

TECHNICAL NOTE

Indices of multilocus linkage disequilibrium

PAUL-MICHAEL AGAPOW and AUSTIN BURT

Department of Biology, Imperial College, Silwood Park, Ascot, Berkshire, SL5 7PY, UK

Abstract

Linkage disequilibrium is an ubiquitous biological phenomenon. However a common metric for disequilibrium – the index of association or I_A – is dependent on sample size. In this paper we present a modification of I_A that removes this dependency. This method has been implemented in a software package.

Keywords: correlation coefficient, index of association, linkage disequilibrium, missing data, MultiLocus, randomization

Received 5 October 2000; revision accepted 9 November 2000

Many interesting biological processes can cause linkage disequilibrium, that is, statistical associations between alleles at different loci. These processes include population differentiation and isolation by distance, asexual reproduction, linkage, and natural selection. Consequently, it is often useful to test for linkage disequilibrium in multilocus data sets. However, there are practical difficulties: for n loci, there are $n(n-1)/2$ pairs of loci between which one can test for disequilibrium. For a survey of 10 loci, there would be 45 different tests, even leaving aside higher order associations (Weir 1996). One common approach is to summarize the data with a single multilocus measure of linkage disequilibrium, the index of association (I_A) (Brown *et al.* 1980; Maynard Smith *et al.* 1993; Haubold *et al.* 1998). This summary statistic is based on the variance of pairwise distances between individuals (i.e. the number of loci at which they differ). Thus, it essentially tests to what extent individuals that are the same (or different) at one locus are more likely than random to be the same (or different) at other loci. Significance tests can be based either on comparisons of the observed value to those for randomized data sets (Burt *et al.* 1996) or using on an analytical approximation (Haubold *et al.* 1998). However, it is an unfortunate property of I_A that its expected value depends upon the sample size of loci, which makes comparisons among populations difficult (Brown *et al.* 1980; Maynard Smith *et al.* 1993). Here we propose a modification of I_A which largely removes this dependency on number of loci. We also suggest how to test for significance when there are missing data.

The index of association is defined as

$$I_A = (V_o/V_e) - 1$$

where V_o is the observed variance of pairwise distances and V_e is the variance expected in the absence of linkage disequilibrium. Pairwise distances can also be calculated for each locus separately (in which case they are either 0, if the isolates are the same, or 1, if they are different). Let the variance of pairwise distances at locus j be var_j . V_e is equal to the sum of this variance over loci, $V_e = \sum \text{var}_j$, and V_o is this plus twice the covariance of distances, summed over all pairs of loci:

$$V_o = \sum \text{var}_j + 2\sum \text{cov}_{j,k}$$

Substituting these expressions into the definition of I_A , we get

$$I_A = \left(\sum \text{var}_j + 2\sum \text{cov}_{j,k} \right) / \sum \text{var}_j - 1 = 2\sum \text{cov}_{j,k} / \sum \text{var}_j$$

This formulation highlights the weakness of the statistic: the number of terms in the numerator increases as n^2 , while the number of terms in the denominator only increases with n . Thus, unless the covariances are 0, I_A will increase with n , the number of loci.

To avoid this problem, we suggest an alternative standardization for the covariances:

$$\bar{r}_d = \sum \text{cov}_{j,k} / \left(\sum \sqrt{\text{var}_j \cdot \text{var}_k} \right)$$

This expression has a form similar to a correlation coefficient, with a maximum value of 1; hence we use the symbol r , with subscript d referring to distances.

To confirm that \bar{r}_d is largely independent of n , we constructed an artificial data set with considerable linkage disequilibrium and calculated I_A and \bar{r}_d for the complete

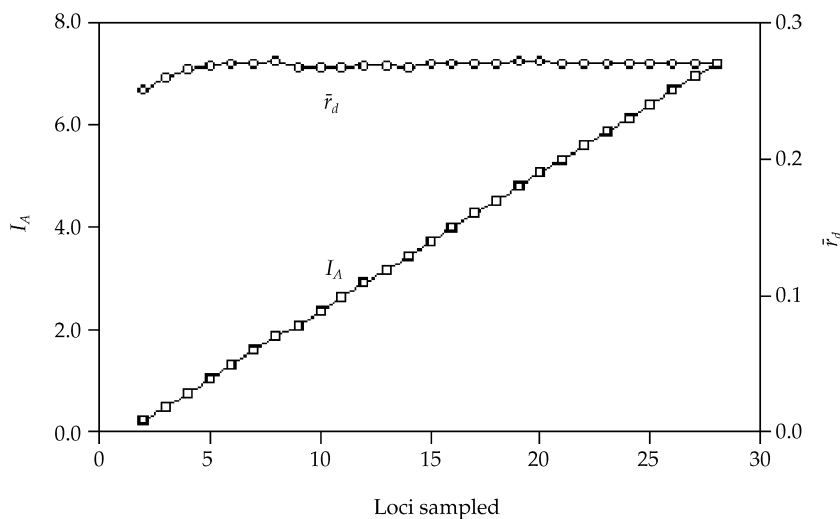


Fig. 1 Indices of multilocus linkage disequilibrium and number of loci. A data set of 32 individuals and 29 loci was constructed by arranging the individuals in a comb-like phylogeny and having each internal branch supported by one locus. I_A and \bar{r}_d were calculated both for the full data set and for random subsets in which varying numbers of loci were included in the analysis. Each sample size was replicated 1000 times. Calculated using MultiLocus v1.2.

data set and for subsamples with varying numbers of loci. The results show a clear dependency of I_A , but not \bar{r}_d , on number of loci (Fig. 1). Use of \bar{r}_d should, therefore, facilitate comparisons among populations.

Tests of significance by randomization are conceptually simple, more robust than analytical approximations (Haubold *et al.* 1998), and easy to perform with modern computers. For any particular data set, I_A and \bar{r}_d are monotonically related across randomizations (because the numerator is the same for the two statistics, and neither denominator changes with randomization). Therefore, *P*-values for the two statistics from randomizations will be identical.

One potential difficulty in tests of significance, is how to deal with missing data. This is particularly problematic if the missing data are nonrandomly distributed such that some individuals have more than others, as is often the case. Under such circumstances, one might conclude there is linkage disequilibrium when in fact none exists. We propose that for tests of significance in the presence of missing data, the observed data set be compared to randomized ones in which the missing data have been fixed in place and the true data shuffled around them. This ensures that any structure in the missing data is also replicated in all randomized data sets: any difference between observed and randomized data sets must, therefore, be due to the differences in the real data.

MultiLocus, a computer program implementing these suggestions is available from <http://www.bio.ic.ac.uk/evolve/software/multilocus/>. Versions are available for Macs and PCs.

Acknowledgements

P-M Agapow was supported by the Natural Environment Research Council through grant GR3/11526.

References

- Brown AHD, Feldman MW & Nevo E (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics*, **96**, 523–536.
- Burt A, Carter DA, Koenig GL, White TJ & Taylor JW (1996) Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. *Proceedings of the National Academy of Sciences of the USA*, **93**, 770–773.
- Haubold B, Travisano M, Rainey PB & Hudson RR (1998) Detecting linkage disequilibrium in bacterial populations. *Genetics*, **150**, 1341–1348.
- Maynard Smith J, Smith NH, O'Rourke M & Spratt BG (1993) How clonal are bacteria? *Proceedings of the National Academy of Sciences of the USA*, **90**, 4384–4388.
- Weir BS (1996) *Genetic Data Analysis II*. Sinauer, Sunderland.