Pairwise Linkage Disequilibrium Estimation for Polyploids

David Gerard

Department of Mathematics and Statistics, American University, Washington, DC, 20016, USA

Abstract

Many tasks in statistical genetics involve pairwise estimation of linkage disequilibrium (LD). The study of LD in diploids is mature. However, in polyploids, the field lacks a comprehensive characterization of LD. Polyploids also exhibit greater levels of genotype uncertainty than diploids, and yet no methods currently exist to estimate LD in polyploids in the presence of such genotype uncertainty. Furthermore, most LD estimation methods do not quantify the level of uncertainty in their LD estimates. Our paper contains three major contributions. (i) We characterize gametic and composite measures of LD in polyploids. These composite measures of LD turn out to be functions of common statistical measures of association. (ii) We derive procedures to estimate gametic and composite LD in polyploids in the presence of genotype uncertainty. We do this by estimating LD directly from genotype likelihoods, which may be obtained from many genotyping platforms. (iii) We derive standard errors of all LD estimators that we discuss. We validate our methods on both real and simulated data. Our methods are implemented in the R package `ldsep`, available on the Comprehensive R Archive Network `https://cran.r-project.org/package=ldsep`.

1 Introduction

Linkage disequilibrium (LD), the statistical association between alleles at different loci, is a fundamental quantity in statistical genetics. Estimates of LD have applications in association mapping [Devlin and Risch, 1995, Jorde, 1995, Xiong and Guo, 1997, Mackay and Powell, 2007, Farh et al., 2015, Gur et al., 2017], analyses using summary statistics from genome-wide association studies (GWAS) [Yang et al., 2012, Benner et al., 2017, Zhu and Stephens, 2018], genomic prediction [Wientjes et al., 2013, Sun et al., 2016], and population genetic studies [Slatkin, 2008, Zhu et al., 2015, Van Wyngaarden et al., 2017, Griffiths et al., 2019], among other tasks [Sved and Hill, 2018].

Many of these LD tasks are now being applied to polyploid organisms. Polyploids, organisms with more than two copies of their genome, are ubiquitous in the plant kingdom [Barker et al., 2016], predominant in agriculture [Udall and Wendel, 2006], and important drivers of evolution [Soltis et al., 2014]. As such, researchers started applying polyploid LD estimates to genotype imputation [Clark et al., 2019, Matias et al., 2019], GWAS [Barreto et al., 2019, Ferrão et al., 2020], genomic prediction [Ramstein et al., 2016, de Bem Oliveira et al., 2019, de C. Lara et al., 2019], and various other applications.

Characterizing and estimating LD in diploids is a mature field. Since the various measures of LD were proposed [Lewontin and Kojima, 1960, Lewontin, 1964, Hill and Robertson, 1968] and the

Keywords and phrases: composite, correlation, gametic, genotype likelihoods, LD, polyploidy.

basic strategies of estimation implemented [see Weir, 1996, for example], researchers have proposed extensions spanning many directions. Methods have been created to estimate LD using genotype data, rather than haplotype data, under the assumption of Hardy-Weinberg equilibrium (HWE) [Hill, 1974, Weir and Cockerham, 1979, Hui and Burt, 2020]. Composite measures of LD have been defined that are estimable using genotype data even when HWE is violated [Cockerham and Weir, 1977, Weir, 1979, Hamilton and Cole, 2004, Zaykin, 2004]. Procedures have been devised to estimate LD in the presence of genotype uncertainty [Li, 2011, Maruki and Lynch, 2014, Bilton et al., 2018, Fox et al., 2019]. Regularization procedures have been suggested to improve LD estimates [Wen and Stephens, 2010].

The research on characterizing and estimating LD in polyploids is much more limited. To date, there have basically been three approaches to estimating LD in polyploids. (i) Researchers assume they have known haplotypes (through phasing of known genotypes or otherwise) and then estimate LD using the empirical haplotype frequencies [Simko et al., 2006, Bradbury et al., 2007, Shen et al., 2016]. (ii) Researchers construct two-way tables of known genotypes and run categorical tests of association, such as Fisher's exact test [Raboin et al., 2008, Julier, 2009, Huang et al., 2020]. (iii) Finally, researchers use standard statistical measures of association, such as the squared Pearson correlation, on the known genotypes between two loci [Björn et al., 2010, Ramstein et al., 2016, Vos et al., 2017, Sharma et al., 2018, de Bem Oliveira et al., 2019].

There are many well-developed methods to obtain phased haplotypes in diploids [Scheet and Stephens, 2006, Browning and Browning, 2007, Li et al., 2010, Swarts et al., 2014], and so method (i) above is an appealing strategy as it allows the study of association directly at the gametic level. However, similar advances in polyploids are relatively infant and are just now emerging in force [Su et al., 2008, Shen et al., 2016, Zheng et al., 2016, Mollinari and Garcia, 2019]. An additional limitation is that these phasing approaches usually require access to a reference genome, which is not always available or necessary in many modern next-generation sequencing pipelines [Lu et al., 2013].

A greater concern is that all of the polyploid LD estimation approaches listed above assume genotypes are known without error. In polyploids, this assumption is incorrect. Even though there have been gains in the accuracy of genotyping methods [Voorrips et al., 2011, Serang et al., 2012, Mollinari and Serang, 2015, Maruki and Lynch, 2017, Schmitz Carley et al., 2017, Blischak et al., 2018, Gerard et al., 2018, Clark et al., 2019, Gerard and Ferrão, 2019, Zych et al., 2019] the issue of genotype uncertainty in polyploids is still severe and much more so than in diploids. In Gerard et al. [2018], we found that genotyping error rates in diploids can be reduced to less than 0.05 at a read-depth of around 5×, but similar error rates could not be achieved for hexaploids in some scenarios until one had read-depths in the many-thousands, an unrealistic scenario for most applied researchers.

In this paper, we provide various methods to estimate LD. After reviewing measures of gametic LD in Section 2.1, in Section 2.2 we derive a procedure to estimate gametic LD in autopolyploids in the presence of genotype uncertainty under the assumption of HWE. We do this by estimating LD directly using genotype likelihoods. In allopolyploids, organisms that exhibit partial preferential pairing, or populations that violate the random-mating assumption, these estimates are not appropriate. Thus, in Section 2.3 we define various "composite" measures of LD that generalize the composite measures proposed in Cockerham and Weir [1977]. These measures turn out to be functions of the statistical moments of the genotypes. In Section 2.4 we provide methods to estimate these composite measures of LD in the presence of genotype uncertainty by directly using genotype likelihoods. To reduce the number of parameters we estimate, we propose a novel and flexible class of distributions over the genotypes. We validate our methods both in simulations (Sections 3.1 and 3.2) and on real data (Sections 3.3 and 3.4). All analyses are reproducible (`https://github.com/dcgerard/ld_simulations`) and all methods are available in the `ldsep` R package on the Comprehensive R Archive Network (`https://cran.r-project.org/package=ldsep`).

2 Methods

Notation: The following contains the notation conventions used throughout this manuscript. We will denote scalars by non-bold letters (a or A), vectors by bold lower-case letters ($\boldsymbol{a}$), and matrices by bold upper-case letters ($\boldsymbol{A}$). The matrix transpose is denoted $\boldsymbol{A}^\intercal$ and the matrix inverse is denoted $\boldsymbol{A}^{-1}$. We let Latin letters denote gametic measures of LD (D, D', and r), while we let Greek letters denote composite measures of LD (Δ, Δ', and ρ). Estimates of population parameters will be denoted with a hat ($\hat{D}$).

2.1 Measures of gametic LD

The original definitions of LD quantify the statistical association between alleles located at different loci on the same gamete. Thus, such associations are sometimes referred to as "gametic LD" [Weir and Cockerham, 1989]. Various measures of gametic LD have been proposed in the literature, each possessing relative strengths and weaknesses for interpreting the association between loci [Hedrick, 1987, Devlin and Risch, 1995]. In this section, we briefly review these common measures of gametic LD.

Perhaps the three most commonly used measures are the LD coefficient D [Lewontin and Kojima, 1960], the standardized LD coefficient D' [Lewontin, 1964], and the Pearson correlation r [Hill and Robertson, 1968]. To define these terms, let A and a be the reference and alternative alleles, respectively, at locus 1. Similarly let B and b be the reference and alternative alleles at locus 2. The LD coefficient is defined to be the difference between the true haplotype frequency and the haplotype frequency under independence:

$$D := p_{AB} - p_A p_B, \tag{1}$$

where p_{AB} denotes the frequency of haplotype AB, p_A denotes the allele frequency of A, and p_B denotes the allele frequency of B. Note that this definition necessarily implies that

$$p_{AB} = p_A p_B + D,\ p_{Ab} = p_A p_b - D,\ p_{aB} = p_a p_B - D,\ \text{and } p_{ab} = p_a p_b + D. \tag{2}$$

The range of D is constrained by the allele frequencies [Lewontin, 1964] by

$$-\min\{p_A p_B, (1-p_A)(1-p_B)\} \leq D \leq \min\{p_A(1-p_B), (1-p_A)p_B\}, \tag{3}$$

and so Lewontin [1964] suggested the standardized LD coefficient:

$$D' := D/D_{max}, \text{ where} \tag{4}$$

$$D_{max} := \begin{cases} \min\{p_A p_B, (1-p_A)(1-p_B)\} & \text{if } D < 0, \\ \min\{p_A(1-p_B), (1-p_A)p_B\} & \text{if } D > 0. \end{cases} \tag{5}$$

The D' coefficient is free to vary between -1 and 1, though it still depends on the loci-specific allele frequencies [Lewontin, 1988]. Finally, the Pearson correlation between loci is

$$r := \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}. \tag{6}$$

It is common to report not D, D', and r, but rather their absolute values or their squares. This is because the sign of the LD between loci depends on the mostly arbitrary labels of the alternative and reference alleles.

2.2 Pairwise gametic LD estimation in autopolyploids under HWE

In this section, we consider estimating gametic LD from a population of autopolyploid individuals under HWE. We will begin by deriving the joint distribution of genotypes at two loci conditional on the LD coefficient and the allele frequencies at both loci. These genotype distributions will then be used, along user-provided genotype likelihoods, to develop a procedure to estimate LD in the presence of genotype uncertainty. Maximum likelihood theory will be used to derive standard errors of these estimators.

Let X_{iAB}, X_{iAb}, X_{iaB}, and X_{iab} be the number of AB, Ab, aB, and ab haplotypes (respectively) in individual i. If each autopolyploid individual is K-ploid, then under HWE we have

$$(X_{iAB}, X_{iAb}, X_{iaB}, X_{iab}) \sim \text{Multinomial}(K; p_{AB}, p_{Ab}, p_{aB}, p_{ab}). \tag{7}$$

Researchers typically do not have access to individual haplotypes as in (7). Rather, most methods assume that researchers have available the dosages of each allele, which we call the genotypes. Let $G_{iA} := X_{iAB} + X_{iAb}$ and $G_{iB} := X_{iAB} + X_{iaB}$ be the number of reference alleles at loci 1 and 2, respectively, in individual i. Then we may sum over the X's to obtain the joint distribution of G_{iA} and G_{iB} (Supplementary Section S1).

$$Pr(G_{iA} = g_A, G_{iB} = g_B | D, p_A, p_B) = \sum_{z=\max(0, g_A+g_B-K)}^{\min(g_A, g_B)} \frac{K! p_{ab}^{K+z-g_A-g_B} p_{Ab}^{g_A-z} p_{aB}^{g_B-z} p_{AB}^{z}}{(K+z-g_A-g_B)!(g_A-z)!(g_B-z)!z!}. \tag{8}$$

Equation (8) is conditional on D, p_A, and p_B due to the relations in (2). Equation (8) contains the distribution of the genotypes under HWE, conditional on D and allele frequencies, and generalizes the formulas in Table 1 of Hill [1974] to autopolyploids.

If the genotypes were known without error, then we could estimate D (along with p_A and p_B)

by maximum likelihood estimation (MLE) using the following log-likelihood:

$$\mathcal{L}(D, p_A, p_B|\mathbf{G}) = \sum_{i=1}^{n} \log Pr(g_{iA}, g_{iB}|D, p_A, p_B), \tag{9}$$

where $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_n)$ and $\mathbf{g}_i = (g_{iA}, g_{iB})$ are the observed genotypes for individual i. Since D' and r are both functions of D, p_A, and p_B (4)–(6), this would also yield the MLEs of D' and r. We have implemented maximizing (9) in our software using gradient ascent. We call the resulting MLEs $\hat{D}_g$, $\hat{D}'_g$, and $\hat{r}_g$ for "genotypes".

However the genotypes are not known without error [Gerard et al., 2018]. Maruki and Lynch [2014] and Bilton et al. [2018] accounted for this in diploid sequencing data by adaptively estimating the sequencing error rate at each locus. However, their models ignore important features concerning polyploid sequencing data, particularly allele bias and overdispersion [Gerard et al., 2018]. Li [2011] and Fox et al. [2019] take the more modular approach of allowing the input of genotype likelihoods for diploids derived from any genotyping software. Letting the input be genotype likelihoods allows the use of different genotyping platforms and software that may account for different features of the data. This also results in greater generalizability since different data types (e.g., microarrays [Fan et al., 2003], next-generation sequencing [Baird et al., 2008, Elshire et al., 2011], or mass spectrometry [Oeth et al., 2009]) can all result in genotype likelihoods that may then be fed into these LD estimation applications. We thus take this approach and consider estimating LD from genotype likelihoods.

We will now describe a procedure to estimate LD while integrating over genotype uncertainty using genotype likelihoods. Let us denote the data for individual i at loci 1 and 2 by y_{iA} and y_{iB}, respectively. Let $p(y_{iA}|g_A)$ and $p(y_{iB}|g_B)$ be the probabilities of the data at loci 1 and 2 given genotypes g_A and g_B. These probabilities are the genotype likelihoods and are assumed to be provided by the user. All of the results in this manuscript use the genotype likelihoods from the `updog` software [Gerard et al., 2018, Gerard and Ferrão, 2019], but we do not require this in the methods below and any genotyping software may be used as long as it returns genotype likelihoods. The log-likelihood of D, p_A, and p_B given the data is:

$$\mathcal{L}(D, p_A, p_B|\mathbf{y}) = \sum_{i=1}^{n} \log \left(\sum_{g_B=0}^{K} \sum_{g_A=0}^{K} p(y_{iA}|g_A) p(y_{iB}|g_B) Pr(g_A, g_B|D, p_A, p_B) \right). \tag{10}$$

We developed an expectation-maximization (EM) algorithm [Dempster et al., 1977] to maximize (10) and estimate haplotype frequencies and, thus, LD (Section S2). For diploids ($K = 2$), this estimation procedure reduces down to that proposed in Li [2011] and implemented in the `ngsLD` software [Fox et al., 2019] (Figure S1). However, we found it more efficient to maximize likelihood (10) using gradient ascent. Optimization was performed over the unit 3-simplex using the unconstrained transformed parameter space used in Betancourt [2012] before back-transforming to the original parameter space. When LD is close to 1, this can cause MLEs on the boundary of the parameter space and, thus, nonsensical standard error estimates (described below). We thus take the approach of [Agresti and Coull, 1998] and add a small penalty on the haplotype frequencies. This penalty is equivalent to placing a Dirichlet(2,2,2,2) prior on the haplotype frequencies and corresponds to the "add two" rule. We call the resulting penalized MLEs $\hat{D}_{gl}$, $\hat{D}'_{gl}$, and $\hat{r}_{gl}$ for "genotype likelihoods".

Standard errors are important for hypothesis testing [Brown, 1975], read-depth suggestions [Maruki and Lynch, 2014], and hierarchical shrinkage [Dey and Stephens, 2018]. However, most LD estimation methods do not return standard error estimates. We now provide a description for how to obtain such estimates using standard maximum likelihood theory. First, we numerically approximated the Hessian matrix of $\mathcal{L}(p_{ab}, p_{Ab}, p_{aB})$ evaluated at the maximum likelihood estimators $(\hat{p}_{ab}, \hat{p}_{Ab}, \hat{p}_{aB})$, $\boldsymbol{H} := (\frac{\partial^2 \mathcal{L}}{\partial p_i \partial p_j})|_{p_{ab}=\hat{p}_{ab}, p_{Ab}=\hat{p}_{Ab}, p_{aB}=\hat{p}_{aB}}$. This likelihood can be either that using genotypes (9) or that using genotype likelihoods (10), both of which are functions of (p_{ab}, p_{Ab}, p_{aB}) since the haplotype frequencies must sum to one. Standard maximum likelihood theory guarantees that for large n, the limiting covariance matrix of $(\hat{p}_{ab}, \hat{p}_{Ab}, \hat{p}_{aB})$ is well approximated by $-\boldsymbol{H}^{-1}$ [Lehmann and Casella, 1998]. The MLEs of the various LD measures ($\hat{D}$, $\hat{D}'$, and $\hat{r}$) are all functions of $(\hat{p}_{ab}, \hat{p}_{Ab}, \hat{p}_{aB})$, and so we can use the δ-method to obtain the limiting variance of these LD estimators. For example, let $\boldsymbol{g} = (\frac{\partial r}{\partial p_{ab}}, \frac{\partial r}{\partial p_{Ab}}, \frac{\partial r}{\partial p_{aB}})^\intercal$ be the gradient of r with respect to (p_{ab}, p_{Ab}, p_{aB}) evaluated at $(\hat{p}_{ab}, \hat{p}_{Ab}, \hat{p}_{aB})$. Then the limiting variance of $\hat{r}$ is $-\boldsymbol{g}^\intercal \boldsymbol{H}^{-1} \boldsymbol{g}$. The gradient calculations are all standard and so have been omitted. We have implemented this procedure for standard error calculation in our software.

2.3 Composite measures of LD

The methods in Section 2.2 were developed under the assumption of HWE in autopolyploids. The goal in this section is to develop measures of LD that are (i) valid measures of association when HWE is violated, (ii) identified even when only genotype information is provided, and (iii) reduced to the gametic measures of LD (Section 2.1) when HWE is satisfied in autopolyploids. These measures will be called "composite" measures of LD, as they account for associations not just at the gametic level.

The easiest composite measure to obtain that satisfies our goals is that which generalizes r (6). Let G_A and G_B be the genotypes of a randomly selected K-ploid individual at loci 1 and 2. Then, quite simply, the composite measure of correlation is the Pearson correlation between genotypes:

$$\rho := \text{cor}(G_A, G_B). \tag{11}$$

We prove in Section S3 that ρ equals r when HWE is satisfied in autopolyploids.

In diploids, Cockerham and Weir [1977] (attributing their result to Peter Burrows' unpublished work) suggested a composite measure of LD which they defined as the sum of gametic LD and non-gametic LD. This sum depends only on genotype frequencies, not the haplotype frequencies, and is thus identified even under general violations from HWE. Under HWE, this composite LD measure was equal to D, and so Weir [1979] suggested its use even when HWE was satisfied. Let q_{ij} denote the probability of a randomly selected individual containing genotype i on locus 1 and j on locus 2. Then the coefficient as defined in Cockerham and Weir [1977] is

$$2q_{22} + q_{12} + q_{21} + \frac{1}{2}q_{11} - 2p_A p_B, \tag{12}$$

where $p_A = \frac{1}{2}\sum_{g_A=0}^{2} g_A(q_{g_A0} + q_{g_A1} + q_{g_A2})$ and $p_B = \frac{1}{2}\sum_{g_B=0}^{2} g_B(q_{0g_B} + q_{1g_B} + q_{2g_B})$. It is easy to show that (12) is actually 1/2 the covariance of the genotypes [Weir, 2008, Rogers and Huff, 2009, Section S4]. This immediately suggests a generalized composite LD coefficient for polyploids:

$$\Delta := \frac{1}{K}\text{cov}(G_A, G_B) = \frac{1}{K}E[G_AG_B] - \frac{1}{K}E[G_A]E[G_B] = \frac{1}{K}\sum_{g_A=0}^{K}\sum_{g_B=0}^{K} g_Ag_Bq_{g_Ag_B} - Kp_Ap_B, \quad (13)$$

where

$$p_A = \frac{1}{K}\sum_{i=0}^{K} i \sum_{j=0}^{K} q_{ij} = \frac{1}{K}E[G_A] \text{ and } p_B = \frac{1}{K}\sum_{j=0}^{K} j \sum_{i=0}^{K} q_{ij} = \frac{1}{K}E[G_B]. \quad (14)$$

In Section S3, we show that under HWE in autopolyploids, Δ (13) equals D (1). We provide a decomposition of Δ in Section 2.3.1 that generalizes the decomposition in Cockerham and Weir [1977].

To define a composite measure that generalizes D' for polyploids there are two approaches that we can take, depending on what properties of the genotype distributions we condition on while finding the maximum value of Δ. D' is found by normalizing D with the maximum value of D while fixing the allele frequencies. The allele frequency at a locus can be considered both a representation of the marginal distribution of an allele and the expected value of an allele. Thus, while normalizing Δ, we can either find the maximum value of Δ while conditioning on the marginal distribution of genotypes, or while conditioning on the expected value of genotypes.

This first approach generalizes that of Zaykin [2004] and Hamilton and Cole [2004] from diploids to polyploids. They found the maximum value of Δ in closed-form for diploids while conditioning on the marginal distributions of genotypes at both loci. For our generalization, we note that we can formulate this maximization problem as a linear program [Nocedal and Wright, 2006] and so can solve it efficiently for any ploidy level. Specifically, we seek to find the s_{ij}'s that solve the following maximization problem:

$$\text{maximize} \sum_{i=0}^{K}\sum_{j=0}^{K} ijs_{ij} \quad (15)$$

$$\begin{aligned}\text{subject to } & \sum_{i=0}^{K}\sum_{j=0}^{K} s_{ij} = 1,\\ & \sum_{j=0}^{K} s_{ij} = c_i \text{ for all } i = 0,1,\ldots,K,\\ & \sum_{i=0}^{K} s_{ij} = d_j \text{ for all } j = 0,1,\ldots,K, \text{ and}\\ & s_{ij} \geq 0 \text{ for all } i = 0,1,\ldots,K \text{ and } j = 0,1,\ldots,K,\end{aligned} \quad (16)$$

where c_i and d_j are the provided marginal probabilities of genotype i at locus 1 and genotype j at locus 2, respectively. Since (15)–(16) formulates a standard linear program, many out-of-the-box optimization software packages may be used to solve it [Berkelaar et al., 2016, e.g.]. The optimal values of the s_{ij}'s may then be used to find the maximum value of Δ using (13), which can then be used to normalize Δ. If Δ is negative then we instead minimize (15) and normalize Δ by the absolute value of the minimum. Specifically, let Δ_m be the maximum of value of (13) when $\Delta > 0$, or the absolute value of the minimum of (13) when $\Delta < 0$, where we use constraints (16). Then we define

$$\Delta_g' := \Delta/\Delta_m, \tag{17}$$

where the subscript is for "genotype frequency".

For diploids, Zaykin [2004] noted that normalizing by the maximum covariance conditional on the marginal distributions of genotypes at each locus did not result in D when HWE was fulfilled. The same is true for polyploids, though we note that under HWE there appears to be a simple piecewise-linear relationship between D' and Δ_g' (Figure S2). As such, Zaykin [2004] in diploids also recommended the second approach of maximizing the covariance conditional on the expected genotype. We could also formulate these bounds as a linear program. However, we can actually represent the bounds on Δ (13) in closed-form (Theorem S2):

$$-K\min((1-p_A)(1-p_B), p_Ap_B) \leq \Delta \leq K\min(p_A(1-p_B), (1-p_A)p_B). \tag{18}$$

Equation (18) would suggest dividing Δ by $K\min((1-p_A)(1-p_B), p_Ap_B)$ when $\Delta < 0$, and dividing Δ by $K\min(p_A(1-p_B), (1-p_A)p_B)$ when $\Delta > 0$. This would result in a composite measure of LD that is bounded within $[-1, 1]$. However, the resulting measure would still not equal D' when HWE is satisfied. Thus, we prefer the following composite measure of LD

$$\Delta_a' := \frac{\frac{1}{K}\operatorname{cov}(G_A, G_B)}{\Delta_e} = \frac{\Delta}{\Delta_e}, \text{ where} \tag{19}$$

$$\Delta_e := \begin{cases} \min\{p_Ap_B, (1-p_A)(1-p_B)\} & \text{if } \Delta < 0, \\ \min\{p_A(1-p_B), (1-p_A)p_B\} & \text{if } \Delta > 0. \end{cases} \tag{20}$$

We prove in Section S3 that this Δ_a' (for "allele frequency") is equal to D' under HWE in autopolyploids but is still estimable when HWE is violated. For the rest of this manuscript, we will only consider Δ_a' and not Δ_g'.

Though a perceived disadvantage of (19) would be that it is constrained to fall within $[-K, K]$ rather than $[-1, 1]$, we find it compelling that Δ_a' can be viewed as a direct generalization of D' and is equal to D' when HWE is satisfied in autopolyploids. However, a researcher may always divide (19) by K to obtain a measure that is constrained to be within $[-1, 1]$.

2.3.1 A decomposition of Δ

Cockerham and Weir [1977] derived their composite LD measure by summing gametic and non-gametic LD. From a different point of view, this can be seen as a decomposition of Δ. We will now derive a generalized decomposition of (13) for polyploids. Let z_{kA} be the indicator variable that equals 1 if an individual has the A allele on locus 1 of chromosome k, and 0 otherwise. Similarly, let z_{kB} be the indicator variable that equals 1 if an individual has the B allele on locus 2 of chromosome k, and 0 otherwise. Then the genotype for an individual equals

$$G_A = \sum_{k=1}^K z_{kA} \text{ and } G_B = \sum_{k=1}^K z_{kB}. \tag{21}$$

Thus,

$$\begin{aligned}\Delta &= \frac{1}{K}\operatorname{cov}(G_A, G_B) &(22)\\ &= \frac{1}{K}\operatorname{cov}(\sum_{k_A=1}^{K} z_{k_A A}, \sum_{k_B=1}^{K} z_{k_B B}) &(23)\\ &= \frac{1}{K}\sum_{k_A=1}^{K}\sum_{k_B=1}^{K}\operatorname{cov}(z_{k_A A}, z_{k_B B}). &(24)\end{aligned}$$

Equation (24) is as simple a decomposition as can be attained for Δ without further assumptions. However, in the special case of auto- and allopolyploidy, we can further reduce (24). In the case of autopolyploidy, we have the following two identities:

$$\begin{aligned}D_{AB} &:= \operatorname{cov}(z_{iA}, z_{iB}) = \operatorname{cov}(z_{i'A}, z_{i'B}) \text{ for all } i \text{ and } i', \text{ and} &(25)\\ D_{A/B} &:= \operatorname{cov}(z_{iA}, z_{jB}) = \operatorname{cov}(z_{i'A}, z_{j'B}) \text{ for all } i \neq j \text{ and } i' \neq j', &(26)\end{aligned}$$

where D_{AB} is gametic LD and $D_{A/B}$ is non-gametic LD. Thus, for autopolyploids, we obtain

$$\Delta = D_{AB} + (K-1)D_{A/B}, \tag{27}$$

which more clearly demonstrates the generalization of the decomposition in Cockerham and Weir [1977] from diploids to autopolyploids. In the case of HWE, $D_{A/B} = 0$ and we obtain $\Delta = D_{AB}$.

In the case of allopolyploidy with an even ploidy level, there are $K/2$ homologous pairs. There are three types of LD that may occur. First, there is gametic LD in homologous pair i, denoted D_{ABi}. Second, there is non-gametic LD in homologous pair i, denoted $D_{A/Bi}$. Third, there is non-gametic LD between a chromosome in homologous pair i and a chromosome in homologues pair j, denoted $D_{A/Bij}$. Thus, in the case of allopolyploids, we obtain

$$\Delta = \frac{2}{K}\left[\sum_{i=1}^{K/2} D_{ABi} + \sum_{i=1}^{K/2} D_{A/Bi} + 4\sum_{i=1}^{K/2-1}\sum_{j=i+1}^{K/2} D_{A/Bij}\right]. \tag{28}$$

In the case of HWE, we have $D_{A/Bi} = 0$ for all i and $D_{A/Bij} = 0$ for all (i,j). This indicates that for allopolyploids, in HWE, the composite measure of LD (24) is the average of gametic LDs for all homologous pairs:

$$\Delta = \frac{1}{K/2}\sum_{i=1}^{K/2} D_{ABi}. \tag{29}$$

2.4 Estimating composite measures of LD

When genotypes are known, it would be appropriate to use the sample moments of genotypes to estimate ρ (11), Δ (13), and Δ'_a (19). When genotypes are not known, it is natural to plug-in a good estimator of the genotypes, such as the posterior means, into the sample moments. The typical methods of covariance and correlation standard errors may then be used (Section S7). This is fast and so the sample correlation of posterior genotypes has been used in the literature [Clark et al., 2019, Fox et al., 2019] (though not with the capability of returning standard errors). We will denote such estimators by $\hat{\rho}_{mom}$, $\hat{\Delta}_{mom}$, and $\hat{\Delta}'_{mom}$ for "moment-based".

However, as we will see in Section 3.1, using the posterior mean in such moment-based estimators results in LD estimates that are biased low. We can explain this by the law of total covariance:

$$\text{cov}(G_A, G_B) = \text{cov}(E[G_A|Y], E[G_B|Y]) + E[\text{cov}(G_A, G_B|Y)], \tag{30}$$

where Y contains the data. The covariance of posterior means is only the first term on the right hand side of (30), and so does not account for the posterior covariance of genotypes. Thus, the covariance of the posterior means is necessarily a biased estimator for Δ.

When genotypes are not known we can still estimate these composite LD measures by first estimating the q_{ij}'s in (13) using maximum likelihood and then using these estimates to obtain the MLEs for ρ, Δ, and Δ'_a. The likelihood to be maximized is

$$\prod_{\ell=1}^{n}\sum_{i=0}^{K}\sum_{j=0}^{K} p(y_{\ell A}|i)p(y_{\ell B}|j)q_{ij} \tag{31}$$

One EM step to maximize (31) consists of

$$w_{\ell ij} = \frac{p(y_{\ell A}|i)p(y_{\ell B}|j)q_{ij}^{(old)}}{\sum_{ij} p(y_{\ell A}|i)p(y_{\ell B}|j)q_{ij}^{(old)}},\ q_{ij} = \frac{1}{n}\sum_{\ell=1}^{n} w_{\ell ij}. \tag{32}$$

We will call the resulting MLEs $\hat{\rho}_{gc}$, $\hat{\Delta}_{gc}$ and $\hat{\Delta}_{gc}$ for "general categorical", as the q_{ij}'s are allowed to vary over the space of general categorical distributions.

To calculate standard errors, it would be possible to take the same approach as in Section 2.2 and derive asymptotic variances using the Fisher information and appealing to the δ-method (Section S6). However, issues arise with ploidies greater than two. As there are $(K+1)^2$ possible genotype conditions ($(K+1)^2 - 1$ free parameters), except for very large n it will not be uncommon for the MLEs to occur on the boundary of the parameter space, thereby violating the regulatory conditions for the asymptotic standard errors of the MLE [Lehmann and Casella, 1998]. In such cases, we appeal to bootstrap standard errors [Efron, 1979] for the composite measures of LD.

Equation (31) uses a general categorical distribution with support over the possible genotype pairs. This distribution over genotype pairs is the most flexible possible, but yields $(K+1)^2 - 1$ parameters to estimate. Adding constraints over the space of possible genotype distributions can improve estimation performance as long as the genotype frequencies follow these constraints. Gerard and Ferrão [2019] introduced the proportional normal distribution, a very flexible class of distributions with support over genotypes at one locus. We now generalize this distribution to include support over the genotypes at two loci. Let $\boldsymbol{\mu} \in \mathbb{R}^2$ and let $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$ be a positive definite matrix. Then the proportional bivariate normal distribution with support over $\{0, 1, \ldots, K\}^2$ is:

$$Pr(g_A, g_B|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{N(\binom{g_A}{g_B}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sum_{i=0}^{K}\sum_{j=0}^{K} N(\binom{i}{j}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}, \tag{33}$$

where $N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the bivariate normal density evaluated at $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix

Estimators	**Type**	**Input**	**Notes**
$\hat{D}_g$, $\hat{D}'_g$, $\hat{r}_g$	Gametic	Genotypes	
$\hat{D}_{gl}$, $\hat{D}'_{gl}$, $\hat{r}_{gl}$	Gametic	Genotype Likelihoods	A Dirichlet$(2, 2, 2, 2)$ prior is placed on the haplotype frequencies.
$\hat{\Delta}_{mom}$, $\hat{\Delta}'_{mom}$, $\hat{\rho}_{mom}$	Composite	Genotypes	Can accept continuous genotypes as input.
$\hat{\Delta}_{gc}$, $\hat{\Delta}'_{gc}$, $\hat{\rho}_{gc}$	Composite	Genotype Likelihoods	Uses the general categorical class of genotype distributions.
$\hat{\Delta}_{pn}$, $\hat{\Delta}'_{pn}$, $\hat{\rho}_{pn}$	Composite	Genotype Likelihoods	Uses the proportional bivariate normal class of genotype distributions.

Table 1: Summary of LD estimators. The $\hat{\Delta}'$ estimators all estimate Δ'_a (19), *not* Δ'_g (17).

$\boldsymbol{\Sigma}$. This distribution, though seemingly *ad hoc*, can be seen as a generalization of the distribution of genotypes under HWE (Figure S3). Though its use for modeling genotypes when HWE is violated can be rationalized by the flexible shapes of genotype distributions it is capable of representing (Figure S4).

Using the proportional bivariate normal distribution, the log-likelihood to be maximized is:

$$\begin{aligned}
&\sum_{\ell=1}^{n} \log\left[\sum_{i=0}^{K}\sum_{j=0}^{K} p(y_{\ell A}|i)p(y_{\ell B}|j)Pr(i,j|\boldsymbol{\mu},\boldsymbol{\Sigma})\right]\\
&= \sum_{\ell=1}^{n} \log\left[\sum_{i=0}^{K}\sum_{j=0}^{K} p(y_{\ell A}|i)p(y_{\ell B}|j)N(\tbinom{i}{j}|\boldsymbol{\mu},\boldsymbol{\Sigma})\right] - n\log\left[\sum_{i=0}^{K}\sum_{j=0}^{K} N(\tbinom{i}{j}|\boldsymbol{\mu},\boldsymbol{\Sigma})\right]
\end{aligned} \tag{34}$$

We can maximize (34) using gradient ascent. We performed this maximization over the space of $\boldsymbol{\mu}$ and lower-triangular matrix $\boldsymbol{L}$ where $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\intercal}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$. Asymptotic standard errors may be obtained using standard results from maximum likelihood theory. That is, we obtain the Hessian of the log-likelihood (34) evaluated at the maximum likelihood estimators $(\hat{\boldsymbol{\mu}}, \text{vec}(\hat{\boldsymbol{L}}))$, where $\text{vec}(\boldsymbol{L})$ is the vectorization of the lower-triangle of $\boldsymbol{L}$. Call this Hessian $\boldsymbol{H}$. The MLEs of ρ, Δ, and Δ'_a are all functions of the MLEs $(\hat{\boldsymbol{\mu}}, \text{vec}(\hat{\boldsymbol{L}}))$. To see this, set $q_{ij} = Pr(i, j|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and substitute the q_{ij}'s in (S73), (S75), and (S84). Thus, they each admit a gradient for the functions mapping from $(\hat{\boldsymbol{\mu}}, \text{vec}(\hat{\boldsymbol{L}}))$ to $\hat{\rho}$, $\hat{\Delta}$, and $\hat{\Delta}'_a$. Call each gradient $\boldsymbol{g}$. Then the asymptotic variance of an estimator of composite LD is $-\boldsymbol{g}^{\intercal}\boldsymbol{H}\boldsymbol{g}^{\intercal}$. We have implemented all gradient calculations numerically. We additionally placed weakly informative priors over $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as we noticed some scenarios resulted in divergent optimization behavior:

$$\boldsymbol{\mu} \sim N(\boldsymbol{0}, (K/2, K/2)^{\intercal}, \text{diag}((2K)^2, (2K)^2)),\ \boldsymbol{\Sigma} \sim Wishart_2(\boldsymbol{I}_2, 2). \tag{35}$$

The induced distribution over $\boldsymbol{L}$ can be found by Bartlett's decomposition [Bartlett, 1934]. We will denote the resulting MLEs by $\hat{\rho}_{pn}$, $\hat{\Delta}_{pn}$ and $\hat{\Delta}'_{pn}$ for "proportional normal".

A summary of the various estimators we have proposed in this paper are presented in Table 1. The $\hat{\Delta}'$ estimators all estimate Δ'_a (19), *not* Δ'_g (17).

3 Results

3.1 Pairwise LD simulations under HWE

In this section we run simulations where genotypes are generated for autopolyploids under HWE. Under HWE for autopolyploids, both the gametic (Section 2.2) and the composite (Section 2.4) estimators are valid estimators of gametic LD.

These are the steps of a single simulation replication. Given the major allele frequencies (p_A and p_B) and Pearson correlation r (6), the haplotype frequencies ($p_{AB}, p_{Ab}, p_{aB}, p_{ab}$) are uniquely identified. For individual $i \in \{1, 2, \ldots, n = 100\}$ at a given ploidy level K, we simulated the number of each haplotype they contained ($X_{iAB}, X_{iAb}, X_{iaB}, X_{iab}$) given the haplotype frequencies using (7). Genotypes were calculated by $G_{iA} := X_{iAB} + X_{iAb}$ and $G_{iB} := X_{iAB} + X_{iaB}$. Given these genotypes, read-counts were simulated using `updog`'s `rflexdog()` function at a specified read-depth, a 0.01 sequencing error rate, no allele bias, and an overdispersion value of 0.01. `Updog` was then used to generate genotype likelihoods, posterior mode genotypes, and posterior mean genotypes. These outputs were fed into `ldsep` to provide the estimators listed in Table 1. The parameters that varied within the simulation were: the read depth $\in \{1, 5, 10, 50, 100\}$, the ploidy $K \in \{2, 4, 6, 8\}$, the major allele frequencies $(p_A, p_B) \in \{(0.5, 0.5), (0.5, 0.75), (0.9, 0.9)\}$, and the Pearson correlation $r \in \{0, 0.5, 0.9\}$. When $p_A = 0.5$ and $p_B = 0.75$, r is constrained to be less than $1/\sqrt{3} \approx 0.58$, and so in this scenario the $r = 0.9$ setting was omitted. Each unique combination of parameters was replicated 200 times.

The conclusions for estimating r^2 and ρ^2 when $p_A = p_B = 0.5$ are presented in Figures 1, S5, and S6. Mean-squared error performance for estimating D and D' when $p_A = p_B = 0.5$ are presented in Figures S7 and S8, respectively. Results for other scenarios are similar and are available on GitHub (`https://github.com/dcgerard/ld_simulations`). The general conclusions are:

- The moment-based estimators of composite LD ($\hat{\Delta}_{mom}$, $\hat{\Delta}'_{mom}$, $\hat{\rho}_{mom}$) have a strong bias toward 0 until a large read-depth is attained (Figure S5). This bias makes these estimators look like they perform very well when LD is close to zero, but they consequently perform very poorly for large levels of LD.
- Maximum likelihood estimation of gametic LD using genotype likelihoods ($\hat{D}_{gl}$, $\hat{D}'_{gl}$, $\hat{r}_{gl}$) generally dominates maximum likelihood estimation of gametic LD using posterior mode genotypes ($\hat{D}_g$, $\hat{D}'_g$, $\hat{r}_g$) in terms of bias and mean squared error (MSE) (Figures S5 and 1).
- Maximum likelihood estimation using the genotype likelihoods is generally unbiased (Figure S5). It pays for this lower bias by having a larger standard error (Figure S6), but it also produces lower MSE in non-zero LD regimes (Figure 1).
- Even though these data were simulated under HWE, estimators of composite measures of LD using genotype likelihoods generally perform as well as the estimators of gametic LD when estimating r^2 (Figure 1). The HWE assumption helps when estimating D and D' (Figures S7 and S8).
- Using genotype likelihoods is mostly important in settings of high ploidy. All methods behave rather similarly in diploids, but the genotype likelihood approaches perform much better at higher ploidy levels in high LD regimes (Figure 1).

Figures 2 and S9 highlight that our standard errors for the LD estimators are generally accurate except when the read-depth is 1. Computation time for each method is presented in Figure S11. We see there that ploidy is the major cause of computation time increases and that genotype likelihood methods tend to be much slower than methods that use estimated genotypes. However, all methods take less than half a second on average.

3.2 Pairwise LD simulations when HWE is violated

In this section, we evaluate the performance of the various LD estimators when HWE is violated. We do this by simulating genotypes directly from the proportional bivariate normal distribution. In this case, since HWE is violated, the composite LD estimators (Section 2.4) are still appropriate measures of association, but the estimators of gametic LD (Section 2.2) are estimators under a misspecified model.

Each replication, given a ploidy K, we generated genotypes for 100 individuals from a proportional bivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^2$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij}) \in \mathbb{R}^{2\times 2}$. Given these genotypes, we simulated read-counts using `updog`'s `rflexdog()` function at a specified read-depth, a 0.01 sequencing error rate, no allele bias, and an overdispersion value of 0.01. `Updog` was then used to generate genotype likelihoods, posterior mode genotypes, and posterior mean genotypes. These outputs were fed into `ldsep` to provide the estimators listed in Table 1. The parameters that varied within the simulation were: the ploidy $K \in \{2, 4, 6, 8\}$, the mean parameter $\boldsymbol{\mu} = (p_1K, p_2K)$ where $(p_1, p_2) \in \{(0.5, 0.5), (0.5, 0.75), (0.9, 0.9)\}$, the scale parameter $\sigma_{11} = \sigma_{22} \in \{K^2/4, K^2\}$, the association parameter $\sigma_{12} \in \{0, 0.5\sqrt{\sigma_{11}\sigma_{22}}, 0.9\sqrt{\sigma_{11}\sigma_{22}}\}$, and the read-depth $\in \{1, 5, 10, 50, 100\}$. Each unique combination of parameter values was replicated 200 times.

The conclusions for estimating ρ^2 when $\mu_1 = \mu_2 = K/2$ and $\sigma_{11} = \sigma_{22} = K^2/4$ are presented in Figures 3, S12, and S13. The results for other scenarios are similar and are available on GitHub (`https://github.com/dcgerard/ld_simulations`). The general conclusions are:

- Composite measures perform the best under high levels of LD. Most methods are unbiased under low levels of LD, but only composite measures are unbiased under high levels of LD.
- The composite measure using genotype likelihoods and the proportional normal genotype distribution class ($\hat{\Delta}_{pn}, \hat{\Delta}'_{pn}, \hat{\rho}_{pn}$) generally had the best performance overall, while using the general categorical class of genotype distributions ($\hat{\Delta}_{gc}, \hat{\Delta}'_{gc}, \hat{\rho}_{gc}$) resulted in higher standard errors.

Computation times for each method are presented in Figure S14, the results being similar to those described in Section 3.1.

3.3 LD estimates using data from Uitdewilligen et al. [2013]

We evaluated all pairwise LD estimators discussed in this paper on the genotyping-by-sequencing data from Uitdewilligen et al. [2013]. These data come from a diversity panel of autotretraploid *Solanum tuberosum* ($2n = 4x = 48$). We obtained pairwise LD estimates from all SNPs on two contigs that have an alternative allele frequency between 0.1 and 0.9. One contig (labeled CONT0988) targets a gene involved in sucrose synthesis, and one contig (labeled CONT1561) targets a zinc finger. These contigs are located on two different super scaffolds. These contigs were chosen somewhat arbitrarily, but the results we present below are robust to the selection of contigs and the reader is encouraged to change the contigs in our reproducible analysis scripts on GitHub (`https://github.com/dcgerard/ld_simulations`).

The main results are presented in Figure 4. We would expect methods to perform well if they exhibit (i) large LD estimates between SNPs within a contig and (ii) small LD estimates between SNPs on different contigs. Both contigs are relatively small (1113 bp and 1585 bp) and

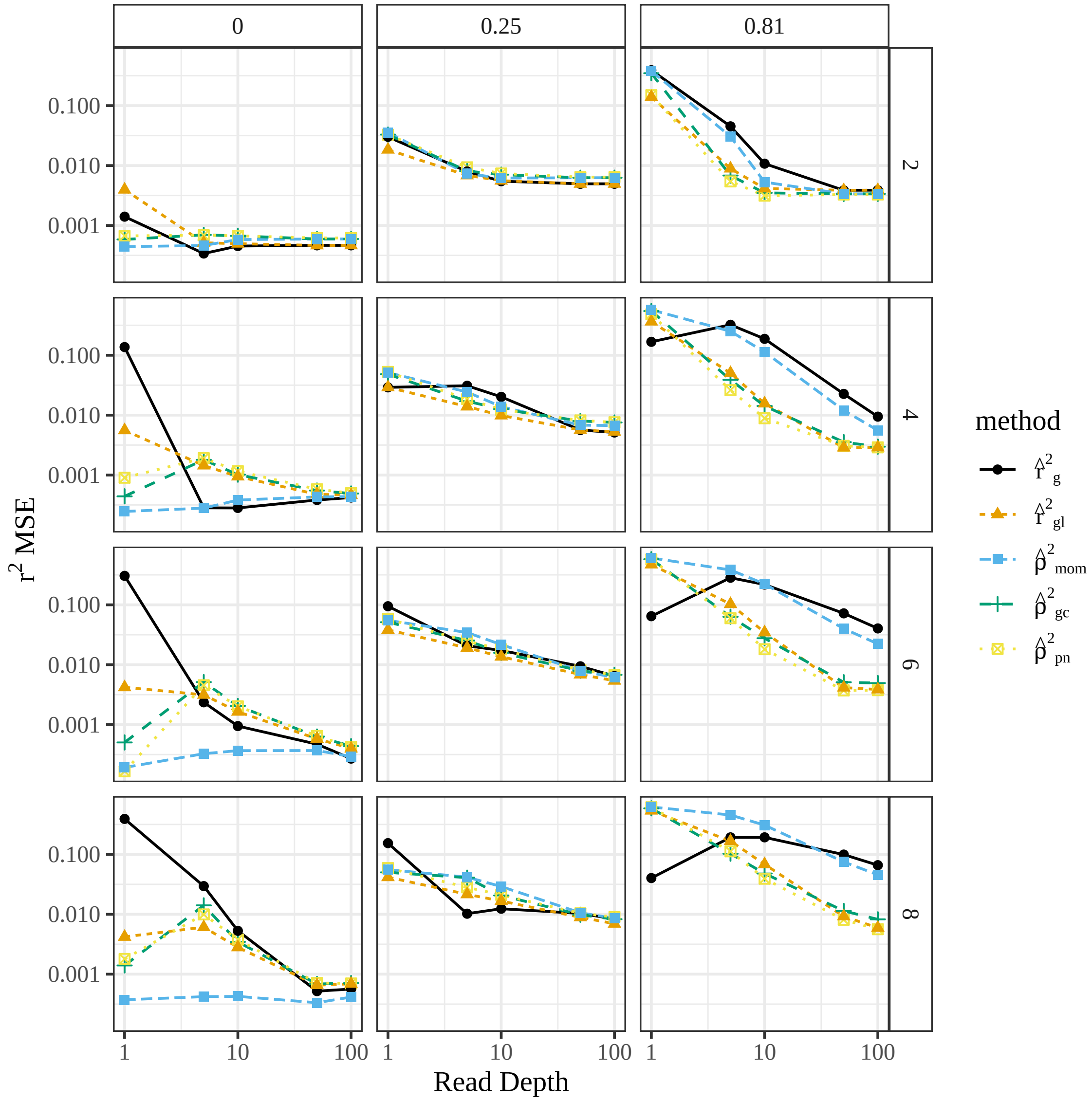

Figure 1: Mean-squared error of r^2 estimators (y-axis) stratified by read-depth (x-axis), estimation method (color), ploidy (row-facets), and true r^2 (column-facets) for the simulations from Section 3.1. Except in low-LD regimes, the MLE using genotype likelihoods has the smallest MSE. The moment-based estimator has a lower MSE in low-LD regimes because of its strong bias toward 0. Simulations were performed with $p_A = 0.5$ and $p_B = 0.5$.

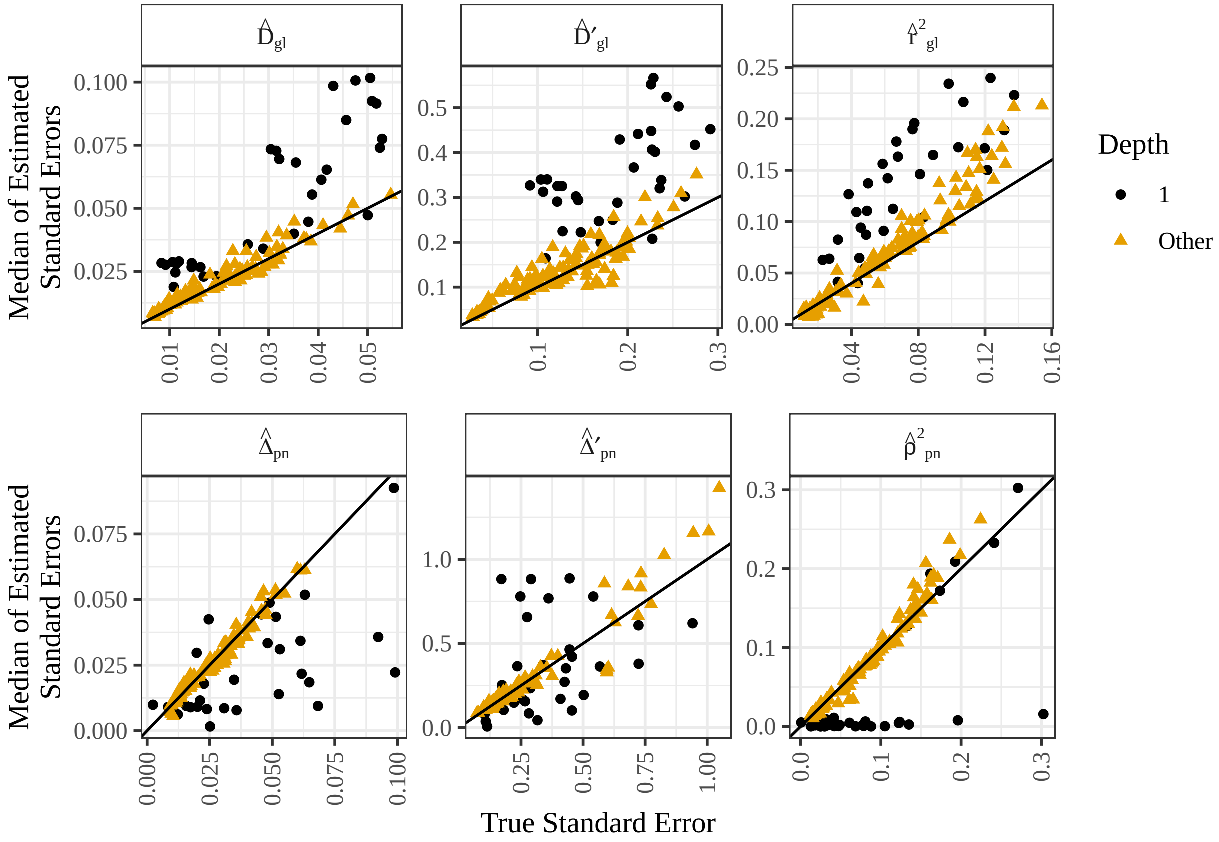

Figure 2: Standard errors (x-axis) of gametic (first row) and composite (second row) LD estimators when using genotype likelihoods from the simulations in Section 3.1. The y-axis contains the median of the estimated standard errors. Each point is a different simulation setting. The line is the $y = x$ line and points above that line indicate that the estimated standard errors are typically larger than the true standard errors. Standard errors are reasonably unbiased except when the read-depth is 1 (color and shape).

so should exhibit large levels of LD within each contig. Generally, the composite estimates of Δ using the proportional bivariate normal genotype class had the largest LD estimates within each contig (Figure 4 **(A)** and **(B)**). For LD estimation between contigs, the methods behaved similarly (Figure 4**(C)**). Heatmaps of LD estimates of r^2 or ρ^2 are presented in Figures S15–S19.

3.4 LD estimates using data from McAllister and Miller [2016]

In this section, we evaluate our LD estimators using the genotyping-by-sequencing data from McAllister and Miller [2016], downloaded from Dryad as a variant call format file [McAllister and Miller, 2017]. These data come from natural populations of *Andropogon gerardii*, where reads were mapped onto the *Sorghum bicolor* genome. *A. gerardii* contains two common cytotypes: hexaploid ($2n = 6x = 60$) and enneaploid ($2n = 9x = 90$). All results in this section use only hexaploid individuals. Unlike the data from Uitdewilligen et al. [2013], the individuals from McAllister and

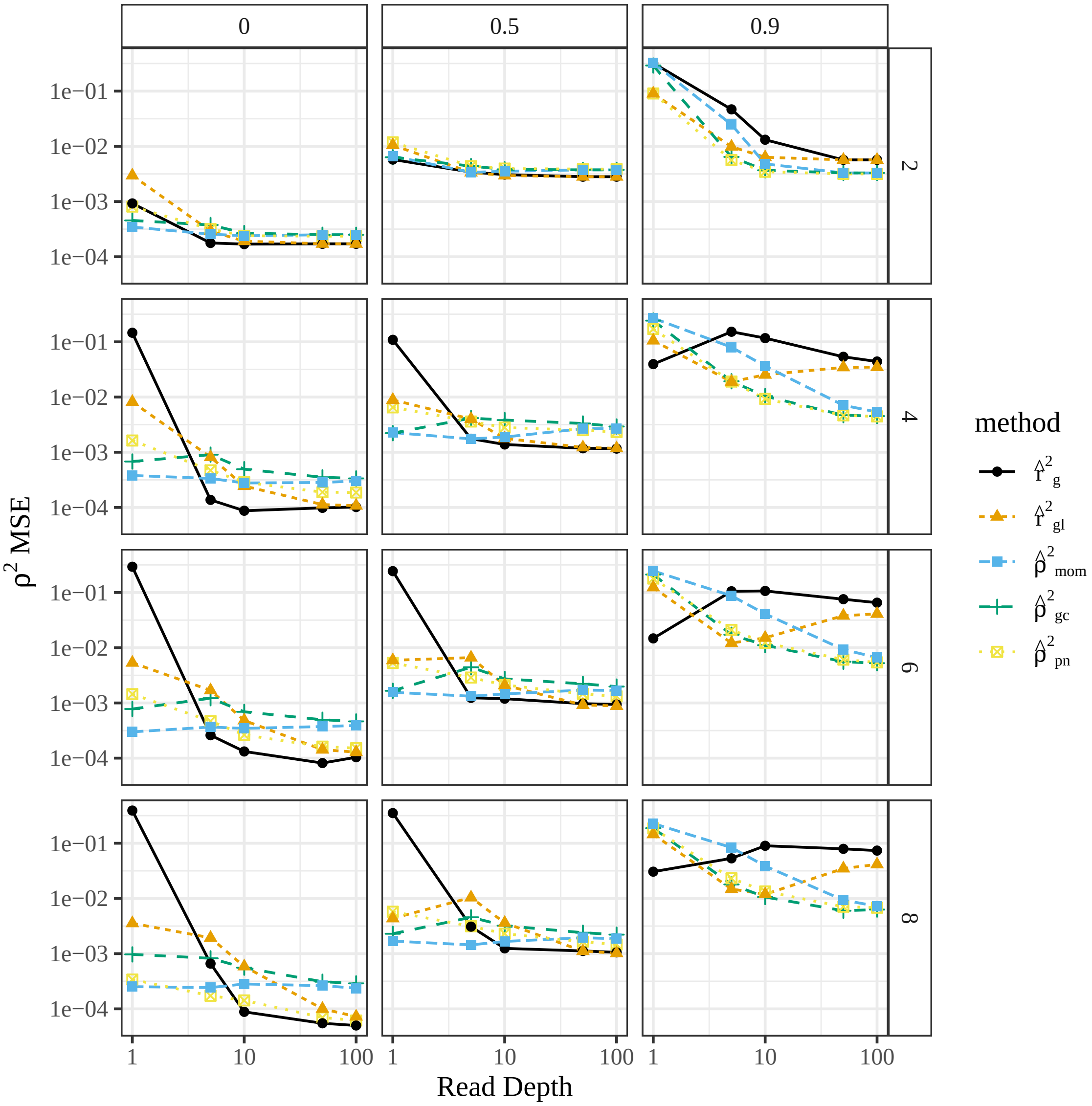

Figure 3: Average mean-squared error (y-axis) stratified by read-depth (x-axis), ploidy (row-facets), and association parameter of the proportional bivariate normal distribution (column-facets) for the simulations from Section 3.2.

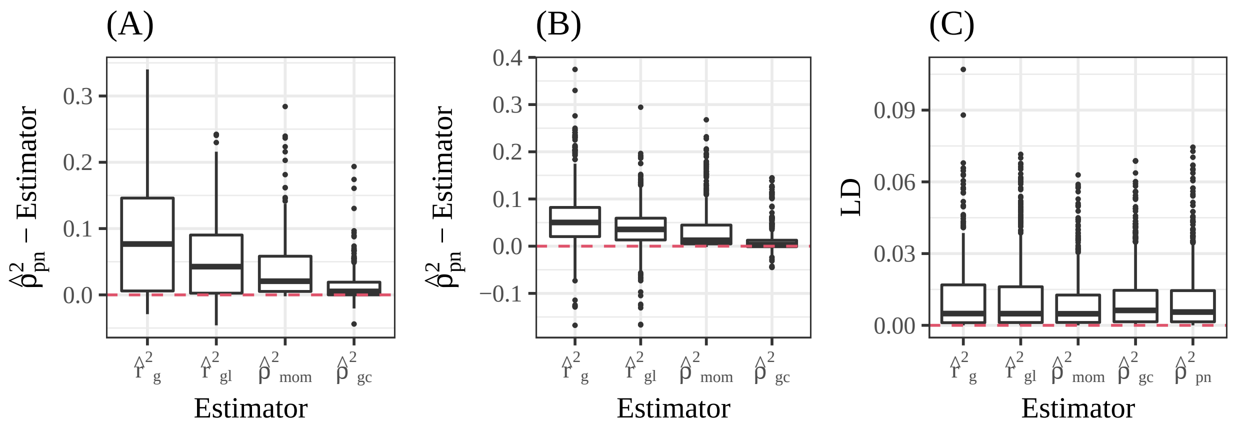

Figure 4: Squared Pearson correlation estimates subtracted from $\hat{\rho}^2_{pn}$ for SNPs on contig CONT0988 **(A)** and on contig CONT1561 **(B)**. A positive value (above the red dashed line) indicates that $\hat{\rho}^2_{pn}$ is larger. **(C)** Squared Pearson correlation estimates on different contigs (CONT0988 and CONT1561) on different super scaffolds where better LD estimates should be closer to 0 (red dashed line).

Miller [2016] were sequenced at relatively low depth (on the order of 10× versus 60×).

As in Section 3.3, we selected two arbitrary regions of the *Sorghum bicolor* genome, located on two different chromosomes, and extracted all biallelic SNPs from these two regions. SNPs were discarded if they contained an alternative allele frequency less than 0.1 or greater than 0.9. SNPs were also discarded if their average read-depth was less than 3. We then estimated genotypes using `updog` [Gerard et al., 2018] using a proportional normal prior class [Gerard and Ferrão, 2019]. The resulting genotype likelihoods were used to estimate pairwise LD between all SNPs.

Heat-maps of all pairwise LD estimates are provided in Figures S21–S25. It at first appears that there is a relatively rapid decay of pairwise LD. However, this is likely because the genomic regions span 100 kb, which is much larger than the regions explored in Section 3.3. The second dominant result is that estimating LD by maximum likelihood using posterior mode genotypes performed very poorly. The other methods performed comparably. Though, there was greater noise in the methods that estimate composite LD using genotype likelihoods. When the LD estimates are shrunk using the hierarchical shrinkage procedure of Stephens [2016] and Dey and Stephens [2018], the results appear to be very close (Figures S26–S29), indicating the signal-to-noise ratio was very similar for the different estimators. This shrinkage was performed on the Fisher-z transformation [Fisher, 1921] of the estimated Pearson correlation, whose distribution in simulations appears to be very well approximated by the normal distribution (Figure S10). The major difference between the shrunken LD heatmaps is that genotype likelihood methods result in a fewer number of non-zero LD estimates, but each are of higher magnitude.

4 Discussion

In this manuscript, we reviewed gametic measures of LD and then derived generalizations of Burrows' composite measures of LD to polyploids. For the composite generalizations of Lewontin's D', this involved deriving novel bounds on the covariance between genotypes. We provided a collection of methods to estimate both gametic LD and composite LD in the presence of genotype uncertainty by directly using genotype likelihoods. For composite LD, this involved developing a novel class of distributions over the genotypes. We validated our methods both in simulations and on real data.

In Section 2.3, though we were able to find closed-form bounds on Δ when conditioning on the genotype expectations, we resorted to the methods of linear programming to numerically find the bounds on Δ when conditioning on the marginal distributions. Under more general conditions, Whitt [1976] characterized the maximum and minimum correlation between two random variables given fixed marginals, corresponding to scenario (15)–(16). Specifically, given random variables X and Y with inverse cumulative distribution functions $F^{-1}(\cdot)$ and $G^{-1}(\cdot)$, Whitt [1976] found that

$$\mathrm{cor}(F^{-1}(U), G^{-1}(1-U)) \leq \mathrm{cor}(X, Y) \leq \mathrm{cor}(F^{-1}(U), G^{-1}(U)), \tag{36}$$

where U is a Uniform$(0, 1)$ random variable. Using (36), Leonov and Qaqish [2020] derived a purpose-built algorithm to find the bounds on the correlation given two random variables that follow general categorical distributions. This algorithm could be used to solve (15)–(16) and might be computationally faster. However, solving the linear program (15)–(16) is not the computational bottleneck we face when estimating LD. The time to solve (15)–(16) for an octoploid species is on the order of a millisecond, whereas the optimization procedures discussed in Section 2.4 take about half a second (Figures S11 and S14). We thus leave improved optimization of (15)–(16) as future work.

We briefly discussed covariance shrinkage in this manuscript. Covariance shrinkage has a long history, dating back at least to James and Stein [1961] and has shown great promise in improving covariance estimates in high dimensional regimes, including in genomic studies [Schäfer and Strimmer, 2005]. Yet, most LD estimators do not use any form of regularization or shrinkage (with the exception of Wen and Stephens [2010]). An off-the-shelf approach that we have used in this manuscript is "adaptive shrinkage" on the Fisher-z transformed correlation matrix [Stephens, 2016, Dey and Stephens, 2018] (Figures S26–S29). This shrinkage estimator is not purpose-built for LD estimation, and so there are many avenues for improvement (e.g., by accounting for known physical locations of SNPs). Research on LD shrinkage estimation is now more accessible with the accurate standard errors that we derived in this manuscript.

Data availability

All methods discussed in this manuscript are implemented in the `ldsep` package, available on the Comprehensive R Archive Network `https://cran.r-project.org/package=ldsep`. Scripts to reproduce the results of this research are available on GitHub `https://github.com/dcgerard/ld_simulations`.

Acknowledgments

We would like to thank Dr. Y. Samuel Wang, at the University of Chicago, for providing a simpler proof to Theorem S1. Our original approach with duality theory was much too cumbersome. We would also like to thank Dr. Luís Felipe Ventorim Ferrão, at the University of Florida, for his insightful comments on a draft of this manuscript.

Most analyses were performed using the R statistical language [R Core Team, 2020]. Many

open-source R packages were used for the analyses, including `corrplot` [Wei and Simko, 2017], `doParallel` [Microsoft and Weston, 2019], `foreach` [Microsoft and Weston, 2020], `ggplot2` [Wickham, 2016], `lpSolve` [Berkelaar et al., 2020], `Rcpp` [Eddelbuettel and François, 2011], `RcppArmadillo` [Eddelbuettel and Sanderson, 2014], `tidyverse` [Wickham et al., 2019], `VariantAnnotation` [Obenchain et al., 2014], and `vcfR` [Knaus and Grünwald, 2017]. We thank these package authors for their contributions.

Supplementary Material

S1 Derivation of Equation (8)

Under Hardy-Weinberg equilibrium, $(X_{iAB}, X_{iAb}, X_{iaB}, X_{iab})$ follows a multinomial distribution (7) with size parameter K and probability parameters $\boldsymbol{p} = (p_{AB}, p_{Ab}, p_{aB}, p_{AB})$. We will denote the multinomial probability mass function by Multinom$(X_{iAB}, X_{iAb}, X_{iaB}, X_{iab}|K, \boldsymbol{p})$. Letting $G_{iA} = X_{iAB} + X_{iAb}$ and $G_{iB} = X_{iAB} + X_{iaB}$, the change of variables results in

$$Pr(G_{iA}, G_{iB}|\boldsymbol{p}) = \sum_{\substack{X_{iAB}, X_{iAb}, X_{iaB}, X_{iab} \text{ s.t.} \\ G_{iA} = X_{iAB} + X_{iAb}, \\ G_{iB} = X_{iAB} + X_{iaB}, \text{ and} \\ X_{iAB} + X_{iAb} + X_{iaB} + X_{iab} = K}} \text{Multinom}(X_{iAB}, X_{iAb}, X_{iaB}, X_{iab}|K, \boldsymbol{p}). \tag{S1}$$

Noting that

$$\begin{aligned} X_{iAb} &= G_{iA} - X_{iAB} && \text{(S2)} \\ X_{iaB} &= G_{iB} - X_{iAB} && \text{(S3)} \\ X_{iab} &= K - X_{iAb} - X_{iaB} - X_{iAB} && \text{(S4)} \\ &= K - G_{iA} - G_{iB} + X_{iAB}, && \text{(S5)} \end{aligned}$$

and then relabeling $z = X_{iAB}$, (S1) becomes

$$Pr(G_{iA}, G_{iB}|\boldsymbol{p}) = \sum_{z} \text{Multinom}(z, G_{iA} - z, G_{iB} - z, K - G_{iA} - G_{iB} + z|K, \boldsymbol{p}). \tag{S6}$$

It remains to find the limits of the summation in (S6). Since each X lies between 0 and K we have

$$\begin{aligned} 0 \leq\, &z \leq K && \text{(S7)} \\ 0 \leq G_{iA} - z \leq K \Rightarrow \quad G_{iA} - K \leq\, &z \leq G_{iA} && \text{(S8)} \\ 0 \leq G_{iB} - z \leq K \Rightarrow \quad G_{iB} - K \leq\, &z \leq G_{iB} && \text{(S9)} \\ 0 \leq K - G_{iA} - G_{iB} + z \leq K \Rightarrow \quad G_{iA} + G_{iB} - K \leq\, &z \leq G_{iA} + G_{iB}. && \text{(S10)} \end{aligned}$$

Taking the intersection of bounds (S7)-(S10), we obtain

$$\max(0, G_{iA} + G_{iB} - K) \leq z \leq \min(G_{iA}, G_{iB}). \tag{S11}$$

Placing the bounds of (S11) in (S6) and substituting in the multinomial probability mass function yields (8).

S2 EM algorithm to estimate haplotype frequencies in autopolyploids under HWE

The following derivation generalizes from diploids to polyploids the EM algorithm described in Li [2011] and later used in Fox et al. [2019]. However, unlike the algorithm in Li [2011] that uses the haplotypes as the latent variable, we use the *number* of each haplotype as the latent variable. This simplifies the EM algorithm derivation for polyploids and significantly reduces the number of summands each iteration from 4^K to $\binom{K+3}{K}$. For example, for an octoploid species like strawberry ($K = 8$), the number of summands reduces from 65536 each iteration to 165 each iteration.

For individual i, let A_{i1} be the number of "00" haplotypes, A_{i2} be the number of "01" haplotypes, A_{i3} be the number of "10" haplotypes, and A_{i4} be the number of "11" haplotypes. Let $\boldsymbol{A}_i = (A_{i1}, A_{i2}, A_{i3}, A_{i4})$. Then $\boldsymbol{A} \sim \text{Multinom}(K, \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, p_2, p_3, p_4)$ are the haplotype frequencies. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ be the data at loci 1 and 2 for individual i. We assume the user provides $p(y_{i1}|g_1)$ and $p(y_{i2}|g_2)$, the genotype likelihoods given genotypes g_1 and g_2 for individual i at loci 1 and 2. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$. Then the complete log-likelihood is:

$$p(\boldsymbol{y}, \boldsymbol{A}|\boldsymbol{p}) = \sum_{i=1}^{n} \sum_{\substack{\boldsymbol{a}\text{ s.t.}\\ a_1+a_2+a_3+a_4=K}} I(\boldsymbol{A}_i = \boldsymbol{a}) \log\left[p(y_{i1}|a_3+a_4)p(y_{i2}|a_2+a_4)Pr(\boldsymbol{a}|\boldsymbol{p})\right]. \tag{S12}$$

The E-step involves calculating the following posterior probabilities:

$$\begin{align}
w_{i\boldsymbol{a}} &:= Pr(\boldsymbol{a}|\boldsymbol{y}_i, \boldsymbol{p}^{(old)}) \tag{S13}\\
&= \frac{p(y_{i1}|a_3+a_4)p(y_{i2}|a_2+a_4)Pr(\boldsymbol{a}|\boldsymbol{p}^{(old)})}{\sum_{\substack{\boldsymbol{a}\text{ s.t.}\\ a_1+a_2+a_3+a_4=K}} p(y_{i1}|a_3+a_4)p(y_{i2}|a_2+a_4)Pr(\boldsymbol{a}|\boldsymbol{p}^{(old)})} \tag{S14}\\
&= \frac{p(y_{i1}|a_3+a_4)p(y_{i2}|a_2+a_4)\,\text{Multinom}(\boldsymbol{a}|K, \boldsymbol{p}^{(old)})}{\sum_{\substack{\boldsymbol{a}\text{ s.t.}\\ a_1+a_2+a_3+a_4=K}} p(y_{i1}|a_3+a_4)p(y_{i2}|a_2+a_4)\,\text{Multinom}(\boldsymbol{a}|K, \boldsymbol{p}^{(old)})} \tag{S15}\\
& \tag{S16}
\end{align}$$

The M-step thus involves maximizing the following objective function:

$$\begin{align}
&\sum_{i=1}^{n} \sum_{\substack{\boldsymbol{a}\text{ s.t.}\\ a_1+a_2+a_3+a_4=K}} w_{i\boldsymbol{a}} \log\left[p(y_{i1}|a_3+a_4)p(y_{i2}|a_2+a_4)Pr(\boldsymbol{a}|\boldsymbol{p})\right] \tag{S17}\\
&= \sum_{i=1}^{n} \sum_{\substack{\boldsymbol{a}\text{ s.t.}\\ a_1+a_2+a_3+a_4=K}} w_{i\boldsymbol{a}} \log\left[Pr(\boldsymbol{a}|\boldsymbol{p})\right] + C, \tag{S18}
\end{align}$$

where C is some constant with respect to $\boldsymbol{p}$. Using Lagrange multipliers, we find that the update is

$$\begin{align}
\eta_\ell &:= \sum_{i=1}^{n} \sum_{\substack{\boldsymbol{a}\text{ s.t.}\\ a_\ell>0 \text{ and}\\ a_1+a_2+a_3+a_4=K}} a_\ell w_{i\boldsymbol{a}} \tag{S19}\\
p_\ell^{(new)} &= \frac{\eta_\ell}{\sum_\ell \eta_\ell} \tag{S20}
\end{align}$$

If one uses a Dirichlet($\boldsymbol{\alpha}$) prior on the haplotype proportions, then (S19) is modified to

$$\eta_\ell := \sum_{i=1}^{n} \sum_{\substack{\boldsymbol{a}\text{ s.t.}\\ a_\ell>0\text{ and}\\ a_1+a_2+a_3+a_4=K}} a_\ell w_{i\boldsymbol{a}} + (\alpha_\ell - 1). \tag{S21}$$

S3 Moments of genotypes

In this section, we derive the moments for the genotypes at two loci under the assumption of HWE. The calculations are simple, but demonstrate that composite measures of LD are equal to gametic measures of LD when HWE is fulfilled. Let

$$(X_1, X_2, X_3, X_4) \sim \text{Multinom}(K, p_1, p_2, p_3, p_4), \tag{S22}$$

where X_1 are the counts of haplotype 00, X_2 are the counts of haplotype 10, X_3 are the counts of haplotype 01, and X_4 are the counts of haplotype 11. Then we have the following moments of multinomial counts:

$$E[X_i] = Kp_i \tag{S23}$$
$$\text{var}(X_i) = Kp_i(1-p_i) \tag{S24}$$
$$\text{cov}(X_i, X_j) = -Kp_ip_j \text{ when } i \neq j. \tag{S25}$$

Let

$$G_1 = X_2 + X_4 \tag{S26}$$
$$G_2 = X_3 + X_4. \tag{S27}$$

Then

$$p_A = (p_2 + p_4) \tag{S28}$$
$$p_B = (p_3 + p_4) \tag{S29}$$
$$E[G_1] = K(p_2 + p_4) = Kp_A \tag{S30}$$
$$E[G_2] = K(p_3 + p_4) = Kp_B \tag{S31}$$
$$\text{var}(G_1) = \text{var}(X_2) + \text{var}(X_4) + 2\,\text{cov}(X_2, X_4) \tag{S32}$$
$$= Kp_2(1-p_2) + Kp_4(1-p_4) - 2Kp_2p_4 \tag{S33}$$
$$= K(p_2+p_4)(1-(p_2+p_4)) \tag{S34}$$
$$= Kp_A(1-p_A) \tag{S35}$$
$$\text{var}(G_2) = K(p_3+p_4)(1-(p_3+p_4)) \tag{S36}$$
$$= Kp_B(1-p_B) \tag{S37}$$

We will now derive the covariance between G_1 and G_2:

$$\begin{aligned}
K\Delta &= \mathrm{cov}(G_1, G_2) && \text{(S38)}\\
&= E[G_1G_2] - E[G_1]E[G_2] && \text{(S39)}\\
&= E[X_2X_3] + E[X_2X_4] + E[X_3X_4] + E[X_4^2] - E[G_1]E[G_2] && \text{(S40)}\\
&= K(K-1)(p_2p_3 + p_2p_4 + p_3p_4) + Kp_4(1-p_4) + K^2p_4^2 - K^2(p_2+p_4)(p_3+p_4) && \text{(S41)}\\
&= K[p_4 - (p_2+p_4)(p_3+p_4)] && \text{(S42)}\\
&= KD && \text{(S43)}
\end{aligned}$$

The correlation between G_1 and G_2 is

$$\begin{aligned}
\rho &= \mathrm{cor}(G_1, G_2) && \text{(S44)}\\
&= \frac{\mathrm{cov}(G_1, G_2)}{\sqrt{\mathrm{var}(G_1)\,\mathrm{var}(G_2)}} && \text{(S45)}\\
&= \frac{KD}{\sqrt{K^2p_A(1-p_A)p_B(1-p_B)}} && \text{(S46)}\\
&= \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} && \text{(S47)}\\
&= r. && \text{(S48)}
\end{aligned}$$

Under HWE, Δ_e defined in (20) is equal to

$$\begin{aligned}
\Delta_e &= \begin{cases} \min\{p_Ap_B, (1-p_A)(1-p_B)\} & \text{if } \Delta < 0,\\ \min\{p_A(1-p_B), (1-p_A)p_B\} & \text{if } \Delta > 0 \end{cases} && \text{(S49)}\\
&= D_{max}. && \text{(S50)}
\end{aligned}$$

Thus,

$$\begin{aligned}
\Delta_a' &:= \frac{\frac{1}{K}\,\mathrm{cov}(G_1, G_2)}{\Delta_e} && \text{(S51)}\\
&= D/D_{max} && \text{(S52)}\\
&= D'. && \text{(S53)}
\end{aligned}$$

S4 Burrows' Δ and genotype covariance

Let G_A and G_B be the genotypes of a diploid at loci 1 and 2. Then the covariance between G_A and G_B is

$$\begin{aligned}
\frac{1}{2}\operatorname{cov}(G_A, G_B) &= \frac{1}{2}(E[G_A G_B] - E[G_A]E[G_B]) && \text{(S54)}\\
&= \frac{1}{2}\sum_{g_A=0}^{2}\sum_{g_B=0}^{2} g_A g_B q_{g_A g_B} - 2p_A p_B && \text{(S55)}\\
&= 2q_{22} + q_{21} + q_{12} + \frac{1}{2}q_{11} - 2p_A p_B && \text{(S56)}\\
&= (12). && \text{(S57)}
\end{aligned}$$

S5 Closed-form bounds for Δ conditional on allele frequencies

Theorem S1. *Suppose X and Y are random variables such that $0 \le X, Y \le K$ almost surely. Furthermore, suppose $E[X] = Kp_A$ and $E[Y] = Kp_B$. Then*

$$E[XY] \le K^2 \min(p_A, p_B). \tag{S58}$$

Proof. The following proof was provided by Dr. Y. Samuel Wang, University of Chicago, via a personal correspondence.

$$\begin{aligned}
E[XY] &= E[|XY|] && \text{(S59)}\\
&\le \min_{\substack{p,q\ge 1\\ 1/p+1/q=1}} E[|X|^p]^{1/p} E[|Y|^q]^{1/q} && \text{(S60)}\\
&\le \min_{\substack{p,q\ge 1\\ 1/p+1/q=1}} \max_{\substack{\tilde{X},\tilde{Y}\\ E[\tilde{X}]=Kp_A, E[\tilde{Y}]=Kp_B\\ 0\le\tilde{X},\tilde{Y}\le K}} E[|\tilde{X}|^p]^{1/p} E[|\tilde{Y}|^q]^{1/q} && \text{(S61)}\\
&= \min_{\substack{p,q\ge 1\\ 1/p+1/q=1}} [K^p p_A]^{1/p}[K^q p_B]^{1/q} && \text{(S62)}\\
&= \min_{\substack{p,q\ge 1\\ 1/p+1/q=1}} K^2 p_A^{1/p} p_B^{1/q} && \text{(S63)}\\
&\le K^2 \min(p_A, p_B). && \text{(S64)}
\end{aligned}$$

Equation (S60) follows by Hölder's inequality. Equation (S62) holds because for any $p \ge 1$, the maximum over $\tilde{X}$ is achieved when $Pr(\tilde{X} = K) = p_A$ and $Pr(\tilde{X} = 0) = 1 - p_A$. Similarly, the maximum over $\tilde{Y}$ is achieved when $Pr(\tilde{Y} = K) = p_B$ and $Pr(\tilde{Y} = 0) = 1 - p_B$. Equation (S64) results by letting $p = \infty$ and $q = 1$ when $p_B \le p_A$, and letting $p = 1$ and $q = \infty$ when $p_B \ge p_A$. □

Theorem S2. *Let X and Y be two random variables, each with support on $\{0, 1, \dots, K\}$. Furthermore, suppose that $E[X] = Kp_A$ and $E[Y] = Kp_B$. Then*

$$-K^2 \min\{p_A p_B, (1-p_A)(1-p_B)\} \le \operatorname{cov}(X, Y) \le K^2 \min\{p_A(1-p_B), (1-p_A)p_B\}, \tag{S65}$$

and these bounds are tight.

Proof.

$$\begin{aligned} \operatorname{cov}(X,Y) &= E[XY] - E[X]E[Y] && \text{(S66)}\\ &\leq K^2 \min(p_A, p_B) - E[X]E[Y] \text{ (Theorem S1)} && \text{(S67)}\\ &= K^2 \min(p_A, p_B) - K^2 p_A p_B && \text{(S68)}\\ &= K^2 \min\{p_A(1-p_B), (1-p_A)p_B\}. && \text{(S69)} \end{aligned}$$

Bound (S69) is tight since it is achieved when $Pr(X = K) = p_A$, $Pr(X = 0) = 1 - p_A$, $Pr(Y = K) = p_B$, and $Pr(Y = 0) = 1 - p_B$.

To prove the lower bound, first set $U := K - Y$. Then $E[U] = K(1 - p_B)$ and we have

$$\operatorname{cov}(X,U) \leq K^2 \min\{p_A p_B, (1-p_A)(1-p_B)\}. \tag{S70}$$

But $\operatorname{cov}(X,U) = -\operatorname{cov}(X,Y)$, and thus

$$\operatorname{cov}(X,Y) \geq -K^2 \min\{p_A p_B, (1-p_A)(1-p_B)\}. \tag{S71}$$

□

S6 MLE standard errors when using the general categorical genotype distribution

The following are the derivatives relating to the log of (31) necessary to derive asymptotic standard errors of the MLEs when using the general categorical genotype distribution to estimate composite measures of LD. The Hessian of the log of (31) can be calculated in closed form:

$$\frac{dL}{dq_{ij}dq_{km}} = -\sum_{\ell=0}^{n} \frac{Pr(D_{\ell 1}|i)Pr(D_{\ell 2}|j)Pr(D_{\ell 1}|k)Pr(D_{\ell 2}|m)}{\left(\sum_{i=0}^{K}\sum_{j=0}^{K} Pr(D_{\ell 1}|i)Pr(D_{\ell 2}|j)q_{ij}\right)^2} \tag{S72}$$

For Δ, we have

$$\Delta = \frac{1}{K}\sum_{i=0}^{K}\sum_{j=0}^{K} ijq_{ij} - \frac{1}{K}\left(\sum_{i=0}^{K} i \sum_{j=0}^{K} q_{ij}\right)\left(\sum_{j=0}^{K} j \sum_{i=0}^{K} q_{ij}\right) \tag{S73}$$

$$\frac{d\Delta}{dq_{\ell m}} = \frac{\ell m}{K} - \frac{\ell}{K}\left(\sum_{j=0}^{K} j \sum_{i=0}^{K} q_{ij}\right) - \frac{m}{K}\left(\sum_{i=0}^{K} i \sum_{j=0}^{K} q_{ij}\right), \tag{S74}$$

For ρ^2, we have

$$\rho^2 = \frac{K^2\Delta^2}{\operatorname{var}(G_A)\operatorname{var}(G_B)} \tag{S75}$$

$$\operatorname{var}(G_A) = \sum_{i=0}^{K} i^2 \sum_{j=0}^{K} q_{ij} - \left(\sum_{i=0}^{K} i \sum_{j=0}^{K} q_{ij}\right)^2 \tag{S76}$$

$$\operatorname{var}(G_B) = \sum_{j=0}^{K} j^2 \sum_{i=0}^{K} q_{ij} - \left(\sum_{j=0}^{K} j \sum_{i=0}^{K} q_{ij}\right)^2 \tag{S77}$$

$$\frac{d\operatorname{var}(G_A)}{dq_{\ell m}} = \ell^2 - 2\ell\left(\sum_{i=0}^{K} i \sum_{j=0}^{K} q_{ij}\right) \tag{S78}$$

$$\frac{d\operatorname{var}(G_B)}{dq_{\ell m}} = m^2 - 2m\left(\sum_{j=0}^{K} j \sum_{i=0}^{K} q_{ij}\right) \tag{S79}$$

$$\frac{d\rho^2}{dq_{\ell m}} = \frac{2K^2\Delta\frac{d\Delta}{dq_{\ell m}}}{\operatorname{var}(G_A)\operatorname{var}(G_B)} - \frac{K^2\Delta^2\frac{d\operatorname{var}(G_A)}{dq_{\ell m}}}{\operatorname{var}(G_A)^2\operatorname{var}(G_B)} - \frac{K^2\Delta^2\frac{d\operatorname{var}(G_B)}{dq_{\ell m}}}{\operatorname{var}(G_A)\operatorname{var}(G_B)^1} \tag{S80}$$

For Δ', we have

$$E[G_A] = \sum_{i=0}^{K} i \sum_{j=0}^{K} q_{ij} \tag{S81}$$

$$E[G_B] = \sum_{j=0}^{K} j \sum_{i=0}^{K} q_{ij} \tag{S82}$$

$$\Delta_e := \begin{cases} \min\{E[G_A]E[G_B], (K - E[G_A])(K - E[G_B])\}/K^2 & \text{if } \Delta < 0, \\ \min\{E[G_A](K - E[G_B]), (K - E[G_A])E[G_B]\}/K^2 & \text{if } \Delta > 0. \end{cases} \tag{S83}$$

$$\Delta' = \Delta/\Delta_e \tag{S84}$$

$$\frac{d\Delta'}{dq_{\ell m}} = \frac{\frac{d\Delta}{dq_{\ell m}}}{\Delta_e} - \frac{\Delta\frac{d\Delta_e}{dq_{\ell m}}}{\Delta_e^2} \tag{S85}$$

$$\frac{d\Delta_e}{dq_{\ell m}} = \begin{cases} \frac{\ell E[G_B] + mE[G_A]}{K^2} & \text{if } \Delta < 0 \\ & \text{and } E[G_A]E[G_B] < (K - E[G_A])(K - E[G_B]) \\ \frac{-\ell(K - E[G_B]) - m(K - E[G_A])}{K^2} & \text{if } \Delta < 0 \\ & \text{and } E[G_A]E[G_B] > (K - E[G_A])(K - E[G_B]) \\ \frac{\ell(K - E[G_B]) - mE[G_A]}{K^2} & \text{if } \Delta > 0 \\ & \text{and } E[G_A](K - E[G_B]) < (K - E[G_A])E[G_B] \\ \frac{-\ell E[G_B] + m(K - E[G_A])}{K^2} & \text{if } \Delta > 0 \\ & \text{and } E[G_A](K - E[G_B]) > (K - E[G_A])E[G_B] \end{cases} \tag{S86}$$

We will now provide an example of how to obtain standard errors for the MLEs. Let $\boldsymbol{q} =$

$(q_{00}, q_{01}, \ldots, q_{ij}, \ldots, q_{KK})$, let $\hat{\boldsymbol{q}}$ be the MLEs of $\boldsymbol{q}$, and let $\boldsymbol{H}$ be the Hessian of the log-likelihood with elements (S72). Then standard maximum likelihood theory states that $\boldsymbol{H}^{-1/2}(\hat{\boldsymbol{q}} - \boldsymbol{q}) \rightarrow N(0, \boldsymbol{I})$. Since the covariance and correlation of genotypes are functions of the q_{ij}'s, we can use the δ-method to obtain the limiting variances of the covariance and correlation of the genotypes. For example, if we set $\boldsymbol{g}$ to contain the elements of (S74), then the asymptotic variance we use for $\hat{\Delta}_{gc}$ is $-\boldsymbol{g}^{\intercal}\boldsymbol{H}^{-1}\boldsymbol{g}$.

The asymptotic standard errors are not valid at points for which the gradient does not exist, which for Δ' occur when $\Delta > 0$ and $E[G_A] = E[G_B]$, when $\Delta < 0$ and $E[G_A] + E[G_B] = K$, or when $\Delta = 0$. These situations occur with Lebesgue measure 0, and so should not invalidate the standard errors.

S7 Standard errors of moment-based estimators

The results in this section can be derived directly from well-known results in the literature [Example 6.6.4 Lehmann and Casella, 1998, e.g.]. These results hold only for multivariate normal random variables, which is not applicable when estimating LD. However, we found that for most estimators the approximations are decent (Figure S9). Improved asymptotic standard errors could be implemented by using the techniques described in Chapter 8 of Ferguson [2002].

Let $\hat{\rho}$ be the sample correlation between genotypes, let $\hat{z} = \text{atanh}(\hat{\rho})$, let $\hat{\Delta}$ be the sample covariance between genotypes divided by K, let $\hat{\Delta}'$ be the sample estimator of Δ', let σ_1^2 be the variance of genotypes at locus 1, let σ_2^2 be the variance of genotypes at locus 2, let μ_1 be the mean genotype at locus 1, and let μ_2 be the mean genotype at locus 2. Then

$$\sqrt{n}(\hat{\rho}_{mom} - \rho) \rightarrow N\left(0, (1-\rho^2)^2\right) \tag{S87}$$

$$\sqrt{n}(\hat{\rho}^2_{mom} - \rho^2) \rightarrow N\left(0, 4\rho^2(1-\rho^2)^2\right) \tag{S88}$$

$$\sqrt{n}(\hat{z}_{mom} - z) \rightarrow N(0, 1) \tag{S89}$$

$$\sqrt{n}(\hat{\Delta}_{mom} - \Delta) \rightarrow N\left(0, \sigma_1^2\sigma_2^2/K^2 + \Delta^2\right) \tag{S90}$$

Equation (S88) follows from the δ-method using (S87). Equation (S90) follows because $K\hat{\Delta}_{mom}$ is the sample covariance of dosages. Equation (S89) is well known [Fisher, 1921, Hotelling, 1953].

For the composite measure Δ' (19), since

$$\hat{\Delta}_e \rightarrow \Delta_e := \begin{cases} \min(\mu_1\mu_2, (K-\mu_1)(K-\mu_2))/K^2 & \text{if } \Delta < 0, \\ \min((K-\mu_1)\mu_2, \mu_1(K-\mu_2))/K^2 & \text{if } \Delta > 0, \end{cases} \tag{S91}$$

we have by Slutsky's theorem that

$$\sqrt{n}(\hat{\Delta}'_{mom} - \Delta') \rightarrow N\left(0, \frac{\sigma_1^2\sigma_2^2/K^2 + \Delta^2}{\Delta_e^2}\right). \tag{S92}$$

S8 Supplementary figures

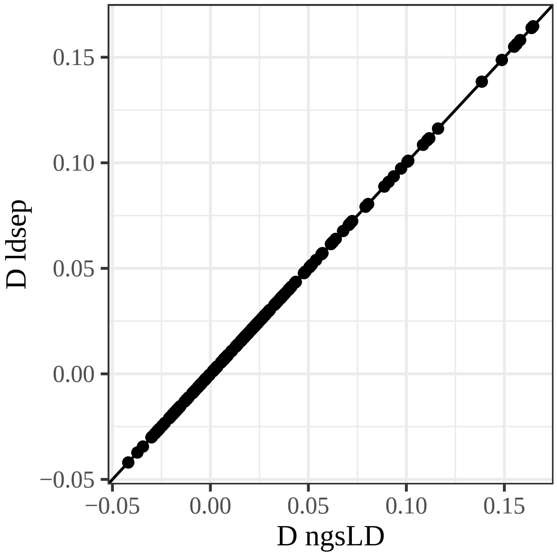

Figure S1: Maximum likelihood estimates of D between 20 simulated loci of 100 diploid individuals in HWE as calculated by `ngsLD` (x-axis) and `ldsep` (y-axis). The line is the $y = x$ line. The estimates are identical.

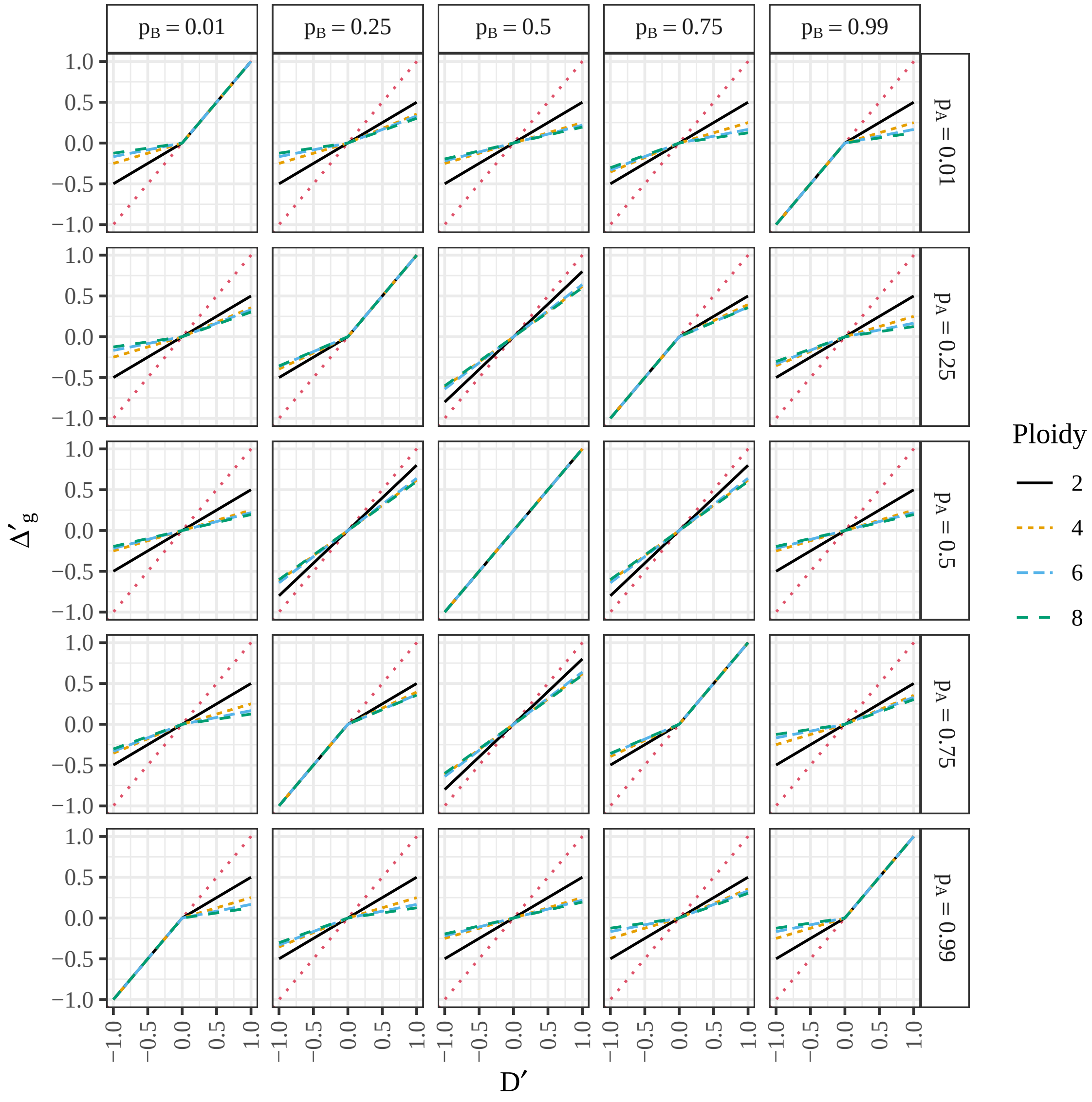

Figure S2: Relationship between D' (4) (x-axis) and Δ'_g (y-axis) (17) for populations with different ploidy (color) under HWE. Row-facets index the allele frequency at the first locus and column-facets index the allele frequency at the second locus. The red dotted line is the $y = x$ line. The functional relationship appears to be piecewise linear with a change of slope at 0.

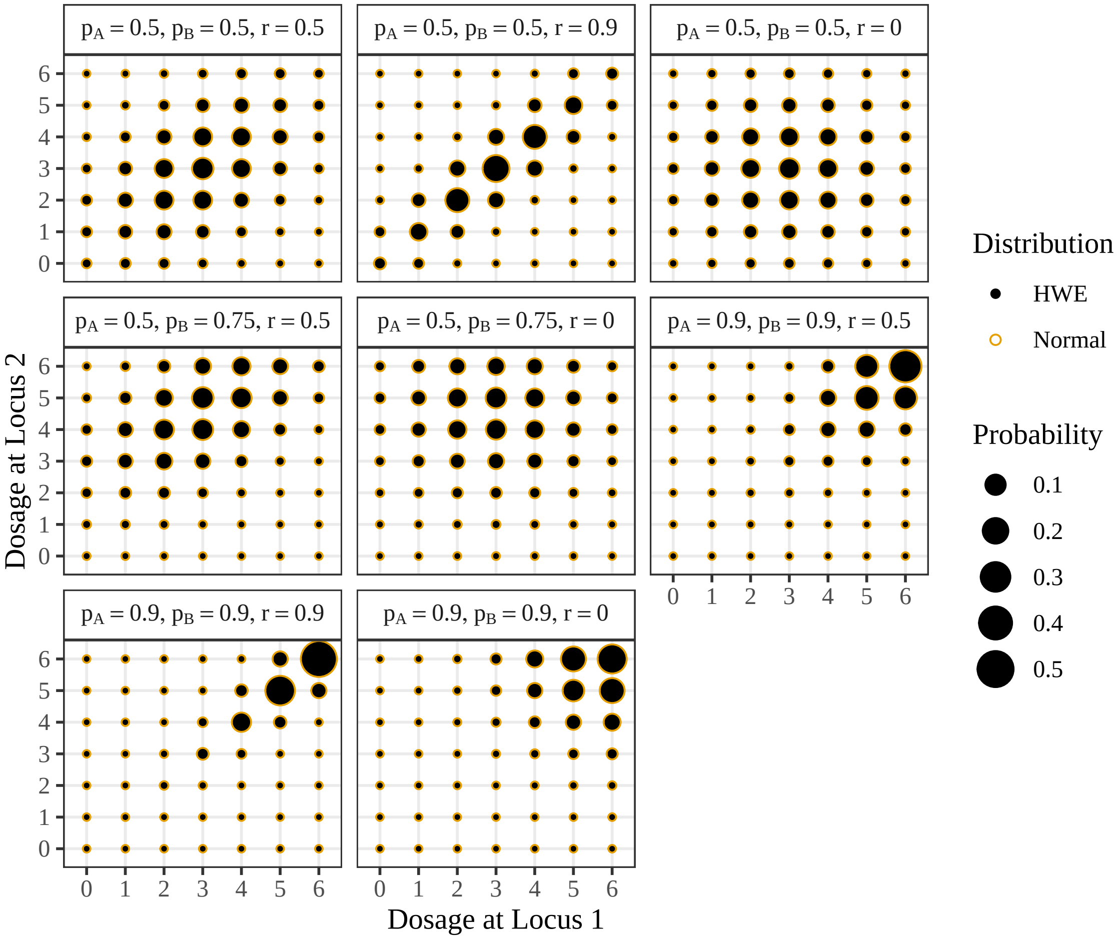

Figure S3: Joint probability distribution of two dosages. Probabilities are denoted by size. The target probability distribution is denoted by solid black circles, the closest (in Kullback-Leibler divergence) probability distribution to the target distribution among the class of proportional bivariate normal distributions is denoted by hollow orange circles. Each facet represents one of the settings of p_A, p_B, and r used in the simulation study in Section 3.1.

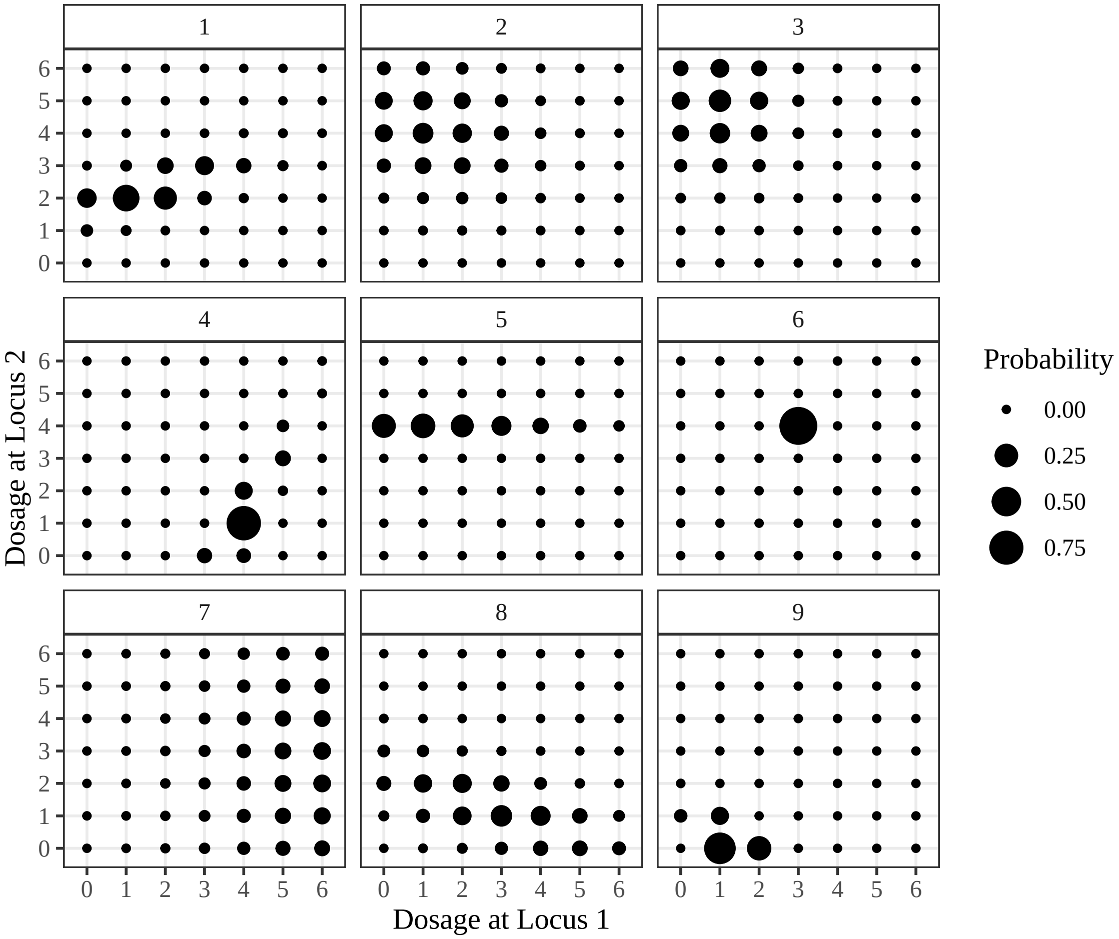

Figure S4: Joint probability distribution of two dosages. Probabilities are denoted by size. These probability distributions were chosen randomly among the class of proportional bivariate normal distributions. The proportional bivariate normal distribution can take on a variety of shapes beyond those when assuming HWE.

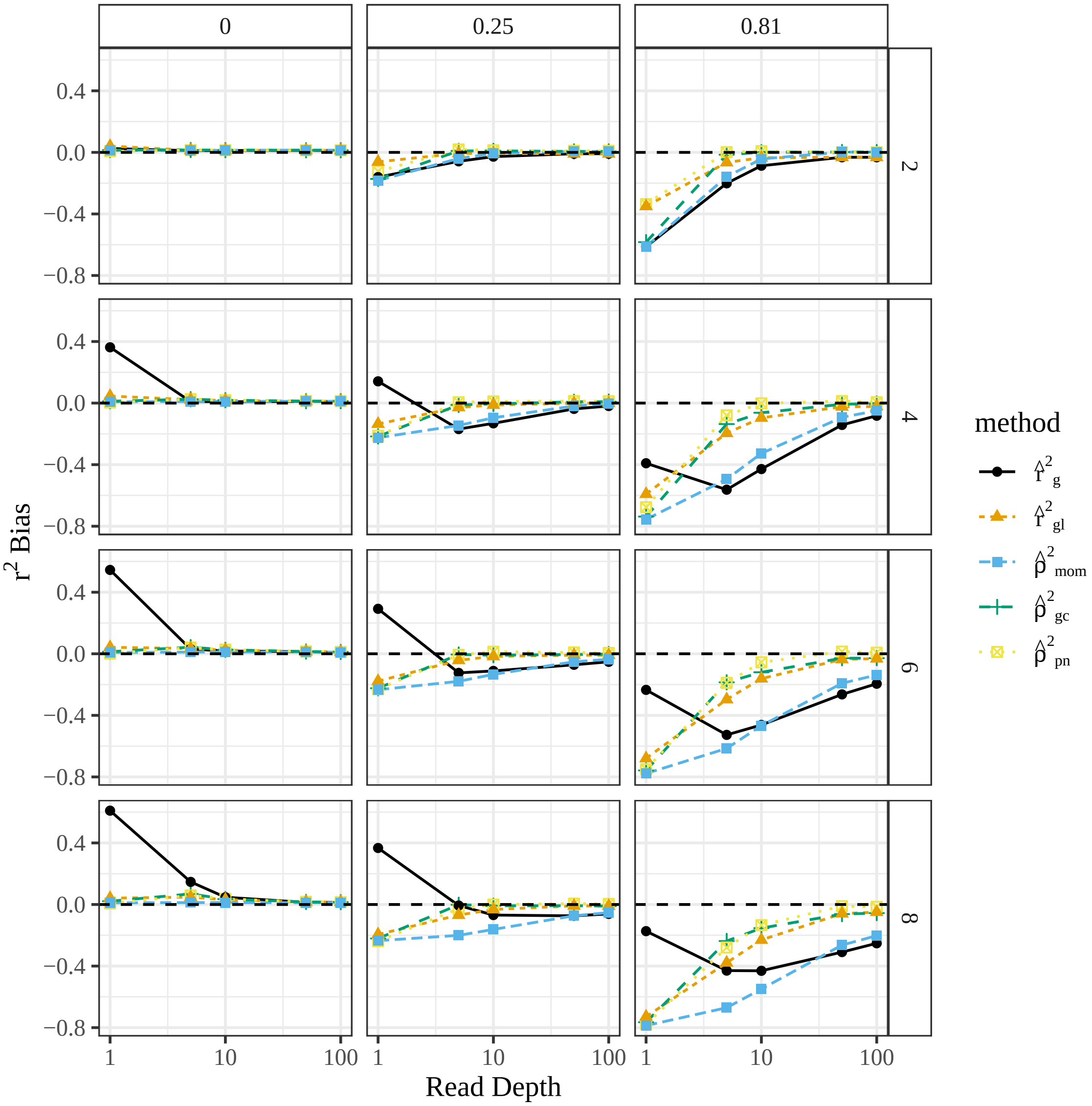

Figure S5: Bias of estimates of r^2 (y-axis) stratified by read-depth (x-axis), estimation method (color), ploidy (row-facets) and true r^2 (column-facets). The moment-based estimator has a strong bias toward zero. The MLE using genotype likelihoods is the least-biased. Simulations were performed with $p_A = 0.5$ and $p_B = 0.5$.

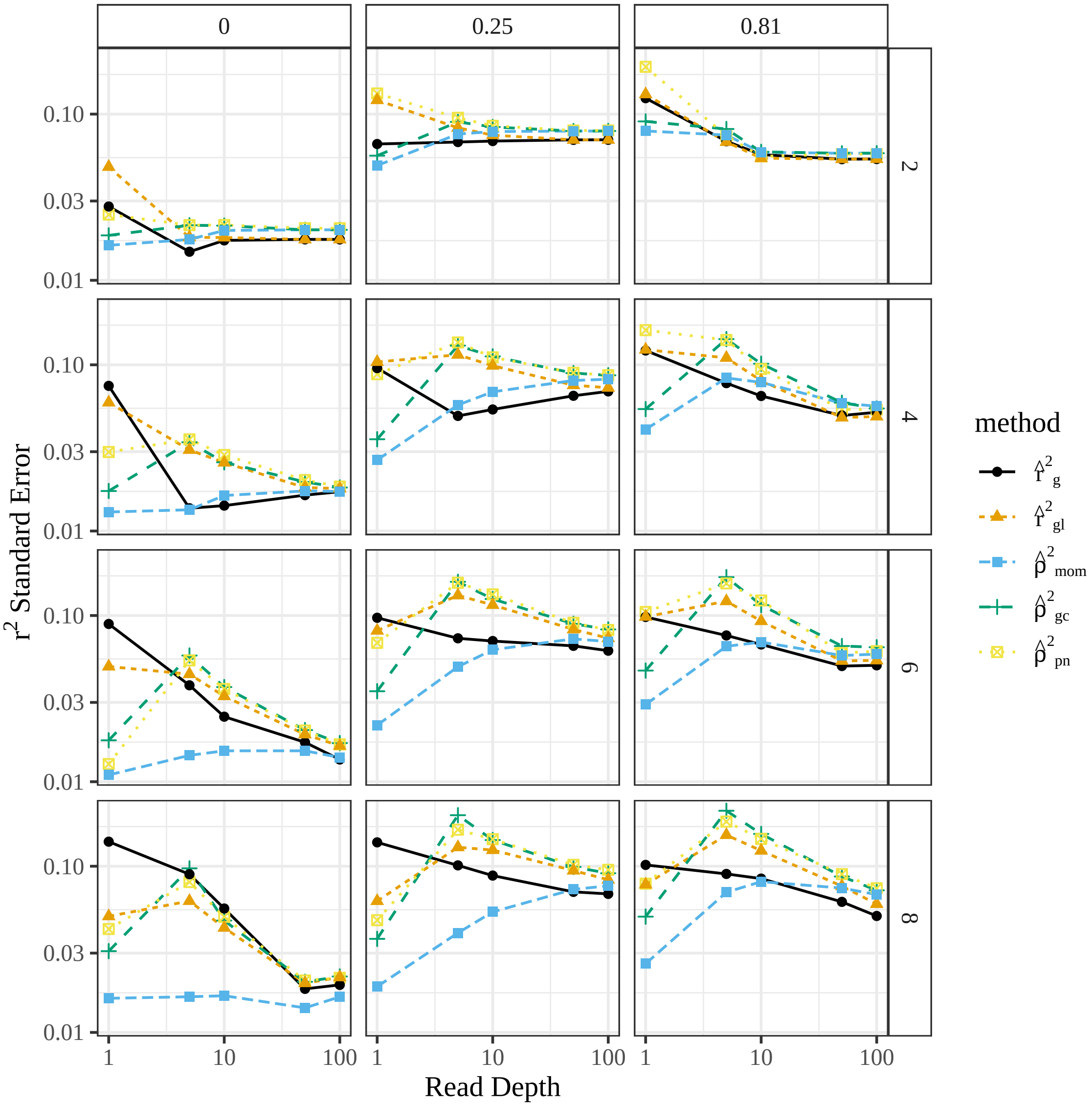

Figure S6: Standard error of r^2 estimators (y-axis) stratified by read-depth (x-axis), estimation method (color), ploidy (row-facets) and true r^2 (column-facets). Methods that use genotype likelihoods have higher standard errors. Simulations were performed with $p_A = 0.5$ and $p_B = 0.5$.

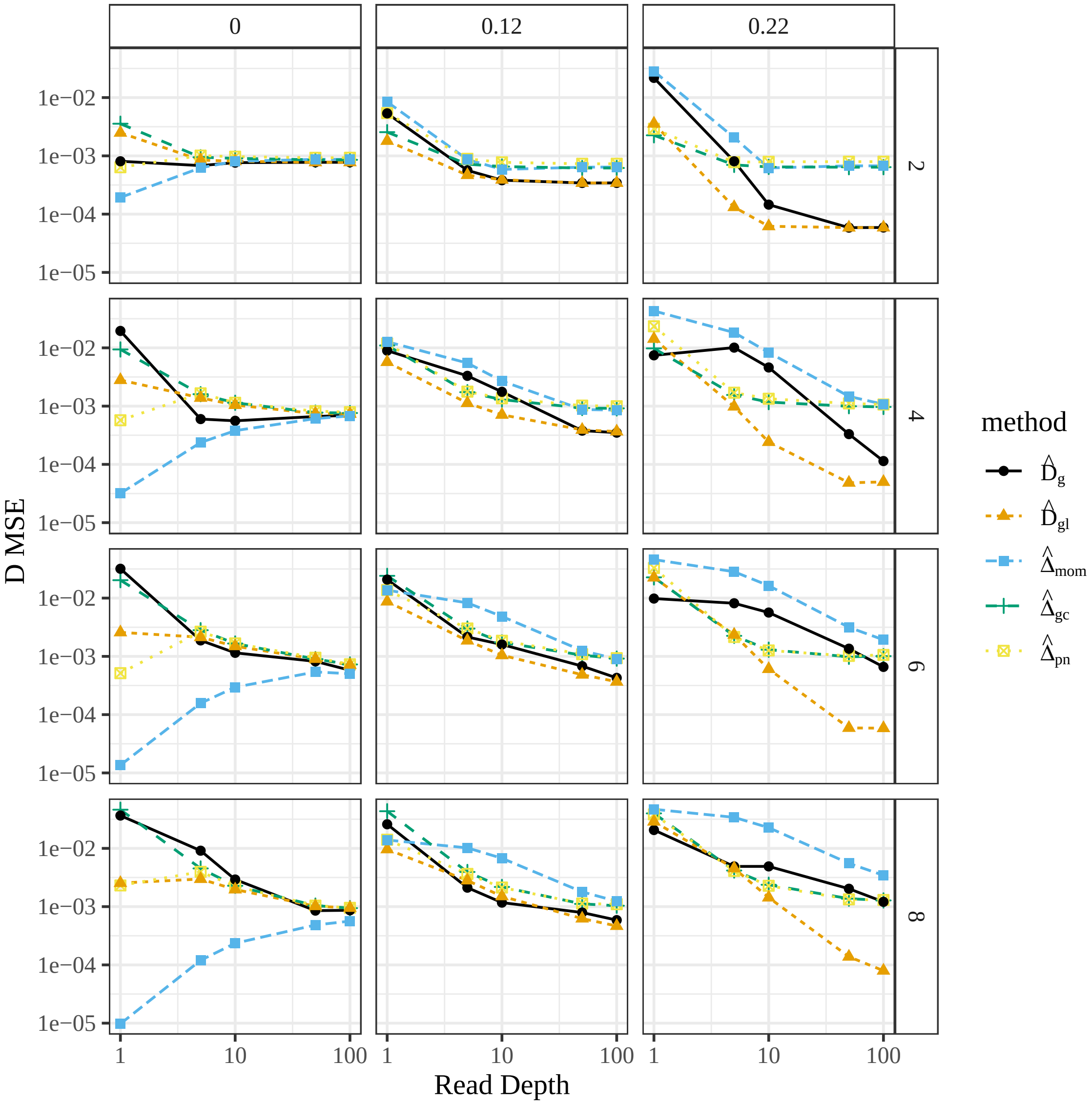

Figure S7: Mean-squared error of D estimators (y-axis) stratified by read-depth (x-axis), estimation method (color), ploidy (row-facets), and true D (column-facets) for the simulations from Section 3.1. Simulations were performed with $p_A = 0.5$ and $p_B = 0.5$.

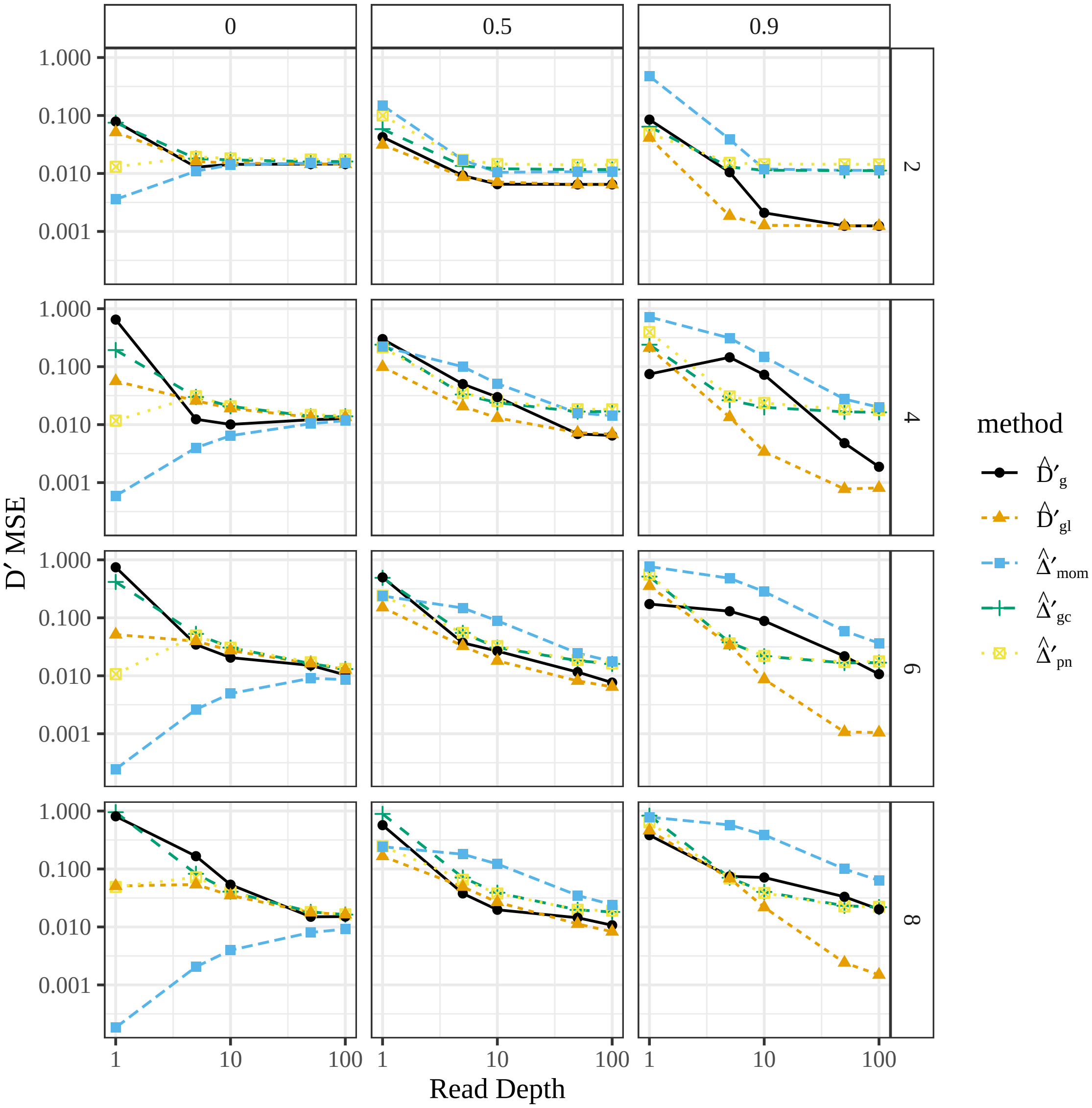

Figure S8: Mean-squared error of D' estimators (y-axis) stratified by read-depth (x-axis), estimation method (color), ploidy (row-facets), and true D' (column-facets) for the simulations from Section 3.1. Simulations were performed with $p_A = 0.5$ and $p_B = 0.5$.

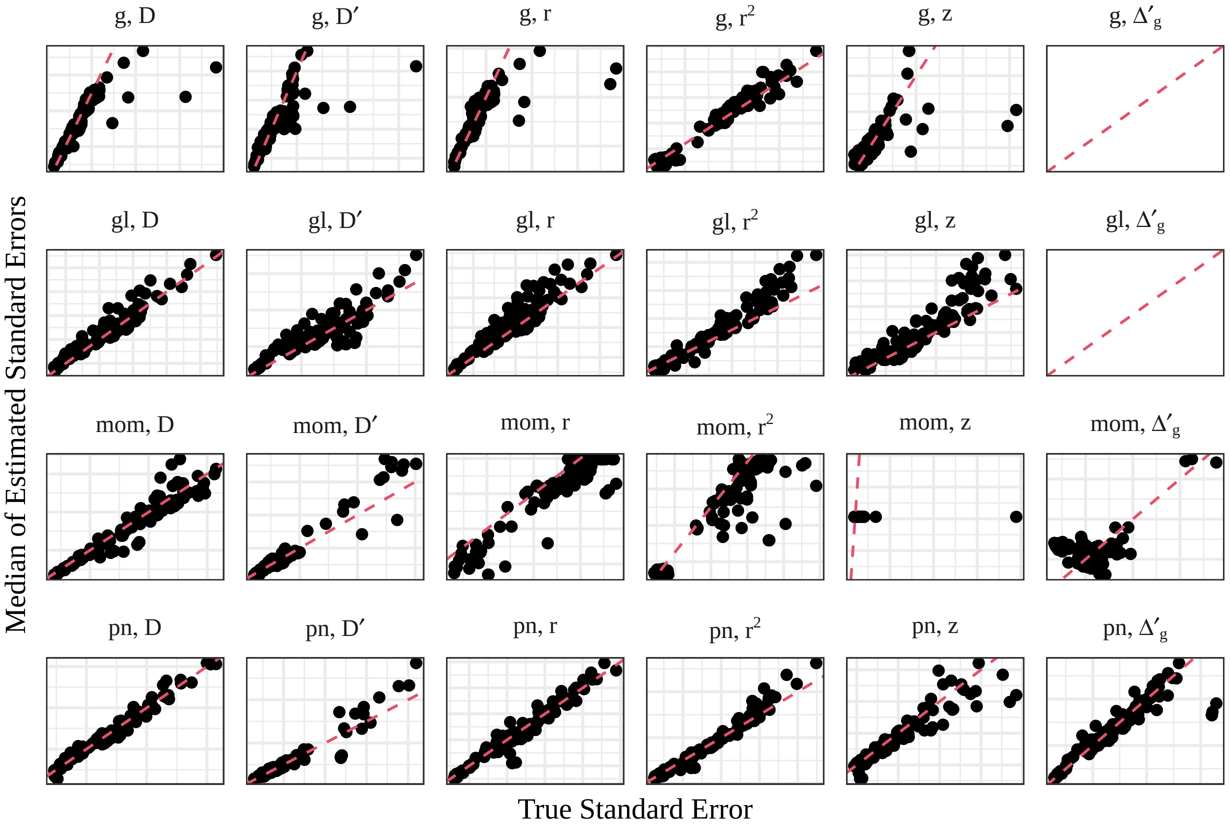

Figure S9: Standard errors of LD estimators (x-axis) versus the median of the estimated standard errors (y-axis) from the simulations in Section 3.1. Scales are freely varying between facets to allow for visualization. Each point is a different simulation setting. The scenarios where the read-depth was 1 were excluded due to poor behavior. The line is the $y = x$ line and any points above that line indicate that the estimated standard errors are typically larger than the true standard errors. Column-facets index the LD measure being estimated: D (1), D' (4), r (6), r^2, $z = \text{atanh}(r)$, and Δ'_g (17). Row-facets index the estimators: "g" for $(\hat{D}_g, \hat{D}'_g, \hat{r}_g)$, "gl" for $(\hat{D}_{gl}, \hat{D}'_{gl}, \hat{r}_{gl})$, "mom" for $(\hat{\Delta}_{mom}, \hat{\Delta}'_{mom}, \hat{\rho}_{mom})$, and "pn" for $(\hat{\Delta}_{pn}, \hat{\Delta}'_{pn}, \hat{\rho}_{pn})$. Two empty facets are present because gametic methods cannot estimate Δ'_g, a purely composite measure.

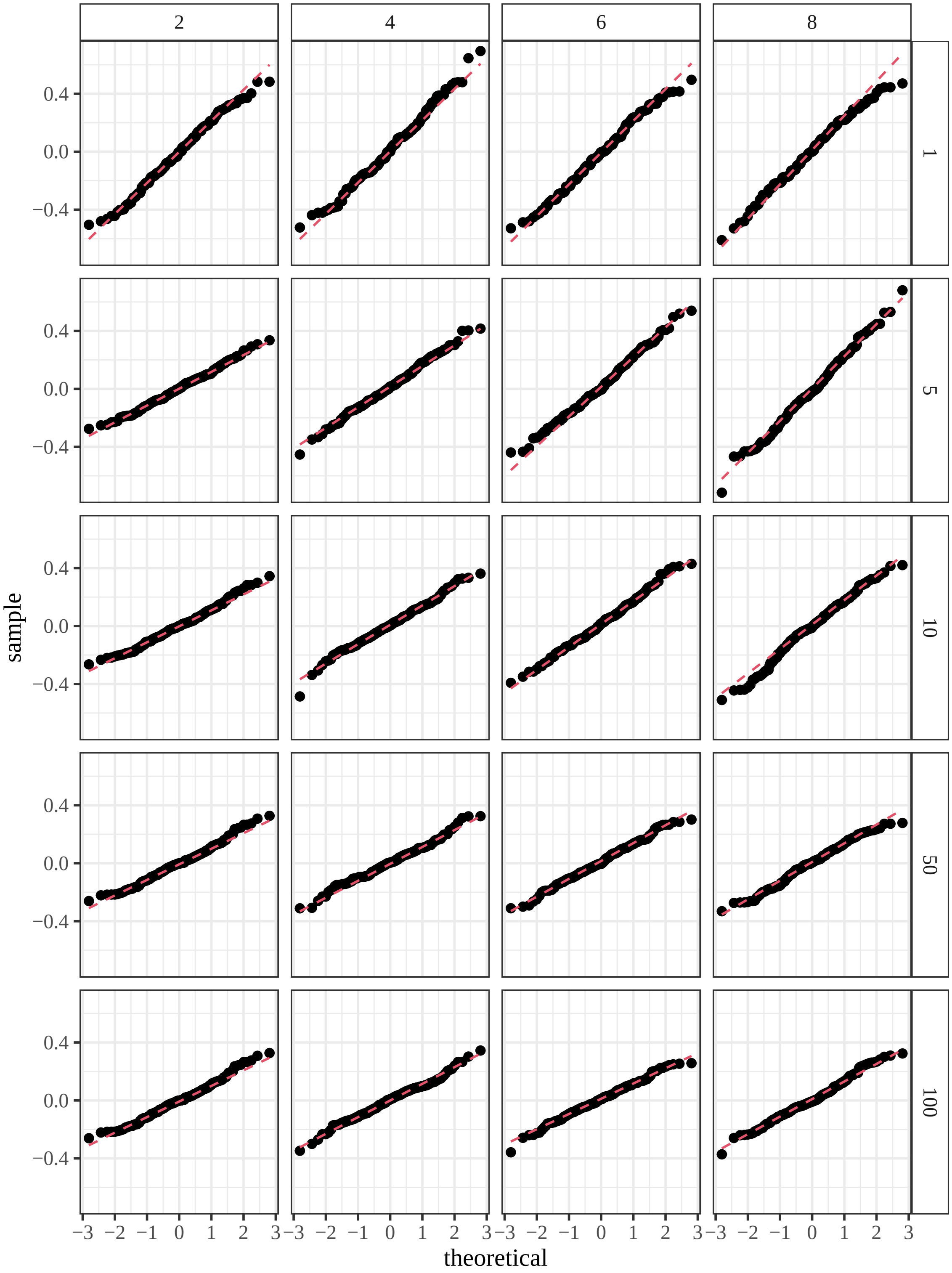

Figure S10: QQ-plots of of the Fisher-z transformation of $\hat{r}_{gl}$ when $p_A = 0.5$ and $p_B = 0.5$ from the simulations in Section 3.1. The estimates appear to approximately follow a normal distribution.

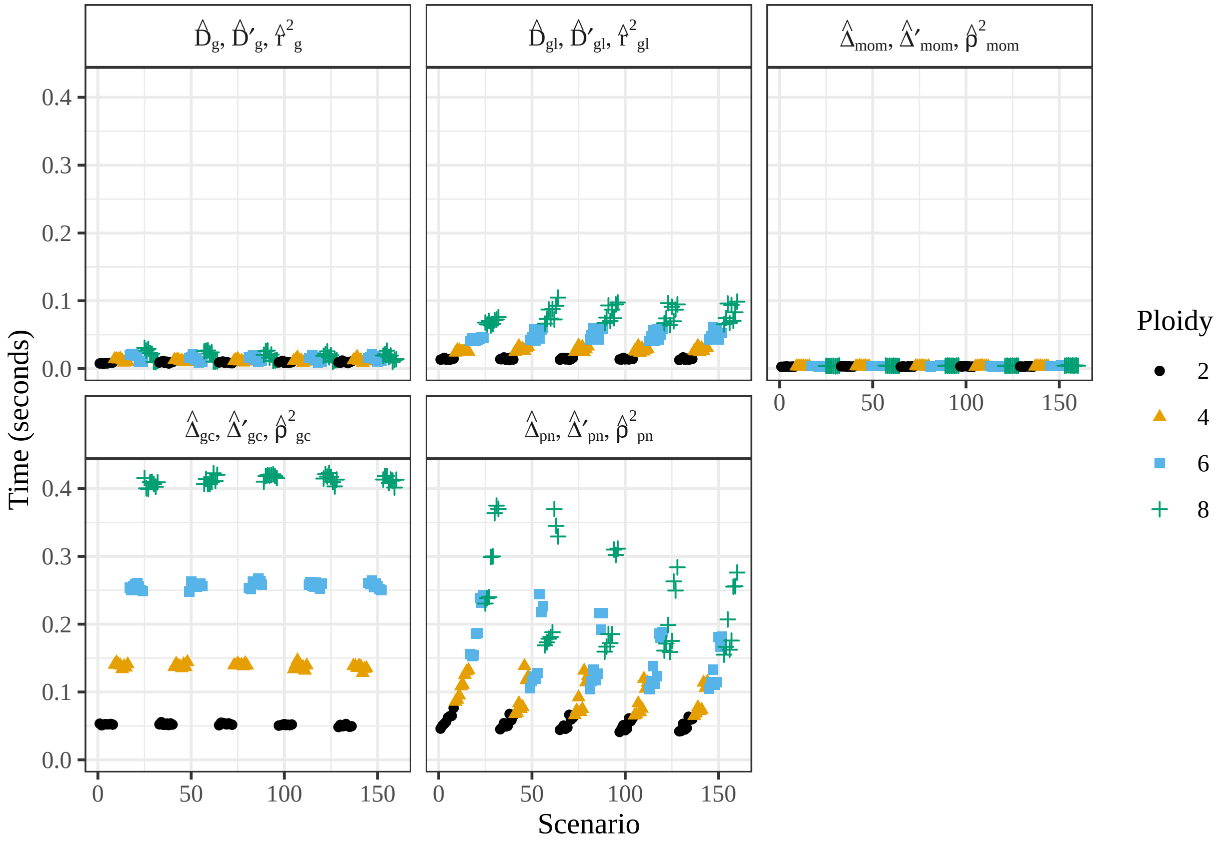

Figure S11: Mean computation time in seconds (y-axis) for each method (facets) stratified by the simulation settings (x-axis) for the simulations in Section 3.1. Methods using genotype likelihoods are generally slower, but all methods take less than half a second on average.

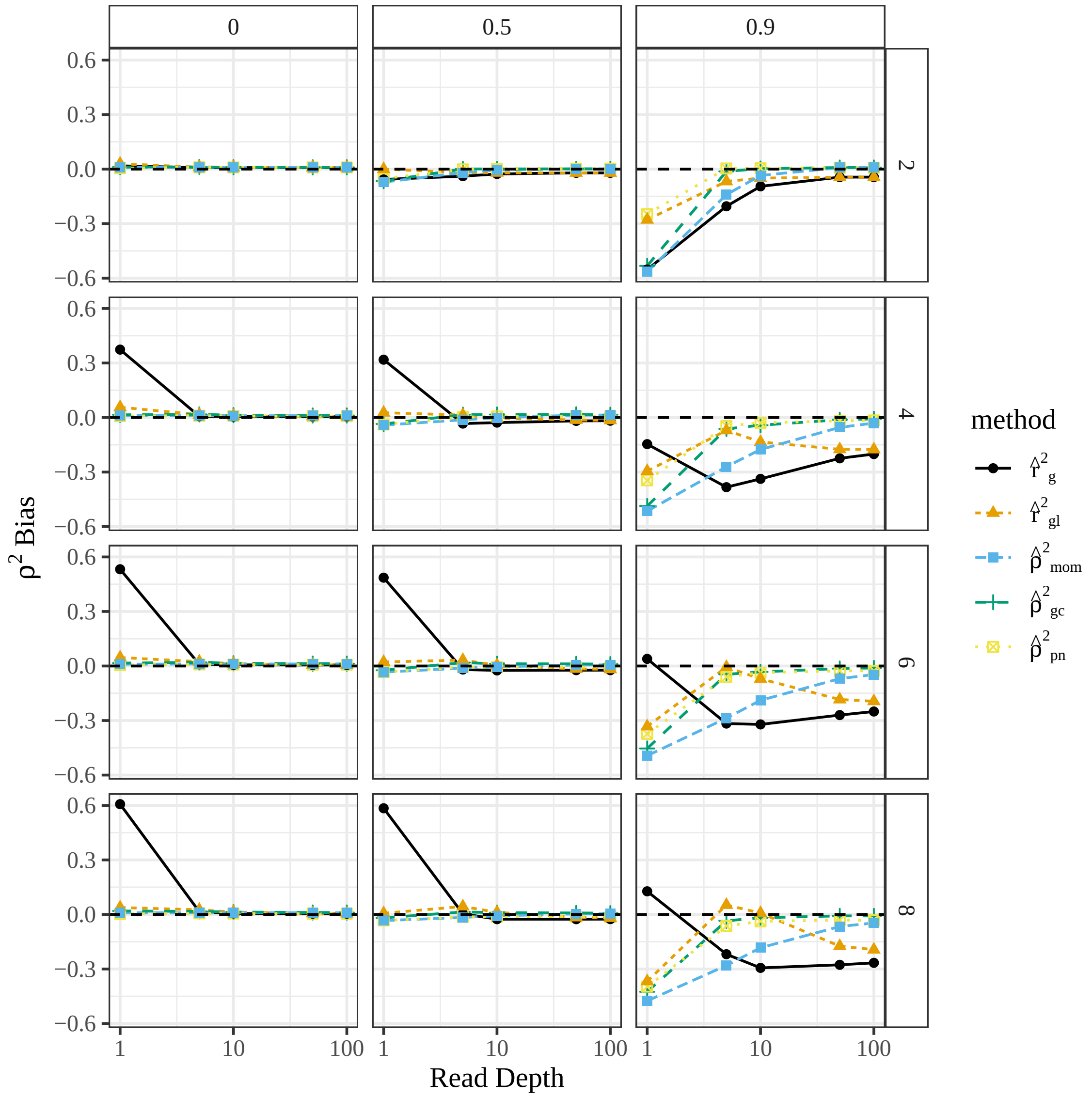

Figure S12: Average bias (y-axis) stratified by read-depth (x-axis), ploidy (row-facets) and association parameter of the proportional bivariate normal distribution (column-facets).

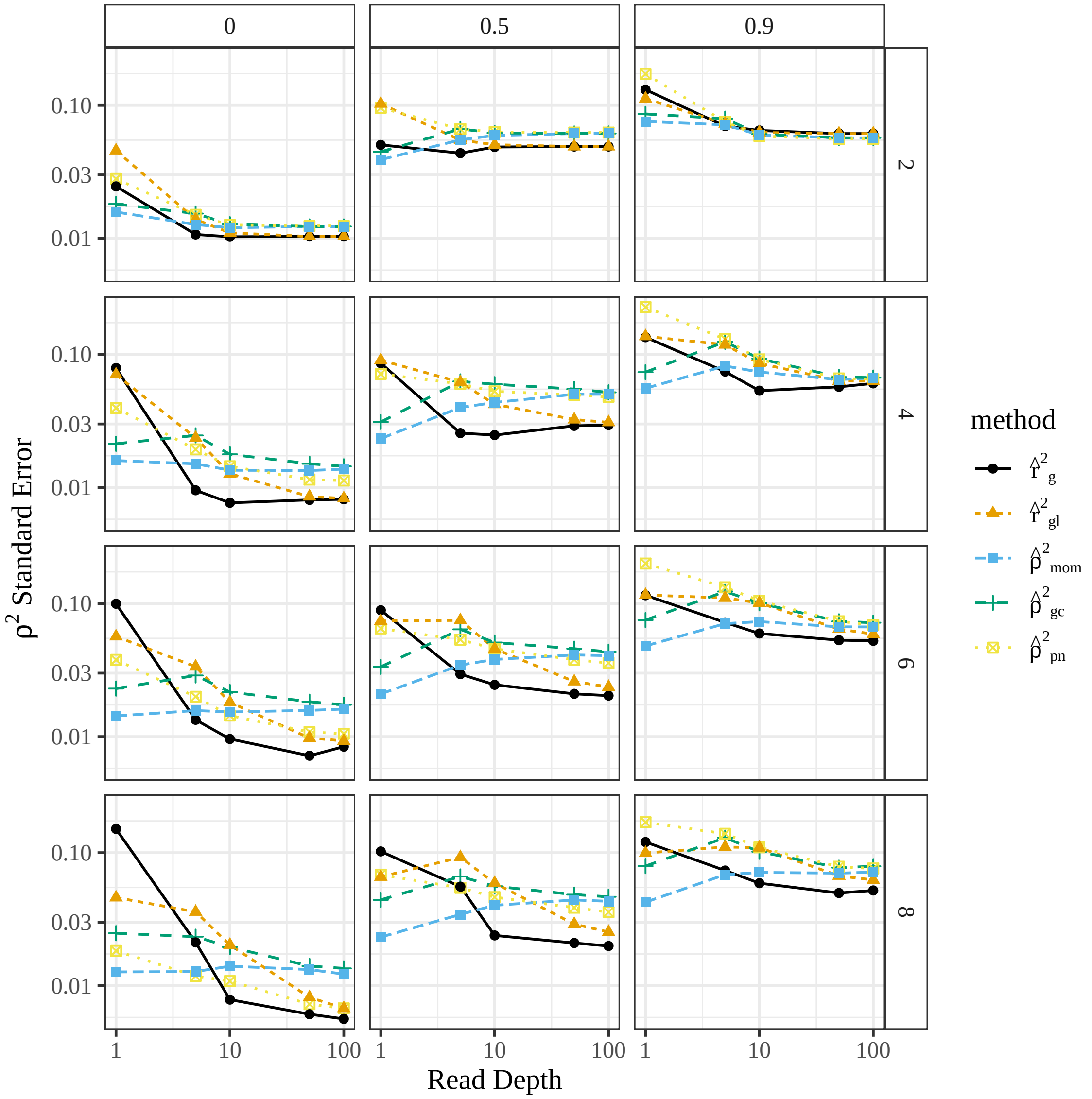

Figure S13: Average standard error (y-axis) stratified by read-depth (x-axis), ploidy (row-facets) and association parameter of the proportional bivariate normal distribution (column-facets).

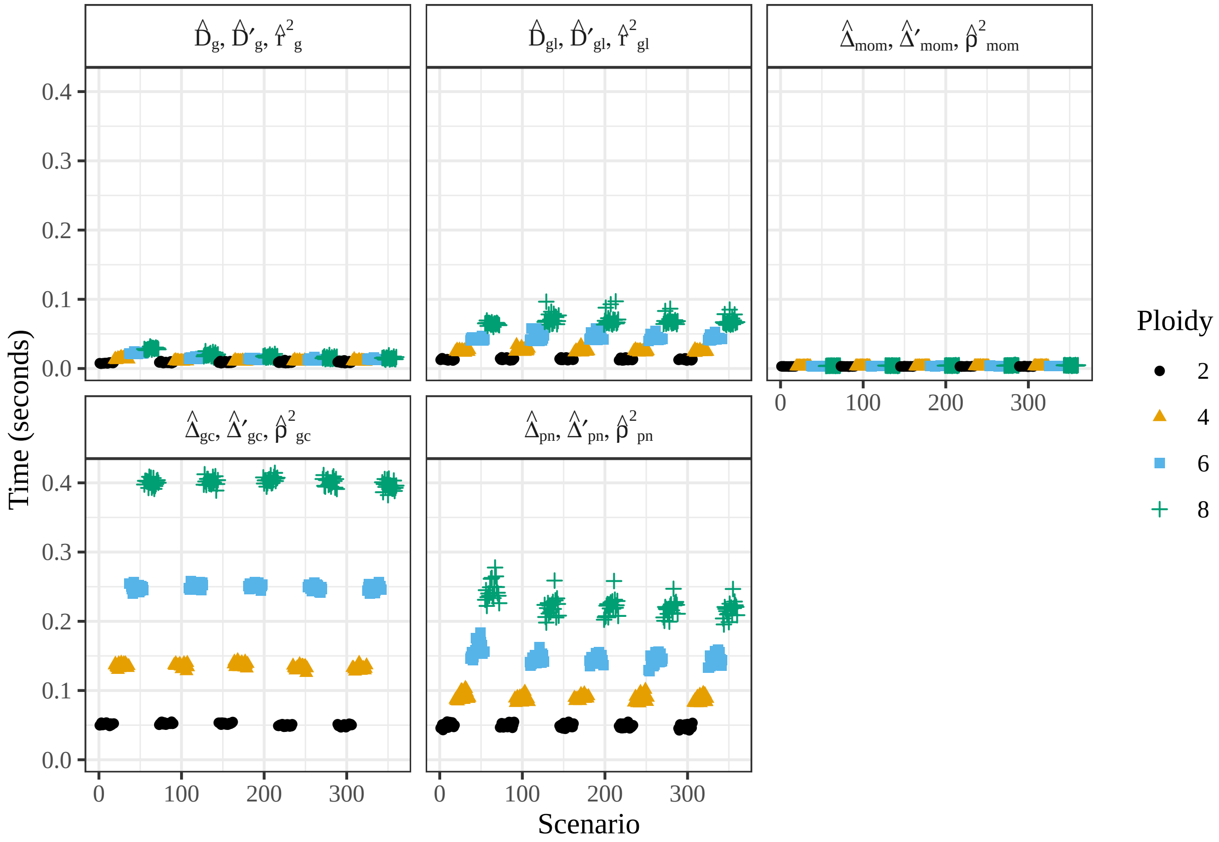

Figure S14: Mean computation time in seconds (y-axis) for each method (facets) stratified by the simulation settings (x-axis) for the simulations in Section 3.2. Methods using genotype likelihoods are generally slower, but all methods take less than half a second on average.

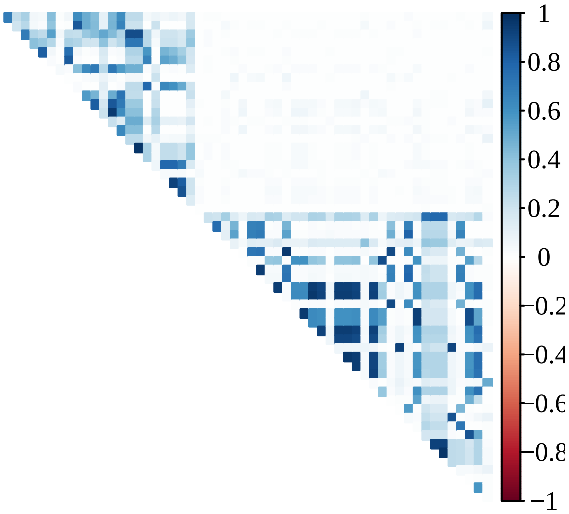

Figure S15: Heatmap of $\hat{r}_g^2$ from Section 3.3 using posterior mode genotypes. The data come from Uitdewilligen et al. [2013].

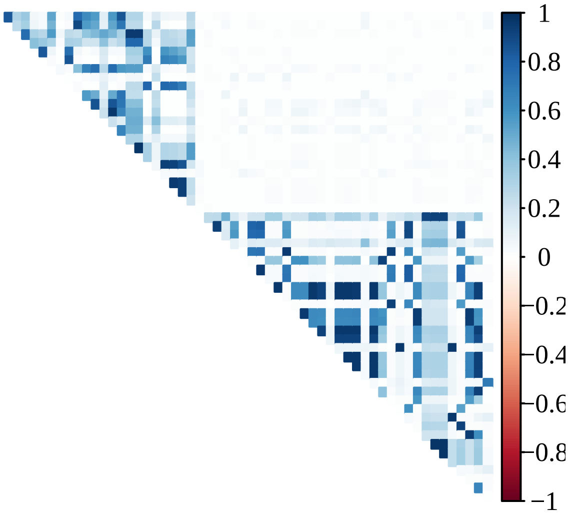

Figure S16: Heatmap of $\hat{r}_{gl}^2$ from Section 3.3. The data come from Uitdewilligen et al. [2013].

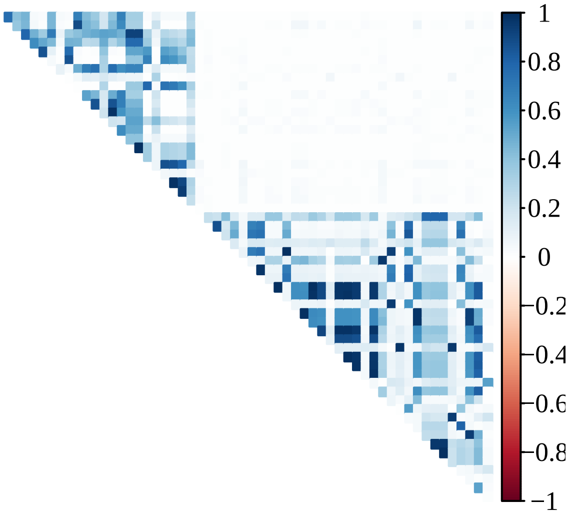

Figure S17: Heatmap of $\hat{\rho}^2_{mom}$ from Section 3.3 using posterior mean genotypes. The data come from Uitdewilligen et al. [2013].

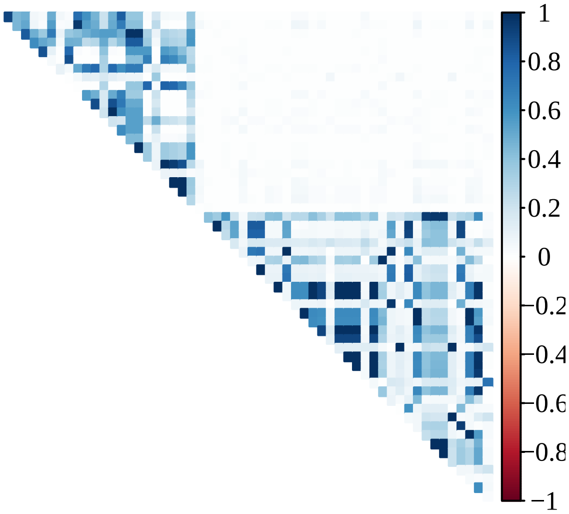

Figure S18: Heatmap of $\hat{\rho}^2_{gc}$ from Section 3.3. The data come from Uitdewilligen et al. [2013].

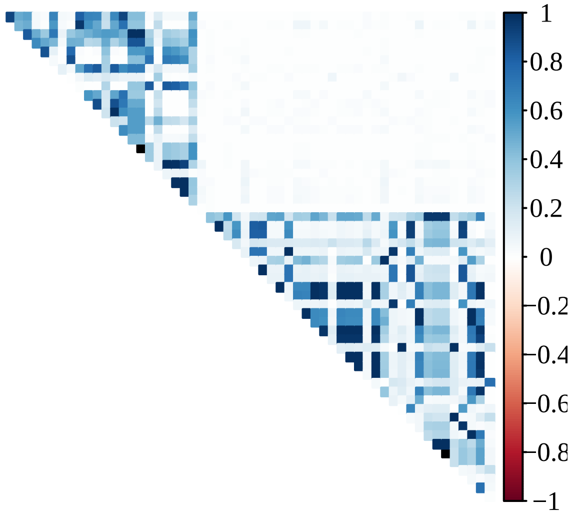

Figure S19: Heatmap of $\hat{\rho}^2_{pn}$ from Section 3.3. The data come from Uitdewilligen et al. [2013].

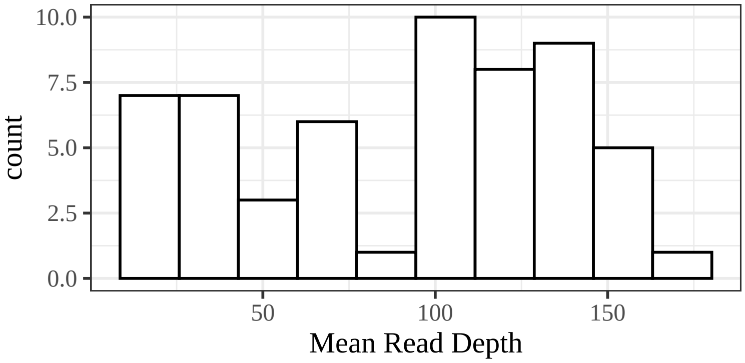

Figure S20: Histogram of mean read-depths of the SNPs used in Section 3.3 from the Uitdewilligen et al. [2013] data.

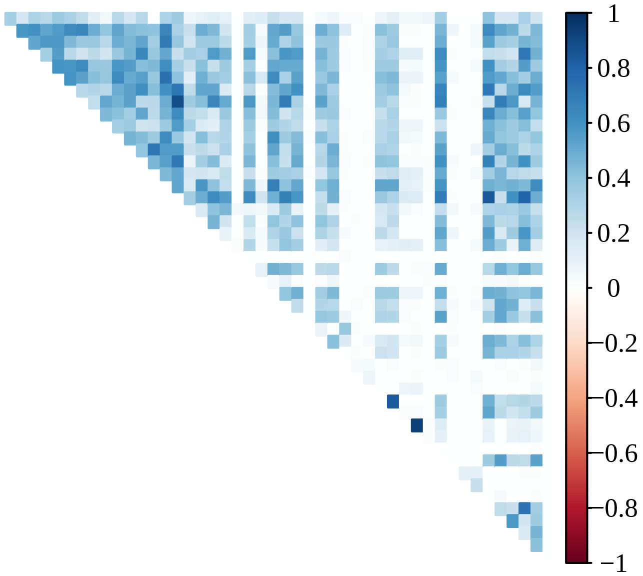

Figure S21: Heatmap of $\hat{r}_g^2$ from Section 3.4 using posterior mode genotypes. The data come from McAllister and Miller [2016].

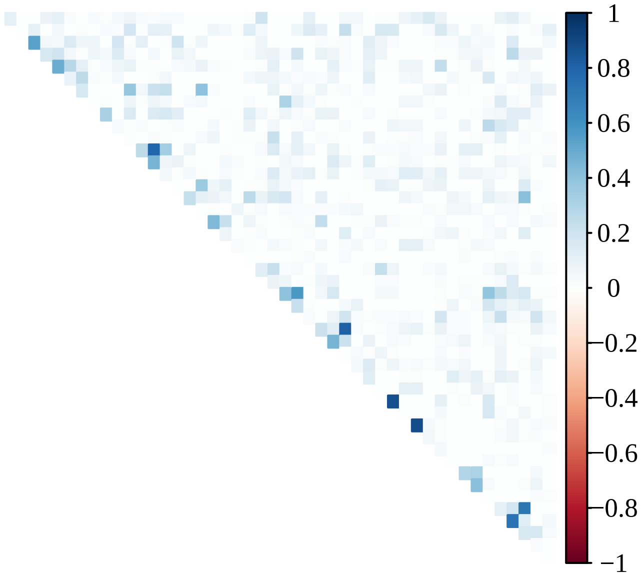

Figure S22: Heatmap of $\hat{r}_{gl}^2$ from Section 3.4. The data come from McAllister and Miller [2016].

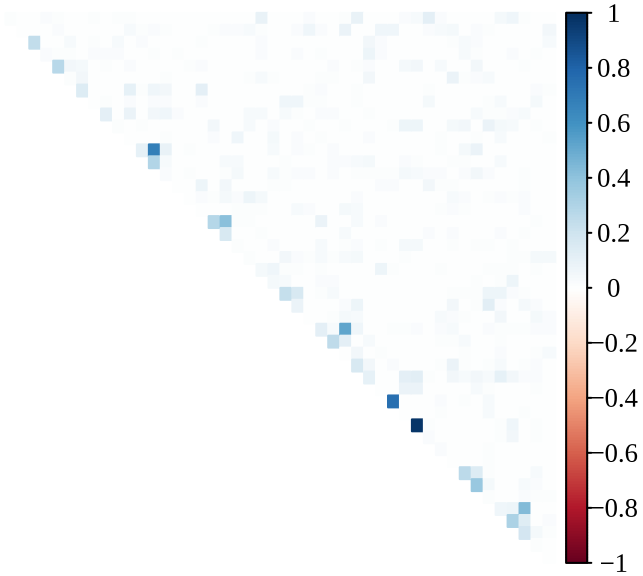

Figure S23: Heatmap of $\hat{\rho}^2_{mom}$ from Section 3.4 using posterior mean genotypes. The data come from McAllister and Miller [2016].

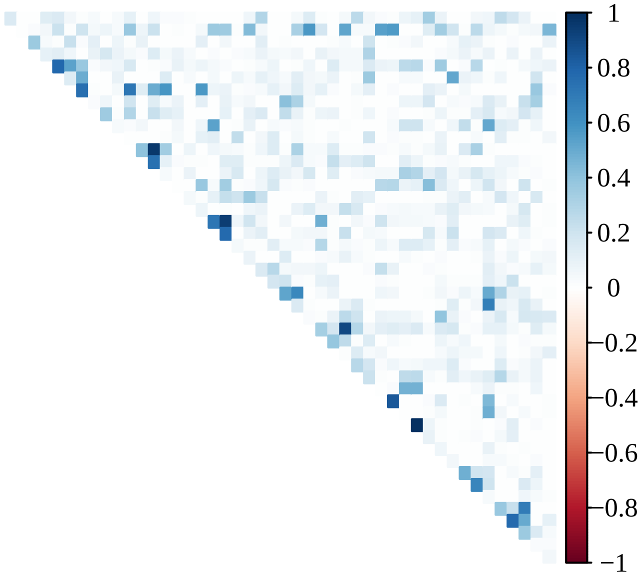

Figure S24: Heatmap of $\hat{\rho}^2_{gc}$ from Section 3.4. The data come from McAllister and Miller [2016].

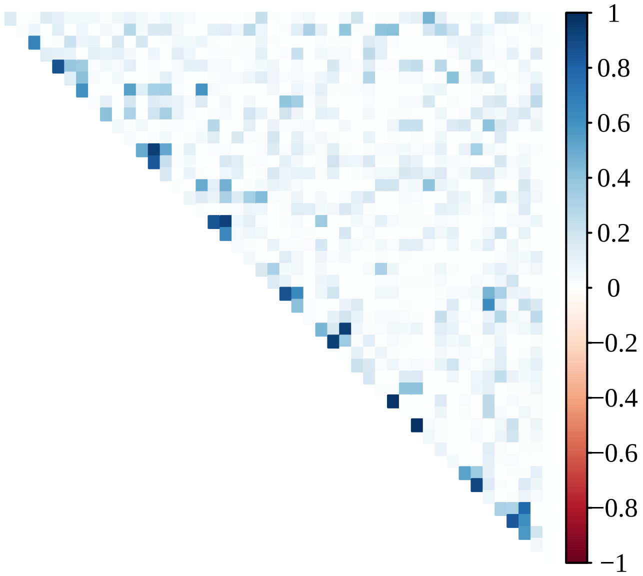

Figure S25: Heatmap of $\hat{\rho}^2_{pn}$ from Section 3.4. The data come from McAllister and Miller [2016].

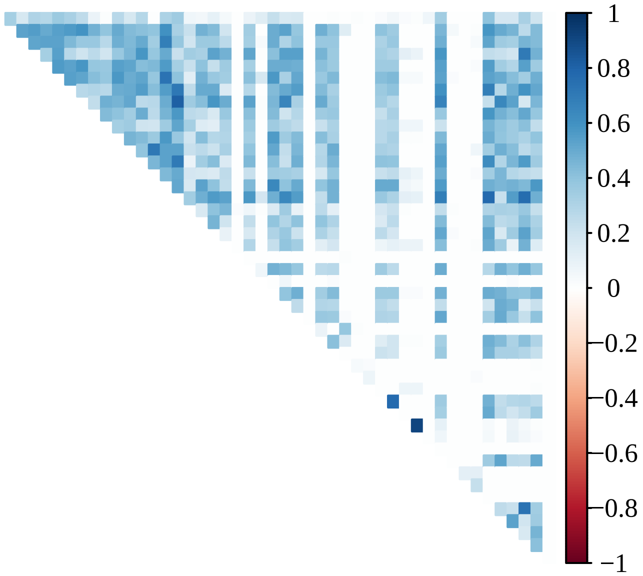

Figure S26: Heatmap of shrunken values of $\hat{r}^2_g$ from Section 3.4 using posterior mode genotypes. The data come from McAllister and Miller [2016]. Shrinkage was done using the methods of Stephens [2016] and Dey and Stephens [2018].

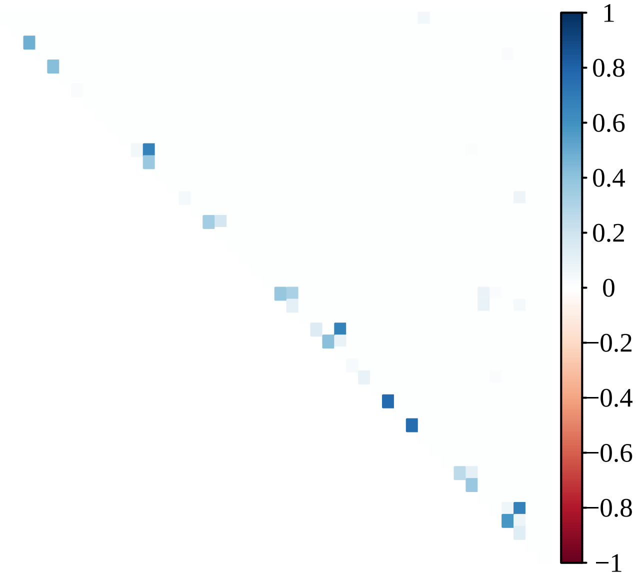

Figure S27: Heatmap of shrunken values of $\hat{r}^2_{gl}$ from Section 3.4. The data come from McAllister and Miller [2016]. Shrinkage was done using the methods of Stephens [2016] and Dey and Stephens [2018].

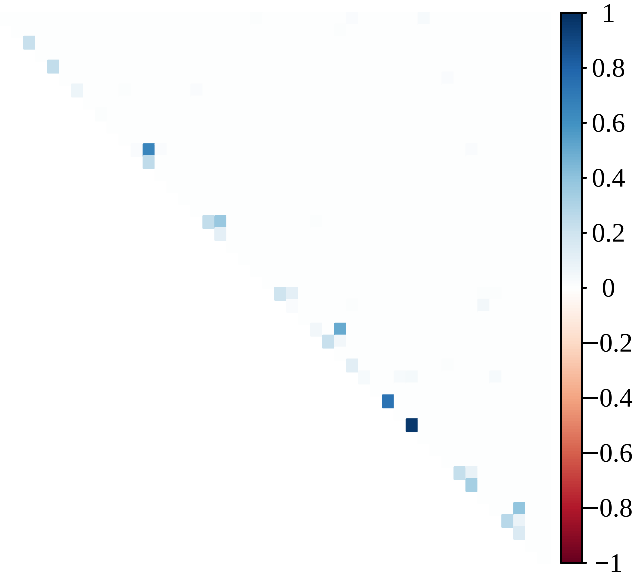

Figure S28: Heatmap of shrunken values of $\hat{\rho}^2_{mom}$ from Section 3.4 using posterior mean genotypes. The data come from McAllister and Miller [2016]. Shrinkage was done using the methods of Stephens [2016] and Dey and Stephens [2018].

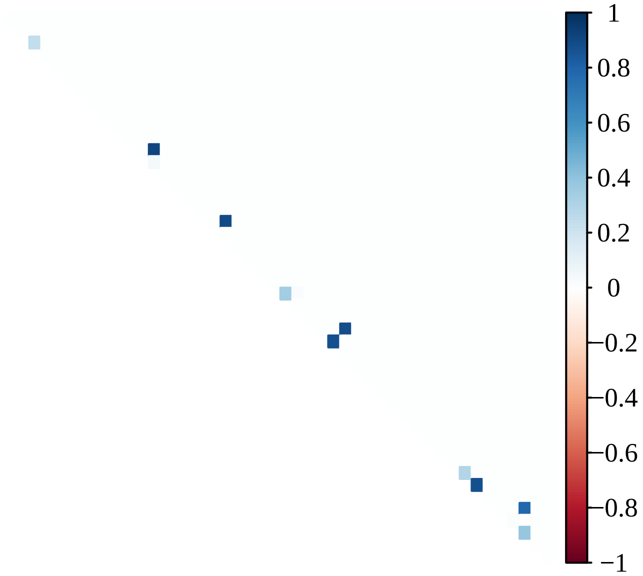

Figure S29: Heatmap of shrunken values of $\hat{\rho}^2_{pn}$ from Section 3.4. The data come from McAllister and Miller [2016]. Shrinkage was done using the methods of Stephens [2016] and Dey and Stephens [2018].

References

A. Agresti and B. A. Coull. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998. doi: 10.1080/00031305.1998.10480550.

N. A. Baird, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE*, 3(10):1–7, 10 2008. doi: 10.1371/journal.pone.0003376.

M. S. Barker, N. Arrigo, A. E. Baniaga, Z. Li, and D. A. Levin. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist*, 210(2):391–398, 2016. doi: 10.1111/nph.13698.

F. Z. Barreto, J. R. B. F. Rosa, T. W. A. Balsalobre, M. M. Pastina, R. R. Silva, H. P. Hoffmann, A. P. de Souza, A. A. F. Garcia, and M. S. Carneiro. A genome-wide association study identified loci for yield component traits in sugarcane (saccharum spp.). *PLOS ONE*, 14(7):1–22, 07 2019. doi: 10.1371/journal.pone.0219843.

M. S. Bartlett. On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, 53: 260–283, 1934. doi: 10.1017/S0370164600015637.

C. Benner, A. S. Havulinna, M.-R. Järvelin, V. Salomaa, S. Ripatti, and M. Pirinen. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *The American Journal of Human Genetics*, 101(4):539 – 551, 2017. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.08.012.

M. Berkelaar, K. Eikland, and P. Notebaert. lp_solve 5.5.2.5, open source (mixed-integer) linear programming system. Software, 2016. URL `http://lpsolve.sourceforge.net/5.5/`. Last accessed July, 6 2020.

M. Berkelaar et al. *lpSolve: Interface to ‘Lp_solve’ v. 5.5 to Solve Linear/Integer Programs*, 2020. URL `https://CRAN.R-project.org/package=lpSolve`. R package version 5.6.15.

M. Betancourt. Cruising the simplex: Hamiltonian Monte Carlo and the Dirichlet distribution. *AIP Conference Proceedings*, 1443(1):157–164, 2012. doi: 10.1063/1.3703631.

T. P. Bilton, J. C. McEwan, S. M. Clarke, R. Brauning, T. C. van Stijn, S. J. Rowe, and K. G. Dodds. Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*, 209(2): 389–400, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.300831.

B. Björn, M. J. Paulo, K. Kowitwanich, M. Sengers, R. G. Visser, H. J. van Eck, and F. A. van Eeuwijk. Population structure and linkage disequilibrium unravelled in tetraploid potato. *Theoretical and Applied Genetics*, 121(6):1151–1170, 2010. doi: 10.1007/s00122-010-1379-5.

P. D. Blischak, L. S. Kubatko, and A. D. Wolfe. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics*, 34(3):407–415, 2018. doi: 10.1093/bioinformatics/btx587.

P. J. Bradbury, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635, 06 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm308.

A. Brown. Sample sizes required to detect linkage disequilibrium between two or three loci. *Theoretical Population Biology*, 8(2):184 – 201, 1975. ISSN 0040-5809. doi: 10.1016/0040-5809(75)90031-3.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007. doi: 10.1086/521987.

L. V. Clark, A. E. Lipka, and E. J. Sacks. polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3: Genes, Genomes, Genetics*, 9(3):663–673, 2019. doi: 10.1534/g3.118.200913.

C. C. Cockerham and B. S. Weir. Digenic descent measures for finite populations. *Genetical Research*, 30 (2):121–147, 1977. doi: 10.1017/S0016672300017547.

I. de Bem Oliveira, M. F. R. Resende, L. F. V. Ferrão, R. R. Amadeu, J. B. Endelman, M. Kirst, A. S. G. Coelho, and P. R. Muñoz. Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3: Genes, Genomes, Genetics*, 9(4): 1189–1198, 2019. doi: 10.1534/g3.119.400059.

L. A. de C. Lara, M. F. Santos, L. Jank, L. Chiari, M. d. M. Vilela, R. R. Amadeu, J. P. R. dos Santos, G. d. S. Pereira, Z.-B. Zeng, and A. A. F. Garcia. Genomic selection with allele dosage in *Panicum maximum* jacq. *G3: Genes, Genomes, Genetics*, 9(8):2463–2475, 2019. doi: 10.1534/g3.118.200986.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322, 1995. doi: 10.1006/geno.1995.9003.

K. K. Dey and M. Stephens. CorShrink: Empirical Bayes shrinkage estimation of correlations, with applications. *bioRxiv*, 2018. doi: 10.1101/368316.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08.

D. Eddelbuettel and C. Sanderson. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, March 2014. doi: 10.1016/j.csda.2013.02.005.

B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979. doi: 10.1214/aos/1176344552.

R. J. Elshire, J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE*, 6(5): 1–10, 05 2011. doi: 10.1371/journal.pone.0019379.

J.-B. Fan, A. Oliphant, R. Shen, B. G. Kermani, F. García, K. L. Gunderson, M. S. T. Hansen, F. Steemers, S. L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. R. Stuelpnagel, and M. S. Chee. Highly parallel SNP genotyping. *Cold Spring Harbor Symposia on Quantitative Biology*, 68:69–78, 2003. doi: 10.1101/sqb.2003.68.69.

K. K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, 2015. doi: 10.1038/nature13835.

T. S. Ferguson. *A course in large sample theory*. CRC Press, 2002. ISBN 0-412-04371-8.

L. F. V. Ferrão, T. S. Johnson, J. Benevenuto, P. P. Edger, T. A. Colquhoun, and P. R. Munoz. Genome-wide association of volatiles reveals candidate loci for blueberry flavor. *New Phytologist*, 226(6):1725–1737, 2020. doi: 10.1111/nph.16459.

R. A. Fisher. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, 1: 3–32, 1921. URL `http://hdl.handle.net/2440/15169`.

E. A. Fox, A. E. Wright, M. Fumagalli, and F. G. Vieira. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, 35(19):3855–3856, 03 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz200.

D. Gerard and L. F. V. Ferrão. Priors for genotyping polyploids. *Bioinformatics*, 36(6):1795–1800, 11 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz852. bioRxiv: 751784.

D. Gerard, L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens. Genotyping polyploids from messy sequencing data. *Genetics*, 210(3):789–807, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.301468.

A. G. Griffiths, R. Moraga, M. Tausen, V. Gupta, T. P. Bilton, M. A. Campbell, R. Ashby, I. Nagy, A. Khan, A. Larking, C. Anderson, B. Franzmayr, K. Hancock, A. Scott, N. W. Ellison, M. P. Cox, T. Asp, T. Mailund, M. H. Schierup, and S. U. Andersen. Breaking free: The genomics of allopolyploidy-facilitated niche expansion in white clover. *The Plant Cell*, 31(7):1466–1487, 2019. ISSN 1040-4651. doi: 10.1105/tpc.18.00606.

A. Gur, G. Tzuri, A. Meir, U. Sa'ar, V. Portnoy, N. Katzir, A. A. Schaffer, L. Li, J. Burger, and Y. Tadmor. Genome-wide linkage-disequilibrium mapping to the candidate gene level in melon (*Cucumis melo*). *Scientific reports*, 7(1):1–13, 2017. doi: 10.1038/s41598-017-09987-4.

D. C. Hamilton and D. E. C. Cole. Standardizing a composite measure of linkage disequilibrium. *Annals of Human Genetics*, 68(3):234–239, 2004. doi: 10.1046/j.1529-8817.2004.00056.x.

P. W. Hedrick. Gametic disequilibrium measures: Proceed with caution. *Genetics*, 117(2):331–341, 1987. ISSN 0016-6731. URL https://www.genetics.org/content/117/2/331.

W. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and applied genetics*, 38 (6):226–231, 1968. doi: 10.1007/BF01245622.

W. G. Hill. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 33(2):229, 1974. doi: 10.1038/hdy.1974.89.

H. Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232, 1953. ISSN 00359246. doi: 10.2307/2983768.

K. Huang, D. W. Dunn, K. Ritland, and B. Li. polygene: Population genetics analyses for autopolyploids based on allelic phenotypes. *Methods in Ecology and Evolution*, 11(3):448–456, 2020. doi: 10.1111/2041-210X.13338.

T.-Y. J. Hui and A. Burt. Estimating linkage disequilibrium from genotypes under Hardy-Weinberg equilibrium. *BMC genetics*, 21(1):1–11, 2020. doi: 10.1186/s12863-020-0818-9.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, Calif., 1961. University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200512173.

L. B. Jorde. Linkage disequilibrium as a gene-mapping tool. *American journal of human genetics*, 56(1):11, 1995.

B. Julier. A program to test linkage disequilibrium between loci in autotetraploid species. *Molecular ecology resources*, 9(3):746–748, 2009. doi: 10.1111/j.1755-0998.2009.02530.x.

B. J. Knaus and N. J. Grünwald. VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1):44–53, 2017. ISSN 757. doi: 10.1111/1755-0998.12549.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, second edition, 1998. ISBN 0-387-98502-6.

S. Leonov and B. Qaqish. Correlated endpoints: simulation, modeling, and extreme correlations. *Statistical Papers*, 61(2):741–766, 2020. doi: 10.1007/s00362-017-0960-2.

R. Lewontin. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49, 1964. URL https://www.genetics.org/content/49/1/49.

R. C. Lewontin. On measures of gametic disequilibrium. *Genetics*, 120(3):849–852, 1988. ISSN 0016-6731. URL https://www.genetics.org/content/120/3/849.

R. C. Lewontin and K.-i. Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4): 458–472, 1960. doi: 10.1111/j.1558-5646.1960.tb03113.x.

H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987, 2011. doi: 10.1093/bioinformatics/btr509.

Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010. doi: 10.1002/gepi.20533.

F. Lu, A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney, M. D. Casler, E. S. Buckler, and D. E. Costich. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLOS Genetics*, 9(1):1–14, 01 2013. doi: 10.1371/journal.pgen.1003215.

I. Mackay and W. Powell. Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science*, 12(2):57 – 63, 2007. ISSN 1360-1385. doi: 10.1016/j.tplants.2006.12.001.

T. Maruki and M. Lynch. Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics*, 197(4):1303–1313, 2014. ISSN 0016-6731. doi: 10.1534/genetics.114.165514.

T. Maruki and M. Lynch. Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics*, 7(5):1393–1404, 2017. doi: 10.1534/g3.117.039008.

F. I. Matias, K. G. Xavier Meireles, S. T. Nagamatsu, S. C. Lima Barrios, C. Borges do Valle, M. F. Carazzolle, R. Fritsche-Neto, and J. B. Endelman. Expected genotype quality and diploidized marker data from genotyping-by-sequencing of *Urochloa* spp. tetraploids. *The Plant Genome*, 12(3):190002, 2019. doi: 10.3835/plantgenome2019.01.0002.

C. A. McAllister and A. J. Miller. Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. *American Journal of Botany*, 103(7):1314–1325, 2016. doi: 10.3732/ajb.1600146.

C. A. McAllister and A. J. Miller. Data from: Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*, 2017. URL `https://doi.org/10.5061/dryad.05qs7`. Dataset.

Microsoft and S. Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2019. URL `https://CRAN.R-project.org/package=doParallel`. R package version 1.0.15.

Microsoft and S. Weston. *foreach: Provides Foreach Looping Construct*, 2020. URL `https://CRAN.R-project.org/package=foreach`. R package version 1.5.0.

M. Mollinari and A. A. F. Garcia. Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden markov models. *G3: Genes, Genomes, Genetics*, 9(10): 3297–3314, 2019. doi: 10.1534/g3.119.400378.

M. Mollinari and O. Serang. Quantitative SNP genotyping of polyploids with MassARRAY and other platforms. In J. Batley, editor, *Plant Genotyping: Methods and Protocols*, pages 215–241. Springer New York, New York, NY, 2015. ISBN 978-1-4939-1966-6. doi: 10.1007/978-1-4939-1966-6_17.

J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006. ISBN 0-387-30303-0.

V. Obenchain, M. Lawrence, V. Carey, S. Gogarten, P. Shannon, and M. Morgan. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14):2076–2078, 2014. doi: 10.1093/bioinformatics/btu168.

P. Oeth, G. del Mistro, G. Marnellos, T. Shi, and D. van den Boom. Qualitative and quantitative genotyping using single base primer extension coupled with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MassARRAY®). In A. Komar, editor, *Single Nucleotide Polymorphisms*, pages 307–343. Humana Press, 2009. ISBN 978-1-60327-411-1. doi: 10.1007/978-1-60327-411-1_20.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

L.-M. Raboin, J. Pauquet, M. Butterfield, A. D'Hont, and J.-C. Glaszmann. Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. *Theoretical and Applied Genetics*, 116(5):701–714, 2008. doi: 10.1007/s00122-007-0703-1.

G. P. Ramstein, J. Evans, S. M. Kaeppler, R. B. Mitchell, K. P. Vogel, C. R. Buell, and M. D. Casler. Accuracy of genomic prediction in switchgrass (*Panicum virgatum* l.) improved by accounting for linkage disequilibrium. *G3: Genes, Genomes, Genetics*, 6(4):1049–1062, 2016. doi: 10.1534/g3.115.024950.

A. R. Rogers and C. Huff. Linkage disequilibrium between loci with unknown phase. *Genetics*, 182(3): 839–844, 2009. ISSN 0016-6731. doi: 10.1534/genetics.108.093153.

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006. doi: 10.1086/502802.

C. A. Schmitz Carley, J. J. Coombs, D. S. Douches, P. C. Bethke, J. P. Palta, R. G. Novy, and J. B. Endelman. Automated tetraploid genotype calling by hierarchical clustering. *Theoretical and Applied Genetics*, 130(4):717–726, 2017. ISSN 1432-2242. doi: 10.1007/s00122-016-2845-5.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. doi: https://doi.org/10.2202/1544-6115.1175.

O. Serang, M. Mollinari, and A. A. F. Garcia. Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLOS ONE*, 7(2):1–13, 02 2012. doi: 10.1371/journal.pone.0030906.

S. K. Sharma, K. MacKenzie, K. McLean, F. Dale, S. Daniels, and G. J. Bryan. Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*, 8(10):3185–3202, 2018. doi: 10.1534/g3.118.200377.

J. Shen, Z. Li, J. Chen, Z. Song, Z. Zhou, and Y. Shi. SHEsisPlus, a toolset for genetic studies on polyploid species. *Scientific reports*, 6:24095, 2016. doi: 10.1038/srep24095.

I. Simko, K. G. Haynes, and R. W. Jones. Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. *Genetics*, 173(4):2237–2245, 2006. ISSN 0016-6731. doi: 10.1534/genetics.106.060905.

M. Slatkin. Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477, 2008. doi: 10.1038/nrg2361.

D. E. Soltis, C. J. Visger, and P. S. Soltis. The polyploidy revolution then…and now: Stebbins revisited. *American Journal of Botany*, 101(7):1057–1078, 2014. doi: 10.3732/ajb.1400178.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 10 2016. ISSN 1465-4644. doi: 10.1093/biostatistics/kxw041.

S.-Y. Su, J. White, D. J. Balding, and L. J. Coin. Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC bioinformatics*, 9(1):1–9, 2008. doi: 10.1186/1471-2105-9-513.

X. Sun, R. Fernando, and J. Dekkers. Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. *Genetics Selection Evolution*, 48(1):77, 2016. doi: 10.1186/s12711-016-0255-4.

J. A. Sved and W. G. Hill. One hundred years of linkage disequilibrium. *Genetics*, 209(3):629–636, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.300642.

K. Swarts, H. Li, J. A. R. Navarro, D. An, M. C. Romay, S. Hearne, C. Acharya, J. Glaubitz, S. E. Mitchell, R. J. Elshire, E. S. Buckler, and P. J. Bradbury. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *The Plant Genome*, 7(3):1–12, 2014. doi: 10.3835/plantgenome2014.05.0023.

J. A. Udall and J. F. Wendel. Polyploidy and crop improvement. *Crop Science*, 46(Supplement_1):S–3, 2006. doi: 10.2135/cropsci2006.07.0489tpg.

J. G. A. M. L. Uitdewilligen, A.-M. A. Wolters, B. B. D'hoop, T. J. A. Borm, R. G. F. Visser, and H. J. van Eck. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLOS ONE*, 8(5):1–14, 05 2013. doi: 10.1371/journal.pone.0062355.

M. Van Wyngaarden, P. V. R. Snelgrove, C. DiBacco, L. C. Hamilton, N. Rodríguez-Ezpeleta, N. W. Jeffery, R. R. E. Stanley, and I. R. Bradbury. Identifying patterns of dispersal, connectivity and selection in the sea scallop, *Placopecten magellanicus*, using RADseq-derived SNPs. *Evolutionary Applications*, 10(1): 102–117, 2017. doi: 10.1111/eva.12432.

R. E. Voorrips, G. Gort, and B. Vosman. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*, 12(1):172, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-172.

P. G. Vos, M. J. Paulo, R. E. Voorrips, R. G. Visser, H. J. van Eck, and F. A. van Eeuwijk. Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics*, 130(1):123–135, 2017. doi: 10.1007/s00122-016-2798-8.

T. Wei and V. Simko. *R package "corrplot": Visualization of a Correlation Matrix*, 2017. URL `https://github.com/taiyun/corrplot`. (Version 0.84).

B. Weir. *Genetic data analysis II*. Sinauer Associates, Inc., 1996. ISBN 0-87893-902-4.

B. Weir. Linkage disequilibrium and association mapping. *Annual Review of Genomics and Human Genetics*, 9(1):129–142, 2008. doi: 10.1146/annurev.genom.9.081307.164347. PMID: 18505378.

B. Weir and C. C. Cockerham. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 42(1):105–111, 1979. doi: 10.1038/hdy.1979.10.

B. S. Weir. Inferences about linkage disequilibrium. *Biometrics*, pages 235–254, 1979. doi: 10.2307/2529947.

B. S. Weir and C. C. Cockerham. Complete characterization of disequilibrium at two loci. In M. W. Feldman, editor, *Mathematical evolutionary theory*, pages 86–110. Princeton University Press, Princeton, NJ, 1989. ISBN 0-691-08502-1.

X. Wen and M. Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158–1182, 2010. ISSN 1932-6157. doi: 10.1214/10-aoas338.

W. Whitt. Bivariate distributions with given marginals. *Ann. Statist.*, 4(6):1280–1289, 11 1976. doi: 10.1214/aos/1176343660.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Y. C. J. Wientjes, R. F. Veerkamp, and M. P. L. Calus. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, 193(2):621–631, 2013. ISSN 0016-6731. doi: 10.1534/genetics.112.146290.

M. Xiong and S.-W. Guo. Fine-scale genetic mapping based on linkage disequilibrium: Theory and applications. *The American Journal of Human Genetics*, 60(6):1513 – 1531, 1997. ISSN 0002-9297. doi: 10.1086/515475.

J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, G. I. of ANthropometric Traits (GIANT) Consortium, D. G. Replication, M. analysis (DIAGRAM) Consortium, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, T. M. Frayling, M. I. McCarthy, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369, 2012. doi: 10.1038/ng.2213.

D. V. Zaykin. Bounds and normalization of the composite linkage disequilibrium coefficient. *Genetic Epidemiology*, 27(3):252–257, 2004. doi: 10.1002/gepi.20015.

C. Zheng, R. E. Voorrips, J. Jansen, C. A. Hackett, J. Ho, and M. C. Bink. Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics*, 203(1):119–131, 2016. doi: 10.1534/genetics.115.185579.

X. Zhu and M. Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications*, 9(1):1–14, 2018. doi: 10.1038/s41467-018-06805-x.

X. Zhu, F. Xu, S. Zhao, W. Bo, L. Jiang, X. Pang, and R. Wu. Inferring the evolutionary history of outcrossing populations through computing a multiallelic linkage–linkage disequilibrium map. *Methods in Ecology and Evolution*, 6(11):1259–1269, 2015. doi: 10.1111/2041-210X.12428.

K. Zych, G. Gort, C. A. Maliepaard, R. C. Jansen, and R. E. Voorrips. FitTetra 2.0–improved genotype calling for tetraploids with multiple population and parental data support. *BMC bioinformatics*, 20(1): 148, 2019. doi: 10.1186/s12859-019-2703-y.