

February 22, 2005

Bio 107/207

Winter 2005

Lecture 15

Linkage disequilibrium and recombination

- in our treatment of population genetics up to this point we have assumed that the transmission of alleles at a given locus across generations are independent of alleles at a second locus.
- we have also assumed that the fitnesses of genotypes at one locus are not affected by genotypes at another locus.
- both of these assumptions are likely to be violated for a reasonable number loci.
- when genetic variation at two or more loci is considered simultaneously, allele frequencies are not sufficient to describe their dynamics in natural populations.
- it becomes necessary to deal with the extent of non-random associations of alleles at different loci.
- the non-random associations of alleles at different loci is called **gametic phase disequilibrium** or, more simply, **linkage disequilibrium**.

What is linkage disequilibrium?

Linkage equilibrium occurs when the genotype present at one locus is independent of the genotype at a second locus.

Linkage disequilibrium occurs when genotypes at the two loci are not independent of another.

- the term linkage disequilibrium is misleading for two reasons.
- first, non-random associations of alleles at two loci can occur even if the two genes are unlinked.
- second, just because two loci are linked this does not mean that they will be in linkage disequilibrium.

Measuring linkage disequilibrium

- consider two loci (A and B), each segregating for two alleles (A_1 , A_2 , B_1 , and B_2)
- there are four possible gametes (or haplotypes) present in the population:

Gamete	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

- allele frequencies can be expressed in terms of the gamete frequencies:

Allele	Frequency
A ₁	p ₁ = x ₁₁ + x ₁₂
A ₂	p ₂ = x ₂₁ + x ₂₂
B ₁	q ₁ = x ₁₁ + x ₂₁
B ₂	q ₂ = x ₁₂ + x ₂₂

- note that $p_1 + p_2 = 1$ and $q_1 + q_2 = 1$, and $\sum x_{ij} = 1.0$

- if alleles at the two loci are randomly associated with one another, then the frequencies of the four gametes are equal to the product of the frequencies of alleles it contains:

$$\begin{aligned}x_{11} &= p_1q_1 \\x_{12} &= p_1q_2 \\x_{21} &= p_2q_1 \\x_{22} &= p_2q_2\end{aligned}$$

- in this situation, there is no linkage disequilibrium and gamete frequencies can be accurately followed using allele frequencies.

- if alleles at the two loci are not randomly associated, then there will a deviation (D) in the expected frequencies:

$$\begin{aligned}x_{11} &= p_1q_1 + D \\x_{12} &= p_1q_2 - D \\x_{21} &= p_2q_1 - D \\x_{22} &= p_2q_2 + D\end{aligned}$$

- this parameter **D is the coefficient of linkage disequilibrium** first proposed by Lewontin and Kojima (1960).

- the most common expression of D is:

$$D = x_{11}x_{22} - x_{12}x_{21}$$

- where x_{11} and x_{22} are referred to as “coupling” gametes and x_{12} and x_{21} the “repulsion” gametes.

- D is thus the difference between these two gamete types.

- table 10.3 in the textbook shows how linkage disequilibrium will change through time.

- for our 2 locus, 2 allele case, there are 10 genotypes (not 9!) because there are 2 types of di-locus heterozygotes: A₁B₁/A₂B₂ and A₁B₂/A₂B₁

- these heterozygotes are important because they are the only genotypes where recombination can create different gametes than found in the parents.

- let **c be the rate of recombination** between the A and B loci.

- c ranges in value between 0 and 0.5.

- the maximum is at 0.5 because with independent assortment of the two loci, one half of the gametes produced will still be the parental type.
- recombination in the two types of double heterozygotes will produce different new genotypes at rate c .
- double heterozygotes will produce offspring with gametes like themselves at rate $(1-c)$.
- table 10.3 shows that the frequencies of the four gametes after one generation are:

$$\begin{aligned}x'_{11} &= x_{11} - cD_0 \\x'_{12} &= x_{12} + cD_0 \\x'_{21} &= x_{21} + cD_0 \\x'_{22} &= x_{22} - cD_0\end{aligned}$$

where D_0 is the extent of linkage disequilibrium present in the preceding generation.

- the magnitude of linkage disequilibrium after one generation is thus:

$$\begin{aligned}D_1 &= x'_{11}x'_{22} - x'_{12}x'_{21} \\&= (x_{11} - cD_0)(x_{22} - cD_0) - (x_{12} + cD_0)(x_{21} + cD_0) \\&= (1 - c)D_0\end{aligned}$$

- this leads to the general relationship:

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

which can be approximated as:

$$D^t = e^{-ct} D_0$$

- therefore, linkage disequilibrium decays each generation at a rate determined by the degree of recombination.

- the coefficient of linkage disequilibrium, D , has two unpleasant properties.
- first, it varies in magnitude between a minimum of -0.25 and a maximum of $+0.25$.
- these maximum values of D occur when there are only repulsion gametes ($x_{11} = x_{22} = 0.50$) or only repulsion gametes ($x_{12} = x_{21} = 0.50$).
- if there is free recombination between the two loci in complete linkage disequilibrium, then it will only take about 7 generations for the disequilibrium to be eliminated.
- if the two loci are tightly linked (say $c = 0.001$) then the decay of linkage disequilibrium will take a substantial period of time.
- it is possible to determine the number of generations for D_0 to decay to certain level D_t from the following equation:

$$t = [\ln (D_t/D_0)]/[\ln (1 - c)]$$

- the other problem with D is that its maximum value changes as a function of allele frequencies at the two loci.
- Lewontin (1964) thus proposed standardizing D to the maximum possible value it can take:

$$D' = D/D_{\max}$$

where D_{\max} is the maximum value of D at the given allele frequencies.

- D_{\max} is equal to the lesser of p_1q_2 or p_2q_1 if D is positive or the lesser of p_1q_1 or p_2q_2 if D is negative.
- D' varies between 0 and 1 and allows to assess the extent of linkage disequilibrium relative to the maximum possible value it can take.

- Hill and Robertson (1968) proposed the following measure of linkage disequilibrium:

$$r^2 = D^2/[p_1p_2q_1q_2]$$

- if allele frequencies are equal, then r^2 varies between 0 and 1.
- as for D, the maximum value of r^2 depends on the allele frequencies and one can determine a r' value in a manner analogous to a D' .

Cytonuclear disequilibrium

- a related form of disequilibrium can arise between nuclear and mitochondrial genes that is called cytonuclear disequilibrium.
- consider the simplest case of two mitochondrial haplotypes (M_1 and M_2) and two nuclear alleles (A_1 and A_2).
- let m_1 and m_2 equal the frequencies of M_1 and M_2 respectively, and p_1 and p_2 be the frequencies of A_1 and A_2 , respectively.
- if x_{11} now equals the frequency of the M_1A_1 gamete, then

$$D_c = x_{11} - m_1p_1$$

- similar to the case for two nuclear genes, D_c decays quickly.
- measures of cytonuclear disequilibrium are an extremely powerful tool to study hybrid zones.
- it is sometime possible to determine asymmetries between interspecies crosses (as discussed for the *Hyla* hybrid zones described in the textbook) and the extent of introgression.

The rate of recombination, c

- the decay of linkage disequilibrium depends primarily on the rate of recombination.
- Haldane (1919) proposed the following mapping function if one assumes that crossover events are Poisson distributed along chromosomes:

$$c = [1 - e^{-2m}]/2$$

where m is the expected number of crossover events.

- the rate of recombination is highly variable in different chromosomal regions within species.
- it also has been found to vary among individuals of a species.
- if some of this variation is genetically controlled, then we would expect that recombination is a process that can evolve over time.
- there are many studies known from *Drosophila* in which recombination rates have changed over the course of selection experiment.
- in *Drosophila* selection experiments, the usual outcome is to observed recombination rates to increase as a consequence of selecting for other characters.
- when one estimates recombination rates between the sexes it is extremely common to find that it occurs at higher rates in females than males.
- in some insects (*Drosophila* being the first identified) there is no recombination in males.
- for human autosomal genes, the rate of recombination is about 60% higher in females.
- why would this be so?

Factors creating linkage disequilibrium

- there are five processes that can produce linkage disequilibrium in a population: epistatic natural selection, mutation, random drift, genetic hitchhiking, and gene flow.
- although natural selection is the most important some of the others (notably gene flow) can create substantial levels of disequilibrium.

Mutation

- similar to its weak effects on allele frequency change, the process of mutation does not lead to any substantial disequilibrium.
- recurrent mutation will certainly produce nonrandom associations between alleles at different loci.
- however, recombination typically occurs at higher levels than mutation and it will break apart any non-random associations of alleles.
- furthermore, recurrent mutations are not expected to be associated with the same alleles at other loci.
- however, in non-recombining regions of the genome (such as part of the human Y chromosome) mutation can form linkage disequilibrium between loci that will not decay and can increase in frequency by drift or selection.

Random drift

- non-random associations between alleles at different loci may be produced by random drift.
- 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.
- this is the same result as the expected allele frequency change for alleles at a single locus.
- however, in any one population, the magnitude of disequilibrium may indeed be substantial.

- in a finite population with effective population size N_e and a rate of recombination, r , between loci the expected value of r^2 is

$$E(r^2) \approx 1/[1 + 4N_e c]$$

- this equation shows that if $4N_e c$ is small, the expected value of r^2 approaches 1.0.
- as $4N_e c$ increases in magnitude, the expected value of disequilibrium approaches 0.
- if $N_e c$ is large, the expectation becomes:

$$E(r^2) \approx 1/4N_e c$$

- here, $4N_e c$ is called the **population recombination rate** (similar to $4N_e \mu$ being the population mutation rate).
- this relationship has attracted the attention of population geneticists because if one knows the rate of recombination for a region in the genome then it is possible to estimate N_e .

Gene flow

- gene flow can produce significant levels of disequilibrium in a population.
- however, this will occur only when the frequencies of alleles at both loci differ between the populations.
- the greater the allele frequency differences, the greater the disequilibrium produced by gene flow.
- if there is linkage disequilibrium between loci in the source populations then this will contribute to even further to the non-random associations of alleles.
- population subdivision has the effect of reducing the rate of decay of linkage disequilibrium.
- if gene flow is small, then the rate of decay is determined by the magnitude of m , the proportion of migrants.
- if gene flow is more extensive, then the decay is determined by the recombination rate.

Inbreeding slows the decay of disequilibrium

- intuition tells us that inbreeding should slow the decay of linkage disequilibrium because it reduces the frequencies of double heterozygotes.
- it is through recombination in these double heterozygotes that disequilibrium is altered.
- the effect of inbreeding on the decay of disequilibrium has been studied for partially selfing species.
- if selfing levels are high, the decay of linkage disequilibrium is slowed substantially simply because inbreeding reduces the conduit (i.e., double heterozygotes) by which disequilibrium is eliminated.

The evolution of supergenes

- natural selection can create linkage disequilibrium between genes by favoring specific combinations of alleles at different loci.

- if particular combinations of alleles function better as a group (than being randomly assembled) then natural selection can act to “push” the different loci physically together into what is called a “**supergene**”.
- the tight linkage acts to reduce recombination between individual genes in the supergene.
- recombination in the region of surrounding the supergene can also be reduced to further maintain the strong linkage disequilibrium between genes.
- one of the best studied examples of a supergene is occurs the land snail *Cepaea nemoralis*.
- this snail is highly polymorphic for its shell coloration and banding patterns.
- shell color ranges from yellow to pink to brown with various shades of each.
- shell color is controlled by several alleles at a single locus.
- brown is dominant over yellow and pink.
- in turn, pink is dominant over yellow.
- the shells of *Cepaea* are also banded to varying degrees.
- the number of bands can range from 0 (dominant) to 5 (or sometimes 6).
- modifier loci also exist that control the number of bands and how they are expressed.
- the loci controlling shell color, banding, and all the modifier genes that act on these traits are tightly linked as a supergene.
- low levels of recombinants are detected in lab crosses but the genes effectively behave as one single locus.
- why did this supergene evolve?
- many studies conducted in both England and France suggest that spatially varying selection favors certain combinations of alleles.
- selection acting on the shell phenotypes is related to predation and thermal biology.
- other good examples of supergenes include the genes for heterostyly in several plants and those controlling mimicry in some butterflies.
- the same principles apply for the coadaptation of genes locked within chromosomal inversions.

Estimating linkage disequilibrium

- although linkage disequilibrium occurs at the gametic level, is it is possible to infer its level from genotype frequency data.
- suppose we scored individuals for their genotypes at two loci, A and B.
- each locus has two alleles (A_1 and A_2 , and B_1 and B_2)
- let p_1 be the frequency of the A_1 allele and p_2 that of the B_1 allele.

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	N_{11}	N_{12}	N_{13}
A_1A_2	N_{21}	N_{22}	N_{23}
A_2A_2	N_{31}	N_{32}	N_{33}

- an estimate of D can be obtained from the following equation:

$$D_{hat} = [2N_{11} + N_{12} + N_{21} + (N_{22}/2)]/N_{total} - 2p_1p_2$$

- here's some data from two restriction sites that span a 5.7 kb region of the pantophysin locus in *G. morhua*.

- this is data from the Newfoundland population.

	B₁B₁	B₁B₂	B₂B₂	Totals
A₁A₁	36	62	27	125
A₁A₂	8	49	34	91
A₂A₂	2	13	29	29
Totals	46	124	75	245

- first, let us estimate allele frequencies at the two loci:

A locus	B locus
p₁ = 0.69592	p₂ = 0.44082
q₁ = 0.30408	q₂ = 0.55918

- note that the single locus genotype frequencies do not differ much from HW expectations.

	Obs.	Exp.		Obs.	Exp.
A₁A₁	125	118.7	B₁B₁	46	47.6
A₁A₂	91	103.7	B₁B₂	124	120.8
A₂A₂	29	22.7	B₂B₂	75	76.6

- let us now estimate D:

$$D_{hat} = [(2 \times 36) + 62 + 8 + (0.5 \times 49)]/245 - [2(0.69592)(0.44082)]$$

$$= 0.0660$$

- what does this value mean?

- here is the problem discussed above – D values vary with allele frequencies so it is not possible to know whether this is substantial or not.
- we need to estimate D'.

- since D is positive, the maximum value of D is the lesser of q_1p_2 or p_1q_2 .
- since q_1p_2 equals 0.13404 and p_1q_2 equals 0.38914, we choose the former.

- therefore, $D' = 0.0660/0.13404 = 0.492$.

- this tells us that D is about half of its maximum possible value.

Why is linkage disequilibrium and recombination important anyway?

- recombination is intimately associated with sex.
- sex is synonymous with **mixis** – the mixing of genes between individuals.
- the evolution of sex has thus been equated with the evolution of genetic recombination.
- an interesting idea put forward by Bernstein et al. (1985) is that mixis is needed to repair damaged DNA.
- many of the enzymes involved in repairing damaged DNA also function in recombination.
- it is possible that the redundancy of DNA (provided by diploidy) acts to facilitate repair by providing a backup copy?
- is recombination simply a serendipitous byproduct of this repair process?

- a considerable amount of genetic variation is produced each generation by the independent assortment of chromosomes.
- consider the human species with 23 pairs of chromosomes.
- the number of different types of gametes produced just by meiotic segregation and independent assortment would be $2^{23} = 8,388,608$.
- now, if one crossover event occurred, on average, between each pair of chromosomes meiosis would produce 4 types of gametes for each pair (rather than 2).
- now the base of the equation becomes 4 and an individual can produce $4^{23} = 7 \times 10^{13}$ different gametes.
- syngamy with a second individual would produce $(4^{23})^2 = 5 \times 10^{27}$ different zygotes!
- so, now you know why you're so different from your siblings!

- in addition to creating an enormous amount of variation for selection to act upon, recombination also increases the efficacy of natural selection.
- let's consider here two benefits that result from recombination – the ability to purge deleterious mutations and the enhanced ability to fix advantageous mutations.

Purging deleterious mutations

- recombination plays a fundamental role in facilitating the purging of deleterious mutations by purifying selection.
- Muller's ratchet predicts that asexual lineages are prone to the accumulation of deleterious mutations.

- the ratchet operates because asexual lineages experience a continued onslaught of mildly deleterious mutations.
 - as lineages accumulate mutations, the least mutated class can occasionally be lost by drift and when this occurs the ratchet has clicked one step forward.
 - there is only one way for asexual populations to evolve – to an ever greater load of deleterious mutations.
 - sex breaks the ratchet because it enables recombination.
 - in the absence of recombination, a deleterious mutation entering a genome is embedded in that genome.
 - it can't be removed unless it happens to undergo a back mutation to the normal wild type allele.
 - recombination acts to disentangle the deleterious mutation from its genetic background.
 - this allows selection to act on the specific deleterious mutation not the entire genome.
- purifying selection acting on deleterious alleles in regions of low recombination will cause a reduction in the levels of neutral polymorphism through an effect called “**background selection**”.
 - a neutral mutation in linkage disequilibrium with a deleterious allele is not free to drift in the population because of the purifying selection acting on the latter.
 - conversely, a deleterious mutation in a region experiencing high levels of recombination will not act to substantially alter levels of linked neutral polymorphism.
 - this is because the footprint of purifying selection will be much more restricted to the region surrounding the deleterious mutation.
 - **background selection is thus expected to generate a positive relationship between levels of neutral polymorphism and the levels of recombination.**
 - this is exactly the pattern that was first described in *Drosophila melanogaster*.

Fixing advantageous mutations

- a similar process occurs when selection acts on advantageous mutations.
 - consider first, a beneficial allele entering an asexual population.
 - if the mutation happens to find itself in a poor genetic background then selection on that mutation could be ineffective.
 - this will occur when the gain in fitness may not be substantial enough to shift the fitness of that clonal lineage above that of others in the population.
 - the beneficial allele is (after Joel Peck) a “ruby in the rubbish”.
- when a beneficial allele enters a population of a sexual species it is not trapped in any one genetic background.
 - recombination acts to move it from one genetic background to another from generation to generation.
 - the overall fate of the mutation is determined by the mean advantage in fitness it confers to individuals.
 - and recombination effectively frees the selective process to affect only the localized region where the beneficial mutation occurs.
 - when natural selection fixes an advantageous mutation, it causes what is called a “**selective sweep**”.

- selective sweeps act to reduce levels of linked silent polymorphism.
- the size of the chromosomal region impacted by a selective sweep is determined by the level of local recombination.
- if the sweep occurs in a region of low recombination, the window of reduced polymorphism can be large (10s or 100s of kilobases).
- in regions of high recombination, the impact of the sweep on linked silent polymorphism is much more restrictive.
- **therefore, selective sweeps are expected to generate a positive relationship between levels of neutral polymorphism and the levels of recombination.**
- does this sound familiar?
- in fact, when this pattern was first described in *Drosophila* by Begun and Aquadro (1991) it was attributed to selective sweeps.
- background selection was an alternative explanation put forward a few years later by Charlesworth et al (1993)