ORIGINAL ARTICLE

An evaluation of a novel estimator of linkage disequilibrium

D Gianola^{1,2}, S Qanbari³ and H Simianer³

The analysis of systems involving many loci is important in population and quantitative genetics. An important problem is the study of linkage disequilibrium (LD), a concept relevant in genome-enabled prediction of quantitative traits and in exploration of marker–phenotype associations. This article introduces a new estimator of a LD parameter (ρ^2) that is much easier to compute than a maximum likelihood (or Bayesian) estimate of a tetra-choric correlation. We examined the conjecture that the sampling distribution of the estimator of ρ^2 could be less frequency dependent than that of the estimator of r^2 , a widely used metric for assessing LD. This was done via an empirical evaluation of LD in 806 Holstein–Friesian cattle using 771 single-nucleotide polymorphism (SNP) markers and of HapMap III data on 21 991 SNPs (chromosome 3) observed in 88 unrelated individuals from Tuscany. Also, 1600 haplotypes over a region of 1 Mb simulated under the coalescent were used to estimate LD using the two measures. Subsequently, a simulation study compared the new estimator with that of r^2 using several scenarios of LD and allelic frequencies. From these studies, it is concluded that ρ^2 provides a useful metric for the study of LD as the distribution of its estimator is less frequency dependent than that of the standard estimator of r^2 .

Heredity (2013) **111**, 275–285; doi:10.1038/hdy.2013.46; published online 7 August 2013

Keywords: linkage disequilibrium; tetra-choric correlation; frequency dependence; Holstein–Friesian; HapMap; coalescent

INTRODUCTION

The analysis of systems involving many loci is relevant in population and quantitative genetics; for example, in studies of gene networks using multi-locus data. An important problem is that of measuring departures of the joint distribution of genotypes at two loci from independence. This is known as ‘linkage disequilibrium’ (LD) and arises, for instance, in genome-enabled prediction of quantitative traits using markers (for example, Meuwissen *et al.*, 2001) or in LD mapping. Association among alleles at various loci stem from physical proximity within a chromosome (linkage), mutation, random drift or selection favoring epistatic combinations. LD can also arise due to correlated allelic frequencies (Ohta, 1982) or when a covariance structure among allelic frequencies results from sub-structure within a population (Gianola *et al.*, 2012). An example of this is in dairy cattle, where there are distinct and large half-sib families. LD has received enormous attention in population genetics (for example, Hill, 1974, 1975; Hedrick, 1987; Lewontin, 1988; McVean, 2007) and has become an increasingly important object of study in the light of genomic data, for example, use of single-nucleotide polymorphism (SNP) markers and sequence information enabling the study of evolution of genomic blocks and of genome-wide associations with quantitative traits.

Here, we address the measurement of the statistical association between allelic states at two loci, that is, ‘gametic’ or ‘linkage disequilibrium’, via the pair-wise correlation measures that are typically used for characterizing this association. The most widely used measure of pair-wise LD is the squared correlation between allelic content or r^2 (Hill and Robertson, 1968); additional estimators are discussed by Hedrick (1987). Lewontin (1988) cautioned about the interpretation and behavior of these correlations (for example,

lack of some ‘invariance’ properties correlations have in a multivariate normal distribution) and warned that the distributions of the standard estimators are frequency dependent. For example, comparisons of the strength of LD structure between populations are confounded by variation in allelic frequencies.

Problems introduced by frequency-dependent estimators of parameters are well-known in quantitative genetic analysis of binary traits with threshold models (Dempster and Lerner, 1950; Gianola, 1982). For example, if a trait is coded as ‘present’ or ‘absent’, the variance of the trait depends on its mean value, and heritability is a function of the fraction of the population having the ‘present’ modality. This problem has been addressed by postulating an underlying (for example, Gaussian) distribution on which gene substitution takes place (Wright, 1934), followed by either transforming estimates from linear models using a formula suggested first by Alan Robertson (Dempster and Lerner, 1950) or by use of generalized linear models (Gianola and Foulley, 1983). The resulting estimators apply to a conceptual Gaussian scale, but their distribution is not completely frequency independent either. For example, Moreno *et al.* (1997), using simulation, found that the performance of Bayesian estimators of variance parameters in a threshold model depended on the extent of sparsity in the data. Although the distribution of estimates in a threshold model is frequency dependent, the parameterization allows for comparisons among populations via use of a scale in which frequencies do not intervene, at least explicitly.

Widely used estimators of LD, such as r^2 , calculated from (0, 1) binary data at each of the two loci (‘1’ denotes the presence of an allele ‘1’ on a locus A, say) are frequency dependent as well. For instance, the disequilibrium parameter $D = P_{ij} - p_i q_j$, where P_{ij} is the population frequency (probability) of haplotype ij and p_i and q_j are

¹Departments of Animal Sciences, Dairy Science, and Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA; ²Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, Norway and ³Animal Breeding and Genetics Group, Georg-August University Goettingen, Goettingen, Germany
Correspondence: Professor D Gianola, Department of Animal Sciences, University of Wisconsin-Madison, 1675 Observatory Dr, Madison, WI 53706, USA.
E-mail: gianola@ansci.wisc.edu

Received 5 October 2012; revised 5 February 2013; accepted 4 April 2013; published online 7 August 2013

the probabilities (allelic frequencies) of observing a '1' at either locus, is frequency dependent itself, by definition. Lewontin (1988) stated:

Our conclusion is that there are no generally gene frequency independent measures of association between loci, and that, indeed, the concept itself is an ill-defined one. In any particular case, we may be able to find a measure of association that is preserved under particular conditions, but the search for a "pure" measure of gametic disequilibrium is doomed to failure.

Even if a frequency-independent measure of LD may not exist, it would be desirable to develop one that is less affected by frequencies than r^2 . An obvious alternative for a 2×2 contingency table, where rows and columns denote allelic states at each of two loci and entries are counts of the observed haplotypes, is a tetra-choric correlation (Pearson, 1901). This is the correlation parameter of a standard bivariate normal latent process representing forces that create association between alleles, for example, selection favoring some combination of alleles. Apart from the threshold-liability models of animal breeding, tetra-choric correlations have been used in human genetics, for example, heritability of migraines and of dependency on opioids (Nyholt *et al.*, 2005; Gelernter *et al.*, 2006). Typically, estimates of tetra-choric correlations are larger than standard correlation statistics (for example, Ueberseax, 2010). However, it seems that the tetra-choric correlation has not been used for LD, at least with genomic data. This correlation can be estimated via maximum likelihood (Tallis, 1962) or by Bayesian methods. Although the calculations are fairly trivial by today's computing standards, this is not the case in the context of genome-wide data. For example, consider a LD analysis using 2.5 million *Drosophila* SNP markers (Ober *et al.*, 2012); a maximum likelihood or Bayesian analysis would require iterating many times over each of a huge number of marker pairs.

The objective of this article is to introduce a new estimator of a LD parameter (ρ^2) based on a metric proposed by Plackett (1965) that is much easier to compute than a maximum likelihood (or Bayesian) estimate of a tetra-choric correlation. Based on arguments given by Plackett (1965), the estimator $\hat{\rho}$ approximates well the tetra-choric correlation, and we conjectured that the sampling distribution of $\hat{\rho}^2$ is less frequency dependent than that of the estimator of r^2 . The manuscript is organized as follows. After this introduction, the 'new' estimator is presented and discussed. The following section presents an empirical evaluation of LD in Holstein-Friesian cattle using 771 SNP markers, followed by an application to a data set from HapMap III representing a sample of 88 unrelated Tuscan people assessed for 21 991 SNPs (chromosome 3). Subsequently, 1600 haplotypes over a region of 1 Mb were simulated under the coalescent, to estimate LD using the two measures. Finally, a modest simulation study contrasted the new estimator with r^2 with respect to statistical properties using several scenarios of LD and allelic frequencies. The paper concludes with the statement that ρ^2 provides a useful metric for the study of LD, especially when comparing populations or groups that differ in allelic frequencies.

A NOVEL ESTIMATOR OF LD

Consider a pair of bi-allelic loci and let (0, 1) denote the two possible allelic states at each of the loci. A standard assumption in LD analysis is that a random sample of size n is drawn from a homogeneous, unstructured population. The n data points (haplotypes) can be organized into the 2×2 contingency table below

Locus1 allele/Locus2 allele	0	1	
0	n_{00}	n_{01}	$n_{00} + n_{01}$
1	n_{10}	n_{11}	$n_{10} + n_{11}$
	$n_{00} + n_{10}$	$n_{01} + n_{11}$	$n = n_{00} + n_{01} + n_{10} + n_{11}$

where n_{00} , n_{01} , n_{10} and n_{11} are haplotype counts. As noted earlier, one possible model for association postulates an underlying Gaussian process creating a correlation (tetra-choric correlation) in a latent scale, as in threshold-liability models. Here, there would be two standard normal liability variates with correlation ρ . Tallis, (1962) presented a maximum likelihood procedure for estimating ρ from data in the contingency table. Although this is computationally fairly simple, calculating the association parameter for each pair of whole-genome markers does not scale well, because iteration must be carried out for each pair of loci. Hence, it would be desirable to produce an approximation that is simple to compute from highly dimensional genomic data. If such procedure leads to a characterization of LD that is less frequency dependent than that provided by the standard estimator of r^2 , this would be desirable as well.

Define the cross-product ratio

$$x = \frac{n_{00}n_{11}}{n_{10}n_{01}}$$

approaching in large numbers

$$x_{\text{Large}} = \frac{P_{00}P_{11}}{P_{10}P_{01}} \quad (1)$$

here, the P 's are true, unknown, gametic frequencies. Plackett (1965) introduced bivariate distributions indexed by a single parameter ψ that, in the case of the 2×2 table, takes the form $\psi = x_{\text{Large}} = P_{00}P_{11}/P_{10}P_{01}$. The relationship between the tetra-choric correlation and ψ is given by

$$\rho = -\cos \left[\pi \frac{\sqrt{\psi}}{1 + \sqrt{\psi}} \right] \quad (2)$$

When x_{Large} is < 1 , negative disequilibrium results; values of x_{Large} between 1/5 and 5 yield a set of values of ρ in the interval $(-0.56, 0.56)$. Rutledge (1977) showed that Equation (2) is precisely the tetra-choric correlation of Pearson (1901).

Let y and z be the frequencies of alleles '0' at each of loci A and B, respectively. In the standard parameterization of LD (for example, Hill, 1974; Lewontin, 1988), one has

$$\begin{aligned} P_{00} &= yz + D \\ P_{01} &= y(1-z) - D \\ P_{10} &= (1-y)z - D \\ P_{11} &= (1-y)(1-z) + D \end{aligned}$$

Using this representation, algebraic manipulation of Equation (1) leads to

$$x_{\text{Large}} = 1 + \frac{D}{[D - y(1-z)][D - z(1-y)]} \quad (3)$$

Now, a standard metric for characterizing LD (for example, Hill and Robertson, 1968; Lewontin, 1988) is the correlation between allelic contents (0, 1) in the discrete scale

$$r = \frac{D}{\sqrt{y(1-y)z(1-z)}}$$

Using this in Equation (3) produces the relationship

$$x_{\text{Large}} = 1 + \frac{r}{\left(r - \sqrt{\frac{\gamma(1-z)}{z(1-\gamma)}}\right) \left(r - \sqrt{\frac{z(1-\gamma)}{(1-z)\gamma}}\right)}.$$

Next, define the allelic odd ratios at loci A and B and parameter γ as

$$\text{odds}_A = \frac{\gamma}{1-\gamma}, \quad \text{odds}_B = \frac{z}{1-z}, \quad \gamma = \sqrt{\frac{\text{odds}_A}{\text{odds}_B}}$$

Then

$$x_{\text{Large}} = 1 + \frac{r}{(r-\gamma)\left(r-\frac{1}{\gamma}\right)} \quad (4)$$

Equation (4) is valid for values of r different from those of γ or from its reciprocal and, of course, provided that x_{Large} is positive; note the discontinuity in the relationship between r and x_{Large} at $r=\gamma$ or $r=\gamma^{-1}$.

In view of Equations (2) and (3), the following relationship between ρ and r can be established

$$\rho = -\cos \left[\frac{\pi \sqrt{1 + \frac{r}{(r-\gamma)\left(r-\frac{1}{\gamma}\right)}}}{\left(1 + \sqrt{1 + \frac{r}{(r-\gamma)\left(r-\frac{1}{\gamma}\right)}}\right)} \right], \quad (5)$$

with a discontinuity at $r=\gamma$ or $r=\gamma^{-1}$. The relationship is illustrated at $\gamma=1, 2$ or 3 in Figure 1. In cases a and b, the relationship is monotonic. Further, variations in r do not translate into the same amount of variation in ρ . For instance, when r is negative, parameter ρ is essentially insensitive with respect to variation in r . When r is positive, ρ is more sensitive, especially for $\gamma=2$. The discontinuity does not show up in cases a and b because ρ reaches 1 before $r=1$ or $r=0.5$, respectively. Also, note that there is always agreement in sign: if r is negative, so is ρ , and likewise if r is positive. Case c ($\gamma=3$) shows the effect of the discontinuity: when r reaches $1/3$, ρ is no longer finite, and at larger values, a positive r translates into negative

values of ρ . This is of course absurd: if Equation (4) is evaluated at $\gamma=3$, it can be verified that there are values of $r > 1/3$ that produce a negative x_{Large} ; this is not possible because this parameter is a ratio of products of probabilities, so it cannot take negative values.

The estimator of ρ suggested by Plackett (1965) is

$$\hat{\rho} = -\cos \left[\pi \frac{\sqrt{x}}{1 + \sqrt{x}} \right], \quad (6)$$

where $x = n_{00}n_{11}/n_{10}n_{01}$, for $n_{10} > 0$ and $n_{01} > 0$, so its calculation is trivial. Rutledge (1977) used simulation to evaluate Equation (6) in the context of repeatability of (0, 1) traits, that is, exactly as in the LD setting, and found that it had a minor upward bias, whereas the standard product-moment correlation (akin to r in LD) gave a sizable upward bias when inferring ρ . Garcia and Toro (1989) used Plackett's estimator in a Monte Carlo evaluation of four estimators of heritability of threshold traits; its performance tended to be better when the number of families and progeny group sizes were large. A derivation of the approximate standard error of the estimates obtained with Equation (6) is given in Appendix 1; Plackett (1965) gives the end results without presenting their derivation.

EMPIRICAL COMPARISONS OF ρ^2 VERSUS r^2

Cattle data

Genotypes from 771 SNPs across Bos taurus autosome 27 (BTA 27) with minor allelic frequencies (MAF) ≥ 0.05 were used; the distribution of SNPs over MAF bins was fairly uniform. To obtain reliable estimates of LD metrics, we used a sample of 806 Holstein-Friesian cattle, including 469 bulls and 341 bull dams; a detailed description of the data is in Qanbari *et al.* (2010). Mean neighbor marker distance was estimated at 72 kb. We reconstructed haplotypes for each chromosome using default options in fastPHASE (Scheet and Stephens, 2006). All 296835 SNP pairs on chromosome 27 were used to estimate pair-wise LD values as conveyed by the ρ^2 and r^2 metrics; densities of the distributions of the estimates are in Figure 2. Irrespective of the estimator used, most pairs of SNPs had a very weak association. However, the tetra-choric correlation (left panel) produced a less sharp distribution of estimates and a higher frequency of values > 0.2 . This is consistent with what is often observed when

Figure 1 Relationship between the tetra-choric correlation (ρ) and the correlation between binary allelic codes at loci A and B at selected values of $\gamma = \sqrt{\frac{\text{odds}_A}{\text{odds}_B}}$ (Gamma). (a) Gamma=1. (b) Gamma=2. (c) Gamma=3.

Figure 2 Cattle data: density of the distribution of estimates of LD parameters: Plackett's estimator at the left, and the standard metric at the right.

Figure 3 Cattle data: LD decay as a function of distance between pairs of SNPs over the entire chromosome (standard metric at the left, Plackett's estimator at the right).

Gaussian latent scale models are used to infer heritability, also a squared correlation (Dempster and Lerner, 1950; Gianola, 1982).

It is well known that there is a decay of LD measures as physical distance between markers increases, with LD expected to be a function of distance. We examined this relationship by plotting all r^2 and ρ^2 estimates against physical distance between marker pairs. As shown in Figure 3, both metrics showed a similar pattern; however, the magnitude of ρ^2 values was markedly larger than that of r^2 for all pair-wise intervals, and the decay of LD was clearly gentler. In order to visualize the decay of LD and the proportion of marker pairs in 'useful' (in the sense of genome-enabled prediction of complex traits) LD, we stacked the estimates and plotted these as a function of inter-marker distance classes (<0.025, 0.025–0.05, 0.05–0.075, 0.075–0.12, 0.12–0.2, 0.2–0.5, 0.5–1.5, 1.5–3 and 3–5 Mbp). The bar plot (Figure 4) illustrates the rate at which LD decays with physical distance and gives a more clear comparison between the two metrics. Figure 4a, based on r^2 , conveys a picture of weaker LD than that resulting from Figure 4b, corresponding to ρ^2 . A mean value of $r^2 = 0.20 \pm 0.32$ was estimated for pair-wise distances of 50–75 kb compared with $\rho^2 = 0.58 \pm 0.36$, which includes the average inter-marker space in this study. Despite the strong relationship between LD and distance, the proportion of physically close marker pairs with low values of r^2 was large. As shown in Figure 4, >70% of marker pairs at 25 kb apart are not in 'sufficient' LD to be useful for association studies as deemed by estimates of r^2 ; Ardlie *et al.* (2002) and Qanbari *et al.* (2010) discuss this issue further. With the ρ^2 metric, a larger proportion of marker pairs displayed strong LD; the

Figure 4 Cattle data: comparison of fraction of marker pairs with different LD levels (<0.1, 0.25, 0.4, 0.6 and >0.6, depicted by different colors) for marker pairs in different distance bins of maximum 5 Mb. Panels (a) and (b) represent r^2 and ρ^2 , respectively.

level at which ρ^2 would be 'useful' for association studies needs to be determined in future research.

When a new mutation occurs in a finite population, LD is created to a degree that depends on the frequency of the allele that is haplotyped together with the new mutation. It has been shown that estimates of LD between SNPs with a low MAF are biased upwards, and that high-frequency polymorphisms are better for accurate estimation of LD (Reich *et al.*, 2001). Although this is probably due to the frequency-dependent nature of the LD estimators (Lewontin, 1988; Pritchard and Przeworski, 2001), it should be kept in mind that low-frequency SNPs have a higher probability of having arisen recently (Nordborg and Tavaré, 2002), so discarding low frequency

SNPs from an analysis may give a distorted evolutionary picture. To examine this issue, we evaluated the decay of LD as a function of the difference in MAF (ΔMAF) between the two loci. Most of the pairs had a ΔMAF that was <0.40 . To explore the dependence of LD on allele frequencies, we calculated the average ρ^2 statistic for all SNP pairs within five bins of ΔMAF and compared results with those obtained by averaging r^2 . As ΔMAF increased, average r^2 decreased. On the other hand, average values of ρ^2 increased with a growing ΔMAF . As shown in Figure 5a, average ρ^2 was 0.065 for SNP pairs of $\Delta\text{MAF} \leq 0.1$, increasing to 0.1 for those in the largest ΔMAF bin. In contrast, r^2 started at 0.016 for frequency-matched SNPs and decreased to 0.009 for the bin with the largest ΔMAF . The emerging picture is that the two estimators are dependent on ΔMAF . Now, as estimates of ρ^2 were larger than those of their r^2 counterparts, a scale-free comparison is more sensible. Co-scaling the LD values as indicated in Figure 5b produced an increase of $\approx 30\%$ in magnitude of ρ^2 between frequency-matched SNPs relative to those with an extreme MAF. In the case of r^2 , co-scaling produced an 80% increase. Hence, the dependency of r^2 on ΔMAF was evidently stronger than that of ρ^2 . Furthermore, the trend of ρ^2 for SNPs of low or moderate ΔMAF was fairly flat, and considering that only 5% of the SNP pairs had extreme ΔMAFs , the dependency of ρ^2 on allele frequency difference is probably negligible (Figure 5, panel b).

Figure 5 Cattle data: (a) Average observed LD as a function of difference in the frequency of minor alleles at each pair of SNPs. (b) Using $x_i = \text{LD}_i / \max(\text{LD})$, for $i = 1, 2, \dots, 5$ ΔMAF bins, co-scaled values for each LD metric were computed as $\max(x) - x_i$. A full color version of this figure is available at the *Heredity* journal online.

In addition, we calculated the average of r^2 estimates within three bins of physical distance at five different ΔMAF bins and compared results with the average of ρ^2 estimates (Figure 6). As shown in Figure 6a, for pairs of SNPs <100 kb apart, a mean value of $r^2 = 0.23$ was observed for frequency-matched SNPs, and this sharply dropped to 0.04 for the pairs with the largest ΔMAF . In contrast, an overall mean value of $\rho^2 = 0.47$ was observed for frequency-matched SNPs <100 kb apart, which rose to 0.57 for those with the largest ΔMAFs . Co-scaling the LD metrics (Figure 6b) again indicated a stronger dependency of r^2 on allele frequencies. In particular, pairs of SNPs at short physical distance had r^2 values that were much more affected by ΔMAF than the corresponding ρ^2 , a trend that was observed for all ΔMAF bins. This dependency of co-scaled r^2 values was almost four times larger than that of ρ^2 . For markers within a distance range of 100 kb, the magnitude of ρ^2 increased from 26% to 39% as ΔMAF augmented, but for SNP pairs that were far from each other, the dependency was negligible. Furthermore, the dependency of co-scaled ρ^2 on allele frequency did not vary much with respect to physical distance between pairs of SNPs.

HapMap data

The Italian Tuscan population sample from the HapMap III data set was used to examine the behavior of the new metric when applied to human genomic data. The Italian Tuscan population involves 88 unrelated individuals, and ρ^2 and r^2 were estimated using the 21 991 SNPs across chromosome 3 (out of a total of 1 387 486). Figure 7 displays the relationships between the estimates of ρ^2 (right panel) and r^2 (left panel) as a function of physical distance and absolute value of the MAF difference between SNP pairs. As expected, the level of LD as conveyed by the two metrics decreased with increasing distance, although more gently for ρ^2 than for r^2 . It is evident in the plot that estimates of ρ^2 were essentially insensitive with respect to ΔMAF , whereas estimates of r^2 were clearly frequency dependent. In the case of r^2 , SNP pairs in short physical distance were more affected by ΔMAF .

This empirical study with cattle and human data indicates that ρ^2 also decays as a function of physical distance, a similar behavior to

Figure 6 Cattle data: panels (a) and (b) present the average and co-scaled measures of LD, respectively, as a function of ΔMAF for pairs of SNPs at three different physical distances. A full color version of this figure is available at the *Heredity* journal online.

Figure 7 Three-dimensional surface plots depicting the behavior of estimates of r^2 (left panel) and ρ^2 (right panel) as a function of inter-marker distance (Mb) and MAF interval (dMAF) for SNPs in chromosome 3 of the Italian Tuscan population in HapMap III. A full color version of this figure is available at the *Heredity* journal online.

that of other LD estimators. However, the new metric is clearly less dependent on the MAF difference between SNP pairs than r^2 and provides a more sensible and potentially useful new metric for analyzing LD across the genome. Although an entirely frequency-independent measure of LD does not seem possible to attain (Lewontin 1988), ρ^2 removes a major source of statistical noise when assessing LD structure. As already noted, further studies are needed to determine the ‘useful’ level of ρ^2 in association studies.

MONTE CARLO COMPARISON OF ρ^2 VERSUS r^2

Statistical properties

We examined small sample ($N=200$ individuals) properties of the estimator of ρ^2 and compared these with those of r^2 via two simulations. In the first simulation, referred to here as ‘ ρ -induced’, the strength of LD was fixed at the latent scale, at values of $\rho^2=0.04, 0.16, 0.36$ and 0.64 . In the second simulation, termed ‘ r -induced’, we took values of $r^2=0.04, 0.16, 0.36$ and 0.64 , but at the observable ($0, 1$) scale. In both simulations, the allelic frequencies used were $(0.1, 0.1)$, $(0.25, 0.25)$ and $(0.50, 0.50)$, where the numbers in each set refer to the frequencies of the ‘0’ allele for the two loci in question; only positive disequilibrium was considered; unequal frequencies were examined as well, leading to similar qualitative pictures, so results are

$\hat{y} = (n_{00} + n_{01} + 0.001)/(2N + 0.002)$. This is equivalent to a Bayesian posterior mean in which the allelic frequency is assigned a Beta $(0.001, 0.001)$ previous distribution. Likewise, the joint frequency in the $(0,0)$ cell was estimated as $\hat{P}_{00} = (n_{00} + 0.001)/(2N + 0.004)$. For the sample size considered ($N=200$), the effect of this shrinkage away from 0 is mild, and essentially negligible at larger sample sizes (results not presented). As the estimator of ρ is not defined when either n_{01} or n_{10} are null, if a simulation produced null values, these cell counts were set to 1.

ρ -induced simulation. As before, let y and z be the true frequencies of the ‘0’ alleles at the two loci and note that $P_{01} = y - P_{00}$, $P_{10} = z - P_{00}$ and $P_{11} = 1 - P_{00} - (y - P_{00}) - (z - P_{00}) = 1 + P_{00} - y - z$. From the definition of x_{Large} given in Equation (1), one can write

$$P_{00} = x_{\text{Large}} \frac{(y - P_{00})(z - P_{00})}{1 + P_{00} - y - z}$$

so that, given y, z and x_{Large} , and assuming equal allelic frequencies at the two loci ($y=z$), the following equation (with the unknown being P_{00}), is satisfied

$$P_{00}^2(1 - x_{\text{Large}}) + P_{00}(1 - 2y + 2x_{\text{Large}}y) - x_{\text{Large}}y^2 = 0.$$

This equation has three P_{00} roots, namely:

$$\begin{cases} \{y^2\} & \text{if } x_{\text{Large}} = 1 \\ \left\{ \frac{1}{2x_{\text{Large}} - 2} \left(1 + 2y(x_{\text{Large}} - 1) + \sqrt{1 + 4y(1 - y)(x_{\text{Large}} - 1)} \right), \frac{1}{2x_{\text{Large}} - 2} \left(1 + 2y(x_{\text{Large}} - 1) - \sqrt{1 + 4y(1 - y)(x_{\text{Large}} - 1)} \right) \right\} & \text{if } x_{\text{Large}} \neq 1 \end{cases}$$

not reported. For each LD by allelic frequency combination, 4000 independent samples were generated, and LD was estimated in each using either $\hat{\rho}^2$ or $\hat{r}^2$ as indicated below.

The rationale for the two different simulations is as follows: we attempted to avoid favoring a given estimator because the resulting haplotype frequencies depend non-trivially on the process that creates association, as this takes place either at the level of the latent scale (ρ -induced) or at that of the observed scale (r -induced). If disequilibrium is due to some latent process, then ρ is a relevant parameter; if there is no such latent process, then ρ lacks meaning. As shown subsequently, the joint distributions of alleles strongly depend on how parameters are elicited, and the same is true of the behavior of the corresponding estimators.

In the two simulations, allelic frequencies were estimated from the observed number of haplotypes in each realization, but with some shrinkage to avoid null frequency estimates. For instance,

The first root is the linkage equilibrium solution, $P_{00} = y^2$, which is trivial. The other two roots can be shown to approach y as x_{Large} goes to infinity; one of these two roots does approach to y from ‘above’ and the other one from ‘below’. As P_{00} is a probability, bounded between 0 and 1, we verified that

$$P_{00} = \frac{1}{2x_{\text{Large}} - 2} \left(1 + 2y(x_{\text{Large}} - 1) - \sqrt{1 + 4y(1 - y)(x_{\text{Large}} - 1)} \right) \quad (7)$$

generated a valid probability distribution, consistent with the allelic frequencies (both equal to y) used as input for solving the equation. Now, from the definition of the tetra-choric correlation

$$\rho = -\cos \left[\pi \frac{\sqrt{\psi}}{1 + \sqrt{\psi}} \right], \text{ it follows that}$$

$$\arccos(-\rho) - \pi \frac{\sqrt{x_{\text{Large}}}}{1 + \sqrt{x_{\text{Large}}}} = 0$$

and, given ρ , this equation can be solved numerically for an x_{Large} value that is consistent with the true level of disequilibrium in the latent scale. Finally, given x_{Large} and y , P_{00} in Equation (7) can be calculated readily, and the probability distribution generated has frequencies $P_{00} = \text{Equation (7)}$, $P_{01} = y - P_{00}$, $P_{10} = z - P_{00}$ and $P_{11} = 1 + P_{00} - y - z$. For example, let $\rho = 0.40$ and $y = 0.10$; then, with $\arccos(-0.4) = 1.9823132$,

$$\left(\frac{\sqrt{x_{\text{Large}}}}{1 + \sqrt{x_{\text{Large}}}} \right) = \frac{1.9823132}{\pi}.$$

Solving the above equation numerically yields $x_{\text{Large}} = 2.9239$, and using this together with y produces the ‘true’ probability distribution (after rounding) $P_{00} = 2.1773 \times 10^{-2}$, $P_{01} = P_{10} = 7.8227 \times 10^{-2}$ and $P_{11} = 0.8218$. It is important to note that while parameter ρ is fixed externally to the allelic (y) and haplotype frequencies (P ’s), the latter depend on ρ . In other words, the simulation input parameters are ρ and y .

From this probability distribution, the induced value of the correlation in the (0,1) scale is $r = (P_{00} - y^2)/y(1 - y)$. For the example given above, $r = 2.1773 \times 10^{-2} - (0.10)^2 / 0.10(0.90) = 0.13$, after rounding, compared with a ‘parental’ value of $\rho = 0.40$.

r-induced simulation. Here, the probability distribution was generated from the parametric definition (assuming equal allelic frequencies at the two loci)

$$r^2 = \frac{P_{00} - y^2}{[y(1 - y)]^2}$$

so that, assuming positive disequilibrium, $P_{00} = y[y + r(1 - y)]$. From this, given r and y one can solve for P_{00} and generate the entire haplotype distribution, as well as x_{Large} . Then, the ρ parameter is assessed from the relationship between ρ and x_{Large} but with x_{Large} evaluated as indicated previously. To illustrate, for $r = 0.40$ and $y = 0.10$, the probability distribution obtained is $P_{00} = 0.046$, $P_{01} = 0.054$, $P_{10} = 0.054$ and $P_{11} = 0.846$, with

$$x_{\text{Large}} = \frac{0.046 \times 0.846}{0.054 \times 0.054} = 13.345679$$

and $\rho = 0.78$, after rounding.

Results. Table 1 gives the results of the ρ -induced simulation. When disequilibrium was simulated at the latent level, the estimators of both ρ^2 and r^2 had an upward bias. However, as r^2 values were much smaller than those of ρ^2 , the relative bias was larger for the estimator of r^2 than for that of ρ^2 . The relative bias decreased as allelic frequency increased and also as disequilibrium became stronger, for example, for $\rho^2 = 0.16$ versus 0.64. When disequilibrium was very weak ($\rho^2 = 0.04$), the empirical bias of the estimator of r^2 did not vary as allelic frequency increased, but the relative bias decreased from 100% of the parameter value (for allele frequency equal to 0.10) to 19% (frequency equal to 0.50). On the other hand, for strong disequilibrium ($\rho^2 = 0.64$), the relative bias of the estimator of r^2 decreased from 3% to 1%. Clearly, the estimator of ρ^2 had a smaller relative bias throughout, which was practically insensitive with respect to allelic frequencies when its true value was 0.36 and 0.64. The inter-quartile ranges were smaller for the estimator of r^2 than for that of ρ^2 , a reflection of the larger values of the latter parameter. However, the relative precision, as measured by the coefficient of variation, was always smaller for the estimator of ρ^2 . Precision was higher as allelic frequencies increased and as latent level of disequilibrium became stronger. This simulation suggests that, to the extent that disequilibrium is caused by forces operating in a latent scale, such as selection (typically, selection operates on phenotypes and not on genotypes), ρ^2 provides a more sensible metric, conceptually and statistically, with the corresponding estimator being fairly insensitive to allelic frequencies except at mild levels of LD.

Results for the r -induced simulation are not presented, and we give a brief summary of the results. Here, disequilibrium was created and parameterized at the (0,1) scale: the true joint frequencies followed from the values of the true-positive disequilibrium at the r -scale and from the allelic frequencies. The bias of estimation of r^2 was negligible, except when LD was weak, whereas that of ρ^2 was very large, except at very strong levels of disequilibrium. Even though the estimator of ρ^2 had better relative precision, its strong bias negated this advantage. The explanation is that disequilibrium is not viewed as following from latent forces, so that ρ^2 does not have a clear mechanistic meaning.

A question is the extent to which the two estimators might lead to a drastically different inference about LD. Not surprisingly, the answer is frequency and sample size dependent. A comparison between the

Table 1 Mean, inter-quartile (third–first) range, coefficient of variation (CV, %) and relative absolute empirical bias (RAB, percentage of ‘true’ parameter value) of estimators of r^2 and ρ^2 at varying allelic frequencies at the two-loci (y, z) and levels of latent level linkage disequilibrium; ρ^2 is the disequilibrium parameter driving the process and r^2 is the induced value

Allelic frequencies (y, z)	Disequilibrium parameter (ρ^2, r^2)	Mean ($\hat{r}^2$)	Range ($\hat{r}^2$) (CV ($\hat{r}^2$))	RAB ($\hat{r}^2$)	Mean ($\hat{\rho}^2$)	Range ($\hat{\rho}^2$) (CV ($\hat{\rho}^2$))	Bias ($\hat{\rho}^2$)
0.1, 0.1	0.04, 0.003	0.006	0.008 (136)	100	0.068	0.071 (96)	70
0.25, 0.25	0.04, 0.010	0.013	0.015 (86)	30	0.049	0.050 (77)	23
0.50, 0.50	0.04, 0.016	0.019	0.017 (70)	19	0.046	0.041 (68)	15
0.1, 0.1	0.16, 0.019	0.021	0.022 (86)	11	0.170	0.143 (60)	11
0.25, 0.25	0.16, 0.048	0.051	0.032 (46)	6	0.164	0.091 (40)	3
0.50, 0.50	0.16, 0.069	0.070	0.034 (36)	1	0.164	0.075 (33)	3
0.1, 0.1	0.36, 0.059	0.063	0.047 (58)	7	0.357	0.142 (35)	1
0.25, 0.25	0.36, 0.130	0.133	0.051 (29)	2	0.364	0.108 (22)	1
0.50, 0.50	0.36, 0.168	0.169	0.051 (22)	1	0.362	0.096 (19)	1
0.1, 0.1	0.64, 0.178	0.183	0.088 (35)	3	0.636	0.135 (16)	0
0.25, 0.25	0.64, 0.298	0.300	0.072 (18)	1	0.641	0.092 (11)	0
0.50, 0.50	0.64, 0.348	0.351	0.065 (13)	1	0.642	0.083 (9)	0

Results are based on 4000 simulated samples of $N=200$ individuals for each frequency–disequilibrium combination.

Figure 8 Relationship between estimates (4000 samples of size $N=200$ or 800 each) of LD parameters r^2 and ρ^2 . 'True' $\rho^2=0.36$. Joint frequencies elicited from the knowledge of allelic frequencies and simulated value of ρ .

Figure 9 Relationship between estimates (4000 samples of size $N=200$ or 800 each) of LD parameters r^2 and ρ^2 . 'True' $r^2=0.25$. Joint frequencies elicited from the knowledge of allelic frequencies and simulated value of r .

Figure 10 Estimates of ρ^2 and of r^2 obtained with a simulation under the coalescent (details are in the text). A full color version of this figure is available at the *Heredity* journal online.

estimates of r^2 and ρ^2 at the three allelic frequencies considered and at sample sizes of $N=200$ or 800 individuals is shown in Figures 8 and 9 for the ρ -induced ($\rho^2=0.36$) and r -induced simulations ($r^2=0.25$), respectively. The estimates varied concomitantly, but the strength of their association depended more on allelic frequencies than on sample size. At intermediate allelic frequencies, the two sets of estimates were almost linearly related; at smaller frequencies, scatter was more patent. As seen throughout this study, to the extent that disequilibrium is the result of forces acting on some latent scale, the r^2 metric conveys a weaker picture of LD than ρ^2 .

Coalescent simulation

We also simulated a sample of 1600 haplotypes over a region of 1 Mb based on the coalescent model in the 'ms' simulator (Hudson, 2002) and estimated all pairwise LD using the two measures. Figure 10 displays the corresponding estimates of the two LD measures plotted against inter-marker distance and ΔMAF . The picture was fairly similar to what it was found with the cattle and HapMap data sets. The simulation and empirical results confirmed that ρ^2 produces a quantitatively and qualitatively different picture of LD from that resulting from r^2 and that the distribution of the estimates of ρ^2 was frequency independent for most practical purposes.

CONCLUSION

A novel metric for assessing pair-wise LD was developed and evaluated in an empirical data set using markers in a sample of dairy cattle, in a Tuscan population available in HapMap III and in small-scale simulation studies. The new estimator targets a parameter (ρ) reflecting an association in a conceptual bivariate normal distribution, the tetra-choric correlation of Pearson (1901); on the other hand, r^2 relates to the chi-square statistic for large samples. The ρ parameter can be estimated in a simple manner, using haplotype frequencies and a cosine transformation. Contrary to the method of maximum likelihood for estimating the tetra-choric correlation, which requires iteration (Tallis, 1962), ρ can be estimated without computational difficulty and it scales well when LD involves an analysis of high-dimensional genomic data. We found that, to the extent that it is sensible to view LD as stemming from forces that operate in a latent scale, the estimator is more precise,

and it is less affected by differences in MAFs than the estimator of r^2 , the most commonly used metric of LD. Although the distribution of the estimates still remains somewhat dependent on frequencies, ρ is an appealing measure and provides a picture of stronger LD. This is similar to what occurs with threshold-liability models of quantitative genetics, which are now common in genetic analysis of binary data: estimates of heritability are larger in a latent scale (Dempster and Lerner, 1950; Gianola and Foulley, 1983; Harville and Mee, 1984; Gilmour *et al.*, 1985). The method does not generalize easily to multiple alleles, but the concept extends naturally to more than two loci by introducing a higher-order latent Gaussian distribution. This would allow to examine LD in many different settings, introducing measures of conditional (given sets of loci) LD. In this case, the estimation procedure would probably have to be maximum likelihood or Bayesian, so computations would be considerably more involved.

In conclusion, our study supports the view that ρ^2 is a useful metric, one that facilitates comparison of levels of LD among populations differing in allelic frequencies; such comparison is distorted by the frequency-dependent nature of r^2 . On the other hand, the expected value of r^2 can be related to genetic parameters, such as effective population size (Sved, 1971). This implies that the distribution of estimates of effective size is frequency dependent as well. Perhaps similar relationships can also be developed for ρ^2 and this may be a subject for future research.

DATA ARCHIVING

Data deposited in the Dryad repository: doi:10.5061/dryad.t9f2g.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Part of this work was carried out while the senior author was a Visiting Professor at Georg-August-Universität, Göttingen (Alexander von Humboldt Foundation Senior Researcher Award). Support to DG by the Wisconsin Agriculture Experiment is acknowledged. SQ and HS acknowledge funding by the German Federal Ministry of Education and Research within the AgroCluster 'Synbreed—Synergistic plant and animal breeding', Grant ID: 0315528C.

- Ardlie KG, Kruglyak L, Seielstad M (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**: 299–309.
- Dempster ER, Lerner IM (1950). Heritability of threshold characters. *Genetics* **35**: 212–236.
- Garcia C, Toro MA (1989). A comparison of five estimators of the heritability of threshold characters. *J Anim Breeding Genet* **106**: 249–253.
- Gelernter J, Panhuysen C, Wilcox M, Hesselbrock V, Rounsaville B, Poling J *et al.* (2006). Genomewide linkage scan for opioid dependence and related traits. *Am J Hum Genet* **78**: 759–769.
- Gianola D (1982). Theory and analysis of threshold characters. *J Anim Sci* **54**: 1079–1096.
- Gianola D, Foulley JL (1983). Sire evaluation for ordered categorical data with a threshold model. *Génét Sélect Evol* **15**: 210–224.
- Gianola D, Manfredi S, Simianer H (2012). On measures of association among genetic variables. *Anim Genet* **43**: 19–35.
- Gilmour AR, Anderson RD, Rae AL (1985). The analysis of binomial data by generalized linear mixed model. *Biometrika* **72**: 593–599.
- Harville DA, Mee RW (1984). A mixed model procedure for analyzing ordered categorical data. *Biometrics* **40**: 393–408.
- Hedrick W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics* **117**: 331–341.
- Hill WG (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**: 229–239.
- Hill WG (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor Popul Biol* **8**: 117–126.
- Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226–231.
- Hudson RR (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.

- Lewontin RC (1988). On measures of gametic disequilibrium. *Genetics* **120**: 849–852.
- McVean G (2007). Linkage disequilibrium, recombination and selection. In: Balding DJ, Bishop M, Cannings C (eds). *Handbook of Statistical Genetics*, 3rd edn, John Wiley & Sons: Chichester, UK.
- Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome wide dense marker maps. *Genetics* **157**: 1819–1829.
- Moreno C, Sorensen D, Garcia-Cortés LA (1997). On biased inferences about variance components in the binary threshold model. *Genét Sév Evol* **29**: 145–160.
- Nyholt DR, Morley KI, Ferreira MAR, Medland SE, Boomsma DI, Heath AC *et al.* (2005). Genomewide significant linkage to migrainous headache on chromosome 5q21. *Am J Hum Genet* **77**: 500–512.
- Nordborg M, Tavaré S (2002). Linkage disequilibrium: what history has to tell us. *Trends Genet* **18**: 83–90.
- Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA *et al.* (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet* **8**: e1002685.
- Ohta T (1982). Linkage disequilibrium with the island model. *Genetics* **101**: 139–155.
- Pearson K (1901). I. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philos Trans R Soc Lond Ser A* **195**: 1–47.
- Plackett RL (1965). A class of bivariate distributions. *J Am Statist Assoc* **60**: 516–522.
- Pritchard JK, Przeworski M (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**: 1–14.
- Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR *et al.* (2010). The pattern of linkage disequilibrium in German Holstein cattle. *Anim Genet* **41**: 346–356.
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter D.J. *et al.* (2001). Linkage disequilibrium in the human genome. *Nature* **411**: 199–204.
- Rutledge JJ (1977). Repeatability of threshold traits. *Biometrics* **33**: 395–399.
- Scheet P, Stephens M.: A fast and flexible method for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629–644 (2006).
- Sved JA (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* **2**: 125–141.
- Tallis GM (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics* **18**: 342–353.
- Ueberseax J (2010). Introduction to the tetrachoric and polychoric correlation coefficients. <http://john-uebersax.com/stat/tetra.htm>
- Wright S (1934). An analysis of variability in number of digits in an inbred strain. *Genetics* **19**: 506–536.

APPENDIX 1

Under multinomial sampling, the variance–covariance matrix of the observed cell numbers in the 2×2 table of gametic type is

$$\text{Var} \begin{pmatrix} n_{00} \\ n_{11} \\ n_{01} \\ n_{10} \end{pmatrix} = n \begin{bmatrix} P_{00}(1-P_{00}) & -P_{00}P_{11} & -P_{00}P_{01} & -P_{00}P_{10} \\ -P_{00}P_{11} & P_{11}(1-P_{11}) & -P_{11}P_{01} & -P_{11}P_{10} \\ -P_{00}P_{01} & -P_{11}P_{01} & P_{01}(1-P_{01}) & -P_{01}P_{10} \\ -P_{00}P_{10} & -P_{11}P_{10} & -P_{01}P_{10} & P_{10}(1-P_{10}) \end{bmatrix} = \mathbf{V}.$$

As given in Equation (6), Plackett's estimator of the tetra-choric correlation is

$$\hat{\rho} = \cos \left(\frac{\pi \sqrt{x}}{(1 + \sqrt{x})} \right)$$

A first-order Taylor series approximation of $\hat{\rho}$ in the vicinity of the observed ratio $x = x_0$ is

$$\hat{\rho} \approx \rho_0 + \left. \frac{d\hat{\rho}}{dx} \right|_{x=x_0} (x - x_0),$$

where ρ_0 is the correlation parameter evaluated at $\psi = x_0$, and

$$\frac{d\hat{\rho}}{dx} = -\frac{d}{dx} \cos \left(\frac{\pi \sqrt{x}}{(1 + \sqrt{x})} \right) = \frac{\pi}{2} \frac{\sin \left(\pi \frac{\sqrt{x}}{\sqrt{x}+1} \right)}{2x + \sqrt{x} + \sqrt{x^3}}.$$

Then

$$\hat{\rho} \approx \rho_0 + \left. \frac{\pi}{2} \frac{\sin \left(\pi \frac{\sqrt{x}}{\sqrt{x}+1} \right)}{2x + \sqrt{x} + \sqrt{x^3}} \right|_{x=x_0} (x - x_0),$$

so that

$$\text{Var}(\hat{\rho}) \approx \left[\left. \frac{\pi}{2} \frac{\sin \left(\pi \frac{\sqrt{x}}{\sqrt{x}+1} \right)}{2x + \sqrt{x} + \sqrt{x^3}} \right|_{x=x_0} \right]^2 \text{Var}(x). \quad (\text{A1})$$

Note that $2x + \sqrt{x} + \sqrt{x^3} = \sqrt{x}(\sqrt{x} + 1)^2$. The next step consists of finding the approximate variance of x . Using a Taylor series approximation near x_0 again yields

$$\text{Var}(x) \approx \begin{bmatrix} \frac{d}{dn_{00}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \\ \frac{d}{dn_{11}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \\ \frac{d}{dn_{01}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \\ \frac{d}{dn_{10}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \end{bmatrix}' \times \mathbf{V} \times \begin{bmatrix} \frac{d}{dn_{00}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \\ \frac{d}{dn_{11}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \\ \frac{d}{dn_{01}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \\ \frac{d}{dn_{10}} \left(\frac{n_{00}n_{11}}{n_{10}n_{01}} \right) \end{bmatrix}, \quad n_{ij} = n_{\text{observed}}$$

with

$$\begin{aligned} \frac{d}{dn_{00}}(x) &= \frac{n_{11}}{n_{01}n_{10}} = \frac{x}{n_{00}}, \\ \frac{d}{dn_{11}}(x) &= \frac{n_{00}}{n_{01}n_{10}} = \frac{x}{n_{00}}, \\ \frac{d}{dn_{01}}(x) &= -\frac{n_{00}n_{11}}{n_{01}^2n_{10}} = -\frac{x}{n_{01}}, \\ \frac{d}{dn_{10}}(x) &= -\frac{n_{00}n_{11}}{n_{01}n_{10}^2} = -\frac{x}{n_{10}}. \end{aligned}$$

Carrying out the algebra and evaluating the derivatives at the observed cell numbers yields

$$\begin{aligned} \text{Var}(x) &\approx \\ &= nx_0^2 \left\{ \frac{P_{00}}{n_{00}} \left(\frac{P_{01}}{n_{01}} + \frac{P_{10}}{n_{10}} - \frac{P_{11}}{n_{11}} - \frac{1}{n_{00}}(P_{00} - 1) \right) \right. \\ &\quad + \frac{P_{01}}{n_{01}} \left(P_{00} \frac{1}{n_{00}} - \frac{P_{10}}{n_{10}} + \frac{P_{11}}{n_{11}} - \frac{1}{n_{01}}(P_{01} - 1) \right) \\ &\quad + \frac{P_{10}}{n_{10}} \left(P_{00} \frac{1}{n_{00}} - P_{01} \frac{1}{n_{01}} + \frac{P_{11}}{n_{11}} - \frac{1}{n_{10}}(P_{10} - 1) \right) \\ &\quad \left. - \frac{P_{11}}{n_{11}} \left(P_{00} \frac{1}{n_{00}} - P_{01} \frac{1}{n_{01}} - P_{10} \frac{1}{n_{10}} + \frac{1}{n_{11}}(P_{11} - 1) \right) \right\} \end{aligned}$$

Replacing the unknown probabilities by their relative frequency estimates $\hat{P}_{00} = n_{00}/n$,

$\hat{P}_{01} = n_{01}/n$, $\hat{P}_{10} = n_{10}/n$, $\hat{P}_{11} = n_{11}/n$, leads to:

$$\text{Var}(x) = x_0^2 \left(\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}} \right) = \frac{4x_0^2}{H}, \quad (\text{A2})$$

where H is the harmonic mean of the cell counts in the 2×2 table. Using this in Equation (A1) produces

$$\text{Var}(\hat{\rho}) \approx \frac{\pi^2}{H} \times \frac{x_0^2 \sin^2 \left(\pi \frac{\sqrt{x_0}}{\sqrt{x_0}+1} \right)}{\left(\sqrt{x_0} + 2x_0 + \sqrt{x_0^3} \right)^2} \quad (\text{A3})$$

Now, the function of x

$$f(x) = \frac{x^2 \sin^2 \left(\pi \frac{\sqrt{x}}{\sqrt{x}+1} \right)}{\left(2x + \sqrt{x} + \sqrt{x^3} \right)^2}$$

is such that

$$\lim_{x \rightarrow 0} \frac{x^2 \sin^2 \left(\pi \frac{\sqrt{x}}{\sqrt{x}+1} \right)}{\left(2x + \sqrt{x} + \sqrt{x^3} \right)^2} = \lim_{x \rightarrow \infty} \frac{x^2 \sin^2 \left(\pi \frac{\sqrt{x}}{\sqrt{x}+1} \right)}{\left(2x + \sqrt{x} + \sqrt{x^3} \right)^2} = 0.$$

A graph of $f(x)$ reveals that this function does not exceed 0.065. Therefore, $\text{Var}(\hat{\rho}) \leq 0.065\pi^2/H$ so that the approximate standard

error (SE) of the estimate has upper bounds at

$$\text{SE}(\hat{\rho}) \approx \sqrt{\text{Var}(\hat{\rho})} \leq \frac{0.64}{\sqrt{H}}.$$

For example, for $H = 100$ the standard error of the estimator of the correlation is $\frac{0.64}{10} = 0.064$.

Using a similar reasoning, one can approximate

$$\text{Var}(\hat{\rho}^2) \approx 4\hat{\rho}^2 \text{Var}(\hat{\rho}),$$

and $\text{SE}(\hat{\rho}^2) \approx 2(\hat{\rho})\sqrt{\text{Var}(\hat{\rho})}$. The approximate coefficient of variation of the estimator $\hat{\rho}$ can be estimated as $\text{CV}(\hat{\rho}) = 2\text{SE}(\sqrt{x}+1)$.