npg

NEWS AND COMMENTARY

Gems from the *Heredity* Archive
Estimating linkage disequilibria

NH Barton

Outcrossing sexual populations are described rather well by 'bean-bag genetics': alleles are typically randomly associated with each other, so that populations can be described simply by listing their allele frequencies (Haldane, 1964). Deviations from random association can be produced by random drift, migration, or selection, and so can be used to estimate the strength of these processes. Moreover, because associations between alleles at different loci—'linkage disequilibria'—are broken down by recombination, they can be used to map recombination rates and to detect the genes responsible for quantitative variation—and in particular, for variation in disease susceptibility. Although linkage disequilibrium (LD) has been seen as a rather obscure aspect of population genetics, it is now central to the analysis of genomic data (Slatkin, 2008).

In a seminal paper, Hill (1974) set out methods for estimating the strength of the association between alleles at two biallelic loci, in a randomly mating diploid population. This would be simple if one could directly observe the two haploid genotypes that make up each diploid individual, but that would usually require tedious crosses. In diploids, the two kinds of double heterozygote cannot be distinguished. However, Hill (1974) showed that this loss of information is counterbalanced by the extra genetic information carried in diploids, relative to haploids. Therefore, estimates of LD made from diploids are just as precise as those made from the same number of haploids, which makes elaborate crosses unnecessary.

Since the paper by Hill (1974), there have been many developments in statistical methodology. For example, Slatkin and Excoffier (1996) extended maximum likelihood estimation to multiple alleles. Rogers and Huff (2009) showed that LD can be estimated from the covariance between diploid genotypes at two loci. This is much simpler than the iterative maximum likelihood algorithm proposed by Hill (1974), and almost as efficient, with the advantage that it extends to allow inbreeding.

Much of population genetics now takes a genealogical view, following the coalescence of ancestral lineages. McVean (2002) showed that under the infinite sites model, there is a very simple relation between LD and coalescence: the variance of LD is proportional to the covariance between coalescence times at the two loci. Nevertheless, although theoretical methods have developed greatly, the basic framework remains that set out by Hill (1974).

Hill's paper was stimulated by the advent of electrophoresis, which for the first time allowed surveys of LD across large numbers of loci. The main aim was to detect selection for particular combinations of alleles. Despite considerable efforts, little LD was found, except between very tightly linked loci (for example within inversions), in asexuals or selfers, or where the population contained cryptic species. The significant LD that was found could be largely attributed to the joint effects of random drift and recombination (Langley *et al.*, 1978); recent surveys of $>10^6$ single nucleotide polymorphism loci have given exceptionally detailed maps of recombination rates across the human genome (Myers *et al.*, 2005). An exception is where distinct populations meet in narrow hybrid zones: then, admixture can generate strong LD even between unlinked loci, and this can give good estimates of the rate of mixing (Szymura and Barton, 1986).

Over the last decade, the availability of genome sequences has given LD a prominent role. Nearby sites do not evolve independently, and so any analysis of genome sequences must take account of LD. In fact, genomes are divided into distinct blocks, each with a different history, and if the rate of mutation is high enough relative to recombination, these haplotype blocks can be seen more or less directly. The traditional population genetic analysis in terms of coefficients of LD is another way to represent this haplotype structure, in the same way that we can follow either the allele frequencies or the genealogy at a single locus.

Analysis of LD has two main applications. First, selection can be detected through reduced diversity in regions of reduced recombination, which is caused by LD between neutral alleles and selected sites. There may be 'background selection' against deleterious mutations, or 'selective sweeps' caused by fixation of favourable mutations (Smith and Haigh, 1974); if they occurred recently, sweeps can be detected through the characteristic pattern of LD that they produce. Second, quantitative trait loci can be mapped through associations between the trait and genetic markers—which must reflect LD with the underlying quantitative trait loci (Weir, 2008). Such association studies have much greater resolution than crosses or pedigree studies, as recombination has acted for very many generations—equal to the typical depth of the genealogy. Their resolution is limited by the extent of LD—in humans, to roughly ~10 kb.

Apart from these practical applications, LD is intimately involved in the evolutionary process. Although populations can often be approximated as 'bean bags' of independently evolving alleles, the random LD generated by drift interferes with adaptation (Hill–Robertson interference), and this most likely generates the selection that maintains recombination and sex (Barton, 2009). The clear methodology for estimating LD that was introduced by Hill (1974) has done much to facilitate our understanding of the role of LD in quantitative genetics and evolution.

Conflict of interest

The author declares no conflict of interest.

NH Barton is at the IST Austria, Klosterneuburg, 3400 Austria or Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK.

e-mail: nick.barton@ist.ac.at

Barton NH (2009). Why sex and recombination? *Cold Spring Harbor Symp Quant Biol* **74**; (doi:10.1101/sqb.2009.74.030).

Haldane JBS (1964). A defense of beanbag genetics. *Perspect Biol Med* **7**: 343–359.

Hill WG (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**: 229–239.

Langley CH, Smith DB, Johnson FM (1978). Analysis of linkage disequilibria between allozyme loci in natural populations of *Drosophila melanogaster*. *Genet Res* **32**: 215–229.

McVean GAT (2002). A genealogical interpretation of linkage disequilibrium. *Genetics* **162**: 987–991.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321–324.

Rogers AR, Huff C (2009). Linkage disequilibrium between loci with unknown phase. *Genetics* **182**: 839–844.

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

Slatkin M, Excoffier L (1996). Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. *Heredity* **76**: 377–383.

Smith JM, Haigh J (1974). The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23–35.

Szymura JM, Barton NH (1986). Genetic analysis of a hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*, near Cracow in Southern Poland. *Evolution* **40**: 1141–1159.

Weir BS (2008). Linkage disequilibrium and association mapping. *Ann Rev Genomics Hum Genet* **9**: 129–142.

Editor's suggested reading

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.

Bellis C, Cox HC, Ovcaric M, Begley KN, Lea RA, Quinlan S (2008). Linkage disequilibrium analysis in the genetically isolated Norfolk Island population. *Heredity* **100**: 366–373.

Li MH, Merilä J (2010). Extensive linkage disequilibrium in a wild bird population. *Heredity* **104**: 600–610.

Servedio MR (2009). The role of linkage disequilibrium in the evolution of premating isolation. *Heredity* **102**: 51–56.