Why Linkage Disequilibrium Helps Us Find Selective Sweeps

Alan R. Rogers

October 17, 2014

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 forty years that selection on a single site cannot generate LD. This result is due to Felsenstein [2], who studied a measure of LD that is seldom used with data. It seems worth exploring the behavior of other measures.

I will suppose that, at one locus, selection favors allele A over its alternative a. At a linked locus, alleles B and b are neutral. In this system, there are four types of gamete, and I will use the following notation for their relative frequencies:

Gamete type	AB	Ab	aB	ab
Frequency	x_1	x_2	x_3	x_4

Alleles A and B have frequencies $p_A = x_1 + x_2$ and $p_B = x_1 + x_3$.

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 allele B. This follows from the fact that B and b are neutral. 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 change because of genetic drift and recombination. But they are not affected by selection, provided that selection acts only on locus A.

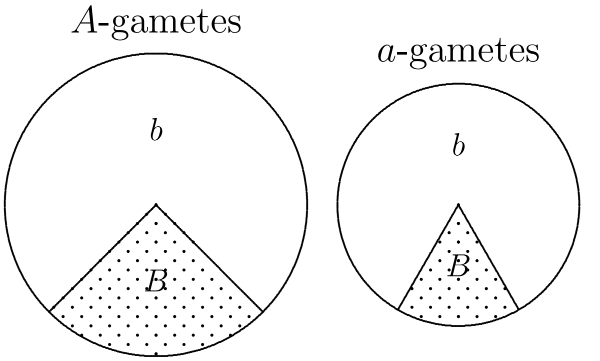

Figure 1: The circles represent the numbers of A-bearing and a-bearing gametes within a population. The shading indicates the relative frequency of B within each of these categories. Because the two shaded fractions are unequal, the diagram illustrates *linkage disequilibrium.*

Let us invent some notation to describe the parts of this system that selection *do not* change. First, 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. Since these quantities 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 $d \neq 0$. Thus, d is a measure of LD.

It is useful to write d in terms of gamete frequencies:

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

where $D = x_1 x_4 - x_2 x_3$ is a conventional measure of LD [3]. Rearranging,

$$D = dH_A/2 \tag{2}$$

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 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. If one were really interested in *this* effect, it would make more sense to study H_A directly. 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, it decays towards 0 under the influence of recombination.

This process is illustrated for a selective sweep in Fig. 2. During the sweep, D rises to a peak when $p_A \approx 1/2$ and then declines to 0. This reflects the fact (shown in Eqn. 2) that D is proportional to the heterozygosity of locus A. D cannot be large unless p_A is near 1/2. Fig. 2 shows that D *can* be large when p_A is near 1/2, but this does not mean that it *must* be 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 remains near 0. We get a different view of this process from d. Its value tends to remain high throughout the selective sweep, but is near 0 during most of the history of the lucky neutral allele.

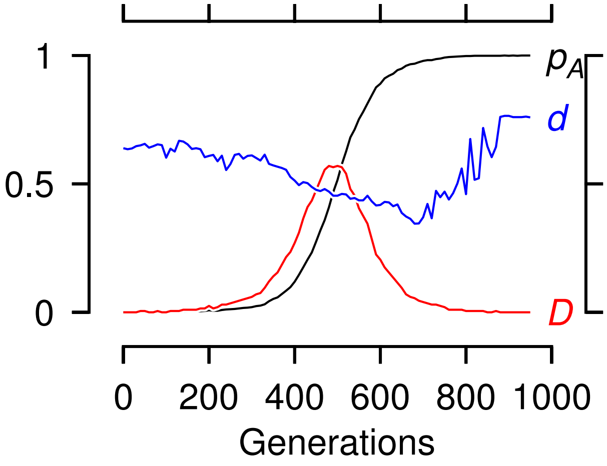

Figure 2: Selective sweep of allele A, which has a selective advantage ($s = 0.02$). Recombination rate is $r = 0.001$, and the haploid population size is $2N = 50,000$.

These examples show how LD can be useful in detecting selection. We find cases of selection by searching for alleles whose frequencies are near 1/2, and which are in strong LD with surrounding loci. Population geneticists are currently exploring different ways to do this. The different approaches mainly involve different measures of LD. To my knowledge, no one has yet used d. I suspect however that it may prove useful because, as Fig. 2 shows, its value tends to remain high throughout a sweep, whereas that of D is only high briefly.

This advantage will be useful mainly in large populations. In small ones, drift within the

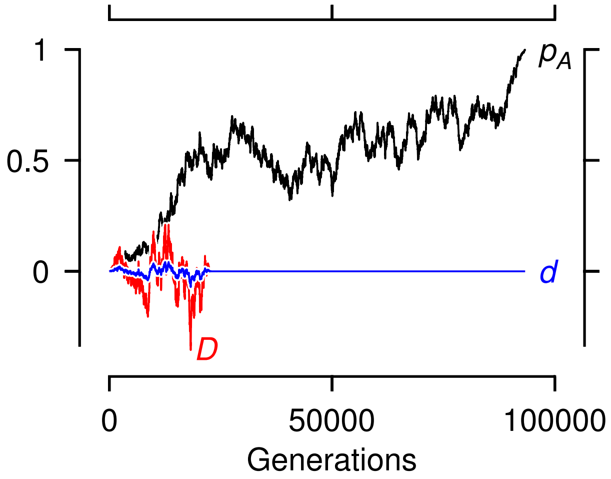

Figure 3: A neutral allele drifting to fixation. Parameters as in Figure 2, except that $s = 0$.

classes of A-gametes and a-gametes causes d to bounce around a lot when p_A is very rare or very common. This effect is shown in Fig. 4 for a population in which $2N = 5000$. In populations this small, d varies wildly except when p_A is near $1/2$.

Regardless of how you measure it, LD decays gradually under the influence of recombination. This gradual decay is more obvious for d than for D, because d is insensitive to the value of p_A. The rate of this decay is the same for a selected allele as for a neutral one. It takes far less time, however, for an allele to reach fixation if it is being urged along by natural selection. Selected alleles are young alleles, and young alleles cannot have experienced much recombination. They retain their initial large values of d and are thus surrounded by blocks of LD.

References

[1] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322, 1995.

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

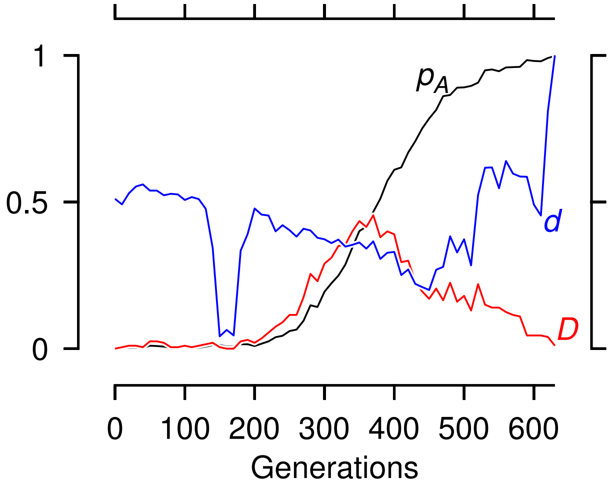

Figure 4: Selective sweep in a small population. Parameters as in Fig. 2, except that the haploid population size is $2N = 5000$.

[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. Nonrandom association between electromorphs and inversion chromosomes in finite populations. *Genetical Research, Cambridge*, 35(1): 65–83, 1980.