Heredity

Heredity (2012) 109, 401–402
© 2012 Macmillan Publishers Limited All rights reserved 0018-067X/12
www.nature.com/hdy

LETTER TO THE EDITOR

Measuring linkage disequilibrium by the partial correlation coefficient

Heredity (2012) **109**, 401–402; doi:10.1038/hdy.2012.54; published online 5 September 2012

Gametic phase disequilibrium, often referred to as linkage disequilibrium (LD), describes the non-independence of alleles at different loci on the same chromosome. There are various measures of LD proposed in the literature (Hedrick, 1987; Devlin and Risch, 1995) for the purposes of inferring population evolutionary history and mapping genes (Slatkin, 2008). In a recent paper in this journal, Mangin *et al.* (2012) proposed a new LD measure r_S^2 aiming to correct the bias due to population structure by taking into account of the population structure matrix. In this letter, we point out that r_S^2 is essentially the square of the partial correlation coefficient between two loci given the population structure, which was not explicitly explained in the paper. We also distinguish between the partial correlation and the conditional correlation, as the latter was ambiguously used in the paper. We further extend the result on the relationship between r_S^2 and power of association tests to generalized linear models and discuss the potential use of r_S^2 in human genetic mapping.

A natural way to measure LD is by the correlation coefficient. Consider two diallelic loci A and B, with alleles A_1 and A_2 at locus A and alleles B_1 and B_2 at locus B. Denote by p_i, where $i \in \{A_1, A_2, B_1, B_2\}$, allele frequencies, and by p_j, where $j \in \{A_1B_1, A_1B_2, A_2B_1, A_2B_2\}$, haplotype frequencies. The widely used LD measure $r_{AB}^2 = D^2/p_{A_1}p_{A_2}p_{B_1}p_{B_2}$, where $D = p_{A_1B_1} - p_{A_1}p_{B_1}$, is the square of Pearson's correlation coefficient that measures the linear dependence between the two loci (Hill and Robertson, 1968). Suppose in a sample there exists population structure that can distort the correlation between the two loci. One way to measure LD controlling for confounding effects is by the partial correlation coefficient. Denote by Y_A and Y_B the random variables of genotypes at loci A and B, respectively, and by S a vector of variables on the population structure. Regress Y_A and Y_B on S by the linear regression models $Y_A = S\beta_A + \varepsilon_A$ and $Y_B = S\beta_B + \varepsilon_B$, respectively, where β_A and β_B are regression coefficients, and ε_A and ε_B are residuals. The partial correlation $r_{AB.S}$ between Y_A and Y_B controlling for S is then defined as Pearson's correlation between the residual variables ε_A and ε_B (Yule, 1907). Alternatively, the partial correlation $r_{AB.S}$ can be calculated as a negative off-diagonal element of the inverse correlation matrix (Whittaker, 1990), which is exactly the square root of formula (1) in Mangin *et al.* (2012). Therefore, the new LD measure they proposed is the square of the partial correlation coefficient—$r_{AB.S}^2$—between two loci controlling for the population structure, which is a direct extension of the original measure r_{AB}^2 that is used in the absence of population stratification.

As the formula of partial covariance was referred to as the one for 'conditional covariance' in the paper (p 286), it is worth pointing out that these two are equivalent only in special situations, such as when variables follow a multivariate normal distribution. The partial correlation $r_{AB.S}$ in general is not equal to the conditional correlation $r_{AB|S}$. The former by definition is independent of S, whereas the latter is not necessarily free of S. Even if $r_{AB|S}$ is free of S, there exists the inequality $r_{AB.S}^2 \leqslant r_{AB\,|\,S}^2$, where the equality holds when both the conditional variances and covariance of Y_A and Y_B given S are free of S (Lawrance, 1976). Below we performed a simulation study to show their subtle difference in case of $r_{AB|S}$ being independent of S. Simulation settings I, III and V mimicked those by Mangin *et al.* (2012) (Table 1), except for replacing $r_{AB}^2 = 0.01$ by 0.1; in settings II, IV and VI, the allele frequencies in the second population were also changed. In these six settings, the two loci were in the same degree of LD in the two populations but with different minor allele frequencies. In each population 1000 haplotypes were simulated and randomly assigned into 500 pairs. The genotypes were then scored in an additive fashion. The crude sample correlation coefficient $\hat{r}_{crude}^2$, the partial correlation coefficient $\hat{r}_{AB.S}^2$ and the conditional correlation coefficient $\hat{r}_{AB\,|\,S}^2$ were estimated based on the genotypic scores. Ten thousand replicates were simulated, and the mean and standard error of the estimates were recorded in Table 1. In all settings, $\hat{r}_{AB.S}^2$ was smaller than $\hat{r}_{AB\,|\,S}^2$, but the difference between them was small. Theoretically, in settings I, III and V they should be equal because the minor allele and the major allele at each diallelic locus were simply flipped between the two populations, and thus both the conditional variances and covariance of Y_A and Y_B given S are free of S; the small differences $(\sim 10^{-4})$ were due to sampling errors. In settings II, IV and VI, the differences between them $(\sim 10^{-3})$ were one order of magnitude greater than that in settings I, III and V.

Measuring LD between loci by $r_{AB.S}^2$ in the case of population stratification is in the same spirit as measuring correlation between covariate-adjusted phenotypes and genotypes in genetic association studies (Price *et al.*, 2006; Xing *et al.*, 2011). Suppose an allele at locus A is the causal variant for a trait. Mangin *et al.* (2012) derived in a linear regression setting that the power to detect association between the trait and locus A would be reduced by a factor of $r_{AB.S}^2$ when locus B was examined instead. As a matter of fact, this conclusion holds in general when modeling phenotype–genotype association by a generalized linear model, as ε_B can be viewed as a surrogate variable for ε_A, and it is well known the asymptotic relative efficiency of a test using ε_B versus using ε_A equals the square of their correlation coefficient (Lagakos, 1988; Tosteson and Tsiatis, 1988).

Characterizing LD structure is instructive in designing genetic association studies, which is a major goal of the International HapMap Consortium (2005) and the 1000 Genomes Project Consortium (2010). These projects focus on genetically homogeneous populations to document population-specific parameters. However, in reality, a study sample can be genetically heterogeneous with

Table 1 Mean (and its s.e.[a]) of correlation coefficient estimates of a mixture of two populations[b]

	I	*II*	*III*	*IV*	*V*	*VI*
Population 1						
r^2_{AB}	0.10	0.10	0.25	0.25	0.50	0.50
p_{A1}	0.90	0.90	0.90	0.80	0.90	0.70
p_{B1}	0.90	0.55	0.90	0.55	0.90	0.55
Population 2						
r^2_{AB}	0.10	0.10	0.25	0.25	0.50	0.50
p_{A1}	0.10	0.70	0.10	0.65	0.10	0.60
p_{B1}	0.10	0.70	0.10	0.65	0.10	0.60
Mixed Population						
$\hat{r}^2_{crude}$	0.7227 (0.0188)	0.0438 (0.0131)	0.7925 (0.0160)	0.1980 (0.0247)	0.8757 (0.0125)	0.4716 (0.0281)
$\hat{r}^2_{AB.S}$	0.1012 (0.0247)	0.0969 (0.0177)	0.2508 (0.0366)	0.2484 (0.0251)	0.5007 (0.0412)	0.4995 (0.0265)
$\hat{r}^2_{AB\|S}$	0.1014 (0.0247)	0.1030 (0.0186)	0.2511 (0.0366)	0.2515 (0.0253)	0.5011 (0.0412)	0.5005 (0.0264)

[a]Based on 10 000 replicates.
[b]Five hundred diploid subjects from each population.

substructure even though the recruiting criterion requires a specific ethnic group. Imagine the diversity of African Americans in a metropolitan area. Considering that a lot of genome-wide association studies have been carried out, it will be valuable to use these available genome-wide genotypes to document ethnic- and geographic-specific $r^2_{AB.S}$ for the purpose of facilitating future genetic studies conducted in the same population.

Finally, we also want to point out that the other LD measure r^2_V proposed by Mangin *et al.* (2012) for the purpose of correcting the bias due to relatedness is the square of the correlation coefficient of two loci modeled by a linear regression—the coefficient of determination—using generalized least squares given the kinship matrix instead of using ordinary least squares and assuming independence between subjects as when calculating the usual correlation coefficient.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Dr Robert Elston for critically reading, commenting and editing the paper. This study is supported by the American Heart Association Scientist Development Grant 10SDG4220051 to CX.

C-Y Lin[1,2], G Xing[3] and C Xing[1,4]
[1]*McDermott Center of Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA;*
[2]*Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan;*
[3]*Bristol-Myers Squibb Company, Pennington, NJ, USA and*
[4]*Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA*
E-mail: chao.xing@utsouthwestern.edu

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Devlin B, Risch N (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311–322.

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.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**: 1299–1320.

Lagakos SW (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Stat Med* **7**: 257–274.

Lawrance AJ (1976). On conditional and partial correlation. *Am Stat* **30**: 146–149.

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

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Slatkin M (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**: 477–485.

Tosteson TD, Tsiatis AA (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika* **75**: 507–514.

Whittaker J (1990). *Graphical Models in Applied Multivariate Statistics*, 1st edn. New York, John Wiley and Sons.

Xing G, Lin CY, Xing C (2011). A comparison of approaches to control for confounding factors by regression models. *Hum Hered* **72**: 194–205.

Yule GU (1907). On the theory of correlation for any number of variables treated by a new system of notation. *Proc Roy Soc A* **79**: 182–193.