Why Linkage Disequilibrium Helps Us Find Selective Sweeps

Alan R. Rogers

November 5, 2018

It is now easy to scan the entire human genome for evidence of natural selection. One class of methods involves linkage disequilibrium (LD), which tends to be inflated in the neighborhood of ongoing selective sweeps. There is no question that these methods work. Yet it is interesting that they do, for we have known for fifty years that selection on a single site cannot generate LD [2].

To explore this puzzle, consider a pair of loci, one neutral and the other selected. At the selected locus, allele A is favored over its alternative a. At a linked locus, alleles B and b are neutral. Because A is favored, selection will tend to increase the frequencies of gametes that carry it (A-gametes) and to decrease those of a-gametes. Graphically, it will increase the size of the circle on the left side of Figure 1 and to decrease that of the other circle. Within the class of A-gametes, however, selection has no effect on the frequency of the neutral allele B. In the figure, the shaded area of each circle represents the fraction of gametes that carry B. If selection were the only force involved, the left circle would grow and the right one would shrink, but the shaded fraction of each circle would remain constant. In the real world, these shaded fractions would change because of genetic drift and recombination. But they are not affected by selection on locus A.

Let us invent some notation to describe the parts of this system that selection *does not* change. The four gamete types have relative frequencies

Gamete type	AB	Ab	aB	ab
Frequency	x_1	x_2	x_3	x_4

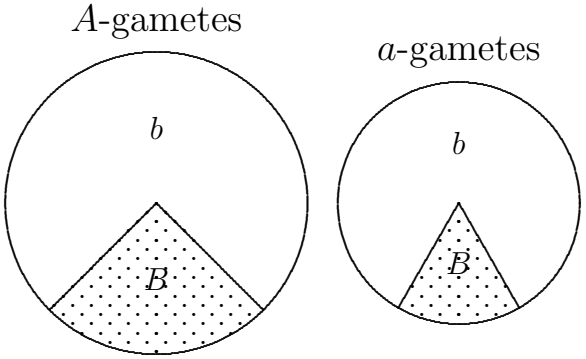

Figure 1: The circles represent the numbers of A-bearing and a-bearing gametes within a population. The shading indicates the relative frequency of B within each of these categories. Because the two shaded fractions are unequal, the diagram illustrates *linkage disequilibrium.*

Alleles A and B have frequencies $p_A = x_1 + x_2$ and $p_B = x_1 + x_3$. Conditional on the allele at locus A, the frequency of allele B is $p_{B|A} = x_1/(x_1 + x_2)$ among A-gametes, but is $p_{B|a} = x_3/(x_3 + x_4)$ among a-gametes. These conditional allele frequencies corresponded to the shaded areas in the two circles in Fig. 1. Since these are not affected by selection, neither is their difference, $d = p_{B|A} - p_{B|a}$. This statistic was introduced by Nei and Li [4] and has been studied by Devlin and Risch [1]. Graphically, d is the difference between in size between the shaded fractions of the two circles in Fig. 1. If these fractions are equal, then the system is at linkage equilibrium and $d = 0$. If the shaded fractions are unequal, we have LD and $d \neq 0$.

Let us re-express d in terms of another mea-

sure of LD:

$$\begin{aligned} d &= \frac{x_1}{x_1+x_2} - \frac{x_3}{x_3+x_4} \\ &= \frac{x_1x_4 - x_2x_3}{(x_1+x_2)(x_3+x_4)} \\ &= \frac{D}{p_A(1-p_A)} \end{aligned} \quad (1)$$

where $D = x_1x_4 - x_2x_3$ is a conventional measure of LD [3]. Rearranging,

$$D = dH_A/2 \quad (2)$$

where $H_A = 2p_A(1-p_A)$ is the heterozygosity at locus A.

Equation 2 shows that D is a product of two factors, of which one (d) is unaffected by selection at a single locus and the other is simply the heterozygosity at locus A. Selection on locus A affects linkage disequilibrium only in the rather uninteresting sense that it changes the heterozygosity of the locus under selection. Why then does D help us detect selective sweeps?

The answer has more to do with initial conditions than with selection. When allele A first arises by mutation, it will exist on a single chromosome, and that chromosome will carry either a single copy of B or a single copy of b. At this early stage, $p_{B|A}$ is either 1 or 0. Meanwhile, the frequency of B among a-gametes equals its frequency in the population as a whole. Thus, d is equal either to $1 - p_B$ or to $-p_B$. Either way, there is every chance that this initial d will be far from 0. Over time, d decays towards 0 under the influence of recombination. But strong selection may outrun recombination, so that d stays far from zero throughout the selective sweep.

This initial LD is not detectable using D, because D is proportional to heterozygosity, which is initially near zero. D rises to an appreciable value only after the sweep is well established, and the selected allele is at an intermediate frequency.

This process is illustrated in Fig. 2. d has a substantial and relatively constant value throughout the selective sweep. By contrast, D starts near zero, rises to a peak when $p_A \approx 1/2$, and then declines again to 0. Although D can be large when p_A is near $1/2$, it is not necessarily large. This is illustrated in Fig. 3, which tracks the history of a neutral allele that happened to drift to fixation. For this lucky neutral allele, evolution is much slower. It does not reach a frequency near $1/2$ until about 30,000 generations have elapsed. By that time there is little LD left, so D and d remain near 0.

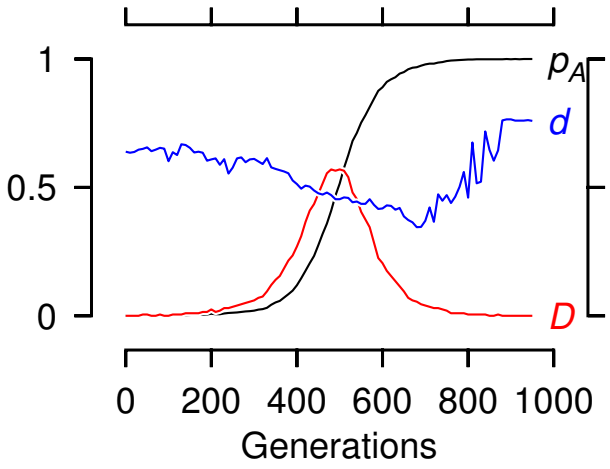

Figure 2: Selective sweep of allele A, which has a selective advantage ($s = 0.02$). Recombination rate is $r = 0.001$, and the haploid population size is $2N = 50,000$.

In summary, LD originates from mutation and then decays gradually under the influence of recombination. This gradual decay is obscured in the time path of D, because of the effect of heterozygosity. It is more obvious in the time path of d, which is insensitive to heterozygosity. Advantageous alleles are young alleles, and young alleles have not experienced much recombination. They retain their initial large values of d and are thus surrounded by blocks of LD. This is why LD tells us about selection.

References

[1] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale

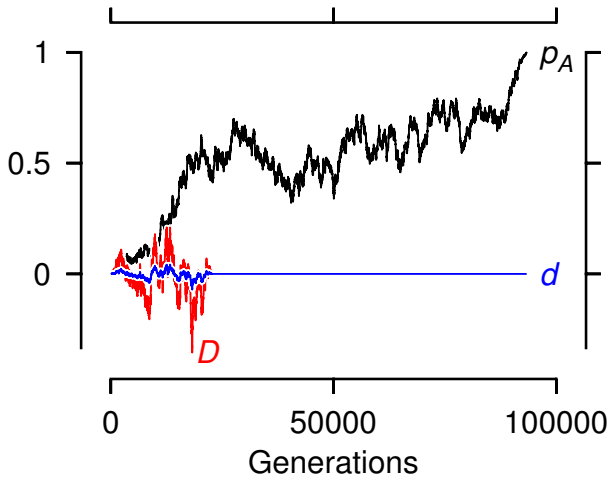

Figure 3: A neutral allele drifting to fixation. Parameters as in Figure 2, except that $s = 0$.

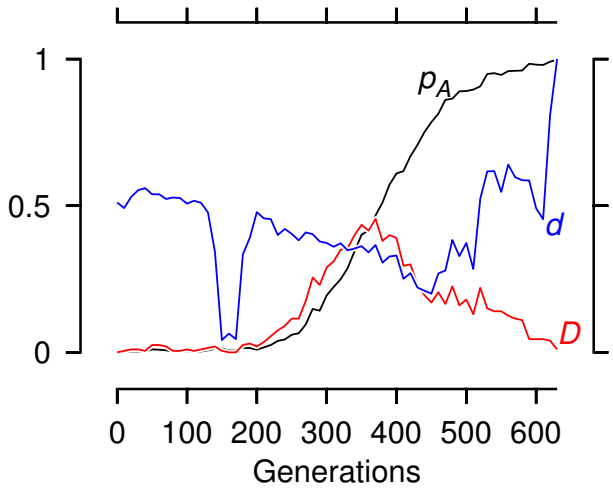

Figure 4: Selective sweep in a small population. Parameters as in Fig. 2, except that the haploid population size is $2N = 5000$.

mapping. *Genomics*, 29(2):311–322, 1995.

[2] Joseph Felsenstein. The effect of linkage on directional selection. *Genetics*, 52:349–363, 1965.

[3] R. C. Lewontin and Ken-ichi Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960.

[4] Masatoshi Nei and Wen-Hsiung Li. Non-random association between electromorphs and inversion chromosomes in finite populations. *Genetical Research, Cambridge*, 35(1): 65–83, 1980.