Linkage Disequilibrium

We have seen that recombination can shuffle combinations of allelic variants among loci and thus generate novel haplotypes. As a simple way to test whether recombination has occurred within a population we used the four-gamete test. While useful, what we really want to obtain is a quantitative measure of how closely genetic variation corresponds across loci. This correspondence or nonrandom association of alleles at two or more loci is defined as **linkage disequilibrium (LD)**. We will see that linkage disequilibrium depends on multiple factors like local recombination rate, non-random mating, mutation rate, genetic drift and population structure. Contrary to normal linkage due to a physical connection of neighboring loci on the same chromosome, LD can even occur between loci on different chromosomes.

To introduce the concept of LD we will start with the simplest case: two bi-allelic loci with alleles *A* and *a* at one locus and alleles *B* and *b* at the second locus. The four possible combinations of alleles (**haplotypes)** are *AB*, *Ab*, *aB* and *ab* segregating at observed frequencies p_{AB}, p_{Ab}, p_{aB} and p_{ab} in the population. By adding the appropriate haplotype frequencies we can obtain allele frequencies at each locus, e.g.

$$p_A = p_{AB} + p_{Ab}$$
$$p_B = p_{AB} + p_{aB}$$

Consider the following example sets:

In population a) it is directly apparent that if there is an *A* on the chromosome it will be associated with the *b* allele of the second locus. On the other hand in population c) it will be associated with allele *B*. In population b) there does not seem to be any obvious association.

To obtain a measure of how likely two alleles are associated among loci we introduce the **coefficient of linkage disequilibrium D**. D simply quantifies the deviation of the observed frequency (p_{AB}) of a haplotype and the expectation under random allelic association (given by the product of the allele frequencies p_A and p_B).

$$D_{AB} = p_{AB} - p_A p_B$$

Example a) shows the case where both capital alleles never occur together in one individual (p_{AB}=0). Both alleles behave as if something keeps them apart.

$$a)\ D_{AB} = p_{AB} - p_A p_B = 0 - (0.5\ x\ 0.5) = -0.25; \qquad \boldsymbol{D \neq 0}$$

In the second case b) one individual carries both capital alleles p_{AB}=0.25 and one neither A nor B. Now the haplotype frequency fits the random association of both alleles.

$$b)\ D_{AB} = p_{AB} - p_A p_B = 0.25 - (0.5\ x\ 0.5) = 0; \qquad \boldsymbol{D = 0}$$

The third case c) indicates strongly coupled loci. The presence of allele A implicates the presence of allele B. The haplotype frequency is p_{AB}=0.5.

$$c)\ D_{AB} = p_{AB} - p_A p_B = 0.5 - (0.5\ x\ 0.5) = 0.25; \qquad \boldsymbol{D \neq 0}$$

The phrase "two alleles are in **linkage disequilibrium**" usually means D≠0, which is the departure from the expected haplotype frequency, based on allele frequencies. D=0 denotes the case in which two alleles segregate independently, they are said to be in **linkage equilibrium**. This terminology is somewhat unfortunate, as it does not refer to any population dynamics. It contains no information on whether a population is in equilibrium relative to an evolutionary process (compare e.g. mutation – drift equilibrium).

We can define a value of D for each possible combination of alleles:

$$\begin{aligned} D_{AB} &= p_{AB} - p_A p_B \\ D_{Ab} &= p_{Ab} - p_A p_b \\ D_{aB} &= p_{aB} - p_a p_B \\ D_{ab} &= p_{ab} - p_a p_b \end{aligned}$$

When there are only two alleles at each locus it turns out that

$$D_{AB} = D_{ab},\ D_{Ab} = D_{aB},\ D_{AB} = -D_{Ab}$$

Consequently, there is only one value of |D|. The sign of *D* is arbitrary and only depends on the order of haplotype pairs one starts with. The *AB* and *ab* haplotypes are often called **coupling gametes** because the same encoding is used for both (here capital/small, often also by subscript A_1B_1 vs. A_2B_2). The haplotypes with differently encoded allelic combinations *Ab* and *aB* are called **repulsion gametes**. Linkage disequilibrium can be then thought of as a measure of the excess of coupling over repulsion gametes. If D > 0 there are more coupling gametes than expected by random assortment; if D < 0 there are more repulsion gametes than expected. The above can then be rewritten as

$$\begin{aligned} p_{AB} &= p_A p_B + D \\ p_{Ab} &= p_A p_b - D \\ p_{aB} &= p_a p_B - D \\ p_{ab} &= p_a p_b + D \end{aligned}$$

Returning to the above examples let us consider what the magnitude of D tells us about the level of linkage disequilibrium? For cases a) and b) linkage disequilibrium appears to be maximal (only 2 of the four possible combinations exist). Hence, does 0.25 represent the maximum possible value for D? Does e.g. D=0.006 reflect a substantially lower level of LD?

The answer is: it depends. Analogous to what we have seen for the fixation indices D is a relative measure depending on allele frequencies. We can find the range of possible values for D in relation to allele frequencies using the above equations. Knowing that haplotype frequencies cannot be negative $p_Ap_b - D > 0$ and $p_ap_B - D > 0$, and consequently $D \leq \min(p_Ap_b, p_ap_B)$. If $D < 0$, then for the same reason $-D \leq \min(p_Ap_B, p_ap_b)$. Consider again the example sets from above. $D \leq \min(p_Ap_b, p_ap_B) = \min(0.5x0.5, 0.5x0.5) = 0.25$. Similarly, $-D \leq \min(p_Ap_B, p_ap_b) = 0.25$. Hence, the range of possible values is $-0.25 \leq D \leq 0.25$ (from exclusive representation of repulsion gametes to a population consisting only of coupling gametes). For $p_A=0.01$ and $p_B=0.4$ D would lie within the range $-0.04 \leq D \leq 0.006$. The largest possible level of LD for this combination of allele frequencies would thus be at D = 0.006. To describe the extent of LD relative to the range of possible values we define D' as

$$D' = \frac{D}{\min(p_Ap_b, p_ap_B)} \; if \; D > 0$$
$$= \frac{D}{\min(p_Ap_B, p_ap_b)} \; if \; D < 0$$

D' is guaranteed to range from 0 for no LD to 1 for maximum levels of LD. A useful property of D' is that we can directly see if one haplotype is missing. At D' = 1 D must be at its maximum for one combination and consequently one of the haplotype frequencies must be 0. It turns out that immediately after a mutation creates a novel allele, D'=1 between that locus and any other polymorphic locus on the same chromosome.

To obtain a definition of LD with well-known statistical properties that is also independent of the sign (see above), we can use the so-called standardized LD r^2:

$$r^2 = \frac{D^2}{p_Ap_ap_Bp_b} = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$$

r is Pearson's correlation coefficient that ranges between 1 (strong positive correlation) and -1 (strong negative correlation). If r^2 is 0 the two segregating variants are not correlated. In addition to intuitive appeal, r^2has another important property that allows for statistical testing of independence between of alleles of locus A and B for a given set of *n* chromosomes in a sample. It can be shown that r^2 is directly related to the Chi-square distribution as

$$nr^2 = \chi^2$$

with 1 degree of freedom.

N	μ	s	m	r
∞	-	-	-	✓

Change ef D over time - the effect of recombination.
Next we consider the effect of recombination on LD in infinitely large populations when there is random mating. Assuming recombination at rate *c* between the two loci we find that

$$D_{t+1} = (1-c)D_t$$

The change in D in a single generation is

$$\Delta D = -cD_t$$

Finally,

$$D_t = (1-c)^t D_0$$

Remember that *c* is the fraction of recombinant gametes (Ab|aB) produced in heterozygotes (AB|ab) and ranges from no recombination (*c*=0) to a maximum at free recombination conforming to Mendel's second law of independent assortment (*c*=0.5). This shows that the ultimate state of the population is at D=0, but LD does not disappear within one generation. That is very different to what happens to genotype frequencies in large populations that immediately settle at frequencies predicted by the HWE after one round of random mating – regardless of how closely linked they are. LD takes substantially longer to be established than HWE.

Using the approximation $e^x \approx 1 + x$ for |x| ≈ 0, i.e. for small *c* we obtain

$$D_t \approx e^{-ct} D_0$$

This nicely illustrates that linkage disequilibrium decays each generation at a rate determined by the degree of recombination. Setting t=1/c we see that it takes about 1/c generations to reduce D to about 37% of its initial value. So even if two loci recombine freely (c=0.5) it takes 2 generations to reduce LD to one third of its initial value, if c=0.01 it takes 100 generations. Remember that the human genetic map was ~30 M long. Assuming 21,000 equally spaced proteins any pair of proteins have an average recombination rate of 30 M / 21,000 = 0.00143 M = c. Hence, LD after 1000 generations is reduced to about ¼ of the initial level. In humans, 1000 generations correspond to about 25,000 years, such that linkage between adjacent genes that existed in our forefathers hunting during the last glacial maximum can still be seen today.

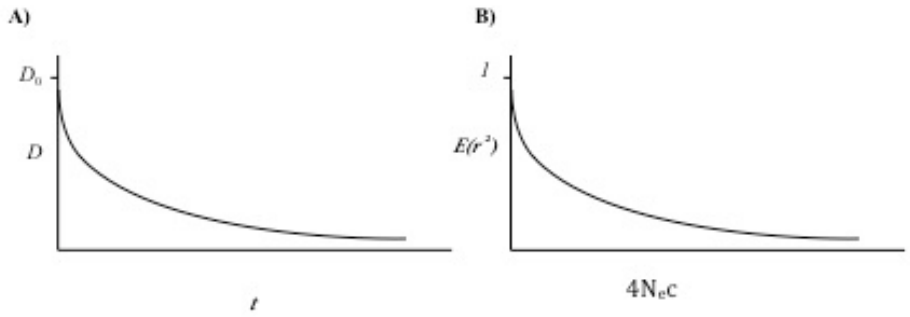

Figure 1: Schematic representation of the decay in LD by recombination as a function of time (A) and of effective population size (B).

Genealogical interpretation of LD

The above result has interesting genealogical implications. We have seen that in populations of finite size and variation in reproductive success a set of homologous genes of remnant individuals traces back to a single common ancestor in the past. Without recombination, in a strictly asexual population, every gene shares the same ancestor. In sexually reproducing

populations, however, **gene genealogies** will differ, as they are being shuffled by recombination. This can be well illustrated with parts of the human genome differing in their mode of inheritance. The mitochondrial genome and the Y-chromosome are exclusively inherited through the mother and father, respectively. All genes located within them thus constitute a single locus (which is an important realization if you construct mitochondrial phylogenies for example). Accordingly, we expect all current copies to coalesce to a single most recent common ancestor (MRCA) – the **mitochondrial Eve** and **Y-chromosomal Adam** (a somewhat awkward, but illustrative analogy to the Biblical reference). mt-MRCA and Y-MRCA did not emerge from the same individual, nor do they coalesce at the same time. Nor is there anything special about them, except that they contributed their mt-DNA and Y-chromosomal copy to all remnant descendants. The situation is very different if we look at any autosomal chromosome. Looking back in time, we see that the genome is divided into many blocks that trace back to a large number of different ancestors – different parts of the genome have different ancestries.

For two linked loci, there are two gene genealogies. The relationship between those gene genealogies depends on the history of recombination between the loci. As time goes on, recombination will chop up associations of alleles, which leads to a decrease of LD as a function of recombination rate. This decrease of LD will happen faster if loci are further apart as the probability of recombination c increases with increasing physical distances (**Figure 1a**). Thus there is a close relationship between the gene genealogies of two loci and their coefficient of LD: closely associated loci will share their evolutionary history for longer (effectively acting as one locus such as mtDNA) than more distant loci (which if far enough apart have entirely independent histories).

Factors affecting LD

LD is influenced by various factors that we shortly touch upon below.

N	μ	s	m	r
finite	-	-	-	✓

Effect of genetic drift and recombination

Assuming infinite population size above, we have seen that the ultimate state of the population is at D = 0. Multi-locus models examining the formation and decay of disequilibrium by drift have shown that the expected value of disequilibrium among replicate populations is zero ($E(D)=0$) - analogous to the result for a single locus that on average allele frequencies do not change due to genetic drift. However in any one population the magnitude of LD may be substantial ($Var(D)>0$). That means, in a finite-size population we observe LD because its variance is larger than zero. Again, the smaller the effective population size the larger is the variance. An extreme reduction in population size, like a very strong bottleneck, can therefore increase LD due to the loss of haplotypes. The associated increase of genetic drift will increase LD. Especially long-distance LD is a sign of population reduction which can be best detected by comparing two populations: one ancestral and one derived. Population expansion, on the other hand, is characterized by a fast decay of LD. New haplotypes are introduced into the population and thin out haplotype combinations.

As we have seen in **Figure 1a**, recombination also affects LD. Thus, if we combine recombination and genetic drift in a finite population, we find the expected equilibrium LD (**Figure 1b**):

$$E(r^2) = \frac{1}{1 + 4N_e c}$$

This shows if $4N_ec$ is small $E(r^2)$ approaches 1, as $4N_ec$ increases $E(r^2)$ the expected value of LD approaches 0. If $4N_ec$ is large the equation can be approximated by

$$E(r^2) \approx \frac{1}{4N_e c}$$

Analogous to the **population mutation rate $\theta = 4N_e\mu$** the quantity $4N_ec$ is called the **population recombination rate $\rho=4N_ec$**. LD thus contains valuable information on the demographic history of populations. If one knows the rate of recombination for a region of the genome it is thus in principle possible to estimate N_e via the amount of LD.

N	μ	s	m	r
∞	-	-	✓	-

Migration and admixture

The mixing of individuals from different subpopulations that have different allele frequencies creates LD. If two subpopulations both fixed unequal allele combinations, such as AB and ab, respectively, any mixture of individuals would create LD, although each subpopulation on its own has none. This can be directly seen with a simple example. Assume two populations each with segregating alleles *A* and *a* at one locus and *B* and *b* at the second locus – within each population the loci are in linkage disequilibrium. Therefore, the frequency of the AB haplotype is equal to the expectation $p_{1A}p_{1B}$ and $p_{2A}p_{2B}$ in both populations. If we now ignore population structure assuming equal population sizes, the average frequency of AB in a mixture of both populations would be:

$$p_{AB} = \frac{1}{2}p_{1A}p_{1B} + \frac{1}{2}p_{2A}p_{2B}$$

The frequency of the alleles are in the mixture are:

$$p_A = \frac{1}{2}p_{1A} + \frac{1}{2}p_{2A};\ p_B = \frac{1}{2}p_{1B} + \frac{1}{2}p_{2B};$$

The coefficient of LD in the mixture is:

$$D = p_{AB} - p_A p_B = \frac{1}{4}\,(p_{1A} - p_{2A})(p_{1B} - p_{2B})$$

Whenever allele frequencies differ for both loci (e.g. by the effect tof genetic drift) D will be non-zero. Population subdivision therefore generates LD even if each population is in linkage equilibrium. This is called the **two-locus Wahlund effect**. As a consequence, migration between divergent populations increases LD in each population. For single major migration bouts, the timing of migration can in principle be estimated by the decay of LD generated during the migration event.

N	μ	s	m	r
∞	-	✓	-	✓

Selection

A novel, advantageous allele under strong positive Darwinian selection will quickly rise in frequency, carrying along neighboring neutral alleles (**selective sweep**). This results in a dramatic loss of local heterozygosity and an unusually strong LD of linked neutral sites with the beneficial site (**Figure 2a**). The LD across the selected site vanishes after the fixation of the beneficial allele, while the LD on either side of the selected locus remains. If recombination occurs during the sweep, chromosomal segments that otherwise would have been lost from the population may recombine with the haplotype carrying the advantageous mutation thus rescuing the present genetic variation (**Figure 2b**). The higher the recombination rate between a locus and the centre of the selective sweep the less its variation will be affected. LD is therefore expected to decline as a function of recombination distance from the locus under selection – a pattern that can be used to screen for sites under positive selection in the genome.

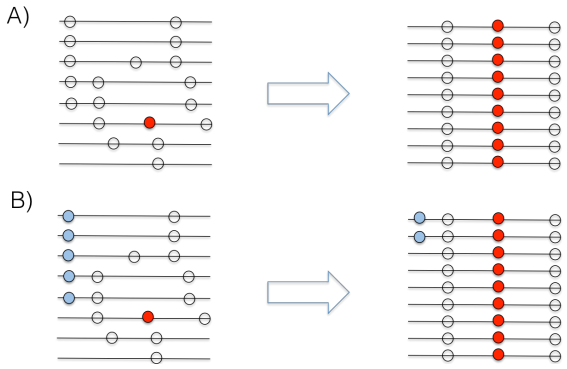

Figure 2: The effect of a selective sweep after the occurence of a beneficial novel mutation (red) on segregating genetic variation A) without recombination B) with recombination where the blue variant represent sites that have recombined onto the chromosome carrying the advantageous allele.

Association mapping: an application making use of LD

Nonrandom associations are not restricted to appear between two genetic loci. They can also be found between phenotypes and a genetic marker. Human disease phenotypes have been the main focus. Genome-wide association studies (GWAS) trying to find an association between the disease and genetic variants that may cause it are generally conducted by contrasting a **case** and **control** group of individuals that should only differ by the phenotype of interest (matched by age, sex, etc.). Then a statistical test is performed for each marker determining whether each marker is significantly associated with the phenotype. For a limited set of genetic markers, we cannot expect that the actual causal variant is included in the marker set. Therefore, GWAS heavily relies on LD between genetic markers and the causative locus – the stronger the level of LD, the more likely it is to find an association. Today, we can in principle sequence every base pair of a genome for thousands of individuals. The causative loci should therefore be directly included without having to rely on LD. However, it is still technically difficult to characterize structural genetic variation (duplications, insertion-

deletion polymorphism, inversions, translocations, etc.) which is consequently hidden from sight. Adjecent SNP in strong LD, however, may still reveal the association.

GWAS has helped identify several diseases with a simple genetic architecture. However, it seems that many, often rare, alleles contribute to a most (disease) quantitative phenotypes. If traits that are shaped by rare alleles with small effect sizes GWAS is limited in power. Studies using hundreds and thousands of makers and thousands of individuals have explained astonishingly little of trait variation considering the effort. Large emerging databases harboring phenotypic and whole-genome sequencing information on millions of humans will empower GWAS studies in humans.

Literature: (Barton et al. 2007; Futuyma 2013; Nielsen and Slatkin 2013)

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH. 2007. Evolution. 1st edition. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press

Futuyma DJ. 2013. Evolution. 3rd ed. Sinauer Associates

Nielsen R, Slatkin M. 2013. An Introduction to Population Genetics: Theory and Applications. Sunderland, Mass: Macmillan Education