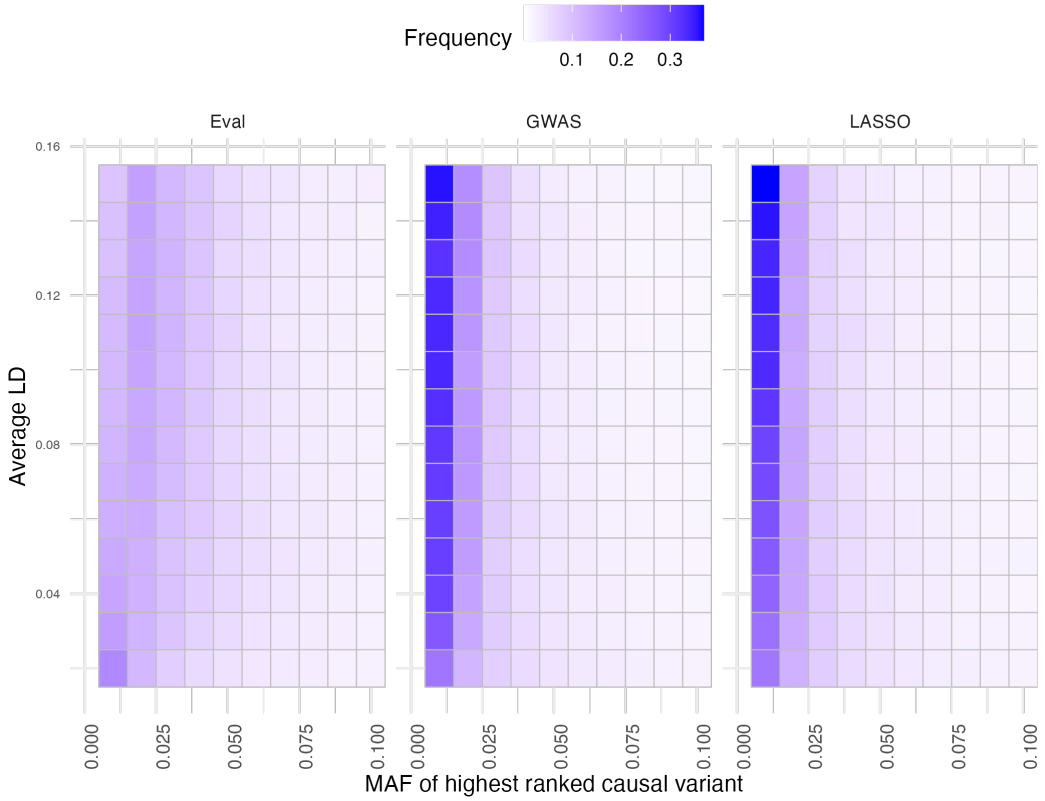


Figure A13: Increasing structure in small populations makes rare variants more visible to GWAS and the LASSO relative to common variants, while the E-value prefers doubletons to singletons. Heat map depicting the frequency of the first variant to be identified by the specified method in populations with increasing amounts of structure. This plot is truncated at MAF = 0.1 for higher resolution of the frequency of rare variants. The simulations were of populations with $n = 100$ individuals and $p = 100$ variants.

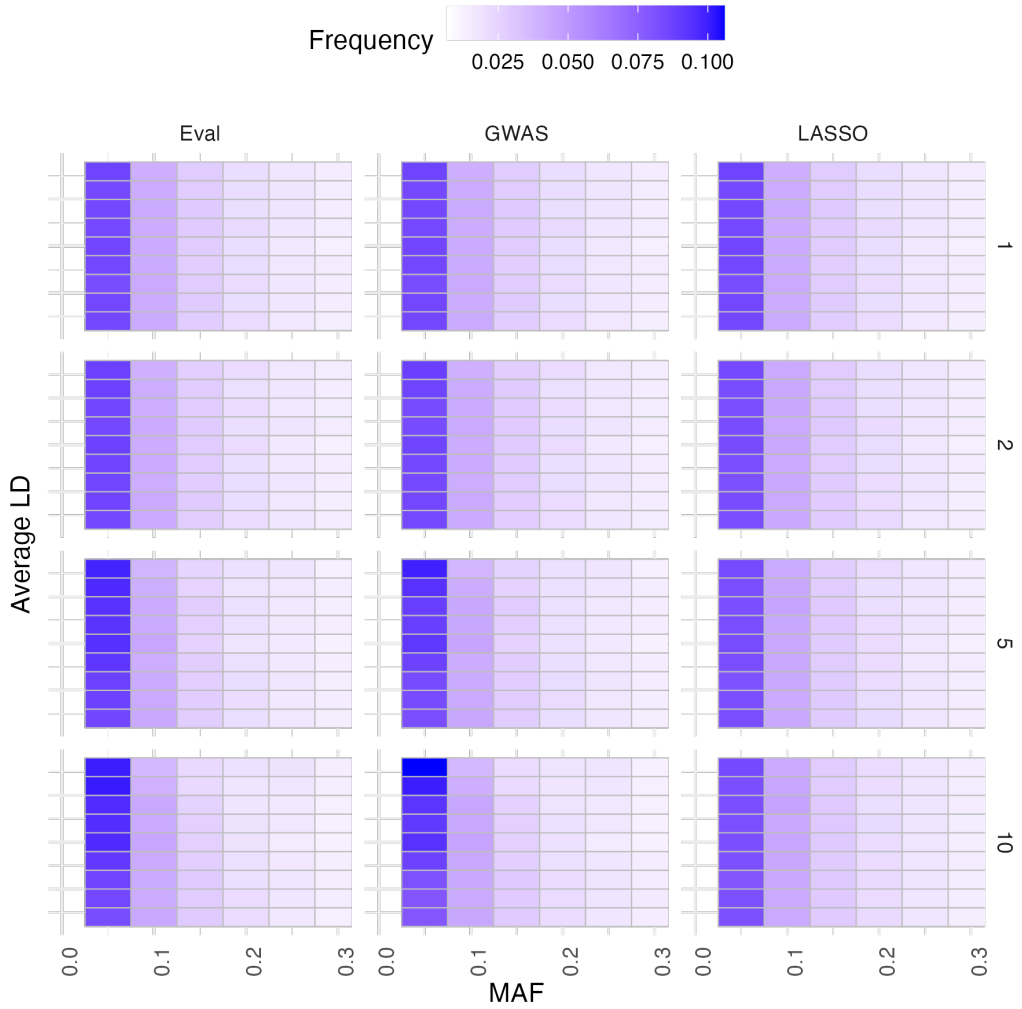


Figure A14: In very small populations, higher numbers of causal variants skew the MAF distribution of identified variants further toward the rare end for GWAS and the E-value. Heat map depicting the frequency of the first variant to be identified by the specified method (shown on the top axis) in populations with increasing structure (shown on the left axis, increasing bottom to top for each panel), simulated with 1, 2, 5 or 10 causal variants (shown on the right axis). This plot is truncated at MAF = 0.3 for higher resolution of the frequency of rare variants. The simulations were of populations with $n = 20$ individuals and $p = 100$ variants.

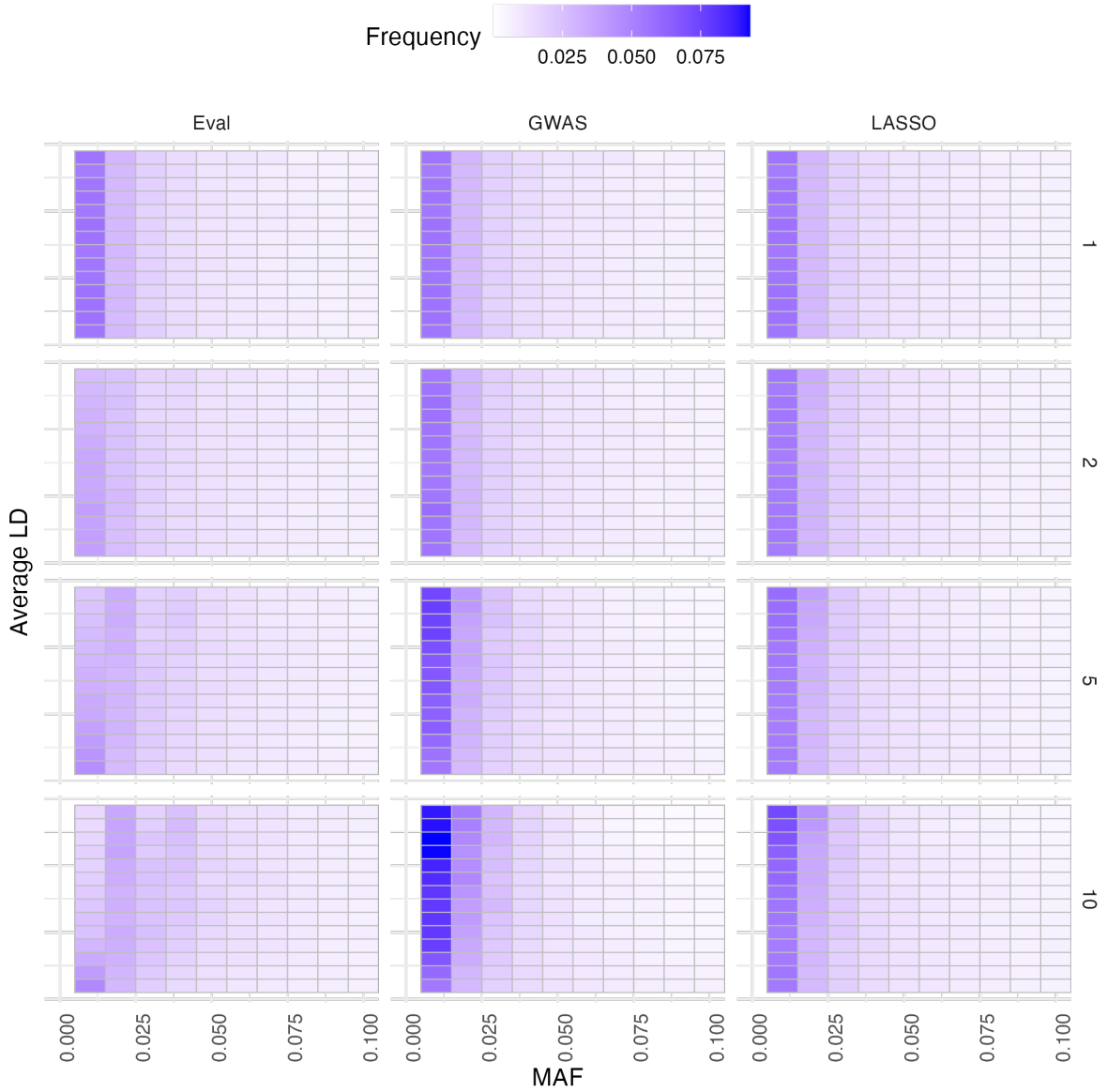


Figure A15: In small populations, higher numbers of causal variants skew the MAF distribution of identified variants further toward singletons for GWAS and the LASSO, while the E-value skews towards doubletons. Heat map depicting the frequency of the first variant to be identified by the specified method (shown on the top axis) in populations with increasing structure (shown on the left axis, increasing bottom to top for each panel), simulated with 1, 2, 5 or 10 causal variants (shown on the right axis). This plot is truncated at MAF = 0.1 for higher resolution of the frequency of rare variants. The simulations were of populations with $n = 100$ individuals, $p = 1000$ variants.