Unbiased estimation of linkage disequilibrium from unphased data

Aaron P. Ragsdale* and Simon Gravel

Department of Human Genetics, McGill University, Montreal, QC, Canada
*aaron.ragsdale@mail.mcgill.ca

April 3, 2019

Abstract

Linkage disequilibrium is used to infer evolutionary history and to identify regions under selection or associated with a given trait. In each case, we require accurate estimates of linkage disequilibrium from sequencing data. Unphased data presents a challenge because the co-occurrence of alleles at different loci is ambiguous. Widely used estimators for the common statistics r^2 and D^2 exhibit large and variable upward biases that complicate interpretation and comparison across cohorts. Here, we show how to find unbiased estimators for a wide range of two-locus statistics, including D^2, for both single and multiple randomly mating populations. These provide accurate estimates over three orders of magnitude in LD. We also use these estimators to construct an estimator for r^2 that is less biased than commonly used estimators, but nevertheless argue for using σ_D^2 rather than r^2 for population size estimates.

Introduction

Linkage disequilibrium (LD), the association of alleles at different loci, is informative about both evolutionary and biological processes. Patterns of LD are used to infer historical demographic events, discover regions under selection, estimate the landscape of recombination across the genome, and identify genes associated with biomedical and phenotypic traits. Such analyses require accurate and efficient estimation of LD statistics from genome sequencing data.

LD between two loci is often described by the covariance or correlation of alleles at the two loci. Estimating this covariance from data is simplest when we directly observe haplotypes (in haploid or phased diploid sequencing), in which case we know which alleles co-occur on the same gamete. However, most whole-genome sequencing of diploids is unphased, leading to ambiguity about which of the two alleles at each locus co-occur.

The statistical foundation for computing LD statistics from unphased data that was developed in the 1970s (e.g. Weir and Cockerham (1979); Cockerham and Weir (1977); Weir (1979)) has led to widely used approaches for their estimation from modern sequencing data (Excoffier and Slatkin, 1995; Rogers and Huff, 2009). While these methods provide accurate estimates for the covariance and correlation (D and r), they do not extend to other two-locus statistics, and they result in biased estimates of r^2 (Waples, 2006). This bias confounds interpretation of r^2 decay curves beyond short recombination distances.

Here, we extend an approach used by Weir (1979) to estimate the covariance D to find unbiased estimators for a large set of two-locus statistics including D^2 and σ_D^2. We show that these estimators accurately compute low-order statistics used in demographic and evolutionary inferences, and we provide an estimator for r^2 with improved qualitative and quantitative behavior over widely used approaches.

As a concrete use case, we consider estimating N_e from LD data, as is commonly done in conservation genomics. Waples (2006) suggested combining an empirical bias correction for estimates of r^2 with an

approximate theoretical result from Weir and Hill (1980) to estimate N_e. We propose an alternative approach to estimate N_e using our unbiased estimator for σ_D^2 and an approximation due to Ohta and Kimura (1969) which avoids many of the assumptions and biases associated with r^2 estimation. For fully unlinked loci, we compute expected σ_D^2 for a haploid model, which matches the Hill-Robertson expectation, and a diploid model under both random mating and full monogamy.

Linkage disequilibrium statistics

For measuring LD between two loci, we assume that each locus carries two alleles: A/a at the left locus and B/b at the right locus. We think of A and B as the derived alleles, although the expectations of statistics that we consider are unchanged if alleles are randomly labeled instead. Allele A has frequency p in the population (allele a has frequency $1-p$), and B has frequency q (b has $1-q$). There are four possible two-locus haplotypes, AB, Ab, aB, and ab, whose frequencies sum to 1.

For two loci, LD is typically given by the covariance or correlation of alleles co-occurring on a haplotype. The covariance is denoted D:

$$\begin{aligned} D = \mathrm{Cov}(A, B) &= f_{AB} - pq \\ &= f_{AB} f_{ab} - f_{Ab} f_{aB}, \end{aligned}$$

and the correlation is denoted r:

$$r = \frac{D}{\sqrt{p(1-p)q(1-q)}}.$$

The expectation of D ($\mathbb{E}[D]$) is zero under general conditions. We think of expectations of quantities as though we average over many realizations of the same evolutionary process, although in reality we have only a single observation for any given pair of loci. In practice, we therefore take this expectation over many distinct pairs of loci across the genome.

Because $\mathbb{E}[D] = 0$, its variance is given by the second moment $\mathbb{E}[D^2]$, and LD is commonly reported as the squared correlation,

$$r^2 = \frac{D^2}{p(1-p)q(1-q)}.$$

r^2 sees wide use in genome-wide association studies to thin data for reducing correlation between SNPs and to characterize local levels of LD (e.g. Speed et al. (2012)). Genome-wide patterns of the decay of r^2 with increasing distance between loci is also informative about demography, as the scale and decay rate of r^2 curves reflect long-term population sizes, while recent admixture will lead to elevated long-range LD.

Results

Estimating LD from data

In the Methods, we present an approach to compute unbiased estimators for a broad set of two-locus statistics, for either phased or unphased data. This includes commonly used statistics, such as D and D^2, the additional statistics in the Hill-Robertson system ($D(1-2p)(1-2q)$ and $p(1-p)q(1-q)$), and, in general, any statistic that can be expressed as a polynomial in haplotype frequencies (f's) or in terms of p, q, and D. We use this same approach to find unbiased estimators for cross-population LD statistics, which we recently used to infer multi-population demographic history (Ragsdale and Gravel, 2018).

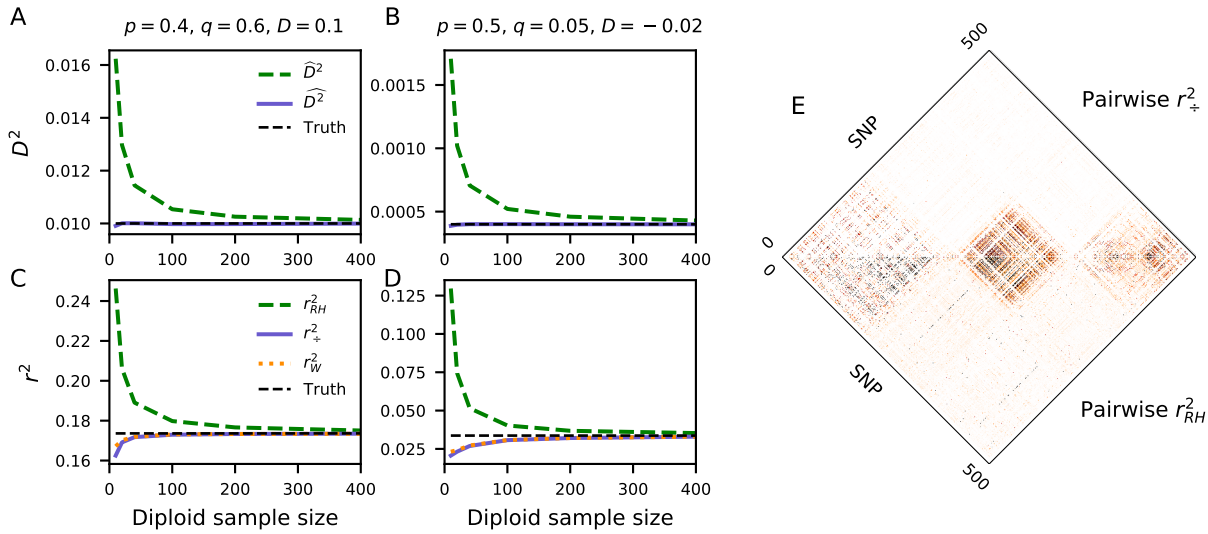


Figure 1: **LD estimation.** A-B: Computing D^2 by taking the square of the covariance overestimates the true value, while our approach is unbiased for any sample size. C-D: Similarly, computing r^2 by estimating r and squaring it (here, via the Rogers-Huff approach, r^2_{RH}) overestimates the true value. Waples proposed empirically estimating the bias due to finite sample size and subtracting this bias from estimated r^2. Our approach, $r^2_{\div}$ is less biased than r^2_{RH} and provides similar estimates to r^2_{W} without the need for ad hoc bias correction. (Additional comparisons found in Figure S3.) E: Pairwise comparison of r^2 for 500 neighboring SNPs in chromosome 22 in CHB from 1000 Genomes Project Consortium et al. (2015). $r^2_{\div}$ (top) and r^2_{RH} (bottom) are strongly correlated, although $r^2_{\div}$ displays less spurious background noise.

We use our estimators for D^2 and $p(1-p)q(1-q)$ to propose an estimator for r^2 from unphased data, which we denote $r^2_{\div} = \widehat{D^2}/\widehat{\pi}_2$, where $\pi_2 = p(1-p)q(1-q)$. $r^2_{\div}$ is a biased estimator for r^2, as we discuss below. However, it performs favorably in comparison to the common approach of first computing $\hat{r}$ (e.g. via Rogers and Huff (2009)) and simply squaring the result.

To explore the performance of this estimator, we first simulated differing diploid sample sizes with direct multinomial sampling from known haplotype frequencies (Figure 1A-D). Estimates of D^2 were unbiased as expected, and $r^2_{\div}$ converged to the true r^2 faster than the Rogers-Huff approach. Standard errors of our estimator were nearly indistinguishable from Rogers and Huff (2009) (Figure S1), and the variances of estimators for statistics in the Hill-Robertson system decayed with sample size as $\sim \frac{1}{n^2}$ (Figure S2). Second, we simulated 1 Mb segments of chromosomes under steady state demography (using `msprime` (Kelleher et al., 2016)) to estimate r^2 decay curves using both approaches. Our estimator was invariant to phasing and displayed the proper decay properties in the large recombination limit (Figure 2A). With increasing distance between SNPs, $r^2_{\div}$ approached zero as expected, while the Rogers-Huff r^2 estimates converged to positive values.

Finally, we computed the decay of r^2 across five population from the 1000 Genomes Project Consortium et al. (2015) (Figure 2B-D). $r^2_{\div}$ shows distinct qualitative behavior across populations, with recently admixed populations exhibiting distinctive long-range LD. However, r^2 as estimated using the Rogers-Huff approach displayed long-range LD in every population, confounding the signal of admixture in the shape of r^2 decay curves.

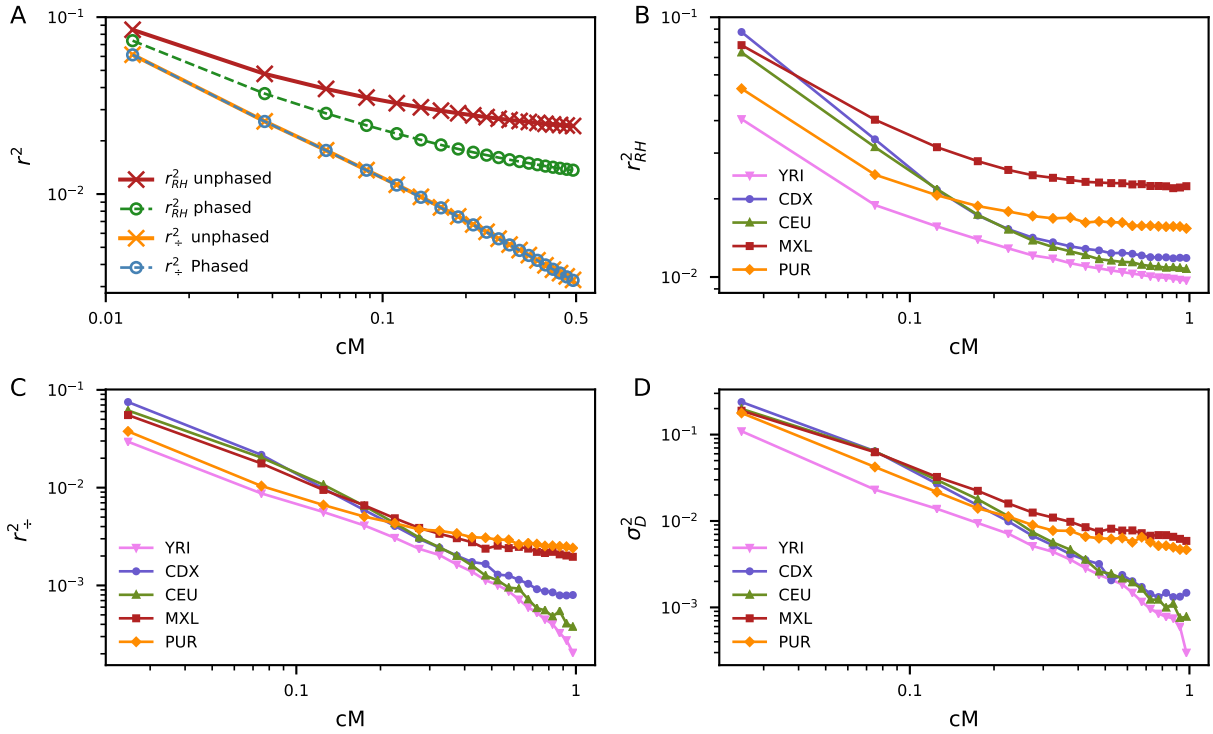


Figure 2: **Decay of r^2 with distance.** A: Comparison between our estimator ($r^2_{\div}$) and Rogers and Huff (2009) (RH) under steady state demography. The $r^2_{\div}$-curve displays the appropriate decay behavior and is invariant to phasing, while the RH approach produces upward biased r^2 and is sensitive to phasing. Estimates were computed from 1,000 1Mb replicate simulations with constant mutation and recombination rates (each 2×10^{-8} per base per generation) for $n = 50$ sampled diploid using `msprime` (Kelleher et al., 2016). B: r^2_{RH} decay for five populations in 1000 Genomes Project Consortium et al. (2015), with two putatively admixed American populations (MXL and PUR), computed from intergenic regions. C: $r^2_{\div}$ decay for the same populations. D: Decay of σ^2_D computed using $\widehat{D^2}/\widehat{\pi_2}$. The r^2_{RH} decay curves show excess long-range LD in each population, while our estimator qualitatively differentiates between populations. Our estimators for both r^2 and σ^2_D are more variable for large recombination distances because they measure a small quantity (population LD) rather than a larger one in the RH case (finite sample bias).

Predicting LD between unlinked loci

It is difficult to compute model predictions for r^2, even in simple evolutionary scenarios. Instead, here we consider $\sigma^2_D = \frac{\mathbb{E}[D^2]}{\mathbb{E}[\pi_2]}$ because we can compute its expectation from models (Hill and Robertson, 1968) and estimate it from data (as described here). For unlinked loci, such as loci on different chromosomes, the probability of the joint inheritance of alleles in a given generation is 1/2, so we can model pairs of unlinked loci by setting $c = 1/2$. We can use the high-order system of equations (Equation 3 in Hill and Robertson (1968)) to find σ^2_D at steady state, while accounting for sampling lineages from the population without replacement (Methods). This gives $\sigma^2_D \approx \frac{1}{6N_e}$, to leading order in $\frac{1}{N_e}$.

There is a subtle problem with this model that only become apparent at large recombination distances. The Hill-Robertson model, like the diffusion approach of Ohta and Kimura (1971), approximates a haploid Wright-Fisher model with recombination, where each offspring lineage in a given generation is formed by the independent pairing of gametes from the parental pool. In this model there is no memory of pairings of gametes as parental diploids, and we end up with an overabundance of quarter-siblings each generation,

which is biologically unrealistic.

In the Methods, we consider different mating systems to compute σ_D^2 for unlinked loci by enumerating the probability of two sampled lineages sharing a common ancestor t generations back and the probability of inheriting pairs of alleles IBD from the shared ancestor. In the 'haploid' model, equivalent to Hill and Robertson (1968), we recover $\sigma_D^2 = \frac{1}{6N}$. In a proper diploid Wright-Fisher model, where siblings are typically half-siblings, we find $\sigma_D^2 = \frac{1}{3N}$. Finally, in a model with full monogamy so that siblings are always full-siblings, we find $\sigma_D^2 = \frac{2}{3N}$.

Discussion

Sources of bias

To make inferences about the evolutionary history or biological processes that shaped observed LD, we must accurately measure LD from the data. Commonly used estimators feature a wide range of biases that make comparisons and interpretation difficult. These biases are due to sequencing and sampling (such as phasing and finite sample sizes), as well as fundamental issues with the statistics we choose to measure (such as comparing expectations of ratios to ratios of expectations).

Further problems in estimation can arise due to unknown population structure or relatedness between samples (Mangin et al., 2012), sequencing or alignment artifacts, low coverage depth, or the choice of minor allele frequency (MAF) cutoff (Hudson, 1985; McVean, 2002). r^2 is particularly sensitive to the choice of MAF cutoff and low coverage data.

Phasing

With phased data, we directly observe gametic haplotypes carried by each diploid genome, so that we can directly estimate haplotype frequencies by counting observations of each type. However, ambiguity arises in unphased data because we cannot directly count haplotypes. For a given pair of loci, an individual may carry AA, Aa, or aa at the left locus, and BB, Bb, or bb at the right locus, so that observed two-locus genotypes will be one of the nine types $\{AABB, AABb, AAbb, AaBB, \ldots, aabb\}$, In the case of the double heterozygote, $AaBb$, we don't know if the underlying gametes are AB/ab or Ab/aB. The RH estimator correctly accounts for phasing uncertainty in the context of finite samples. However, Figure 2 shows large residual biases in estimating r^2_{RH} in finite samples.

Finite sample

For any summary statistic, we compute estimates from a finite sample of the much larger population, which can lead to inflated or deflated estimates the true quantity. For illustration, heterozygosity (H) is the probability that any two sampled chromosomes differ at a given site, which is given by $2p(1-p)$, where p is the frequency of the derived allele A in the population. To estimate H over a region of length L, we can estimate $\tilde{p} = n_A/n$ for each variable position in this region and compute $\widetilde{H} = \frac{1}{L}\sum_i 2\tilde{p}_i(1-\tilde{p}_i)$. This is a biased estimator for H, and an unbiased estimator is $\widehat{H} = \frac{n}{n-1}\widetilde{H}$. LD estimates, as with other diversity statistics, may also be biased by finite samples. Correcting for this bias is a main focus of this article.

Computing the squared correlation

Computing estimates and expectations of ratios is challenging, and sometimes intractable. One commonly used approach to estimate r^2 is to first compute $\hat{r}$ via an EM algorithm (Excoffier and Slatkin, 1995) or

genotype covariances (Rogers and Huff, 2009), and then square the result. While we can compute unbiased estimators for r from either phased or unphased data, this approach gives inflated estimates of r^2 because it does not properly account for the variance in $\hat{r}$. In general, the expectation of a function of a random variable is not equal to the function of its expectation, or in our case, for given haplotype frequencies

$$r^2 \neq \mathbb{E}\left[\hat{r}^2\right].$$

For large enough sample sizes, this error will be practically negligible, but for small to moderate sample sizes, the estimates will be upwardly biased, sometimes drastically (Figures 1 and 2 show that there can be large residual biases).

Our approach is to instead compute estimators for D^2 and π_2 and compute their ratio for each pair of loci. Even though our estimates of both the numerator and denominator are unbiased, the ratio is still a biased estimator for r^2, since

$$r^2 \neq \mathbb{E}\left[\frac{\widehat{D^2}}{\widehat{\pi}_2}\right].$$

However, this estimator performs favorably to the Rogers-Huff approach (Figure 1C-D) and displays the appropriate decay behavior in the large recombination limit (Figure 2).

The limits of r^2

There are fundamental problems with using genome-wide patterns of r^2 in evolutionary inferences. r^2 is sensitive to the chosen MAF cutoff and low-coverage data, while σ_D^2 is more robust to low coverage data (Rogers, 2014; Ragsdale and Gravel, 2018). Estimates of r^2 also vary with sample size, hindering comparisons across studies and cohorts.

Compounding the issues of estimating r^2 from data, model-based predictions for $\mathbb{E}[r^2]$ are unavailable or difficult to compute in even the simplest scenarios (McVean, 2002; Song and Song, 2007; Rogers, 2014). This hinders any comparison between model and data. Motivated by this, Hill and Robertson (1968) introduced a system of ordinary differential equations to compute $\mathbb{E}[D^2]$, $\mathbb{E}[D(1-2p)(1-2q)]$, and $\mathbb{E}[p(1-p)q(1-q)]$. They used these estimates to compute an alternate measure of LD,

$$\sigma_D^2 = \frac{\mathbb{E}[D^2]}{\mathbb{E}[p(1-p)q(1-q)]},$$

which has the advantage that we can accurately compute the numerator and denominator from both data and models (Hill and Robertson, 1968; McVean, 2002; Rogers, 2014; Ragsdale and Gravel, 2018). While σ_D^2 is not generally a good approximation for r^2, it is itself informative about LD, is interpretable, and has more convenient statistical and computational properties.

Estimating N_e from LD

To illustrate the difference between estimates based on r^2 and σ_D^2, we consider the inference of N_e based on linkage disequilibrium. While analytic solutions for $E[r^2]$ are unavailable, Weir and Hill used a ratio of expectations to approximate

$$\mathbb{E}[r^2] \approx \frac{c^2 + (1-c)^2}{2N_e c(2-c)},$$

where c is the per generation recombination probability between two loci (Equation 3 in Weir and Hill (1980) due to Avery (1978)). Rearranging this equation provides an estimate for N_e if we can estimate r^2 from data. However, as pointed out by Waples (2006), failing to account for sample size bias when estimating r^2 leads to strong downward biases in $\hat{N}_e$. Waples used Burrows' Δ to estimate $\hat{r}^2_\Delta$ (again following Weir

and Hill (1980)) and used simulations to empirically estimate the bias $\mathrm{Var}(\hat{r}_\Delta)$ due to finite sample size. Subtracting the estimated bias from observed $\hat{r}^2_\Delta$ gives an empirically corrected estimate for r^2,

$$\hat{r}^2_{\mathrm{W}} \approx \hat{r}^2_\Delta - \mathrm{Var}(\hat{r}_\Delta).$$

Waples showed that $\hat{r}^2_{\mathrm{W}}$ removes much of the bias in N_e estimates (Figure 1A). Bulik-Sullivan et al. (2015) introduced a similar ad-hoc bias correction that appears to perform comparably to r^2_{W} and $r^2_{\div}$.

Using our estimators for $\widehat{D^2}$ and $\widehat{\pi}_2$, we can instead use σ^2_D to estimate N_e, which is more straightforward, requires fewer assumptions and approximations, and removes any need for ad hoc empirical bias correction. Ohta and Kimura (1969) (Equation 18) showed that at steady state, σ^2_D can be approximated as

$$\sigma^2_D \approx \frac{1}{3 + 4N_e c - 2/(2.5 + N_e c)}, \tag{1}$$

which is accurate for small mutation rates and for both large and small population sizes (Figure 3A). Because our estimators for D^2 and π_2 are unbiased, we can accurately estimate σ^2_D from the data. Rearranging Equation 1 provides a direct estimate for N_e (Figure 3B). Whereas the popular approach of Waples (2006) requires filtering out low-frequency variants because it uses approximations that are uncontrolled for rare variants, the σ^2_D approach requires neither MAF filtering nor empirical bias correction. Our estimator is also valid over a wide range of recombination distances separating loci ($0 \leq c \leq 0.5$).

Do we need unbiased estimators?

In inference, it is often more convenient to include the bias in the model than to derive unbiased statistics from the data. This is the approach taken by Waples (2006), Rogers (2014), and Ragsdale and Gravel (2018). This is also a convenient approach for simulation work. For example, Gutenkunst et al. (2009) verified demographic models inferred from the allele frequency spectrum by comparing equally biased r^2 estimates from coalescent simulations and data.

However, because bias dominates signal for all but the shortest recombination distances, comparisons using biased statistics miss relevant patterns of linkage disequilibrium (Figure 2). Furthermore, unbiased statistics are needed for comparisons between studies or cohorts with different sample sizes or with different patterns of missing data.

Finally, the benefits of using unbiased ratios of expectations (such as σ^2_D) rather than biased expectations of ratios (such as r^2) has also been pointed out for F_{ST} estimation where differences in biases across studies have led to incorrect conclusions (Bhatia et al., 2013). Bhatia et al. recommended replacing the "classical" F_{ST} by a ratio of expectations (confusingly also referred to as F_{ST}), because the latter is less sensitive to sample size and frequency cutoff differences. This better behavior across cohorts is a natural consequence of using unbiased estimators.

Limitations

Our estimators, particularly for unphased data, often contain many terms. For example, expanding $\mathbb{E}[D^2]$ as a monomial series in genotype frequencies results in nearly 100 terms. The algebra is straightforward, but writing the estimator down by hand would be a tedious exercise, and we used symbolic computation to simplify terms and avoid algebraic mistakes. This might explain why such estimators were not proposed for higher orders than D in the foundational work of LD estimation in 1970s and 80s. Deriving and computing estimators poses no problem for an efficiently written computer program that operates on observed genotype counts.

For very large sample sizes, the bias in the Rogers-Huff estimator for r^2 is weak, and it may be preferable to use their more straightforward approach. Additionally, it is worth noting that like many unbiased estimators,

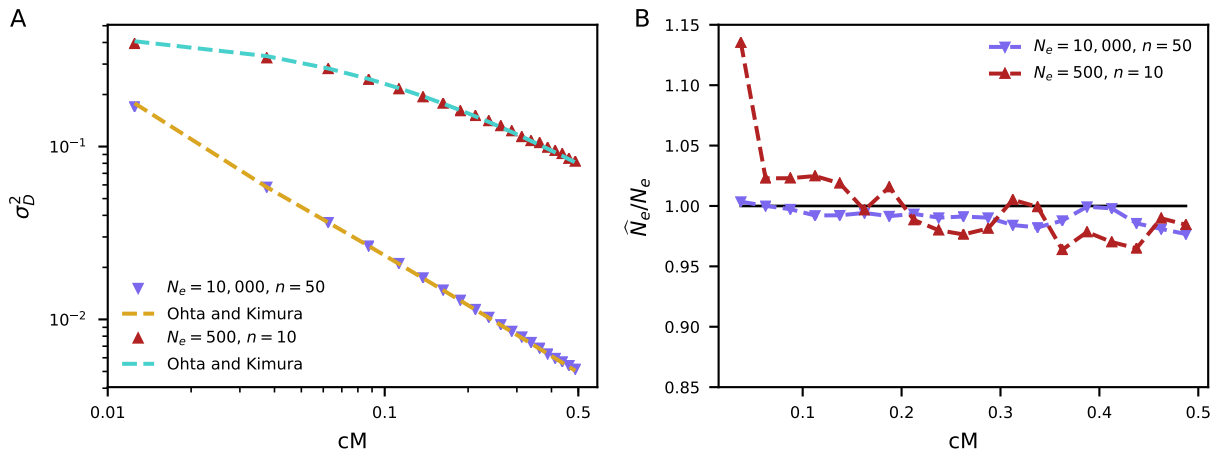


Figure 3: **Using σ_D^2 to estimate N_e.** A: The estimation due to Ohta and Kimura (1969) provides an accurate approximation for σ_D^2 for both large and small sample sizes. Here, we compare to the same simulations used in Figure 2A for N_e =10,000 with sample size $n = 50$ and $N_e = 500$ with sample size $n = 10$. B: Using σ_D^2 estimated from these same simulations and rearranging Equation 1 provides an estimate for N_e for each recombination bin. The larger variance for $N_e = 500$ is due to the small sample size leading to noise in estimated σ_D^2.

$r_{\div}^2$ can take values that exceed the range of r^2, so that for a given pair of loci $r_{\div}^2$ may be slightly negative or greater than one.

Throughout, we assumed populations to be randomly mating. Under inbreeding, there are multiple interpretations of D depending on whether we consider the covariance between two randomly drawn haplotypes from the population or consider two haplotypes within the same diploid individual (Cockerham and Weir, 1977) (In randomly mating populations these quantities are expected to be equal). Rogers and Huff (2009) proposed an estimator for D and r from genotype data with a known inbreeding fraction by considering genotype covariances when two gametes are identical by descent. We see no reason why our approach cannot be extended to account for inbreeding, although we leave those developments to future work.

Methods

Notation

Variables without decoration (hats/tildes) represent quantities estimated from the true population haplotype freuencies. We use tildes to represent statistics estimated by taking maximum likelihood estimates for allele frequencies from a finite sample: e.g. $\widetilde{p} = n_A/n$, $\widetilde{f}_{AB} = n_{AB}/n$, $\widetilde{\pi} = 2\widetilde{p}(1-\widetilde{p})$, etc. Hats represent unbiased estimates of quantities: e.g. $\widehat{\pi} = \frac{n}{n-1}\widetilde{\pi}$. f's denote haplotype frequencies in the population (f_{AB}, etc), while g's denote genotype frequencies (Table 1).

Estimating statistics from phased data

Suppose that we observe haplotype counts $(n_{AB}, n_{Ab}, n_{aB}, n_{ab})$, with $\sum n_j = n$, for a given pair of loci. Estimating LD in this case is straightforward. An unbiased estimator for D is

$$\widehat{D} = \frac{n}{n-1}\left(\frac{n_{AB}}{n}\frac{n_{ab}}{n} - \frac{n_{Ab}}{n}\frac{n_{aB}}{n}\right).$$

We can interpret the genome-wide $\mathbb{E}[D] = \mathbb{E}[f_{AB}f_{ab} - f_{Ab}f_{aB}]$ as the probability of drawing two chromosomes from the population and observing haplotype AB in the first sample and ab in the second, minus the probability of observing Ab followed by aB. This intuition leads us to the same estimator $\widehat{D}$:

$$\begin{aligned}\widehat{D} &= \frac{1}{2}\frac{\binom{n_{AB}}{1}\binom{n_{ab}}{1}}{\binom{n}{2}} - \frac{1}{2}\frac{\binom{n_{Ab}}{1}\binom{n_{aB}}{1}}{\binom{n}{2}} \\ &= \frac{n_{AB}}{n}\frac{n_{ab}}{n-1} - \frac{n_{Ab}}{n}\frac{n_{aB}}{n-1}.\end{aligned}$$

In this same way we can find an unbiased estimator for any two-locus statistic that can be expressed as a polynomial in haplotype frequencies. For example, the variance of D is

$$\begin{aligned}\mathbb{E}[D^2] &= E\left[(f_{AB}f_{AB} - f_{Ab}f_{aB})^2\right] \\ &= \mathbb{E}[f_{AB}^2 f_{AB}^2] + \mathbb{E}[f_{Ab}^2 f_{aB}^2] - 2\mathbb{E}[f_{AB}f_{Ab}f_{aB}f_{AB}],\end{aligned}$$

and Strobeck and Morgan (1978) and Hudson (1985) showed that each term can be interpreted as the probability of sampling the given ordered haplotype configuration in a sample of size four. An unbiased estimator for D^2 is then

$$\widehat{D^2} = \frac{1}{\binom{4}{2,0,0,2}}\frac{\binom{n_{AB}}{2}\binom{n_{ab}}{2}}{\binom{n}{4}} + \frac{1}{\binom{4}{0,2,2,0}}\frac{\binom{n_{Ab}}{2}\binom{n_{aB}}{2}}{\binom{n}{4}} - \frac{2}{\binom{4}{1,1,1,1}}\frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{1}\binom{n_{aB}}{1}\binom{n_{ab}}{1}}{\binom{n}{4}}.$$

The multinomial factors in front of each term account for the number of distinct orderings of the sampled haplotypes. We similarly find unbiased estimators for the other terms in the Hill-Robertson system, $\mathbb{E}[D(1-2p)(1-2q)]$ and $\mathbb{E}[p(1-p)q(1-q)]$ (shown in the appendix), or any other statistic that we compute from haplotype frequencies.

Estimating statistics from unphased data

Estimating two-locus statistics from genotype data is more involved because the underlying haplotypes are ambiguous in a double heterozygote, $AaBb$. Without direct knowledge of haplotype counts in a sample, we must instead turn to estimators from sample genotype counts $(n_1, \ldots, n_9)$ from underlying population genotype frequencies $(g_1, \ldots, g_9)$.

	BB	Bb	bb	Freq.
AA	g_1	g_2	g_3	p^2
Aa	g_4	g_5	g_6	$2p(1-p)$
aa	g_7	g_8	g_9	$(1-p)^2$
Freq.	q^2	$2q(1-q)$	$(1-q)^2$	1

Table 1: Genotype frequencies under random mating

Hill's iterative method

Weir and Cockerham (1979) proposed a maximum likelihood approach to estimate $\widetilde{f}_{AB}$, and thus compute $\widetilde{D} = \widetilde{f}_{AB} - \widetilde{p}\widetilde{q}$. Assuming random mating, a double heterozygote $AaBb$ individual carries underlying haplotypes AB/ab with probability $\frac{f_{AB}f_{ab}}{f_{AB}f_{ab}+f_{Ab}f_{aB}}$, and Ab/aB with probability $\frac{f_{Ab}f_{aB}}{f_{AB}f_{ab}+f_{Ab}f_{aB}}$. Hill wrote

$$\widetilde{f}_{AB} = \widetilde{g}_1 + \frac{1}{2}\widetilde{g}_2 + \frac{1}{2}\widetilde{g}_4 + \frac{1}{2}\widetilde{g}_5\frac{f_{AB}f_{ab}}{f_{AB}f_{ab} + f_{Ab}f_{aB}}.$$

Each f is unknown, but we can estimate $\widetilde{f}_{Ab} = \widetilde{p} - \widetilde{f}_{AB}$, $\widetilde{f}_{aB} = \widetilde{q} - \widetilde{f}_{AB}$, and $\widetilde{f}_{ab} = 1 - \widetilde{p} - \widetilde{q} + \widetilde{f}_{AB}$, to get an equation with $\widetilde{f}_{AB}$ as the only unknown:

$$\widetilde{f}_{AB} = \widetilde{g}_1 + \frac{1}{2}\widetilde{g}_2 + \frac{1}{2}\widetilde{g}_4 + \frac{1}{2}\widetilde{g}_5 \frac{\widetilde{f}_{AB}(1 - \widetilde{p} - \widetilde{q} + \widetilde{f}_{AB})}{\widetilde{f}_{AB}(1 - \widetilde{p} - \widetilde{q} + \widetilde{f}_{AB}) + (\widetilde{p} - \widetilde{f}_{AB})(\widetilde{q} - f_{AB})}.$$

This is a cubic equation, which can be solved numerically or iteratively, and gives the maximum likelihood estimate for f_{AB} under a multinomial sampling likelihood, and thus estimate D and r. The EM algorithm implemented in Excoffier and Slatkin (1995) follows this approach and sees wide use.

Weir's composite measure and Rogers and Huff's covariance method

Weir described Burrows' "composite" measure of linkage disequilibrium (see Weir (1996), page 126, or Weir (1979)), which is denoted Δ. Given whole-population genotype frequencies Table 1, Weir defines

$$\Delta = \left(2g_1 + g_2 + g_4 + \frac{1}{2}g_5\right) - 2pq. \tag{2}$$

Under random mating, Δ equals D, which we can see by noting that, since $g_5 = 2f_{AB}f_{ab} + 2f_{Ab}f_{aB}$ and $D = f_{AB}f_{ab} - f_{Ab}f_{aB}$, we can write

$$f_{AB}f_{ab} = \frac{1}{4}g_5 + \frac{1}{2}D.$$

Then we can list all possible ways that the AB haplotype occurs,

$$f_{AB} = g_1 + \frac{1}{2}g_2 + \frac{1}{2}g_4 + \left(\frac{1}{4}g_5 + \frac{1}{2}D\right). \tag{3}$$

Using this and $D = f_{AB} - pq$, we can therefore rewrite Equation 2 as

$$\Delta = 2\left(f_{AB} - \frac{1}{2}D\right) - 2pq = D.$$

Weir (2008) and Rogers and Huff (2009) point out that Δ can by found by directly computing the covariance of observed genotype counts between two loci. Set Y as genotype values at the left locus and Z as genotype values at the right locus ($Y, Z \in \{0, 1, 2\}$). Then, since each observed $Y = y_1 + y_2$, where y_i are the gametic values (0 or 1), and $Z = z_1 + z_2$,

$$\mathrm{Cov}(Y, Z) = \mathrm{Cov}(y_1, z_1) + \mathrm{Cov}(y_1, z_2) + \mathrm{Cov}(y_2, z_1) + \mathrm{Cov}(y_2, z_2).$$

Again, under random mating, $\mathrm{Cov}(y_1, z_2) = \mathrm{Cov}(y_2, z_1) = 0$, and $\Delta = \frac{1}{2}\mathrm{Cov}(Y, Z)$, which Rogers and Huff (2009) then use to estimate r. This is the approach taken by Loh et al. (2013) and used in their program ALDER to estimate the parameters of recent admixture events, while Moorjani et al. (2011) uses the correlation coefficient r.

Unbiased estimation of two-locus statistics

We take a similar approach to Weir's composite estimate. We first express expected statistics as polynomials in genotype frequencies g_i and then compute a finite-sample estimate of each monomial in terms of the observed genotype counts n_i. We define "naive" estimators of haplotype frequencies f as if the g_5 genotype was equally likely to be composed of the AB and ab haplotypes as of the Ab and aB haplotypes:

$$x_{AB} = g_1 + \frac{1}{2}g_2 + \frac{1}{2}g_4 + \frac{1}{4}g_5.$$

The naive frequencies x_{Ab}, x_{aB}, and x_{ab} can be defined in a similar way. These are biased by a factor $\pm\frac{D}{2}$:

$$f_\star = x_\star \pm D/2,$$

as in Equation 3. Then with a bit of algebra we can recover Weir's result,

$$\mathbb{E}[D] = \mathbb{E}[\Delta] = 2\mathbb{E}[x_{AB}x_{ab} - x_{Ab}x_{aB}].$$

Using $p = x_{AB} + x_{Ab}$, $q = x_{AB} + x_{aB}$, we can similarly write the Hill-Robertson statistics as

$$\begin{aligned}
\mathbb{E}[D^2] &= 4\mathbb{E}[(x_{AB}x_{ab} - x_{Ab}x_{aB})^2], \\
\mathbb{E}[D(1-2p)(1-2q)] &= 2\mathbb{E}[(x_{AB}x_{ab} - x_{Ab}x_{aB})(x_{aB} + x_{Ab} - x_{AB} - x_{Ab})(x_{Ab} + x_{ab} - x_{AB} - x_{aB})], \\
\mathbb{E}[p(1-p)q(1-q)] &= \mathbb{E}[(x_{AB} + x_{Ab})(x_{aB} + x_{ab})(x_{AB} + x_{aB})(x_{Ab} + x_{ab})].
\end{aligned}$$

Given a statistic written as a polynomial in the $x_{..}$, $S = \mathbb{E}[h(x_{AB}, x_{Ab}, x_{aB}, x_{ab})]$, we can expand the expectation as a monomial series in genotype frequencies g_j, $j = 1, \ldots, 9$:

$$\mathbb{E}[S] = \sum_i E\left[a_i \prod_{j=1}^{9} {g_{j,i}}^{k_{j,i}}\right].$$

Each term of the form $a_i \prod {g_{j,i}}^{k_{j,i}}$ can be interpreted as the probability of drawing $k = \sum k_j$ diploid samples, and observing the ordered configuration of k_1 of type g_1, k_2 of type g_2, and so on. Then, from a diploid sample size of $n \geq k$, this term has the unbiased estimator

$$a_i \frac{1}{\binom{k_i}{k_{1,i},\ldots,k_{9,i}}} \frac{\binom{n_{1,i}}{k_{1,i}} \cdots \binom{n_{9,i}}{k_{9,i}}}{\binom{n_i}{k_i}}.$$

Summing over all terms gives us an unbiased estimator for S:

$$\widehat{S} = \sum_i a_i \frac{1}{\binom{k_i}{k_{1,i},\ldots,k_{9,i}}} \frac{\binom{n_{1,i}}{k_{1,i}} \cdots \binom{n_{9,i}}{k_{9,i}}}{\binom{n_i}{k_i}}. \tag{4}$$

For example, D has the unbiased estimator

$$\begin{aligned}
\widehat{\Delta} = \frac{1}{n(n-1)} \Bigg[&\left(n_1 + \frac{n_2}{2} + \frac{n_4}{2} + \frac{n_5}{4}\right)\left(\frac{n_5}{4} + \frac{n_6}{2} + \frac{n_8}{2} + n_9\right) \\
&- \left(\frac{n_2}{2} + n_3 + \frac{n_5}{4} + \frac{n_6}{2}\right)\left(\frac{n_4}{2} + \frac{n_5}{4} + n_7 + \frac{n_8}{2}\right)\Bigg],
\end{aligned}$$

which simplifies to the known Burrows (1979) estimator,

$$\widehat{\Delta} = \frac{n}{n-1}\widetilde{\Delta}.$$

For statistics of higher order than D, such as those in the Hill-Robertson system, expanding these statistics often involves a large number of terms. In practice, we use symbolic computation software to compute our estimators. In some cases the estimators simplify into compact expressions, although in other cases they may remain expansive. However, even when there are many terms, the sums do not consist of large terms of alternating sign, and so computation is stable.

Expected LD between unlinked loci

Via Hill-Robertson statistics and hypergeometric sampling

Equations 2 and 3 from Hill and Robertson (1968) give a linear recursion on the statistics $\mathbb{E}[D^2]$, $\mathbb{E}[D(1-2p)(1-2q)]$, and $\mathbb{E}[p(1-p)q(1-q)]$, which can be analytically solved at equilibrium (with an influx of mutation, here under the infinite sites model). In Ragsdale and Gravel (2018), we interpreted these two-locus statistics as sampling probabilities, similar to the identity coefficients considered by Hudson (1985) and McVean (2002). Specifically, each term in $\mathbb{E}[D^2] = \mathbb{E}[f_{AB}^2 f_{ab}^2] - 2\mathbb{E}[f_{AB} f_{Ab} f_{aB} f_{ab}] + \mathbb{E}[f_{Ab}^2 f_{aB}^2]$ can be thought of as drawing a particular number of each haplotype, in order. For example, the first term is the probability of drawing two haploid samples that are each AB followed by two haploid samples that are ab.

As written in the Hill-Robertson equations, these are probabilities of sampling lineages *with replacement*. In practice we don't sample and sequence the same individual multiple times and consider each instance to be an independent sample. Instead we sample *without replacement* from the full population.

One approach to account for this difference in sampling strategies is to adjust the expected Hill-Robertson statistics via hypergeometric sampling with known census population size. To do this, we expand each of the Hill-Robertson statistics as expectations over monomial terms in haplotype frequencies, rewrite each term via the hypergeometric distribution with a factor to account for the sampling order, and then simplify again in terms of the Hill-Robertson statistics. For D^2, for example, we expand and rewrite the sampling probability as

$$\frac{1}{\binom{4}{2,0,0,2}} \frac{\binom{2Nf_{AB}}{2}\binom{2Nf_{ab}}{2}}{\binom{2N}{4}} - \frac{2}{\binom{4}{1,1,1,1}} \frac{\binom{2Nf_{AB}}{1}\binom{2Nf_{Ab}}{1}\binom{2Nf_{aB}}{1}\binom{2Nf_{ab}}{1}}{\binom{2N}{4}} + \frac{1}{\binom{4}{0,2,2,0}} \frac{\binom{2Nf_{Ab}}{2}\binom{2Nf_{aB}}{2}}{\binom{2N}{4}}.$$

All together, writing

$$\mathbf{y} = \begin{pmatrix} \mathbb{E}[D^2] \\ \mathbb{E}[D(1-2p)(1-2q)] \\ \mathbb{E}[p(1-p)q(1-q)] \end{pmatrix},$$

we have

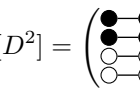
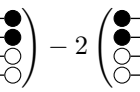
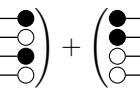
$$\mathbf{y}_{\text{w/o repl.}} = \frac{1}{(2N-1)(2N-3)(N-1)} \begin{pmatrix} 4N^3 - 6N^2 + 2N & -2N^2 + N & -2N^2 + 2N \\ -8N^2 + 4N & 4N^3 - 2N^2 + 2N & 4N \\ 2N & -2N^2 + N & 4N^3 - 8N^2 + 2N \end{pmatrix} \mathbf{y}_{\text{H-R}}.$$

From this, we can compute σ_D^2 for unlinked loci ($c = 1/2$) under the Hill-Robertson equations assuming sampling without replacement, which gives

$$\sigma_D^2 = \frac{1}{6N} - \frac{1}{6N^2} + o\left(\frac{1}{N^2}\right). \tag{5}$$

Haploid inheritance model

Here, we will approximate σ_D^2 for unlinked loci through an argument based on the probability that two sampled haplotypes are IBD through a recent common. This approach will rely on a few assumptions. First consider $\mathbb{E}[D^2] = \mathbb{E}[f_{AB}^2 f_{ab}^2] - 2\mathbb{E}[f_{AB} f_{Ab} f_{aB} f_{ab}] + \mathbb{E}[f_{Ab}^2 f_{aB}^2]$, which we illustrate as

$E[D^2] =$  $- 2$  $+$  . (6)

Again, the terms can be thought of as sampling probabilities within a sample of four haplotypes from the population.

In a sample of four lineages, if they do not share a recent common ancestor we assume that recombination splits each lineage before any coalescent event so that the left and right loci coalesce independently. In this case, each term is equally probable, and $\mathbb{E}[D^2] = 0$.

Alternatively, two of the sampled lineages may share a recent common ancestor (we neglect the case of more than two lineages sharing very recent common ancestors). If two lineages share a recent common ancestor, the sample are expected to contribute nonzero D^2 only if the two loci were inherited identical-by-decent (IBD) from the common ancestor. In the case that two lineages are IBD at both loci, we can think of drawing three haplotypes at random (without replacement) from the population, and then copying one of them. The middle term in Equation 6 is impossible if two of the haplotypes are IBD, so we can write

$$\begin{aligned}
\mathbb{E}[D^2] = \Bigg[&\left(\begin{smallmatrix}\bullet\!-\!\bullet\\ \circ\!-\!\circ\\ \circ\!-\!\circ\end{smallmatrix}\right) \cdot P(\text{copy } \bullet\!-\!\bullet \text{ IBD}) + \left(\begin{smallmatrix}\bullet\!-\!\bullet\\ \bullet\!-\!\bullet\\ \circ\!-\!\circ\end{smallmatrix}\right) \cdot P(\text{copy } \circ\!-\!\circ \text{ IBD}) \\
&+ \left(\begin{smallmatrix}\bullet\!-\!\circ\\ \circ\!-\!\bullet\\ \circ\!-\!\bullet\end{smallmatrix}\right) \cdot P(\text{copy } \bullet\!-\!\circ \text{ IBD}) + \left(\begin{smallmatrix}\bullet\!-\!\circ\\ \bullet\!-\!\circ\\ \circ\!-\!\bullet\end{smallmatrix}\right) \cdot P(\text{copy } \circ\!-\!\bullet \text{ IBD}) \Bigg] \\
= \Bigg[&\left(\begin{smallmatrix}\bullet\!-\!\bullet\\ \circ\!-\!\circ\\ \circ\!-\!\circ\end{smallmatrix}\right) + \left(\begin{smallmatrix}\bullet\!-\!\bullet\\ \bullet\!-\!\bullet\\ \circ\!-\!\circ\end{smallmatrix}\right) + \left(\begin{smallmatrix}\bullet\!-\!\circ\\ \circ\!-\!\bullet\\ \circ\!-\!\bullet\end{smallmatrix}\right) + \left(\begin{smallmatrix}\bullet\!-\!\circ\\ \bullet\!-\!\circ\\ \circ\!-\!\bullet\end{smallmatrix}\right) \Bigg] \\
&\cdot P\left(\text{IBD} \mid 2 \text{ lineages share ancestry}\right) \cdot P\left(\text{Shared ancestry}\right).
\end{aligned}$$

Since the two loci are unlinked, we can assume that population-wide haplotype frequencies are equal to the product of the allele probabilities of the two alleles at the two loci, i.e. $P\left(\bullet\!-\!\bullet\right) = P\left(\bullet\!-\ \right) \cdot P\left(\ -\!\bullet\right)$. First we observe

$$\begin{aligned}
\left[\left(\begin{smallmatrix}\bullet\!-\!\bullet\\ \circ\!-\!\circ\\ \circ\!-\!\circ\end{smallmatrix}\right) + \left(\begin{smallmatrix}\bullet\!-\!\bullet\\ \bullet\!-\!\bullet\\ \circ\!-\!\circ\end{smallmatrix}\right) + \left(\begin{smallmatrix}\bullet\!-\!\circ\\ \circ\!-\!\bullet\\ \circ\!-\!\bullet\end{smallmatrix}\right) + \left(\begin{smallmatrix}\bullet\!-\!\circ\\ \bullet\!-\!\circ\\ \circ\!-\!\bullet\end{smallmatrix}\right) \right] &= p(1-p)^2 q(1-q)^2 + p^2(1-p)q^2(1-q) \\
&\quad + p(1-p)^2 q^2 (1-q) + p^2(1-p)q(1-q)^2 \\
&= p(1-p)q(1-q) \\
&= \pi_2.
\end{aligned}$$

We now ask what the probability is that two lineages share a common ancestor some time t in the past. In the haploid model, there is no memory of diploid pairings for parents, so that gametes are drawn independently from the population ($2N$ total) to form offspring. In the recent past, a gamete has 2^t genealogical ancestors. The second gamete also has 2^t ancestors, so each of those ancestors has a $\frac{2^t}{2N}$ chance of being the same as one of the ancestors to the first gamete. Thus, the probability of sharing a common ancestor t generations ago is approximately

$$P(\text{shared ancestor} \mid t) = \frac{2^{2t}}{2N}.$$

Finally, we compute the probability that both loci were inherited IBD, given that two lineages shared a genealogical ancestor t generations ago. In a single generation, there is a 1/4 chance that both loci are inherited from the same parental gamete. Thus, the probability that two lineages are IBD from a common ancestor t generations ago is

$$P(\text{IBD} \mid \text{shared ancestor } t \text{ gen. ago}) = \frac{1}{4^{2t}}.$$

We take this all together, summing over generations, to find

$$\begin{aligned}
\mathbb{E}[D^2] &= \pi_2 \sum_{t=1}^{\infty} \frac{2^{2t}}{2N} \frac{1}{4^{2t}} \\
&= \frac{\pi_2}{2N} \sum_{t=1}^{\infty} \frac{1}{2^{2t}} \\
&= \frac{\pi_2}{6N}.
\end{aligned}$$

From this, in a model in which we do not explicitly create diploid parents,

$$\sigma_D^2 = \frac{1}{6N},$$

matching to leading order the result of Hill and Robertson (1968) with sampling without replacement (Equation 5).

Diploid inheritance model

In the diploid Wright-Fisher model, we assume that parents are randomly mating diploid individuals. The difference between this model and the haploid model from the previous section is that while parents are randomly drawn (from N total individuals), parents have fixed gametic states. This changes the calculation of the probability of sharing a genealogical ancestor, as we now draw two diploids instead of four haploids to generate offspring. It also changes the probability of sharing two unlinked loci IBD.

From the decomposition of $\mathbb{E}[D^2]$ for unlinked loci,

$$\mathbb{E}[D^2] = \pi_2 \cdot P\left(\text{IBD} \mid \text{2 lineages share ancestry}\right) \cdot P\left(\text{Shared ancestry}\right),$$

we first compute the probability of sharing a common ancestor t generations ago, and then compute the probability of inheriting a pair of unlinked loci IBD from that common ancestor.

First, each sampled haplotype has 2^{t-1} ancestors, so any two given samples share a common ancestor in the recent past with probability $\frac{2^{2t-2}}{N}$. Two offspring of the common ancestor are IBD at two unlinked loci with probability $\frac{1}{4}$. Additional generations reduce the probability of inheriting both loci IBD by $\left(\frac{1}{4}\right)^2$, so that the probability of begin IBD at the two loci from a common ancestor t generations ago is $\frac{1}{2^{4t-2}}$.

Combining terms gives us

$$\begin{aligned}
\mathbb{E}[D^2] &= \pi_2 \sum_{t=1}^{\infty} \frac{2^{2t-2}}{N} \frac{1}{2^{4t-2}} \\
&= \frac{\pi_2}{N} \sum_{t=1}^{\infty} \frac{1}{2^{2t}} \\
&= \frac{\pi_2}{3N},
\end{aligned}$$

so that

$$\sigma_D^2 = \frac{1}{3N}.$$

Monogamy

The model for monogamy is similar to the randomly mating diploid inheritance model in the previous section. Parental pairs are fixed, so siblings always share both parents, and the probability of two randomly drawn haplotypes sharing ancestral (grand)parents t generations ago is $\frac{2^{2(t-1)}}{N/2} = \frac{2^{2t-1}}{N}$. The probability of being IBD at both loci, given that (grand)parents were shared t generations ago, is $\frac{1}{2^{2(2t-1)}} = \frac{1}{2^{4t-2}}$. Taken together, we find

$$\begin{aligned}
\sigma_D^2 &= \frac{2}{N} \sum_{t=1}^{\infty} \frac{1}{2^{2t}} \\
&= \frac{2}{3N}.
\end{aligned}$$

Software

Code to compute two-locus statistics in the Hill-Robertson system is packaged with our software `moments.LD`, a python program that computes expected LD statistics with flexible evolutionary models and performs likelihood-based demographic inference (`https://bitbucket.org/simongravel/moments`). Code used to compute and simplify unbiased estimators and python scripts to recreate analyses and figures in this manuscript can be found at `https://bitbucket.org/aragsdale/estimateld`.

Acknowledgements

We thank Lounès Chikhi, Mandy Yao, Alex Diaz-Papkovich, and Chris Gignoux for useful discussions. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program, the NSERC discovery grant, and CIHR MOP-136855.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Avery, P. J. (1978). The effect of finite population size on models of linked overdominant loci. *Genetical Research*, 31(3):239–254.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23(9):1514–1521.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.

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

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921–927.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695.

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231.

Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, 109(3):611–631.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5):1–22.

Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–1254.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, 108(3):285–291.

McVean, G. A. T. (2002). A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–991.

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. L., and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics*, 7(4):e1001373.

Ohta, T. and Kimura, M. (1969). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, 63(1):229–238.

Ohta, T. and Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, 68(4):571–580.

Ragsdale, A. P. and Gravel, S. (2018). Models of archaic admixture and recent history from two-locus statistics. *bioRxiv*, doi: 10.1101/489401:doi:10.1101/489401.

Rogers, A. R. (2014). How population growth affects linkage disequilibrium. *Genetics*, 197(4):1329–1341.

Rogers, A. R. and Huff, C. (2009). Linkage disequilibrium between loci with unknown phase. *Genetics*, 182(3):839–844.

Song, Y. S. and Song, J. S. (2007). Analytic computation of the expectation of the linkage disequilibrium coefficient r2. *Theoretical Population Biology*, 71(1):49–60.

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, 91(6):1011–1021.

Strobeck, C. and Morgan, K. (1978). The effect of intragenic recombination on the number of alleles in a finite population. *Genetics*, Vol. 88(4 (I)):829–844.

Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, 7(2):167–184.

Weir, B. (2008). Linkage Disequilibrium and Association Mapping. *Annual Review of Genomics and Human Genetics*, 9(1):129–142.

Weir, B. S. (1979). Inferences about linkage disequilibrium. *Biometrics*, 35(1):235–254.

Weir, B. S. (1996). *Genetic Data Analysis II.* Sinauer Associates, Inc., Sunderland, MA, 2 edition.

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

Weir, B. S. and Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics*, 95(2):477–88.

Appendix

Unbiased estimators from haplotype data

In Methods, we showed how to estimate D^2 from phased data. We can similarly estimate the Hill-Robertson statistics $Dz = D(1-2p)(1-2q)$ and $\pi_2 = p(1-p)q(1-q)$:

$$\begin{aligned}
\mathbb{E}[Dz] &= E\left[(f_{AB}f_{ab} - f_{Ab}f_{aB})(f_{AB} + f_{Ab} - f_{aB} - f_{ab})(f_{AB} + f_{aB} - f_{Ab} - f_{ab})\right] \\
&= \mathbb{E}[f_{AB}^3 f_{ab}] - \mathbb{E}[f_{AB}^2 f_{Ab} f_{aB}] - 2\mathbb{E}[f_{AB}^2 f_{ab}^2] - \mathbb{E}[f_{AB} f_{Ab}^2 f_{ab}] \\
&\quad + 4\mathbb{E}[f_{AB} f_{Ab} f_{aB} f_{ab}] - \mathbb{E}[f_{AB} f_{aB}^2 f_{ab}] + \mathbb{E}[f_{AB} f_{ab}^3] + \mathbb{E}[f_{Ab}^3 f_{aB}] \\
&\quad - 2\mathbb{E}[f_{Ab}^2 f_{aB}^2] + \mathbb{E}[f_{Ab} f_{aB}^3] - \mathbb{E}[f_{Ab} f_{aB} f_{ab}^2]
\end{aligned}$$

and

$$
\begin{aligned}
\mathbb{E}[\pi_2] &= E\left[(f_{AB}+f_{Ab})(f_{aB}+f_{ab})(f_{AB}+f_{aB})(f_{Ab}+f_{ab})\right] \\
&= \mathbb{E}[f_{AB}^2 f_{Ab} f_{aB}] + \mathbb{E}[f_{AB}^2 f_{Ab} f_{ab}] + \mathbb{E}[f_{AB}^2 f_{aB} f_{ab}] + \mathbb{E}[f_{AB}^2 f_{ab}^2] \\
&\quad + \mathbb{E}[f_{AB} f_{Ab}^2 f_{aB}] + \mathbb{E}[f_{AB} f_{Ab}^2 f_{ab}] + \mathbb{E}[f_{AB} f_{Ab} f_{aB}^2] + 2\mathbb{E}[f_{AB} f_{Ab} f_{aB} f_{ab}] \\
&\quad + \mathbb{E}[f_{AB} f_{Ab} f_{ab}^2] + \mathbb{E}[f_{AB} f_{aB}^2 f_{ab}] + \mathbb{E}[f_{AB} f_{aB} f_{ab}^2] + \mathbb{E}[f_{Ab}^2 f_{aB}^2] \\
&\quad + \mathbb{E}[f_{Ab}^2 f_{aB} f_{ab}] + \mathbb{E}[f_{Ab} f_{aB}^2 f_{ab}] + \mathbb{E}[f_{Ab} f_{aB} f_{ab}^2]
\end{aligned}
$$

Unbiased estimators for both of these statistics are

$$
\begin{aligned}
\widehat{Dz} &= \frac{1}{\binom{4}{3,0,0,1}} \frac{\binom{n_{AB}}{3}\binom{n_{ab}}{1}}{\binom{n}{4}} - \frac{1}{\binom{4}{2,1,0,1}} \frac{\binom{n_{AB}}{2}\binom{n_{Ab}}{1}\binom{n_{ab}}{1}}{\binom{n}{4}} - 2\frac{1}{\binom{4}{2,0,0,2}} \frac{\binom{n_{AB}}{2}\binom{n_{ab}}{2}}{\binom{n}{4}} \\
&\quad - \frac{1}{\binom{4}{1,2,0,1}} \frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{2}\binom{n_{ab}}{1}}{\binom{n}{4}} + 4\frac{1}{\binom{4}{1,1,1,1}} \frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{1}\binom{n_{aB}}{1}\binom{n_{ab}}{1}}{\binom{n}{4}} \\
&\quad - \frac{1}{\binom{4}{1,0,2,1}} \frac{\binom{n_{AB}}{1}\binom{n_{aB}}{2}\binom{n_{ab}}{1}}{\binom{n}{4}} + \frac{1}{\binom{4}{1,0,0,3}} \frac{\binom{n_{AB}}{1}\binom{n_{ab}}{3}}{\binom{n}{4}} + \frac{1}{\binom{4}{0,3,1,0}} \frac{\binom{n_{Ab}}{3}\binom{n_{aB}}{1}}{\binom{n}{4}} \\
&\quad - 2\frac{1}{\binom{4}{0,2,2,0}} \frac{\binom{n_{Ab}}{2}\binom{n_{aB}}{2}}{\binom{n}{4}} + \frac{1}{\binom{4}{0,1,3,0}} \frac{\binom{n_{Ab}}{1}\binom{n_{aB}}{3}}{\binom{n}{4}} - \frac{1}{\binom{4}{0,1,1,2}} \frac{\binom{n_{Ab}}{1}\binom{n_{aB}}{1}\binom{n_{ab}}{2}}{\binom{n}{4}} \\
&= \frac{1}{n(n-1)(n-2)(n-3)} \big((n_{AB}n_{ab} - n_{Ab}n_{aB})(n_{aB}+n_{ab}-n_{AB}-n_{Ab})(n_{Ab}+n_{ab}-n_{AB}-n_{aB}) \\
&\qquad - n_{AB}n_{ab}(n_{AB}+n_{ab}-n_{Ab}-n_{aB}-2) - n_{Ab}n_{aB}(n_{Ab}+n_{aB}-n_{AB}-n_{ab}-2)\big)
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{\pi}_2 &= \frac{1}{\binom{4}{2,1,1,0}} \frac{\binom{n_{AB}}{2}\binom{n_{Ab}}{1}\binom{n_{aB}}{1}}{\binom{n}{4}} + \frac{1}{\binom{4}{2,1,0,1}} \frac{\binom{n_{AB}}{2}\binom{n_{Ab}}{1}\binom{n_{ab}}{1}}{\binom{n}{4}} + \frac{1}{\binom{4}{2,0,1,1}} \frac{\binom{n_{AB}}{2}\binom{n_{aB}}{1}\binom{n_{ab}}{1}}{\binom{n}{4}} \\
&\quad + \frac{1}{\binom{4}{2,0,0,2}} \frac{\binom{n_{AB}}{2}\binom{n_{ab}}{2}}{\binom{n}{4}} + \frac{1}{\binom{4}{1,2,1,0}} \frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{2}\binom{n_{aB}}{1}}{\binom{n}{4}} + \frac{1}{\binom{4}{1,2,0,1}} \frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{2}\binom{n_{ab}}{1}}{\binom{n}{4}} \\
&\quad + \frac{1}{\binom{4}{1,1,2,0}} \frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{1}\binom{n_{aB}}{2}}{\binom{n}{4}} + 2\frac{1}{\binom{4}{1,1,1,1}} \frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{1}\binom{n_{aB}}{1}\binom{n_{ab}}{1}}{\binom{n}{4}} \\
&\quad + \frac{1}{\binom{4}{1,1,0,2}} \frac{\binom{n_{AB}}{1}\binom{n_{Ab}}{1}\binom{n_{ab}}{2}}{\binom{n}{4}} + \frac{1}{\binom{4}{1,0,2,1}} \frac{\binom{n_{AB}}{1}\binom{n_{aB}}{2}\binom{n_{ab}}{1}}{\binom{n}{4}} + \frac{1}{\binom{4}{1,0,1,2}} \frac{\binom{n_{AB}}{1}\binom{n_{aB}}{1}\binom{n_{ab}}{2}}{\binom{n}{4}} \\
&\quad + \frac{1}{\binom{4}{0,2,2,0}} \frac{\binom{n_{Ab}}{2}\binom{n_{aB}}{2}}{\binom{n}{4}} + \frac{1}{\binom{4}{0,2,1,1}} \frac{\binom{n_{Ab}}{2}\binom{n_{aB}}{1}\binom{n_{ab}}{1}}{\binom{n}{4}} \\
&\quad + \frac{1}{\binom{4}{0,1,2,1}} \frac{\binom{n_{Ab}}{1}\binom{n_{aB}}{2}\binom{n_{ab}}{1}}{\binom{n}{4}} + \frac{1}{\binom{4}{0,1,1,2}} \frac{\binom{n_{Ab}}{1}\binom{n_{aB}}{1}\binom{n_{ab}}{2}}{\binom{n}{4}} \\
&= \frac{1}{n(n-1)(n-2)(n-3)} \big((n_{AB}+n_{Ab})(n_{aB}+n_{ab})(n_{AB}+n_{aB})(n_{Ab}+n_{ab}) \\
&\qquad - n_{AB}n_{ab}(n_{AB}+n_{ab}+3n_{Ab}+3n_{aB}-1) - n_{Ab}n_{aB}(n_{Ab}+n_{aB}+3n_{AB}+3n_{ab}-1)\big).
\end{aligned}
$$

Multiple populations

We recently presented an extension to the Hill-Robertson system for D^2 to compute expected two-locus statistics across multiple populations, which can be computed rapidly for many related populations with

complex demography (Ragsdale and Gravel, 2018). The natural multi-population extension to the Hill-Robertson system (D^2, Dz, π_2) is on the basis,

$$\mathbf{y} = \begin{pmatrix} \mathbb{E}[D_i D_j] \\ \mathbb{E}[D_i z_{jk}] \\ \mathbb{E}[\pi_{2ijkl}] \end{pmatrix},$$

where $1 \leq i, j, k, l \leq P$ index populations, $z_{jk} = (1 - 2p_j)(1 - 2q_k)$, and

$$\pi_{2ijkl} = \begin{cases} p_i(1-p_i)q_k(1-q_k), & i = j, k = l \\ \frac{1}{2}p_i(1-p_j)q_k(1-q_k) + \frac{1}{2}p_j(1-p_i)q_k(1-q_k), & i \neq j, k = l \\ \frac{1}{2}p_i(1-p_i)q_k(1-q_l) + \frac{1}{2}p_i(1-p_i)q_l(1-q_k), & i = j, k \neq l \\ \frac{1}{4}p_i(1-p_j)q_k(1-q_l) + \frac{1}{4}p_j(1-p_i)q_k(1-q_l) & \\ \quad + \frac{1}{4}p_i(1-p_j)q_l(1-q_k) + \frac{1}{4}p_j(1-p_i)q_l(1-q_k), & i \neq j, k = l \end{cases}.$$

In order to perform inference using these statistics, we estimate them from data and then compare them to their computed expectations. We therefore seek unbiased estimators for each of these statistics from pairs of variable loci across the genome, for either phased or unphased data.

Our approach here is directly analogous to that for computing unbiased estimates of single-population statistics given above, with an additional index over each population. For phased data, we suppose we sample n_i haplotypes from population i $(1 \leq i \leq P)$, and from each population we observed haplotype counts $(n_{i1}, n_{i2}, n_{i3}, n_{i4})$, $\sum_j n_{ij} = n_i$. For any statistic S as a function of f_{ij}'s, we expand its expectation into a series of monomials, with each term taking the form

$$E\left[a \prod_{i=1}^{P} \prod_{j=1}^{4} {f_{ij}}^{k_{ij}}\right].$$

Each term has the interpretation as the probability that we sample $\sum_j k_{ij}$ haplotypes from each population, and observe the ordered configuration of k_{ij} of type j in population i.

From our sampled haplotypes, an unbiased estimator for each term is

$$a \prod_{i=1}^{P} \frac{1}{\binom{\sum_j k_{ij}}{k_{i1}, \ldots, k_{i4}}} \frac{\prod_{j=1}^{4} \binom{n_{ij}}{k_{ij}}}{\binom{n_i}{\sum_j k_{ij}}}.$$

We then sum over each term to obtain our unbiased estimator $\widehat{S}$.

For example, the covariance of D between populations i and j $(i \neq j)$ is given by

$$\begin{aligned} \mathrm{Cov}(D_i, D_j) = \mathbb{E}[D_i D_j] &= \mathbb{E}[(f_{i1}f_{i4} - f_{i2}f_{i3})(f_{j1}f_{j4} - f_{j2}f_{j3})] \\ &= \mathbb{E}[f_{i1}f_{i4}f_{j1}f_{j4}] - \mathbb{E}[f_{i1}f_{i4}f_{j2}f_{j3}] - \mathbb{E}[f_{i2}f_{i3}f_{j1}f_{j4}] + \mathbb{E}[f_{i2}f_{i3}f_{j2}f_{j3}], \end{aligned}$$

which has the unbiased estimator

$$\begin{aligned} \widehat{D_i D_j} =& \frac{1}{\binom{2}{1,0,0,1}} \frac{\binom{n_{i1}}{1}\binom{n_{i4}}{1}}{\binom{n_i}{2}} \frac{1}{\binom{2}{1,0,0,1}} \frac{\binom{n_{j1}}{1}\binom{n_{j4}}{1}}{\binom{n_j}{2}} - \frac{1}{\binom{2}{1,0,0,1}} \frac{\binom{n_{i1}}{1}\binom{n_{i4}}{1}}{\binom{n_i}{2}} \frac{1}{\binom{2}{0,1,1,0}} \frac{\binom{n_{j2}}{1}\binom{n_{j3}}{1}}{\binom{n_j}{2}} \\ & - \frac{1}{\binom{2}{0,1,1,0}} \frac{\binom{n_{i2}}{1}\binom{n_{i3}}{1}}{\binom{n_i}{2}} \frac{1}{\binom{2}{1,0,0,1}} \frac{\binom{n_{j1}}{1}\binom{n_{j4}}{1}}{\binom{n_j}{2}} + \frac{1}{\binom{2}{0,1,1,0}} \frac{\binom{n_{i2}}{1}\binom{n_{i3}}{1}}{\binom{n_i}{2}} \frac{1}{\binom{2}{0,1,1,0}} \frac{\binom{n_{j2}}{1}\binom{n_{j3}}{1}}{\binom{n_j}{2}} \\ =& \frac{(n_{i1}n_{i4} - n_{i2}n_{i3})}{n_i(n_i - 1)} \frac{(n_{j1}n_{j4} - n_{j2}n_{j3})}{n_j(n_j - 1)}. \end{aligned}$$

For unphased data, we take the exact same approach, but use genotype frequencies g_{ij} instead of haplotype frequencies (Table 1), and use the "composite" haplotype frequency estimates in each population x_{ij}, $1 \leq j \leq 4$, as we defined above in the single population case.

Expected statistics on genotype data can be expanded as before as a series of monomials in g_{ij}, with each term taking the form

$$E\left[a\prod_{i=1}^{P}\prod_{j=1}^{9} {g_{ij}}^{k_{ij}}\right],$$

and we again interpret this as the probability of observing a certain genotype configuration in a sample of diploids from each population. Then, just as before, if we sample n_i diploids from each population i, with genotype sampling configurations $(n_{i1}, \ldots, n_{i9})$, an unbiased estimator for any given term is

$$a\prod_{i=1}^{P}\frac{1}{\binom{\sum_j k_{ij}}{k_{i1},\ldots,k_{i4}}}\frac{\prod_{j=1}^{9}\binom{n_{ij}}{k_{ij}}}{\binom{n_i}{\sum_j k_{ij}}}.$$

We then sum over terms and simplify to find an unbiased estimator for S.

Supplementary figures

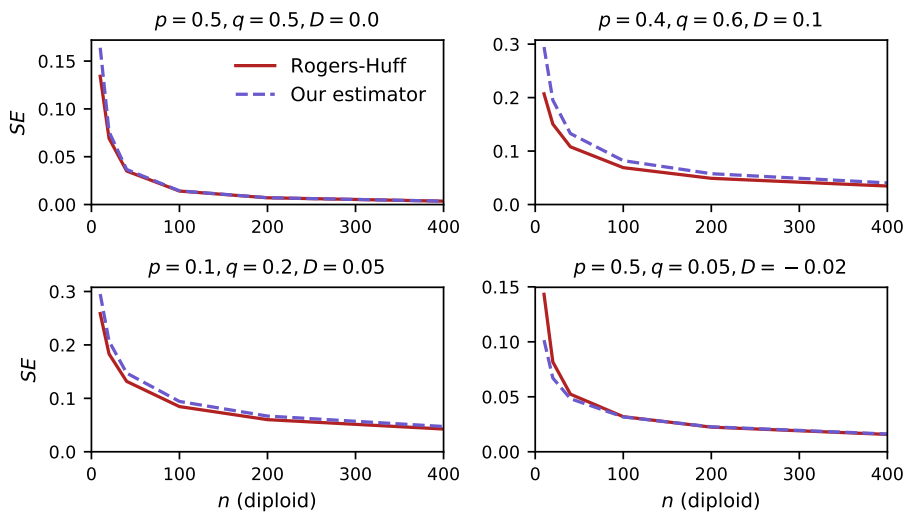


Figure S1: **Standard error of r^2 estimators.** The Rogers and Huff (2009) estimator for r^2 and our estimator display similar standard errors. However, our estimator is more accurate, especially for small sample sizes (Figure 1).

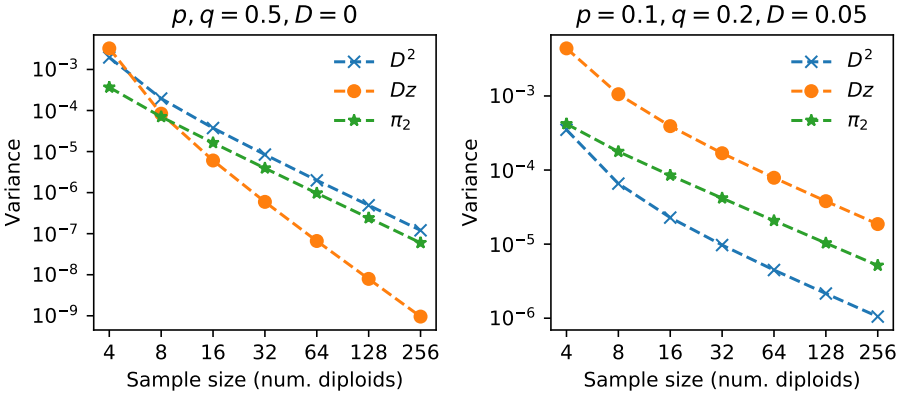


Figure S2: **Variance of estimators decays with sample size.** Variances decay as $\sim \frac{1}{n^2}$ with diploid sample size n. These were computed from one million replicates sampled with the given sample size from known haplotype frequencies.

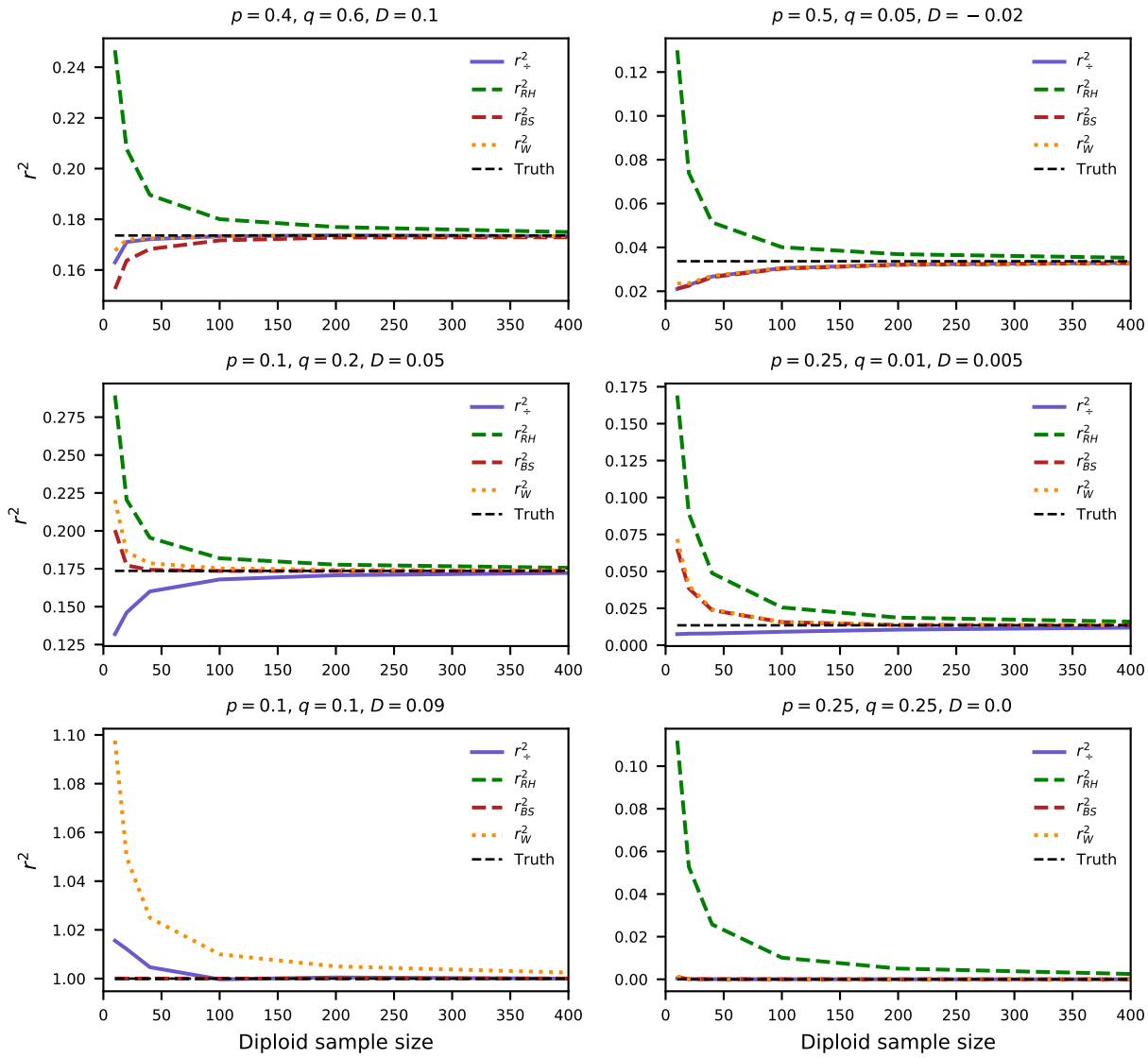


Figure S3: **Comparison of estimators for r^2.** As in Figure 1C-D, and also including the Bulik-Sullivan et al. (2015) estimator: $r^2_{\mathrm{BS}} = \hat{r}^2 - \frac{1-\hat{r}^2}{n-2}$, where n is the diploid sample size. The Rogers-Huff estimator consistently overestimates r^2 by a large amount, except for the case of perfect LD (lower-left), while the other estimators display variable biases depending on allele frequencies and underly LD.