Alemu *et al. Genetics Selection Evolution* (2022) 54:12
https://doi.org/10.1186/s12711-022-00703-z

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Comparison of linkage disequilibrium estimated from genotypes versus haplotypes for crossbred populations

Setegn Worku Alemu[1*], Piter Bijma[2], Mario P. L. Calus[2], Huiming Liu[3], Rohan L. Fernando[4] and Jack C. M. Dekkers[4]

Abstract

Background: Linkage disequilibrium (LD) is commonly measured based on the squared coefficient of correlation (r^2) between the alleles at two loci that are carried by haplotypes. LD can also be estimated as the r^2 between unphased genotype dosage at two loci when the allele frequencies and inbreeding coefficients at both loci are identical for the parental lines. Here, we investigated whether r^2 for a crossbred population (F1) can be estimated using genotype data. The parental lines of the crossbred (F1) can be purebred or crossbred.

Methods: We approached this by first showing that inbreeding coefficients for an F1 crossbred population are negative, and typically differ in size between loci. Then, we proved that the expected r^2 computed from unphased genotype data is expected to be identical to the r^2 computed from haplotype data for an F1 crossbred population, regardless of the inbreeding coefficients at the two loci. Finally, we investigated the bias and precision of the r^2 estimated using unphased genotype versus haplotype data in stochastic simulation.

Results: Our findings show that estimates of r^2 based on haplotype and unphased genotype data are both unbiased for different combinations of allele frequencies, sample sizes (900, 1800, and 2700), and levels of LD. In general, for any allele frequency combination and r^2 value scenarios considered, and for both methods to estimate r^2, the precision of the estimates increased, and the bias of the estimates decreased as sample size increased, indicating that both estimators are consistent. For a given scenario, the r^2 estimates using haplotype data were more precise and less biased using haplotype data than using unphased genotype data. As sample size increased, the difference in precision and biasedness between the r^2 estimates using haplotype data and unphased genotype data decreased.

Conclusions: Our theoretical derivations showed that estimates of LD between loci based on unphased genotypes and haplotypes in F1 crossbreds have identical expectations. Based on our simulation results, we conclude that the LD for an F1 crossbred population can be accurately estimated from unphased genotype data. The results also apply for other crosses (F2, F3, Fn, BC1, BC2, and BCn), as long as (selected) individuals from the two parental lines mate randomly.

Background

Linkage disequilibrium (LD) is the non-random association of alleles at different loci within haplotypes. LD plays an important role in both population and quantitative genetics. In population genetics, LD can for example be used to detect selection [1]. In quantitative genetics, LD has been used to map quantitative

*Correspondence: s.w.alemu@massey.ac.nz
[1] AL Rae Centre for Genetics and Breeding, Massey University, 10 Bisley Drive, Hamilton 3240, New Zealand
Full list of author information is available at the end of the article

trait loci [1–3] and for marker-assisted selection [4] and genomic selection [5]. Thus, knowledge of LD is required for diverse applications in genetics.

LD is traditionally measured based on the comparison of the observed haplotype frequencies with the expected haplotype frequencies under linkage equilibrium. A common statistical measure of LD is the covariance between loci, D, which is equal to the excess of coupling phase haplotypes, $D_{ij} = P_{ij} - P_iP_j$, where P_{ij} refers to the frequency of gametes (haplotypes) that carry the pair of alleles i and j at the two loci, P_i and P_j refer to the frequency at locus i and locus j, respectively, and P_iP_j is the expected frequency of this haplotype under linkage equilibrium [6]. Another common measure is the squared coefficient of correlation (r^2) between the alleles at the two loci within haplotypes, $r^2_{ij} = \frac{D^2_{ij}}{P_i(1-P_i)P_j(1-P_j)}$ [7].

To calculate D and r^2 using the expressions given above, the haplotypes carried by the individuals must be known. However, Rogers and Huff [8] showed that LD can also be estimated by correlating unphased genotype dosages at the two loci, which makes the computation simple and fast. They demonstrated that LD estimated from unphased genotypes yields very similar results to LD estimated from haplotypes. In their derivation, however, they assumed equal inbreeding coefficients for the two loci and equal allele frequencies for the paternal and maternal gametes that created the population. In this context, the inbreeding coefficient measures the departure from Hardy–Weinberg equilibrium and, thus, can take positive or negative values. However, for crossbred individuals inbreeding coefficients can differ between the two loci, and paternal and maternal allele frequencies can differ because the two parents come from different lines.

Here, we investigated whether LD in crossbred populations can be estimated using unphased genotype data. We assumed that sires and dams of the crossbreds originate from two distinct lines but are otherwise mated to each other at random. We address this question in three steps. First, we derive the inbreeding coefficients of crossbreds, showing that they take negative values that typically differ between loci. As a result, the derivation of Rogers and Huff [8] cannot be used to demonstrate the equivalence of genotype-based LD to haplotype-based LD for a crossbred population. Second, we show theoretically that LD computed from genotype frequencies has the same expected value for a given dataset as LD computed from haplotype frequencies, even for a crossbred population. Finally, we investigate the precision and potential bias of LD estimated from unphased genotype data versus haplotype data, using stochastic simulation.

Methods

Inbreeding coefficients for a crossbred population

Consider two outbred lines, A and B. We want to investigate the inbreeding coefficients for two bi-allelic loci, M and N, in the F1 crossbred offspring that result from the crossing of random individuals from two parental lines. With alleles denoted *0* and *1*, p_{AM} is the frequency of allele *1* at locus M in line A, and p_{BM} is the frequency of allele *1* at locus M in line B. The expected frequency of allele *1* at locus M in the crossbreds then is $p_M = \frac{p_{AM}+p_{BM}}{2}$. With random mating between individuals from the two parental lines, the frequency of genotype *11* in the crossbreds is $p_{AM}p_{BM}$. The deviation of this frequency from Hardy–Weinberg equilibrium follows from [6, 9].

$$p_{AM}p_{BM} = p_M^2 + p_M(1-p_M)f_M,$$

where f_M is the inbreeding coefficient at locus M in the crossbreds.

The inbreeding coefficient follows from solving this expression for f_M, substituting $p_M = \frac{p_{AM}+p_{BM}}{2}$, and simplifying the expression, giving:

$$f_M = \frac{-(p_{AM} - p_{BM})^2}{(p_{AM} + p_{BM})(2 - p_{AM} + p_{BM})}.$$

Similarly, $f_N = \frac{-(p_{AN}-p_{BN})^2}{(p_{AN}+p_{BN})(2-p_{AN}+p_{BN})}$.

Note that the numerators of f_M and f_N are always negative, except when $p_{AM} = p_{BM}$ and $p_{AN} = p_{BN}$, while the denominators are always positive. This shows that the inbreeding coefficients of crossbreds are negative, meaning that heterozygosity is greater than would be expected under Hardy–Weinberg equilibrium (for example $p_{AM} = 0.05$, $p_{BM} = 0.09$, $p_{AN} = 0.25$, and $p_{BN} = 0.29$ yields $f_M = -0.0056$ and $f_N = -0.0015$).

We investigated under which conditions the inbreeding coefficients at the two loci are equal by solving the expression $f_N = f_M$ for the allele frequencies, using Wolfram Mathematica (www.wolfram.com). Apart from the trivial solutions of $p = 0$, $p = 1$, and equal allele

Table 1 Haplotype frequencies and marginal allele frequencies for line A[a]

Alleles at locus *M*	Alleles at locus *N*		
	0	**1**	**Marginal frequency**
0	r	s	$r+s$
1	t	u	$t+u$
Marginal frequency	$r+t$	$s+u$	$r+s+t+u=1$

[a] Corresponding symbols for line B are denoted by $'$

frequencies at both loci, we found only three solutions (see Appendix 1). Hence, this result demonstrates that the inbreeding coefficients at two arbitrary loci in a crossbred population will usually be different. This implies that the derivation of Rogers and Huff [8] cannot be used to demonstrate the equivalence of genotype-based LD to haplotype-based LD for a crossbred population.

Haplotype-based linkage disequilibrium

In this section, we show that the expected LD based on r^2 computed from the genotype frequencies of the crossbred population is identical to the true r^2 based on haplotype frequencies, even when the inbreeding coefficients differ between the two loci. Note that we consider the true (*i.e.,* population) value of r^2 here, rather than an estimate from a sample. As we consider bi-allelic loci, we have four haplotype frequencies for each line, denoted r, s, t, and u for line A, and using $'$ to refer to frequencies for line B, we have haplotype frequencies r', s', t', and u' for line B. Table 1 shows expressions for the marginal frequency for each of the alleles. Although the expressions for the marginal frequencies in Table 1 can be simplified by formulating them in terms of allele frequencies, we stick to the haplotype frequencies to facilitate comparison with results for the genotype-based r^2.

Crossbred genotypes consist of two sets of haplotypes, one from each parental line, which may have a different r^2. By definition, the r^2 in the crossbreds depends on the (co)variances between loci in the crossbred population, so we cannot simply average the r^2 of the two parental lines. From the definitions of correlation, variance, and covariance, it follows that the r^2 for the crossbred population equals the square of the average of the covariances between haplotypes for each of the two lines, divided by the product of the average variance across the two lines at each locus. For line A, the covariance between haplotypes (i.e. *D)* follows from Table 1 as $u-(t+u)(s+u)$, where u is the expectation of the cross product of the allele frequencies at each locus, while $(t+u)(s+u)$ is the cross product of the expectations of these allele frequencies (expected haplotype frequency in line A under linkage equilibrum). Hence, this result follows immediately from the definition of a covariance. The covariance (D) for line B is analogous, using symbols denoted by $'$. The variance in allele count follows from the binomial distribution with $n=1$ for haplotypes and are thus equal to $p(1-p)$, p denoting the allele frequency. For line A the variance equals $(s+u)(r+t)$ for locus N, and $(t+u)(r+s)$ for locus M, with analogous equations for line B. Using these values in the haplotype-based r^2 for the crossbred population yields the following true r^2 in the crossbred population:

$$r^2_{hap}=\frac{\left[(u-(t+u)(s+u))+\left(u'-\left(t'+u'\right)\left(s'+u'\right)\right)\right]^2}{[(s+u)(r+t)+(s'+u')(r'+t')][(t+u)(r+s)+(t'+u')(r'+s')]}, \quad (1)$$

where the numerator is the square of the average of the covariances for the two parental lines, while the denominator is the product of the average of the variances. Note that the constant 2^2 in the numerator of Eq. (1) and 2^2 in the demoninator of Eq. (1) (2 for each variance) cancelled out in the derivation of the equation.

Genotype-based squared correlation

The following inputs are required to derive the genotype-based r^2 in crossbreds: genotype frequencies and the expectations of squares and cross products of genotype dosage, 0, 1, and 2, in crossbreds. Using the haplotype frequencies in Table 1 and the assumption that individuals of line A mate at random to individuals of line B, we find the genotype frequencies in the crossbred population as shown in Table 2. Next, using these genotype frequencies, Table 3 shows the expectations of squares and cross products of genotype dosages. Computations of the expectations of combinations of genotypic values are in Appendix 1.

Using the values in Table 3, the covariance of genotype dosage at the two loci follows from $cov(M_gN_g) = E(M_gN_g) - E(M_g)E(N_g)$, where M_g and N_g are the genotype dosages at loci M and N, and the variances of genotype dosage follow from $var(M_g) = E\left(M_g^2\right) - E^2(M_g)$ and the corresponding expression for locus N. Substituting the resulting expressions into the

Table 2 Expected genotype frequencies in the crossbred offspring when individuals from lines *A* and *B* are mated at random to each other

Line *A* haplotype	Line *B* haplotype			
	00 r'	**01** s'	**10** t'	**11** u'
00 r[a]	$\frac{00}{00}r'r$[b]	$\frac{00}{01}s'r$	$\frac{00}{10}t'r$	$\frac{00}{11}u'r$
01 s	$\frac{01}{00}r's$	$\frac{01}{01}s's$	$\frac{01}{10}t's$	$\frac{01}{11}u's$
10 t	$\frac{10}{00}r't$	$\frac{10}{01}s't$	$\frac{10}{10}t't$	$\frac{10}{11}u't$
11 u	$\frac{11}{00}r'u$	$\frac{11}{01}s'u$	$\frac{11}{10}t'u$	$\frac{11}{11}u'u$

[a] Marginal frequencies are haplotype frequencies

[b] Joint frequencies are the genotype frequencies

Table 3 Unordered genotypes, their genotype dosages, frequencies, and expectations of genotype dosages, and squares and cross products of genotype dosages, for locus *M* and *N*, in the crossbred offspring from random mating between lines A and B

Genotype dosage		Frequency	Expectations				
M_g	N_g	f	$E(M_g^2)$[a]	$E(M_g)$	$E(N_g^2)$	$E(N_g)$	$E(M_gN_g)$
0	0	rr'	0	0	0	0	0
0	1	$r's + rs'$	0	0	$r's + rs'$	$r's + rs'$	0
0	2	$s's$	0	0	$4s's$	$2s's$	0
1	0	$r't + rt'$	$r't + rt'$	$r't + rt'$	0	0	0
1	1	$r'u + ru' + s't + st'$	$r'u + ru' + s't + st'$	$r'u + ru' + s't + st'$	$r'u + ru' + s't + st'$	$r'u + ru' + s't + st'$	$r'u + ru' + s't + st'$
1	2	$s'u + su'$	$s'u + su'$	$s'u + su'$	$4s'u + 4su'$	$2s'u + 2su'$	$2s'u + 2su'$
2	0	$t't$	$4t't$	$2t't$	0	0	0
2	1	$t'u + tu'$	$4t'u + 4tu'$	$2t'u + 2tu'$	$t'u + tu'$	$t'u + tu'$	$2t'u + 2tu'$
2	2	$u'u$	$4u'u$	$2u'u$	$4u'u$	$2u'u$	$4u'u$
		1	$(t+u) + (t'+u') + 2(t'+u')(t+u)$	$(t+u) + (t'+u')$	$(s+u) + (s'+u') + 2(s'+u')(s+u)$	$(s+u) + (s'+u')$	$u'(1+t+u) + u(1+t'+u') + s'(t+u) + s(t'+u')w$

[a] $E(M_g^2)$ refers to the expectation of the squared genotype dosage for locus *M*, and similar definitions apply for the $E(M_g)$, $E(N_g^2)$, $E(N_g)$, $E(M_gN_g)$. The derivation of the expectations of genotype dosages is given in Appendix 2

expression for the correlation coefficient yields the following expectation of the genotype-based r^2:

$$r^2_{geno} = \frac{\left[(u - (t+u)(s+u)) + \left(u' - \left(t'+u'\right)\left(s'+u'\right)\right)\right]^2}{[(s+u)(r+t) + (s'+u')(r'+t')][(t+u)(r+s) + (t'+u')(r'+s')]}. \quad (2)$$

This expression is identical to the expression for the true haplotype-based r^2 (Eq. (1)). Thus, when two lines (the lines can be pure or crossbred) are crossed but individuals from the two lines are mated at random to each other, expectations of the genotype-based and the haplotype-based r^2 in the crossbreds (F1, F2, Fn) and in other cross types (BC1, BC2, BCn) are identical, irrespective of differences in the inbreeding coefficients at the two loci. Note that our derivation also applies to other measures of LD, i.e. D and D'. For example, measures of D based on genotypes and haplotypes are the numerators of Eqs. (1) and (2), which are identical. Furthermore, using Eqs. (1) and (2), the r in the crossbred population can be predicted if the haplotype and genotype frequencies of the two parental lines are known.

Note that Eq. (2) refers to the expected r^2 between the genotype dosage at the two loci, not to an estimate thereof. Hence, although the expected values of r^2_{geno} and r^2_{hap} are identical, their estimates for a given data set may differ depending on sampling bias and the sampling errors of the estimates. This will be investigated using a simulation study in the next section.

Simulation

The objective of the simulation was to investigate and compare the bias and precision of the genotype-based and haplotype-based estimates of r^2 for a crossbred population. We investigated the bias and the precision for different sets of allele frequencies, levels of LD as measured by r^2, and sample sizes. To limit computation time, we directly sampled haplotypes according to their probability distribution, rather than simulating a population of individuals. The haplotype probability distribution follows from the allele frequencies at the two loci and the level of LD. Using the haplotype frequencies and sample size, haplotypes were sampled from a multinomial distribution for each of the two parental lines. The genotypes of the crossbred individuals were obtained by random sampling of one haplotype from each line. Next, the genotype-based and haplotype-based estimates of r^2 were computed from the genotypes and haplotypes, respectively, of the crossbred offspring. The parameter values (allele frequencies, r^2 for each line, and sample size) that were used for simulation were used to compute the true r^2 in the crossbreds, using Eq. (1), which was used as a benchmark to evaluate the precision and bias of the two estimates of r^2. Thus, there were three measures of r^2:

the true r^2 calculated from the parameter values used for simulation, the haplotype-based estimate of r^2, and the genotype-based estimate of r^2. For each set of parameters, results were based on 1000 replicates. We used the R software [10] to simulate the data and analyse the results. The source code for the simulation is available at the following GitHub repository. https://github.com/setegnworku/Simulation-code-for_LD_crossbred_pop.

Scenarios investigated

We considered only biallelic loci at two loci in crossbreds resulting from the random mating of two outbred lines (*A* and *B*). We varied three parameters: (i) allele frequencies and (ii) r^2 in the parental lines, and (iii) the sample size. For the allele frequencies, we considered a range from 0.05 to 0.45, incremented by 0.10, for both lines. To limit the number of scenarios, we used equal allele frequencies at the two loci for most scenarios. Note that there is no true difference between the major and the minor allele, e.g., $p_A = 0.05$ is equivalent to $p_A = 0.95$, such that results for allele frequencies ranging from 0.55 to 0.95 are identical to those for 0.05 to 0.45. For r^2 in the parental lines, we considered values of 0.2, 0.4, 0.6, and 0.8. To reduce the number of scenarios, r^2 was the same in both lines. We considered sample sizes of 900, 1800, and 2700. This resulted in a total of 180 scenarios with equal allele frequencies at the two loci within each line, of which 120 had different allele frequencies between the two lines, and all had equal r^2 in the two lines (Table 4). In addition to those 180 scenarios, we investigated a few scenarios where allele frequencies differed between loci within the parental lines and for which r^2 differed between the parental lines.

Table 4 Combinations of minor allele frequencies for lines *A* and *B* investigated in the simulation[a]

Line *B*	Line *A*				
	0.05	**0.15**	**0.25**	**0.35**	**0.45**
0.05	X	X	X	X	X
0.15		X	X	X	X
0.25			X	X	X
0.35				X	X
0.45					X

[a] Allele frequencies were equal for the two loci (*M* and *N*) within a line. Apart from the diagonal elements, the allele frequencies differed between the two lines. Scenarios in this Table were replicated for sample sizes of 900, 1800, and 2700, and r^2 in the parental lines equal to 0.2, 0.4, 0.6, and 0.8 (equal for both lines), yielding a total of 3*4*15 = 180 scenarios

Results and discussion

The full results for all 180 simulated scenarios, including bias, ratio of precision (ratio of standard deviation for the r^2 estimates using unphased genotype and haplotype data), correlation of the standard deviation, of the r^2 estimate using unphased genotype and haplotype data is given in the following R shiny App (https://setegnmaths.shinyapps.io/LD_App/). The source code for the Shiny App is available in the following github repository:

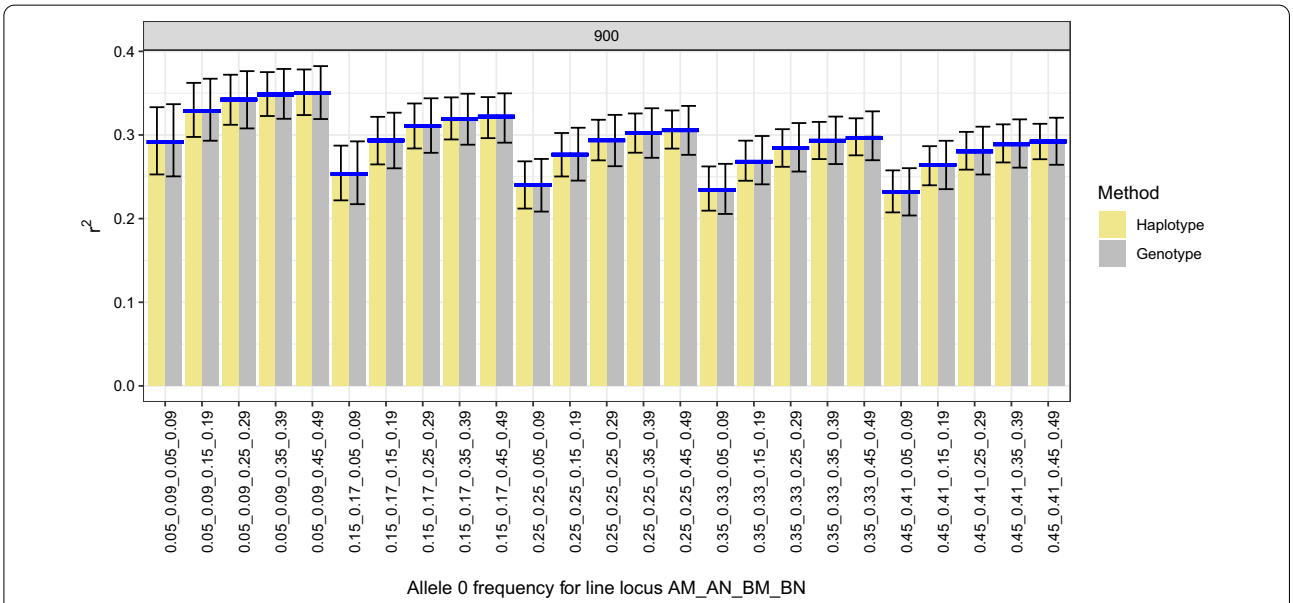

Fig. 1 Comparison of estimates of linkage disequilibrium ($r^2 \pm SD$) based on unphased genotype and haplotype data for scenarios where allele frequencies differed between loci and between lines, with $r^2 = 0.2$ for line *A* and $r^2 = 0.4$ for line *B*. Sample size was 900 (1000 replicates)

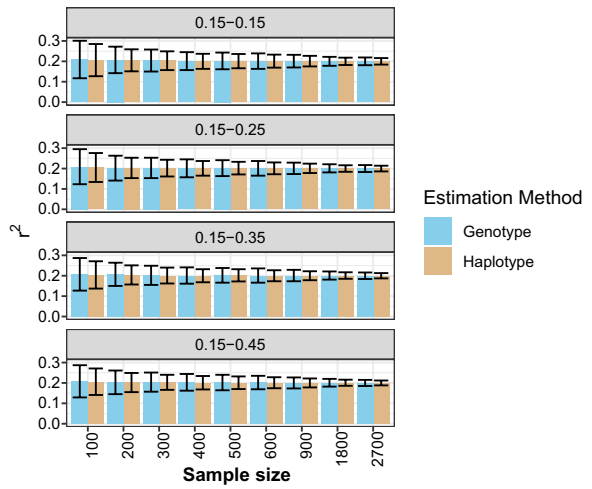

Fig. 2 Comparison of linkage disequilibrium estimated from unphased genotype and haplotype data, for different sample sizes

https://github.com/setegnworku/Linkage_disequilibrium_crossbred_ShinyApp.

Results showed that the estimates of r^2 for 180 scenarios were unbiased, both for the haplotype-based and the unphased genotype-based estimates of r^2. Moreover, simulation results also confirmed our theoretical finding that unphased genotype-based and haplotype-based r^2 on average are the same for a given dataset, irrespective of differences in inbreeding coefficients between the two loci (Fig. 1).

As shown in Fig. 1, r^2 for a given dataset was unbiased for scenarios where allele frequencies differed between loci (i.e., inbreeding coefficients differed between the two loci) and between lines, and when the r^2 differed between the lines (0.2 and 0.4). We also tested the bias of LD estimates using unphased genotype and haplotype data for different sample sizes (Fig. 2). As shown in Fig. 2, for all scenarios, both estimators were unbiased for a sample size above 300. However, with sample size of 300 or less (100, 200, and 300), we found a small downward bias for both the unphased genotype- and haplotype-based estimates (the independent sample t-test showed the bias was significant for some of the scenarios for both the unphases genotype- and haplotype-based estimates). It is well known that the estimator of the correlation coefficient is known to be biased, and more so for smaller samples [11], which may explain the bias we found in small samples.

Bias

For all scenarios (180), the estimates of the r^2 using unphased genotype and haplotype data were both unbiased. We ran an independent sample t-test to test the bias of the estimates of r^2 using unphased genotype and haplotype data from the true r^2. For all 180 scenarios, the bias of the estimates was not significantly different from zero for both methods (p value > 0.05). The average absolute bias across 180 scenarios was 0.0004 when using unphased genotype data, and 0.0003 when using haplotype data (Table 5). The maximum absolute bias across the 180 scenarios was 0.003 when using unphased genotype data and 0.002 when using haplotype data. As expected, the bias decreased as sample size increased. For example, with unphased genotype data, the average absolute bias was 0.0005 for a sample size of 900 and 0.0001 for a sample size of 2700. Corresponding values for haplotype data were 0.0003 and 0.0001. These results show that the estimators of r^2 are consistent for both unphased genotype data and haplotype data, because the bias of the r^2 estimates decreased as sample size increased.

Table 5 Summary of estimates of bias and precision (standard deviation) of r^2 using unphased genotype and haplotype data

Parameter	Haplotype	Genotype
Average absolute bias across 180 scenarios	0.0003	0.0004
Maximum absolute bias	0.002	0.003
Average absolute bias sample size 900	0.0003	0.0005
Average absolute bias sample size 2700	0.0001	0.0001
Standard deviation (SD) across 180 scenarios	0.021	0.023
Maximum SD	0.055	0.057
Average SD with sample size of 900	0.027	0.031
Average SD with sample size of 2700	0.016	0.018

Precision

For all scenarios, estimates of LD based on haplotype data were more precise than estimates based on unphased genotype data, although the differences were small. For example, the mean standard deviation of the estimates of r^2 across all scenarios was 0.023 when using unphased genotype data and 0.021 when using haplotype data. The maximum standard deviation for estimates of r^2 across all scenarios was 0.057 using unphased genotype data and 0.055 using haplotype data. The precision of the estimates of r^2 increased as sample size increased, both with unphased genotype and with haplotype data. For example, the average standard deviation across all scenarios with a sample size of 900 was 0.031 with unphased genotype data and 0.027 with haplotype data. The corresponding values for a sample size of 2700 were 0.018 and 0.016. This result was as expected because the standard error of the estimate of a correlation coefficient decreases as sample size increases [12]. Thus, with

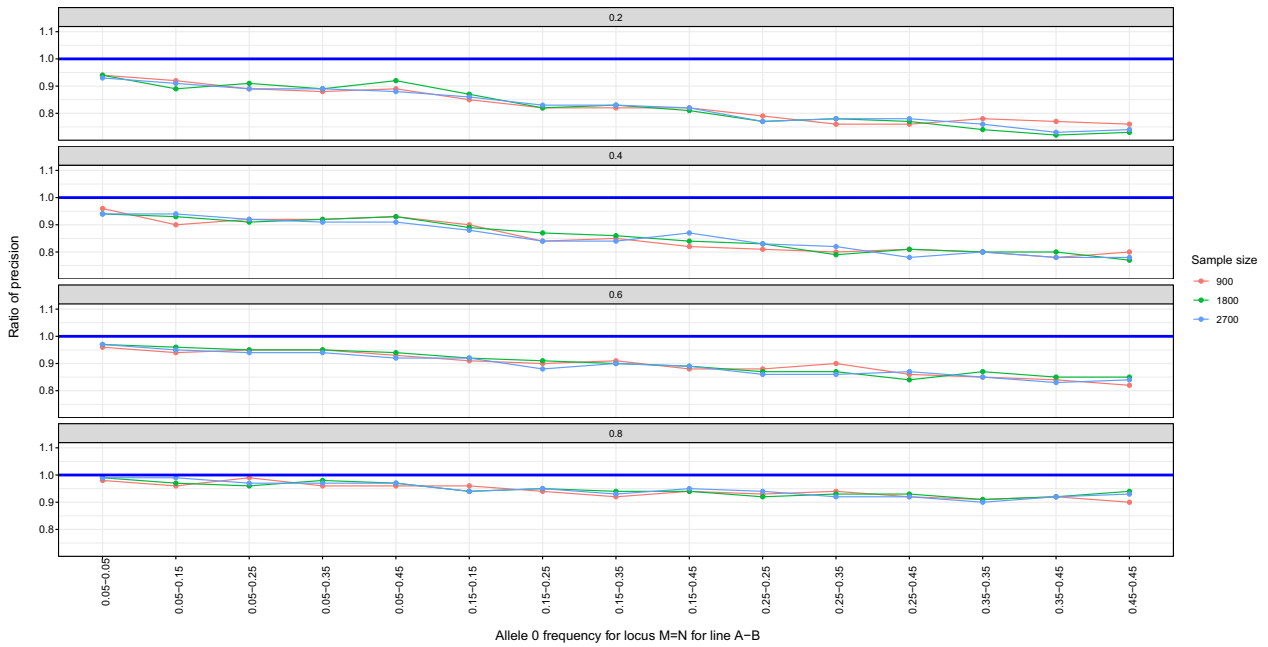

Fig. 3 Ratio of precision for all scenarios investigated for r^2 values of 0.2, 0.4, 0.6, and 0.8

a sufficient sample size, r^2 in crossbreds can be estimated accurately based on unphased genotype data.

We further investigated in which scenarios the difference in precision for the estimates of r^2 using unphased genotype versus haplotype data was the largest. We investigated this by computing the ratio of the standard deviations of the estimates of r^2 using haplotype data and unphased genotype data. Thus, smaller values of this ratio indicate a greater superiority of estimates based on haplotypes. As shown in Fig. 3, the ratio of precision was less than 1 for all scenarios, indicating that the estimate based on haplotype data was more precise than that based on unphased genotype data. The ratio of the precision increased as the level of LD increased. For example, for an r^2 of 0.2, the ratio of precision ranged from 0.75 to 0.9, while with an r^2 of 0.8, the ratio ranged from 0.92 to 0.98. The difference between the estimates of r^2 based on unphased genotype vs. haplotype data originates solely from the double heterozygotes (00/11 for coupling phase, or 01/10 for repulsion phase). As r^2 increases, the frequencies of the coupling phase haplotypes 00 and 11 or of the repulsion phase haplotypes 01 and 10, increase, which reduces the opportunity for the haplotype method to provide extra information by distinguishing between them. As a result, at larger r^2, the precision of the estimates of r^2 using unphased genotype and haplotype data are expected to be closer to each other. On the other hand, at low r^2, all haplotypes (00, 01, 10, 11) are possible and the haplotype-based method provides additional

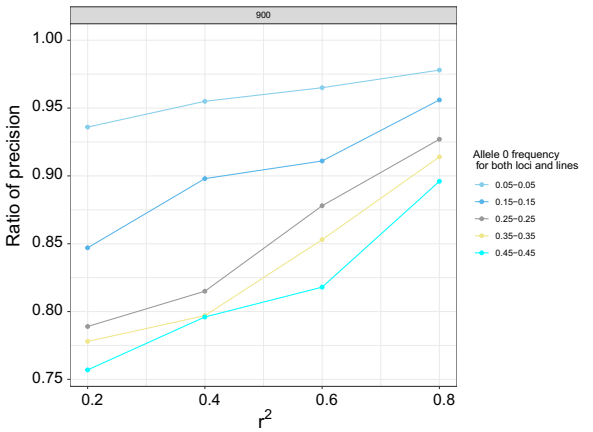

Fig. 4 Ratio of precision for sample size of 900 for selected scenarios

information. For this reason, the estimate of r^2 based on haplotype data is more precise than the estimate based on unphased genotype data, in particular when the true r^2 is small.

The ratio of precision decreased when the minor allele frequencies for the two loci increased (Figs. 3 and 4). For example, for allele frequencies of 0.05 and 0.05 at the two loci, the ratio of precision ranged from 0.93 to 0.99, while it ranged from 0.73 to 0.94 for allele frequencies of 0.45 and 0.45. This is because the proportion of the double heterozygotes in the population decreases when

the minor allele frequencies at the two loci decrease, which reduces the extra information provided by the haplotype-based method. This is in agreement with [13]. There was also an interaction between the level of LD and the minor allele frequency, with the ratio of precision increasing when the level of LD increased but this increase was larger for higher values of the minor allele frequency (Fig. 4). The ratio of precision at allele frequencies of 0.05 and 0.05 was 0.91 when r^2 was 0.2 and 0.99 for an r^2 of 0.9. However, the corresponding values for allele frequencies of 0.45 and 0.45 were 0.70 when r^2 was 0.2 and 0.94 for an r^2 of 0.9. When the minor allele frequencies at the two loci decrease, the proportion of double heterozygotes decreases, which reduces the extra information provided by the haplotype-based method. Thus, with extreme allele frequencies at the loci (e.g. 0.05 and 0.05), both methods yielded similar results, irrespective of the level of LD. On the other hand, at intermediate allele frequencies, such as 0.45 and 0.45, the proportion of double heterozygotes in the population increases, which increases the extra information provided by the haplotype-based method, particularly when LD is weak.

In real applications, the true r^2 is unknown and the r^2 computed using haplotype data would serve as the reference value. In that case, the comparison would be between the r^2 computed using unphased genotype data relative to the estimate based on haplotype data. In this case, the average absolute bias across the 180 scenarios using unphased genotype was very close to zero (0.00017) and the average standard deviation of estimates based on unphased genotype data across all scenarios relative to haplotype data was 0.0026. In addition, the haplotype-based method assumes that the haplotype can be determined without error for each individual, which means that in reality the absolute bias may be lower than the above value of 0.00017, depending on the error of haplotype estimation. Thus, estimates of r^2 computed using unphased genotype and haplotype data are indistinguishable in terms of both bias and precision in practice, particularly with sufficient sample size.

This paper extends the work of Rogers and Huff [8] and Weir [14], who showed that LD can be estimated from unphased genotype data when the allele frequency in line *A* and line *B* is the same, and when the inbreeding coefficient is identical for the two loci. Here, we showed that LD can also be estimated using unphased genotype data when the allele frequencies differ between lines *A* and *B* and the inbreeding coefficients differ between the two loci. This is particularly relevant for hybrids in plant breeding [15] and for crossbreds in animal breeding [16, 17].

Conclusions

This work shows that the expectation of estimates of linkage disequilibrum (LD) between loci based on unphased genotypes and haplotypes in F1 crossbreds are identical. Estimates of LD, i.e. r^2, are more precise and less biased when based on haplotype data compared to unphased genotype data. For both unphased genotype and haplotype data, the precision of r^2 increases and the bias of the estimates decreases as sample size increases. More importantly, the difference in precision and bias between estimates of r^2 using haplotype and unphased genotype data decreases as sample size increases. Thus, LD in a crossbred population can be estimated using unphased genotyped data with little bias and good precision, particularly with sufficient sample size.

Appendix 1

This appendix shows under which conditions the inbreeding coefficients at the two loci (*M* and *N*) are equal when two outbred lines (*A* and *B*) mate randomly. The alleles are denoted *0* and *1*, p_{AM} is the frequency of allele *1* at locus *M* in line *A*, and p_{BM} is the frequency of allele *1* at locus *M* in line *B*.

To simplify the symbols notation let $a = p_{AM}$, $b = p_{BM}$, $c = p_{AN}$, and $d = p_{BN}$. The inbreeding coefficient for locus *M* is $f_M = \frac{-(\mathrm{a}-\mathrm{b})^2}{(a+b)(2-a+b)}$ and for locus *N* is $f_N = \frac{-(c-d)^2}{(\mathrm{c}+\mathrm{d})(2-c+d)}$. By solving the expression $f_N = f_M$, using Wolfram Mathematica (www.wolfram.com), we get the following complex solutions:

$$\left\{d \to \frac{a^2 - 2ab + b^2 + 2ac - 2a^2c + 2bc - 2b^2c - (a-b)\sqrt{a^2 - 2ab + b^2 + 8ac - 4a^2c + 8bc - 8abc - 4b^2c - 8ac^2 + 4a^2c^2 - 8bc^2 + 8abc^2 + 4b^2c^2}}{2(a+b-2ab)}\right\},$$

$$\left\{a \to \frac{b}{-1+2b}, d \to \frac{c}{-1+2c}\right\},$$
$$\{a \to 0, b \to 0\}, \{a \to 1, b \to 1\}$$

Conclusion: there are two trivial solutions, i.e. p = 0($\{a \to 0, b \to 0\}$), and p = 1($\{a \to 1, b \to 1\}$); two simple solutions ($\left\{a \to \frac{b}{-1+2b}, d \to \frac{c}{-1+2c}\right\}$,), and two rather complex solutions; thus, in general the two inbreeding coefficients are different.

Appendix 2

This appendix shows the computation of the expectation for different linear combination of genotypic values M_g and N_g at loci M and N, respectively, as indicated in Table 3.

Computation of $E(M_g)$

$$\begin{aligned}E(M_g) =& r't + rt' + r'u + ru' + s't \\ &+ st' + s'u + su' + 2t't \\ &+ 2u't + 2ut' + 2u'u\end{aligned}$$

$$\begin{aligned}E(M_g) =& r'(t+u) + t'(r+s) + u'(r+s) \\ &+ s'(t+u) + 2t(t'+u') \\ &+ 2u(t'+u')\end{aligned}$$

$$\begin{aligned}E(M_g) =& (r'+s')(t+u) + (t'+u')(r+s) \\ &+ 2(t+u)(t'+u').\end{aligned}$$

Simplifying this yields:

$$E(M_g) = (t+u) + (t'+u'). \tag{3}$$

Computation of $E\left(M_g^2\right)$

$$E\left(M_g^2\right) = r't + rt' + r'u + ru' + s't + st' + s'u + su' + 4t't + 4u't + 4ut' + 4u'u$$

$$E\left(M_g^2\right) = r'(t+u) + t'(r+s) + u'(r+s) + s'(u+t) + 4t'(t+u) + 4u'(t+u)$$

$$E\left(M_g^2\right) = r'(t+u)+t'(r+s)+u'(r+s)+s'(u+t)+2(t'+u')(t+u)+2(t'+u')(t+u).$$

Simplifying this yields:

$$E\left(M_g^2\right) = (t+u) + (t'+u') + 2(t+u)(t'+u'). \tag{4}$$

Computation of $E(N_g)$

$$\begin{aligned}E(N_g) =& r's + rs' + 2s's + r'u + ru' + s't + st' \\ &+ 2s'u + 2su' + tu' + t'u + 2u'u\end{aligned}$$

$$\begin{aligned}E(N_g) =& r'(s+u) + s'(r+t) + u'(r+t) + t'(s+u) \\ &+ 2s'u + 2s's + 2su' + 2u'u\end{aligned}$$

$$\begin{aligned}E(N_g) =& (r'+t')(s+u) + (s'+u')(r+t) \\ &+ 2(s'+u')(s+u)\end{aligned}$$

$$\begin{aligned}E(N_g) =& (r'+t')(s+u) + (s'+u')(r+t) \\ &+ (s'+u')(s+u) + (s'+u')(s+u).\end{aligned}$$

Simplifying this yields

$$E(N_g) = (s+u) + (s'+u'). \tag{5}$$

Computation of $E\left(N_g^2\right)$

$$\begin{aligned}E\left(N_g^2\right) =& r's + rs' + r'u + ru' + s't \\ &+ st' + t'u + tu' + 4s's \\ &+ 4s'u + 4su' + 4u'u\end{aligned}$$

$$\begin{aligned}E\left(N_g^2\right) =& r's + r'u + rs' + s't + ru' \\ &+ st' + t'u + tu' \\ &+ 4s'(s+u) + 4u'(s+u)\end{aligned}$$

$$\begin{aligned}E\left(N_g^2\right) =& r'(s+u) + s'(r+t) + u'(r+t) \\ &+ t'(s+u) + 4(s+u)(s'+u')\end{aligned}$$

$$\begin{aligned}E\left(N_g^2\right) =& r'(s+u) + s'(r+t) + u'(r+t) \\ &+ t'(s+u) + 2(s+u)(s'+u') \\ &+ 2(s+u)(s'+u').\end{aligned}$$

Simplifying this yields:

$$E\left(N_g^2\right) = (s+u) + \left(s'+u'\right) + 2(s+u) + \left(s'+u'\right). \tag{6}$$

Computation of $E\left(M_g N_g\right)$

$$\begin{aligned} E\left(M_g N_g\right) =& r'u + ru' + t's + ts' \\ &+ 2s'u + 2su' + 2t'u \\ &+ 2u't + 4u'u \end{aligned}$$

$$\begin{aligned} E\left(M_g N_g\right) =& u\left(r'+s'\right) + 2u\left(t'+u'\right) \\ &+ u'(r+s) + 2u'(t+u) \\ &+ s\left(t'+u'\right) + s'(t+u). \end{aligned}$$

Simplifying this yields:

$$\begin{aligned} E\left(M_g N_g\right) =& u\left(1+t'+u'\right) + u'(1+t+u) \\ &+ s\left(t'+u'\right) + s'(t+u). \end{aligned} \tag{7}$$

Acknowledgements
SWA thanks Bernt Guldbrandtsen and Dorian Garrick for helpful discussions on this topic.

Authors' contributions
SWA, PB, JCMD and RF conceived the study. SWA derived the equations, wrote the simulation script, and drafted the manuscript. PB, HL and MPLC involved in simulation. PB, JCMD and MPLC edited the drafted manuscript. PB, RF and JCMD involved in the derivation of the equations. All authors read and approved the final manuscript.

Funding
Iowa State University, Wageningen University and Research Centre.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]AL Rae Centre for Genetics and Breeding, Massey University, 10 Bisley Drive, Hamilton 3240, New Zealand. [2]Animal Breeding and Genomics, Wageningen University and Research, 6700 AH Wageningen, The Netherlands. [3]Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark. [4]Department of Animal Science, Iowa State University, Ames, IA 50011, USA.

Received: 4 February 2021 Accepted: 21 January 2022
Published online: 08 February 2022

References

1. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum Genet. 2001;69:1–14.
2. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10:381–91.
3. Dekkers JCM, Hospital F. The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet. 2002;3:22–32.
4. Fernando RL, Grossman M. Marker assisted selection using best linear unbiased prediction. Genet Sel Evol. 1989;21:467–77.
5. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
6. Lynch M, Walsh B. Genetics and analysis of quantitative traits. 1st ed. Sunderland: Sinauer Associates; 1998.
7. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet. 1968;38:226–31.
8. Rogers AR, Huff C. Linkage disequilibrium between loci with unknown phase. Genetics. 2009;182:839–44.
9. Falconer D, Mackay T. Introduction to quantitative genetics. Harlow: Pearson Education Limited; 1996.
10. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computin; 2021. http://www.R-project.org/. Accessed 01 Nov 2021.
11. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915;10:507–21.
12. Stuart A, Ord JK. Kendall's advanced theory of statistics: distribution theory. 6th ed. London: Halsted Press; 1994.
13. Berger S, Schlather M, de los Campos G, Weigend S, Preisinger R, Erbe M, et al. A scale-corrected comparison of linkage disequilibrium levels between genic and non-genic regions. PLoS One. 2015;10:e0141216.
14. Weir BS. Linkage disequilibrium and association mapping. Ann Rev Genomics Hum Genet. 2008;9:129–42.
15. Breeding AG, Cultivars H. In Principles of plant genetics and breeding. Chichester: John Wiley & Sons Ltd; 2012. p. 355–73.
16. Dekkers JCM, Mathur PK, Knol EF. Genetic improvement of the pig. In: Rothschild MF, Ruvinsky A, editors. The genetics of the pig. Wallingford: CABI Publishing; 2011. p. 390–425.
17. Arthur JA, Albers GAA. Industrial perspective on problems and issues associated with poultry breeding. In: Muir WM, Aggrey SE, editors. Poultry genetics, breeding and biotechnology. Wallingford: CABI Publishing; 2003. p. 1–12.

Publisher's Note