Heredity

Heredity (2012) 108, 285–291
© 2012 Macmillan Publishers Limited All rights reserved 0018-067X/12
www.nature.com/hdy

ORIGINAL ARTICLE

Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness

B Mangin[1], A Siberchicot[1], S Nicolas[2], A Doligez[2], P This[2] and C Cierco-Ayrolles[1]

Among the several linkage disequilibrium measures known to capture different features of the non-independence between alleles at different loci, the most commonly used for diallelic loci is the r^2 measure. In the present study, we tackled the problem of the bias of r^2 estimate, which results from the sample structure and/or the relatedness between genotyped individuals. We derived two novel linkage disequilibrium measures for diallelic loci that are both extensions of the usual r^2 measure. The first one, r_S^2, uses the population structure matrix, which consists of information about the origins of each individual and the admixture proportions of each individual genome. The second one, r_V^2, includes the kinship matrix into the calculation. These two corrections can be applied together in order to correct for both biases and are defined either on phased or unphased genotypes. We proved that these novel measures are linked to the power of association tests under the mixed linear model including structure and kinship corrections. We validated them on simulated data and applied them to real data sets collected on *Vitis vinifera* plants. Our results clearly showed the usefulness of the two corrected r^2 measures, which actually captured 'true' linkage disequilibrium unlike the usual r^2 measure.
Heredity (2012) **108,** 285–291; doi:10.1038/hdy.2011.73; published online 31 August 2011

Keywords: linkage disequilibrium; population structure; kinship coefficients; r^2 measure

INTRODUCTION

Linkage disequilibrium, the non-independence of alleles at different loci, has been widely used to infer important historical population events or detect the effect of selection pressure (Barton, 2011 and references therein). Two loci on a chromosome, with alleles *A*, *a* and *B*, *b* respectively, are said to be in linkage equilibrium, or gametic phase equilibrium, when each haplotype frequency is equal to the product of the corresponding allelic frequencies. Linkage disequilibrium refers to the deviation from this equilibrium and, when linkage phase is known, is measured as a function of the statistics $D=p_{AB}-p_Ap_B$ where p_{AB} is the frequency of the haplotype *AB* and p_A (p_B) is the frequency of the *A* (*B*) allele.

Various measures have been published in the literature for assessing statistical association between alleles at different loci (see for instance Hedrick, 1987 for a comparative study of six measures for diallelic loci and Cierco-Ayrolles *et al.*, 2004 for an investigation of their statistical properties). Most of these measures can be expressed as functions of covariances or correlations between loci, and, among them, one of the most commonly used for diallelic loci is the r^2 measure, the square of the loci correlation.

For association genetics, the extent to which measures of linkage disequilibrium reflect the true association between a marker and quantitative trait loci is an issue of considerable interest. Association mapping analysis requires careful design of experiments, since population structure and relatedness between individuals can lead to spurious associations (Myles *et al.*, 2009; Thornsberry *et al.*, 2001). The presence of individuals from different genetic origins within the sample produces linkage disequilibrium between unlinked loci, simply because of differences in allele frequencies. Consequently, such a structured sample can lead to a biased estimate of linkage disequilibrium, which may increase the rate of false positives statistically associated to the trait without actually being involved in its variation. Altshuler *et al.* (2008) and references therein mentioned this problem and recently, Mezmouk *et al.* (2011) illustrated the strong effect of structure corrections on the association mapping results using a real maize data set. Moreover, samples for linkage disequilibrium estimation are usually composed of unrelated individuals or individuals related through a pedigree so complex that it is ignored. A biased estimate is also obtained when genotyped individuals are not independent.

The r^2 linkage disequilibrium measure is used at different levels to design whole genome association mapping experiments. First, Pritchard and Przeworski (2001) showed that to achieve the same power of detection with a marker locus linked to the causal polymorphism, the sample size had to be increased by a factor of $1/r^2$. The r^2 measure was therefore extensively studied in order to evaluate the power of experiments to capture the genetic effects of causal polymorphisms with genotyped markers (Pe'er *et al.*, 2006). In particular, r^2 decay with physical and/or genetic distance between loci has often been analyzed to determine the density of markers to use in whole genome association scans (Stram, 2004). The long-range linkage disequilibrium present in admixed populations makes this task much more complicated. Also, analysis of the structure of r^2 is performed following association mapping to determine the physical window size in linkage equilibrium flanking the causal polymorphism and thereby the precision of its location. All these approaches rely on the assumption

[1]INRA, UR875 Unité de Biométrie et Intelligence Artificielle, Chemin de Borde Rouge, Castanet-Tolosan, France and [2]INRA IRD Montpellier-SupAgro Univ-Montpellier-2, Montpellier, France
Correspondence: Dr B Mangin, INRA, UR875 Unite de Biometrie et Intelligence Artificielle, Chemin de Borde Rouge, BP 62527, Castanet-Tolosan F-31320, France.
E-mail: brigitte.mangin@toulouse.inra.fr
Received 27 January 2011; revised 30 June 2011; accepted 11 July 2011; published online 31 August 2011

that the extent of r^2 around the causal polymorphism depends only on a drift-recombination process in a random-mating population without selection. This assumption is hardly ever true in real life. Therefore structure and relatedness were corrected for in association mapping tests to take into account the non-independence of loci due to both population differentiation and uneven levels of relatedness (Yu *et al.*, 2006). No corresponding correction for the r^2 measure itself has ever been developed to take into account the population differentiation and relatedness present within the samples although a biased estimation of linkage disequilibrium could lead to inappropriate choice of marker density and therefore in many cases to a reduced power of association tests.

In this study, our main objective was to tackle the problem of the bias of linkage disequilibrium estimate, resulting from the structure of the sample or the relatedness of genotyped individuals. We worked with a measure based on genotypes only. Thus we avoided the necessary preliminary step of inference of the haplotype phase. Among the various methods measuring linkage disequilibrium between loci with unknown phase (for example, Weir, 1996; Schaid, 2004), we focused on the one studied by Rogers and Huff (2009). Following Weir's pioneer work on a composite measure based only on genotypic data to evaluate linkage disequilibrium (Weir, 1979), Rogers and Huff (2009) validated this composite measure. We derived two new measures of linkage disequilibrium for diallelic loci, extending both the r^2 measure and the work of Rogers and Huff (2009). We proved that our structure-corrected measure is particularly useful for fine mapping purposes as it is the factor by which the sample size has to be increased in order to achieve the same power at a single nucleotide polymorphism (SNP) locus in linkage disequilibrium with the diallelic causal polymorphism as the one at the causal polymorphism itself. After validating both measures by simulation studies, we applied them to a real data set from *Vitis vinifera* plants.

MATERIALS AND METHODS

Theory

The r^2 measure of linkage disequilibrium. The standard measures of linkage disequilibrium, D and r^2, are respectively equivalent to the covariance and the correlation between alleles at two different loci l and m (Hill and Robertson, 1968). Consider two diallelic loci l and m, with alleles A and a at the l locus and alleles B and b at the m locus. There are four possible haplotypes, AB, Ab, aB and ab with respective frequencies p_{AB}, p_{Ab}, p_{aB} and p_{ab}. The linkage disequilibrium measure D is equal to:

$$D = p_{AB} - p_A p_B$$

where p_A and p_B denote the respective frequencies of alleles A and B.

Let X^l (respectively X^m) denote the dummy random variable equaling 1 when an individual carries the A allele at locus l (respectively B allele at locus m) and 0 otherwise. Then $D=Cov(X^l, X^m)$. By definition $r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$, thus it is equal to $Cor^2(X^l, X^m)$ as $Var(X^l)=p_A(1-p_A)$ and $Var(X^m)=p_B(1-p_B)$.

The case of sample structure. As in Ohta (1982), our aim is the decomposition of the r^2 measure of linkage disequilibrium into a first part due to the sample structure and a second part clear of the structure effect. We prove that the part clear of the structure effect has two major properties. First, it is unbiased for unlinked loci, unlike the usual r^2 measure of linkage disequilibrium. Second, it is the proportion factor of increase of the sample size necessary to achieve the same power when testing association at a SNP locus in linkage disequilibrium with the causal locus, under the linear Q model of association (Yu *et al.*, 2006).

Let us illustrate the idea leading to the decomposition of the r^2 measure of linkage disequilibrium by considering a simple situation. Assuming that gametes are sampled from two different populations, we denote by S the Bernoulli random variable that equals 1 when the gamete is sampled from the first population and 0 otherwise. If S is known, a natural statistical way of using this information, is to consider the conditional random variables $X^l|S$ and $X^m|S$. We propose the square of their correlation as a novel measure of linkage disequilibrium and we denote this novel measure r_S^2.

When allele frequencies are different between populations, the r^2 measure is not equal to zero for independent loci, as the ideal measure of linkage disequilibrium should be. Based on the following property of the conditional covariance,

$$Cov(X^l, X^m|S) = Cov(X^l, X^m) - Cov(X^l, S)Var^{-1}(S)Cov(S, X^m)$$

it can be proved that r_S^2 is equal to 0 for independent loci and thus that it corrects the bias of the usual r^2 measure (see APPENDIX A in Supplementary Information).

As defined, the novel measure allows a straightforward generalization to cases of samples composed of gametes coming from K different populations. The only changes needed are to set $S=(S_1, \dots, S_k, \dots, S_{K-1})^T$ where S_k is the Bernoulli random variable that equals 1 when the gamete is sampled from the kth population and 0 otherwise, and to use the *Cov* and *Var* matrix extensions. The generalization to the case of admixed populations, which is modeled by a known mixture $\sum_k p_{ik}S_k$ on K ancestral groups for each individual i, raises no further problem.

We now derive the way to compute r_S^2 sample value. Having sampled N individuals, we consider the $2N$ derived gametes and the population structure matrix S of the gametes in K populations. The S matrix is composed of $2N$ rows and $K-1$ columns. It can be obtained from Bayesian model-based clustering methods, as in STRUCTURE (University of Chicago, IL, USA; Pritchard *et al.*, 2000) and BAPS (Abo Akademi University, Turku, Finland; Corander *et al.*, 2008) software applications, by removing the last column of the output structure matrix, or from dimension-reduction methods, such as principal component analysis (Patterson *et al.*, 2006), non-metric multidimensional scaling (Zhu and Yu, 2009) or sparse factor analysis (Engelhardt and Stephens, 2010) by considering $K-1$ dimensions.

Let X^l, X^m be the 0–1 column vectors of the observed gametic alleles at loci l and m. We denote by $\Sigma_{X^l,X^m,S}$ the sample variance-covariance matrix of X^l, X^m and S. It consists in 3×3 blocks, namely

$$\begin{pmatrix} \Sigma_{X^l,X^l} & \Sigma_{X^l,X^m} & \Sigma_{X^l,S} \\ \Sigma_{X^m,X^l} & \Sigma_{X^m,X^m} & \Sigma_{X^m,S} \\ \Sigma_{S,X^l} & \Sigma_{S,X^m} & \Sigma_{S,S} \end{pmatrix}$$

Then $\hat{r}_S^2(l, m)$ the estimate of the $r_S^2(l, m)$ measure is equal to

$$\frac{\left(\Sigma_{X^l,X^m} - \Sigma_{X^l,S}\Sigma_{S,S}^{-1}\Sigma_{S,X^m}\right)^2}{\left(\Sigma_{X^l,X^l} - \Sigma_{X^l,S}\Sigma_{S,S}^{-1}\Sigma_{S,X^l}\right)\left(\Sigma_{X^m,X^m} - \Sigma_{X^m,S}\Sigma_{S,S}^{-1}\Sigma_{S,X^m}\right)} \quad (1)$$

The above equation is well defined if $\Sigma_{S,S}$ is not singular. Otherwise, it can be proved with the help of linear algebraic theorems (Searle, 1971 (chapter 2, p. 37)) that the number of ancestral groups can be reduced without loss of information.

The measure on unphased genotypes (Weir, 1979; Rogers and Huff, 2009) is defined by the correlation between X^l and X^m where X^l and X^m are now multinomial random variables taking values 0, 1 or 2 according to the number of A alleles in the genome at locus l (respectively m). As the population structure matrix can be obtained for genotypes as well as for gametes, it is straightforward to extend the r_S^2 measure to phase-unknown genotypes.

We give here the main arguments to prove that the power of the detection test at a SNP locus linked to the causal polymorphism is equal to the power at the causal polymorphism if the sample size is increased by $1/r_S^2$. The power of a t-test is an increasing function of its expectation, so we studied this expectation both at the causal locus and at a SNP locus in linkage disequilibrium. First, we show that $\hat{r}_S^2(l, m)$ is the factor that reduces the power when testing the association at the linked marker locus using the Q model and then we deal with two samples of different sizes. The complete proof is given in appendix (see APPENDIX B in Supplementary material).

The case of relatedness of sampled individuals. When the sample is composed of N related gametes, the sample correlation $\hat{r}^2(l, m)$ is not a good estimate of the $r^2(l, m)$ measure of linkage disequilibrium as the independence assumption is violated. The same problem occurs for the sample value of the correlation

conditionally to the structure, $r_S^2(l, m)$. However, if the covariance between the pairs of gametes is known, then the observations can be decorrelated by premultiplying the data by $V^{-1/2}$ if it exists, where V is a matrix built from the V_{ij} covariances for all the (i, j) pairs of gametes. Assuming that V is invertible, we denote $\Sigma^V_{X^l,X^m,S}$ the sample variance-covariance matrix of $V^{-1/2}[X^l, X^m, S]$ which is equal to

$$\left([X^l, X^m, S] - \frac{1_N 1_N^T V^{-1}}{1_N^T V^{-1} 1_N}[X^l, X^m, S]\right)^T V^{-1}\left([X^l, X^m, S] - \frac{1_N 1_N^T V^{-1}}{1_N^T V^{-1} 1_N}[X^l, X^m, S]\right).$$

We derive the estimate of the novel measure of linkage disequilibrium given the covariance between gametes

$$\hat{r}_V^2(l, m) = \frac{(\Sigma^V_{X^l,X^m})^2}{\Sigma^V_{X^l,X^l}\Sigma^V_{X^m,X^m}} \tag{2}$$

and the estimate of the novel measure of linkage disequilibrium given both the covariance between gametes and the sample structure

$$\hat{r}_{VS}^2(l, m) = \frac{\left(\Sigma^V_{X^l,X^m} - \Sigma^V_{X^l,S}(\Sigma^V_{S,S})^{-1}\Sigma^V_{S,X^m}\right)^2}{\left(\Sigma^V_{X^l,X^l} - \Sigma^V_{X^l,S}(\Sigma^V_{S,S})^{-1}\Sigma^V_{S,X^l}\right)\left(\Sigma^V_{X^m,X^m} - \Sigma^V_{X^m,S}(\Sigma^V_{S,S})^{-1}\Sigma^V_{S,X^m}\right)} \tag{3}$$

where $\Sigma^V_{\cdot,\cdot}$ denotes the blocks in the matrix $\Sigma^V_{X^l,X^m,S}$.

When the V matrix is not invertible, we propose to replace V^{-1} in equations (2) and (3) by V^{-}, the generalized Moore–Penrose inverse which is always defined.

As in the case of sample structure, extension to genotypes of unknown phase is straightforward by replacing all matrices and data related to gametes by the corresponding genotypic information, with V becoming the covariance matrix between genotypes.

The kinship coefficient which represents the degree of genetic covariance among individuals is the genetic parameter that should be used in the V matrix. It has already been proposed in association mapping methods to remove the bias of the type I error of association tests due to different degrees of relatedness between individuals and it was shown to be adequate (Yu *et al.*, 2006).

Computer simulations

The aim was to show that these novel measures actually corrected the bias of the usual r^2 estimate. Simulation process was conducted differently for the r_S^2 and the r_V^2 measures.

For the r_S^2 measure, each population was described by two parameters, p^l, p^m, the A (respectively B) allele frequency of the first (respectively the second) diallelic locus. The r^2 measure of linkage disequilibrium was assumed to be the same in both populations. In each population, a N sample was simulated using a multinomial distribution $M(1, p_{AB}, p_{Ab}, p_{aB}, p_{ab})$ where p_{AA}, p_{Ab}, p_{aB} and p_{ab} were the frequencies of the four allelic combinations linked to the above parameters.

For consistency with our real data set on grapevine, simulations were conducted with a sample size of 100 genotypes per population. First, we studied three levels of linkage disequilibrium by setting r^2 to 0.01, 0.25 and 0.5. The allelic frequencies were chosen in order to show a strong differentiation between the two populations either for both loci or for one locus only. Then, we explored the whole parameter space for r^2, p^l and p^m (see Supplementary material).

For the r_V^2 measure, two series of simulations were carried out. The first one involved in simulating a sample composed of both independent and cloned genotypes. A total of 80 genotypes were obtained by randomly associating two gametes that were simulated as described above with the multinomial distribution. A single genotype, homozygous for the most common allele at both loci, was chosen and repeated 19 times in order to create a set of 20 cloned genotypes within the sample of 100 genotypes. The V matrix was the kinship matrix with 1 diagonal. The cloned genotypes were assigned a kinship coefficient equal to 1, whereas all other kinship coefficients were set to 0. We studied three levels of linkage disequilibrium (r^2=0.05, 0.25, 0.4) and the allelic frequencies at both loci were chosen as either high (0.5) or medium (0.2).

The second series of simulations focused on a sample composed of full and half-sibs. A sample of 20 genotypes was simulated with a multinomial distribution. These 20 genotypes were used as the parents of a complete half-diallel of 190 F_2 families with 200 members each. For a given genetic distance between the two loci, the genotype of an offspring was obtained from the genotypes of its parents using Mendelian laws. The allelic transmission from parents to sampled descendants was preserved across replicates in order to limit the randomness to the multinomial draw. Thus, the parent genotypes varied across the replicates and the genotypes of descendants were rebuilt given the allelic transmission and their parent genotypes. Two random samplings of the descendants were used: either 100 individuals were randomly sampled in the whole population (simple sampling) or 5 individuals were randomly sampled in 20 randomly drawn families (hierarchical sampling). The V matrix was the kinship matrix with 1 diagonal. The kinship coefficients between the descendants were set to 1/4, 1/8 or 0 depending on the number of parents they shared. We set r^2 to 0.4, 0.25 and 0.05 and, based on a previous study on grapevine (Barnaud *et al.*, 2010, Figure 1), we linked these values to 0.1, 1 and 5 cM, respectively.

For both scenarii, we had to limit the number of replicates compared with the structure simulations because computing the inverse of a matrix drastically slowed down the computation.

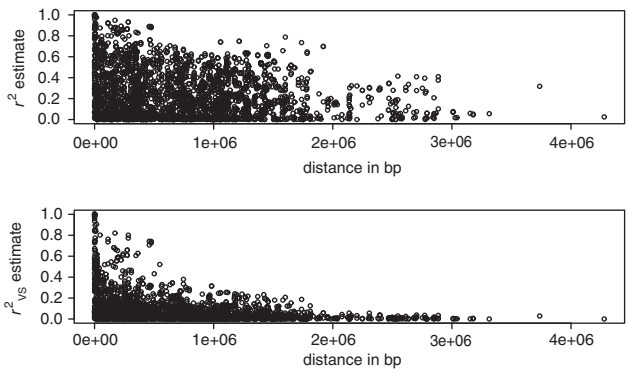

Figure 1 r^2 and r_{VS}^2 estimates for a sample composed of 92 wild and 91 cultivated *Vitis vinifera* accessions as a function of physical distance (in bp) between pairs of loci.

Plant material

To illustrate the behavior of the novel linkage disequilibrium measures, we also analyzed a real stratified population obtained by mixing cultivated and wild grapevine accessions with different kinship levels. Respectively 92 and 91 plants were selected among the wild and cultivated accessions of the *V. vinifera* grapevine germplasm collection of the vassal INRA experimental station. The 91 cultivated accessions (CU sample) were selected in the 'table east' sub-population (table grape cultivars from Eastern Europe). This sub-population was one of the three ones obtained by stratification analysis with STRUCTURE software application, using 20 microsatellite markers spanning the 19 chromosomes on the 2276 distinct genotypes of the cultivated *V. vinifera* germplasm. The 92 wild accessions selected (WI sample) came mainly from Western Europe and most particularly from France.

Molecular analysis

DNA was extracted from these 183 accessions and genotyped at 176 SNP markers using VeraCode technology (Illumina, San Diego, CA, USA). A total of 109 SNPs belonging to 50 different genes within a 2-Mb region were used to estimate linkage disequilibrium. In all, 67 additional SNPs originating from 67 different genes scattered on the whole genome were used together with the 20 scattered SSRs to measure relatedness between the 183 accessions. All SNPs were physically localized on the sequence of grapevine genome by blasting their flanking sequence on the 12×reference grapevine genome sequence (Jaillon *et al.*, 2007).

Data analysis

Linkage disequilibrium analysis was performed for the 80 SNPs out of 109 with a minimum allele frequency >10%. We checked stratification between wild

and cultivated grapevine accessions by performing a phylogenetic analysis based on the dissimilarity matrix estimated from the 87 markers scattered on the whole genome (see above). We obtained a clear classification into two genetic pools corresponding to wild and cultivated accessions (data not shown). The structure matrix was obtained from the same 87 scattered markers with STRUCTURE assuming no admixture. The kinship matrix was obtained with the same 87 scattered markers with ML-Relate (Montana state University, Bozeman, MT, USA), a software application that uses maximum likelihood to estimate relatedness between individuals (Milligan, 2003).

RESULTS

Simulation study

Validation of the bias correction for structure. To validate the r_S^2 measure, simulations were conducted on a two-population structured sample.

Table 1 presents the mean (and its standard error) of the estimates of r_S^2 and r^2, computed over 5000 replicates. The r^2 measure was estimated both in the first population (100 genotypes) and in the whole sample (200 genotypes).

As expected, the higher the differentiation of the loci, the higher was the bias of the r^2 estimate. In all cases, the r_S^2 measure corrected the bias of the r^2 estimate in the whole sample. Depending on allele frequencies and linkage disequilibrium values, the r^2 estimate either under or overestimated the r^2 value but in all cases $\hat{r}_S^2$ was nearly unbiased. For the variability of the estimates, as shown by the standard error, the trends were opposite. The difference in $\hat{r}^2$ variance between the one and two-differentiated locus situations decreased with increasing linkage disequilibrium whereas the opposite was observed for $\hat{r}_S^2$. However, the trends for the variability of $\hat{r}_S^2$ and $\hat{r}^2$ in the first population, were the same, suggesting that this is the usual trend of a correct estimate of linkage disequilibrium based on correlation across gametes. $\hat{r}_S^2$ showed less unbiasedness and a smaller variance than $\hat{r}^2$ in the first population because the sample size was twice as large.

The r_S^2 estimate was almost unbiased over the whole parameter space, as the absolute difference between its mean over the 500 replicates and the r^2 value was always <0.03 (see Supplementary material).

Validation of the bias correction for relatedness. To validate the r_V^2 measure, two scenarii were simulated, a clone scenario and a half-diallel design. Tables 2 and 3 present the mean (and its standard error)

Table 1 Mean (and its standard error) of the r^2 and the r_S^2 estimates (over 5000 replicates) of a two-population sample of size (2×100)

r^2	0.01		0.01		0.25		0.25		0.5		0.5	
pop 1 p^l	0.9		0.9		0.9		0.8		0.9		0.7	
pop 2 p^l	0.1		0.1		0.1		0.2		0.1		0.3	
pop 1 p^m	0.9		0.55		0.9		0.55		0.9		0.55	
pop 2 p^m	0.1		0.45		0.1		0.45		0.1		0.45	
pop 1 $\hat{r}^2$	0.018	(0.0003)	0.015	(0.0002)	0.258	(0.0014)	0.251	(0.0007)	0.504	(0.0017)	0.501	(0.0008)
whole $\hat{r}^2$	0.460	(0.0007)	0.022	(0.0002)	0.673	(0.0007)	0.211	(0.0005)	0.801	(0.0006)	0.469	(0.0007)
$\hat{r}_S^2$	0.014	(0.0002)	0.012	(0.0001)	0.253	(0.0010)	0.250	(0.0005)	0.502	(0.0012)	0.500	(0.0006)

p^1 is the *A* allele frequency at the first locus in the first (pop 1) and the second (pop 2) population, p^m is the *B* allele frequency at the second locus in the first (pop 1) and the second (pop 2) population, the r^2 measure was estimated with the sample of the first population (pop 1) and with the whole sample (both populations).

Table 2 Mean (and its standard error) of the r^2 and the r_V^2 estimates (over 5000 replicates) of a sample composed of 80 independent genotypes and 20 clones

r^2	0.4		0.4		0.25		0.25		0.05		0.05	
p^l	0.5		0.2		0.5		0.2		0.5		0.2	
p^m	0.5		0.2		0.5		0.2		0.5		0.2	
$\hat{r}^2$	0.543	(0.0035)	0.450	(0.0049)	0.414	(0.0038)	0.308	(0.0047)	0.209	(0.0031)	0.098	(0.0028)
$\hat{r}_V^2$	0.396	(0.0043)	0.407	(0.0051)	0.248	(0.0042)	0.261	(0.0047)	0.060	(0.0022)	0.063	(0.0024)

p^l is the *A* allele frequency at the first locus, p^m is the *B* allele frequency at the second locus.

Table 3 Mean (and its standard error) of the r^2 and the r_V^2 estimates (over 500 replicates) for two different samplings of 100 genotypes from a complete half-diallel of 190 families of 200 full-sibs

r^2	0.4		0.4		0.25		0.25		0.05		0.05	
p^l	0.5		0.2		0.5		0.2		0.5		0.2	
p^m	0.5		0.2		0.5		0.2		0.5		0.2	
dist (cM)	0.1		0.1		1		1		5		5	
	Hierarchical sampling of 5 full-sibs in 20 families											
$\hat{r}^2$	0.403	(0.0113)	0.466	(0.0128)	0.276	(0.0094)	0.331	(0.0115)	0.083	(0.0045)	0.133	(0.0073)
$\hat{r}_V^2$	0.399	(0.0116)	0.407	(0.0131)	0.258	(0.0097)	0.264	(0.0110)	0.063	(0.0040)	0.083	(0.0055)
	Simple sampling of 100 genotypes											
$\hat{r}^2$	0.400	(0.0108)	0.450	(0.0117)	0.239	(0.0088)	0.305	(0.0106)	0.075	(0.0044)	0.119	(0.0068)
$\hat{r}_V^2$	0.400	(0.0107)	0.431	(0.0117)	0.240	(0.0088)	0.285	(0.0105)	0.072	(0.0044)	0.0115	(0.0064)

p^l is the *A* allele frequency at the first locus, p^m is the *B* allele frequency at the second locus.

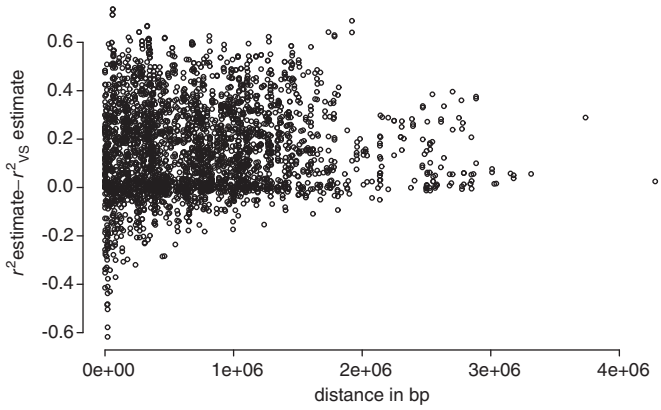

Figure 2 Difference between estimated r^2 and r^2_{VS} for a sample composed of 92 wild and 91 cultivated *Vitis vinifera* accessions as a function of physical distance (in bp) between pairs of loci.

of the estimates of r^2_V and r^2, over 500 replicates. In the clone scenario, the large overestimation of linkage disequilibrium when using the r^2 measure was expected since the cloned genotype was the most common haplotype. In the half diallel design, as parents could have different haplotypes and they all produced the same numbers of offspring, the impact on the bias of r^2 estimate was not predictable. This bias was largely positive for medium allele frequencies whereas it was positive but nearly null for common alleles.

$\hat{r}^2_V$ was always less biased than $\hat{r}^2$. Bias correction was nearly perfect for common alleles but bias was only partially removed for medium allelic frequencies in the worst situation of simple population sampling with low r^2. Variability was very close between r^2 and r^2_V estimates. It was substantially larger in the half-diallel design than in the clone scenario, most probably because the effective population size was much lower in the half-diallel (only 20 parents) than in the clone scenario (80).

Grapevine data

r^2 and r^2_{VS} estimates are presented in Figure 1 as a function of the physical distance between loci. With the r^2 measure, long-range linkage disequilibrium was questionable and could be due to either the structure of the sample and/or the high mean level of relatedness (0.32) within the wild sample as compared with the cultivated one (0.10). With the r^2_{VS} measure, lower values overall were obtained, as well as an expected exponential decline of linkage disequilibrium with distance, which clearly demonstrated the efficiency of our novel measure in correcting bias.

Figure 2 shows the difference between the estimates of r^2 and r^2_{VS}, plotted as a function of the distance between loci. The positive bias was removed all along the chromosome segment. However, for some close loci, r^2_{VS} estimate was larger than $\hat{r}^2$, leading to a negative removed bias.

DISCUSSION

We proposed two novel measures of linkage disequilibrium that correct for the bias of the usual r^2 measure due to the structure of the sample or the relatedness of the genotyped individuals. These two measures can be combined in order to correct for both sources of bias and can be computed either on phased or unphased genotypes. Estimation of these novel measures of linkage disequilibrium was implemented in a R package (LDcorSV) and integrated in the SNiPlay pipeline, which is freely available at http://sniplay.cirad.fr/cgi-bin/home.cgi. These measures are based on formulae involving the covariance between gametes or diploid genotypes. They make full use of the added information given by the population structure or kinship matrices.

We proved, in the case of a two-population structured sample, that the estimate of the measure corrected for the structure of the sample, r^2_S, is unbiased for unlinked loci. We showed by simulation of a two-population structured sample that the bias of the usual estimate $\hat{r}^2$ increased with the differentiation of loci and with decreasing linkage disequilibrium, as already shown by Ohta (1982). This bias was always removed by the use of the r^2_S estimate. Our simulations were limited to two populations having different levels of linkage disequilibrium only because of different allele frequencies. Both populations thus have similar r^2 values and an unbiased estimate of r^2 can be clearly defined as an estimate with expectation r^2. On the contrary, in the case where the two populations would have both different r^2 values and different allele frequencies, it is not clear what should be the correct measure of linkage disequilibrium, and consequently what would be the bias of the usual r^2 estimate. If two populations diverged a long time ago and underwent very different evolutions, it may be wrong to assume that they differ only in allele frequencies. However, for geographically close populations, for recently admixed populations, or for genotypes pooled to form a panel for association studies, assuming the r^2 value is the same is not unsoundness. Moreover, if we assume equal effective population sizes and panmixia in each population, the approximation of the expectation of r^2 along the evolution process depends on Nc, where N is the effective population size and c the per-generation rate of recombination (Ohta and Kimura, 1971). Therefore, r^2 may be linked to the recombination rate, which is expected to be similar in two close populations.

Another extreme clone scenario could be useful to show that the r^2 measure may also underestimate the true linkage disequilibrium value, in which case the r^2_V estimate would be much larger than the usual r^2 estimate. Let n_{AB} (respectively n_{ab}) denote the number of haplotypes carrying A allele at the first locus and B allele at the second locus (a and b alleles, respectively). If there were only one haplotype carrying Ab and one haplotype carrying aB, the r^2 estimated value of the sample would be large, nearly equal to 1 if n_{AB} and n_{ab} were large. However, if the single Ab haplotype and the single aB haplotype were both cloned, then the r^2 estimated value of this new sample would decrease to zero with an increasing number of clones. Unlike the r^2 estimate, the r^2_V estimate would remain equal to the r^2 estimate of the original sample and would thus be nearly unbiased for a large sample size. By simulating a diallel design of full-sibs, which is less extreme and thus more likely to reflect real situations, we showed that the usual measure was always more biased than the novel one. However, we observed a remaining bias that was probably due to the small effective size of our sample as observed by Yan *et al.* (2009) with a real maize data set when the sample size was <50.

Analyses of the real data set revealed a number of close loci with a linkage disequilibrium underestimated by the usual measure. This negative bias appeared even when we performed exclusively the structure correction (Supplementary Figure S1 in Supplementary material), although it was never found in the simulation study. The reason is probably that we performed simulations with exactly the same r^2 value in both populations, as our goal was to estimate the part of linkage disequilibrium linked to the recombination rate within populations in panmixia. In our real data set, these pairs of close loci

showed a significant difference of r^2 estimates between populations. However, among these pairs, those with the highest differences between r^2 and r_S^2, had low allelic frequency for one locus within at least one population (Supplementary Figure S2 in Supplementary material). Given the small sample size within each population (<100 accessions), one must proceed with caution when drawing conclusions but this suggests that those loci had probably undergone selection.

Recently, Zaykin *et al.* (2008) proposed an extension of the r^2 measure to handle multiallelic loci. Their T^2 statistic, which is the sum of the r^2 measures over all pairs of alleles multiplied by a constant term, was shown to be efficient for testing for independence between loci in phase-known genotypes. The transformations using population structure and kinship matrices could be applied to the T^2 statistics to correct for bias effects.

Kinship matrices are rarely known and there are several methods and software applications for obtaining an estimated kinship matrix based on molecular marker data (Loiselle *et al.*, 1995; Lynch and Ritland, 1999; Milligan, 2003). However, they all have the limited goal of estimating pairwise kinship coefficients, none of them estimates a whole kinship matrix. The estimates of pairwise kinship coefficients of the whole sample, even when all of them are constrained to be positive or null, do not provide a valid matrix for the r_V^2 measure (that is, a positive semi-definite matrix). The simplest method of building a positive semi-definite matrix from a non-positive one, is to perform a single value decomposition and set all negative eigenvalues to 0. We used this method to analyze our real data set. However, as highlighted by Saïdou *et al.* (2009), this projection method does not necessarily lead to a valid kinship matrix (that is, a matrix with elements >0 and <1). Which is the best method when computing the r_V^2 measure, either to use a valid kinship matrix or simply a valid variance-covariance matrix, remains an open question. In the association mapping context with a mixed model approach, Kang *et al.* (2008) compared three different variance-covariance matrices of the random genotype background effect: a SPAGeDi-based kinship estimate, a genotype similarity estimate and a phylogeny control estimate. Based on both the BIC criterion, which measures the goodness of fit of the model, and the *P*-value distribution, which shows the correction needed for inflated false positive tests, the three estimates gave comparable results. For linkage disequilibrium estimation, we do not have a clear criterion, such as the BIC or the *P*-value distribution, that could help us to compare different variance-covariance matrices on real data sets. However, we may rely on Kang *et al.* (2008) comparisons to suggest that comparable results would probably be obtained.

We proved that the sample size has to be increased by $1/r_S^2$ to achieve the same power at a SNP locus in linkage disequilibrium with a causal locus when using the association test corrected by the sample structure (the Q model of Yu *et al.*, 2006). If V is the variance-covariance matrix of the K+Q model (Yu *et al.*, 2006), then the r_{VS}^2 measure can be proved to be the proportional factor of the association test power. The proof follows the same steps as for the r_S measure, after premultiplying the K+Q model equation by $V^{-1/2}$. The variance-covariance matrix of this mixed linear model is generally unknown, as it depends on the heritability of the observed trait. However, the r_{VS}^2 measure could be computed for a range of heritabilities to help to select the marker density, or could be estimated after the association tests in order to assess the precision around the detected SNPs.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Stéphane Nicolas was supported by a postdoctoral fellowship of the Institut National de la Recherche Agronomique. This work was carried out with the financial support of the ANR—Agence Nationale de la Recherche, The French National Research Agency and the CNIV—Comité National des Interprofessions des Vins à Appellation d'Origine, under the Programme Génomique des Plantes project ANR-08-GENM-02, DLVitis. We thank the INRA experimental station at Vassal for plant management; Thierry Lacombe for in natura collection of wild accessions and his valuable help with the delimitation of the wild and cultivated study populations; the genomic platform of Genotoul in Toulouse for technical support in the acquisition of genotypic data with the Illumina VeraCode technology; Audrey Weber, Muriel Latreille and Sylvain Santoni for DNA extraction. Resequencing for SNP detection was supported by the INRA project SNPGrapeMap.

Altshuler D, Daly MJ, Lander ES (2008). Genetic mapping in human disease. *Science* **322**: 881–888.
Barnaud A, Laucou V, This P, Lacombe T, Doligez A (2010). Linkage disequilibrium in wild french grapevine, vitis vinifera l. subsp. silvestris. *Heredity* **104**: 431–437.
Barton NH (2011). Estimating linkage disequilibria. *Heredity* **106**: 205–206.
Cierco-Ayrolles C, Abdallah J, Boitard S, Chikhi L, de Rochambeau H, Tsitrone A *et al.* (2004). *Recent Research Developments in Genetics and Breeding*. Research Signpost: Kerala, India.
Corander J, Marttinen P, Sirén J, Tang J (2008). Enhanced bayesian modelling in baps software for learning genetic structures of populations. *BMC Bioinformatics* **9**: 539.
Engelhardt BE, Stephens M (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* **6**: e1001117.
Hedrick PW (1987). Gametic disequilibrium measures: proceed with caution. *Genetics* **117**: 331–341.
Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226–231.
Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.
Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al.* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**: 1709–1723.
Loiselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, psychotria officinalis (rubiaceae). *Am J Bot* **82**: 1420–1425.
Lynch M, Ritland K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1753–1766.
Mezmouk S, Dubreuil P, Bosio M, Décousset L, Charcosset A, Praud S *et al.* (2011). Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. *Theor Appl Genet* **122**: 1149–1160.
Milligan BG (2003). Maximum-likelihood estimation of relatedness. *Genetics* **163**: 1153–1167.
Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich D *et al.* (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**: 2194–2202.
Ohta T (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc Natl Acad Sci* **79**: 1940–1944.
Ohta T, Kimura M (1971). Behaviour of neutral mutants influenced by associated over-dominant loci in finite populations. *Genetics* **69**: 247–260.
Patterson N, Price AL, Reich D (2006). Population structure and eigen analysis. *PloS Genet* **2**: 2074–2093.
Pe'er I, de Bakker PIW, Maller J, Yelensky R, Altshuler D, Daly MJ (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* **38**: 663–667.
Pritchard JK, Przeworski M (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**: 1–14.
Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
Rogers AR, Huff C (2009). Linkage disequilibrium between loci with unknown phase. *Genetics* **182**: 839–844.
Saïdou AA, Mariac C, Luong V, Pham JL, Bezançon G, Vigouroux Y (2009). Association studies identify natural variation at PHYC linked to flowering time and morphological efficient control of population structure in model organism association mapping variation in pearl millet. *Genetics* **182**: 899–910.
Schaid DJ (2004). Linkage disequilibrium testing when linkage phase is unknown. *Genetics* **166**: 505–512.
Searle SR (1971). *Linear models*. John Wiley & Sons Inc: New York, USA.
Stram D (2004). Tag snp selection for association studies. *Genet Epidemiol* **27**: 365–374.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* **28**: 286–289.
Weir BS (1979). Inferences about linkage disequilibrium. *Biometrics* **35**: 235–254.
Weir BS (1996). *Genetic Data analysis*. Sinauer associates Inc: Sunderland, USA.
Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using snp markers. *PLoS ONE* **4**: e8451.
Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF *et al.* (2006). A unified mixed model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 1061–4036.
Zaykin DV, Pudovkin A, Weir BS (2008). Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics* **180**: 533–545.
Zhu C, Yu J (2009). Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* **182**: 875–888.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)