Linkage and Linkage Disequilibrium

Summer Institute in Statistical Genetics 2013

Module 8

Topic 3

Linkage

Linkage in a simple genetic cross

In the early 1900's Bateson and Punnet conducted genetic studies using sweet peas. They studied two characters: petal color (P purple dominates p red) and pollen grain shape (L elongated is dominant to l disc-shaped).

PPLL X ppll

↓

PpLl

Plants in the F_1 generation were intercrossed: PpLl X PpLl. According to Mendel's second law, "during gamete formation, the segregation of one gene pair is independent of other gene pairs."

The table helps us calculate the expected relative frequencies of the four types of plants in the F_2 generation according to Mendel's second law.

	PL	Pl	pL	pl
PL	Purple Long	Purple Long	Purple Long	Purple Long
Pl	Purple Long	Purple disc-shaped	Purple Long	Purple disc-shaped
pL	Purple Long	Purple Long	red Long	red Long
pl	Purple Long	Purple disc-shaped	red Long	red disc-shaped

Here are the expected relative frequencies of the four phenotypes of plants:

	Elongated	disc-shaped
Purple	9	3
red	3	1

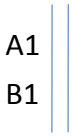
Here are the observed data:

	Elongated	disc-shaped
Purple	284	21
red	21	55

The observed data clearly do not fit what is expected under the model. The reason is that the loci are linked.

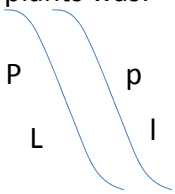
Linkage

Genes are physically arranged in linear strands of DNA and grouped into chromosomes. When a gamete is formed, chunks of a chromosome are passed on. Suppose we have two genes, one with alleles A1 and A2 and another with alleles B1 and B2, that are physically close on a chromosome. Suppose an individual is heterozygous at both loci and, furthermore, the **phase** is as follows:



If the genes are closely linked, a gamete is much more likely to contain (A1,B1) or (A2,B2) - "non- recombinants." If there is recombination, a gamete contains (A1, B2) or (A2,B1), but these two possibilities are less likely. (In contrast Mendel's second law says that all four possibilities are equally likely.)

In sweet peas the loci controlling petal color and pollen grain shape are physically close together on the same chromosome. In the F_1 generation, the **phase** of the PpLl plants was:



Both dominant alleles were located on the same chromosome. A sweet pea was equally likely to pass on a gamete containing P as a gamete containing p, and equally likely to pass on a gamete containing L as a gamete containing l. However, two events such as "pass on P" and "pass on L" were NOT independent because of linkage. A gamete is much more likely to contain PL or pl than to contain Pl or pL. A gamete contains Pl or pL only if there is a recombination event between the two loci when the gamete is formed.

Genotypes and Haplotypes

Haplotype: A sequence of alleles, or of DNA bases, that are on the same chromosome and thus were inherited together.

Depending on the context, haplotypes may or may not refer to adjacent bases/alleles.

Terminology	
Genotype	**Haplotype**
AaBb	AB/ab or Ab/aB
"phase is unknown"	"phase is known"

Recombination Fraction

When two loci follow Mendel's Second law, recombinants and non-recombinants are produced with equal frequency. When loci are physically close to one another on a chromosome, there is a deviation from this relationship. This deviation is summarized by the **recombination fraction c** (sometimes denoted by θ).

$$c = P(\text{recombinant gamete})$$

When loci are unlinked, c=1/2.

When loci are completely linked, c=0.

For c in between 0 and 1/2, the loci are said to be "linked" or "in genetic linkage."

How might we estimate the recombination fraction from data from Bateson and Punnet's sweet pea intercross?

P(recombinant gamete)=c

		½ (1-c)	½ c	½ c	½ (1-c)
		PL	Pl	pL	pl
½ (1-c)	PL	Purple Long	Purple Long	Purple Long	Purple Long
½ c	Pl	Purple Long	Purple disc-shaped	Purple Long	Purple disc-shaped
½ c	pL	Purple Long	Purple Long	red Long	red Long
½ (1-c)	pl	Purple Long	Purple disc-shaped	red Long	red disc-shaped

We can form a likelihood for the data that is a function of the recombination fraction c. We can find the value of c that maximizes this likelihood.

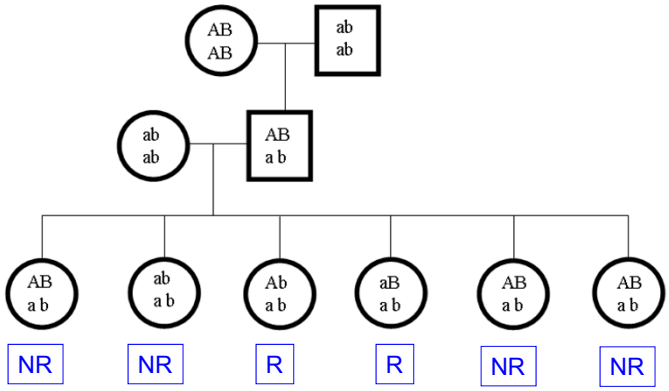
P(red, disc-shaped)=1/4 $(1-c)^2$

P(red, long)=(½ c)(½ c) + (½ c)(½ (1−c)) + (½ c)(½ (1−c))

et cetera

Exercise 1

two loci and a human pedigree



Which children are **recombinants**? Which are **non-recombinants**?
How could you estimate the recombination fraction for loci A and B?

In the previous example, we have all the genotypes for both markers and know the phase for the genotypes. That is, we know how the alleles are coupled on each chromosome for the third generation. We could therefore tell without ambiguity which offspring have a paternal chromosome that is a recombinant.

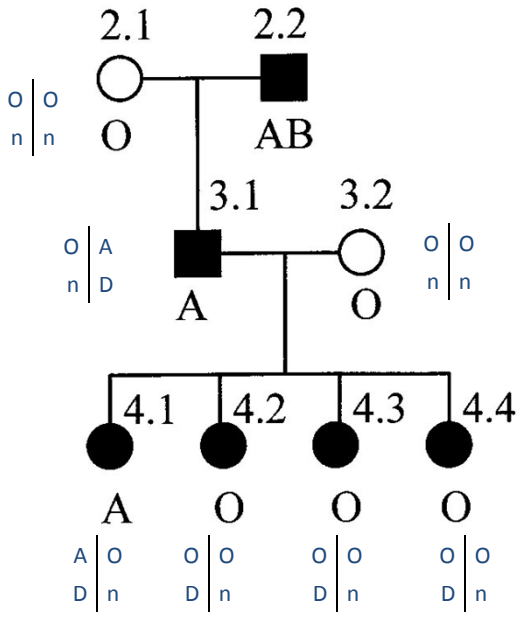
Example

simple linkage analysis: phase unknown

Charcot-Marie-Tooth neuropathy (CMT) is a disease with autosomal dominant inheritance. Consider the following pedigree (Ott J (1999) Analysis of human genetic linkage, 3rd edition. Johns Hopkins University Press). Shading represents an individual with CMT and the ABO blood types are given below each individual. As we know, ABO blood types are determined by a single gene with three alleles. We might be interested in estimating the recombination fraction between the ABO blood type locus and the gene for CMT.

The first thing that we need to do is determine all genotypes with phase.

Phenotypes



In the example we were able to directly count the number of recombinants in the offspring generation. This is an idealized situation. Note that the grandparent generation was necessary for establishing phase, i.e. a nuclear family would not have been sufficient to determine recombinants directly.

Note that we assumed a very specific mode of inheritance: dominant single allele, fully penetrant. We also used a single recombination fraction for males and females.

Suppose we had a set of CMT/ABO families for which we counted N non-recombinants and R recombinants. Then we would have a likelihood of:

$$L(c) = \Pr(data \mid c) = \binom{N+R}{R}(1-c)^N c^R$$

Under the null hypothesis of no linkage, i.e. H_0: c=0.5:

$$L(c = \frac{1}{2}) = \mathrm{P}(data \mid c = \frac{1}{2}) = \binom{N+R}{R}\frac{1}{2}^{(N+R)}$$

LOD Scores

In linkage analysis it is conventional to use LOD scores. We get a LOD score by taking $\log_{10}$ of the likelihood ratio

$$LOD = \log_{10}\left(\frac{\Pr(data \mid alternative)}{\Pr(data \mid null)}\right)$$

The LOD scores for this example is:

$$LOD(c) = N\log(1-c) + R\log c - (N+R)\log\left(\frac{1}{2}\right)$$

Linkage Analysis

In linkage analysis on human pedigrees, we find evidence for the location of a gene for a trait of interest if we find a gene (or marker) of known location with which it is tightly linked. Linkage analysis on human pedigrees is based on the kind of analysis we just did. However, real-life linkage analysis is hugely more complicated because much more is unknown.

- missing genotypes through the pedigree
- phase may not be able to be determined without ambiguity
- decreased penentrance: P(disease | disease genotype) < 1
- Misspecified relationships, e.g. false paternity
- Inbreeding, creating “loops” in the pedigree

Allelic Association

"The excessive co-occurrence of certain combinations of alleles in the same gamete because of tight linkage, or for other reasons, is known as allelic association"

-Pak Sham, Statistics in Human Genetics 1998, p. 145

Consider two loci A and B, with alleles $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ occurring at frequencies $p_1, p_2, \ldots, p_m$ and $q_1, q_2, \ldots, q_n$ in the population. We can consider an individual's two **haplotypes** with respect to these loci, one of maternal origin and one of paternal origin.

How many possible haplotypes are there? mn

The haplotypes can be denoted

$A_1B_1, A_1B_2, \ldots, A_mB_n$ with frequencies $h_{11}, h_{12}, \ldots, h_{mn}$.

In a population, if the occurrence of allele A_i and the occurrence of allele B_j in a haplotype are independent events, then, by definition $h_{ij} = p_i q_j$.

Example of LD: Lactase gene

- Samples from three populations
  - CEU – Eurpean ancestry, living in Utah
  - YRI – Yoruban origin from Nigeria
  - CHB/JPT – Chinese and Japanese, living in Beijing and Tokyo
- Lactase gene: a particular mutation appears to allow the ability to digest milk into adulthood. The high frequency of the mutation in thought to be the result of positive selection in Europeans; the mutation is much less common in non-pastorialist populations.

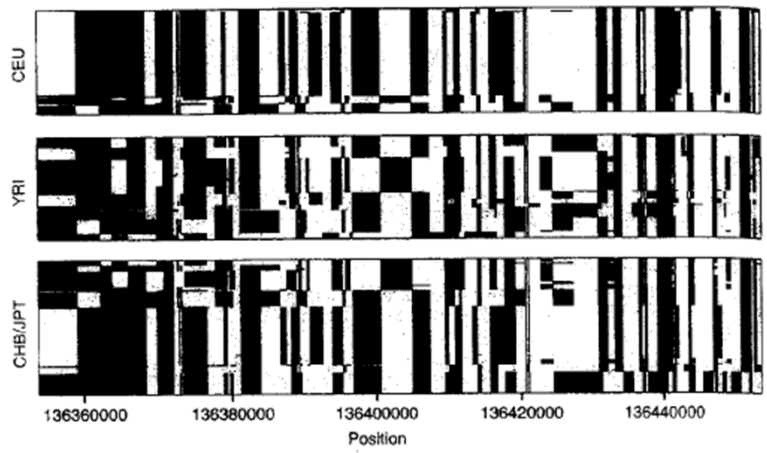


Figure 27.2 Haplotype structure in a 100-kb region surrounding the *Lactase* gene on human chromosome 2 for the four HapMap populations (The International HapMap Consortium, 2005). CEU = Individuals of European origin from Utah; YRI = Yoruba from Nigeria; CHB/JPT = Han Chinese from Beijing and Japanese from the Tokyo region (60 unrelated individuals in each of CEU and YRI, 90 in CHB/JPT). The CEU panel is dominated by a single haplotype that extends over the entire region. Much higher haplotype diversity is found in the other populations (see also Table 27.1).

Terminology

"independent alleles"	$h_{ij} = p_i q_j$
"positively associated alleles"	$h_{ij} > p_i q_j$
"negatively associated alleles"	$h_{ij} < p_i q_j$

Exercise 2

allelic associations

Consider two biallelic loci A and B. There are four possible haplotypes: A_1B_1, A_1B_2, A_2B_1, A_2B_2. Suppose that the frequencies of these four haplotypes in a large population are 0.4, 0.1, 0.2, and 0.3, respectively. Are there any allelic associations between these loci? What are they?

Exercise 2

Allele frequencies:

A_1: 0.4+0.1=0.5
A_2: 0.2+0.3=0.5
B_1: 0.4+0.2=0.6
B_2: 0.1+0.3=0.4

If there were no allelic associations, we should have:
$P(A_1B_1)=(0.5)(0.6)=0.3 < 0.4$ – positive association
$P(A_1B_2)=(0.5)(0.4)=0.2 > 0.1$ – negative association
$P(A_2B_1)=(0.5)(0.6)=0.3 > 0.2$ – negative association
$P(A_2B_2)=(0.5)(0.4)=0.2 < 0.3$ – positive association

Linkage Disequilibrium (LD)

"The term linkage disequilibrium (LD) is broadly used to refer to the non-random sharing (or lack thereof) of combinations of [genetic] variants. "

-G. McVean, "Linkage Disequilibrium, Recombination, and Selection" , in Handbook of Statistical Genetics, Eds. Balding, Bishop, Cannings)

Linkage Disequilibrium (LD)

Consider two loci A and B, with alleles A_1, A_2, ... , A_m and B_1, B_2, ... , B_n. If there is allelic association between any pair of alleles in the different loci, then the loci are said to be in **linkage disequilibrium**. Notice that linkage disequilibrium is a population-level characteristic.

We will discuss three measures of Linkage Disequilibrium, D, D', and R^2.

Linkage Disequilibrium Coefficient D

For ease of notation, we define D for two biallelic loci with alleles A and a at locus 1; B and b at locus 2:

$$D_{AB}=P(AB)-P(A)P(B)$$

Note that $D_{aB} = P(aB)-P(a)P(B)$

$= P(aB)-(1-P(A))P(B)$

$= P(aB)-P(B)+P(A)P(B)$

$= P(aB)-P(aB) -P(AB)+P(A)P(B)$

$= P(A)P(B) -P(AB)$

$= -D_{AB}$

Thus, the magnitude of the coefficient is important, not the sign.

Linkage Disequilibrium Coefficient D

LD is a property of two loci, not their alleles. Therefore it is good that the magnitude of D does not depend on the choice of alleles.

The range of values the linkage disequilibrium coefficient can take on varies with allele frequencies. Allele frequencies cannot be negative, and we know, for example, that P(AB) must be less than both P(A) and P(B). This gives us relations such as:

$$(1)\ 0 \le p_{AB} = p_A p_B + D_{AB} \le p_A, p_B$$
$$(2)\ 0 \le p_{aB} = p_a p_B - D_{AB} \le p_a, p_B$$
$$(3)\ 0 \le p_{Ab} = p_A p_b - D_{AB} \le p_A, p_b$$
$$(4)\ 0 \le p_{ab} = p_a p_b + D_{AB} \le p_a, p_b$$

For example, to derive expression (1):

From the definition: $p_{AB} = p_A p_B + D_{AB}$

Since p_{AB} is a probability, we have $0 \le p_{AB}$

From the rules of probability: $p_{AB} \le p_A, p_B$

So putting these together we have (1):

$$0 \le p_{AB} = p_A p_B + D_{AB} \le p_A, p_B$$

For (2)-(4), use relationships $D_{AB} = -D_{aB}$, etc.

Take the first inequality from (1) through (4):

(1) $0 \leq p_A p_B + D_{AB}$
(2) $0 \leq p_a p_B - D_{AB}$
(3) $0 \leq p_A p_b - D_{AB}$
(4) $0 \leq p_a p_b + D_{AB}$

(1) $-p_A p_B \leq D_{AB}$
(2) $D_{AB} \leq p_a p_B$
(3) $D_{AB} \leq p_A p_b$
(4) $-p_a p_b \leq D_{AB}$

Linkage Disequilibrium Coefficient D

The inequalities on the previous slide lead to bounds for D_{AB} :

$$-p_A p_B, -p_a p_b \leq D_{AB} \leq p_a p_B, p_A p_b$$

Normalized LD Coefficient D'

We have just seen that the possible values of D depend on allele frequencies. This makes D difficult to interpret. For reporting purposes, the **normalized linkage disequilibrium coefficient D'** is often used.

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\max(-p_A p_B, -p_a p_b)} & \text{if } D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(p_a p_B, p_A p_b)} & \text{if } D_{AB} > 0 \end{cases}$$

Exercise 3

Consider two biallelic loci, A,a and B,b. What is the theoretical range of the linkage disequilibrium coefficient D_{AB} and its absolute value $|D_{AB}|$ under the follow scenarios?

(a)P(A)=1/2, P(B)=1/2

All frequencies are 0.5;

$-0.25 \leq D_{AB} \leq 0.25$

$0 \leq |D_{AB}| \leq 0.25$

(b) P(A)=.95, P(B)=.95

$$-(0.95)(0.95),-(0.05)(0.05) \le D_{AB} \le (0.95)(0.05)$$
$$-0.9025,-0.0025 \le D_{AB} \le 0.0475$$
$$-0.0025 \le D_{AB} \le 0.0475$$
$$0 \le |D_{AB}| \le 0.0475$$

(c) P(A)=.95, P(B)=.05

$$-(0.95)(0.05) \le D_{AB} \le (0.95)(0.95),(0.05)(0.05)$$
$$-0.0475 \le D_{AB} \le 0.0025$$
$$0 \le |D_{AB}| \le 0.0475$$

(d) P(A)=1/2, P(B)=.95

$$-(0.5)\ (0.95),(0.5)\ (0.05) \le D_{AB} \le (0.5)\ (0.95),(0.5)\ (0.05)$$
$$-0.025 \le D_{AB} \le 0.025$$
$$0 \le |D_{AB}| \le 0.025$$

Exercise 4

Under what circumstances might D_{AB} reach its theoretical maximum value? Suppose D_{AB} =P(a)P(B). What does this imply? Why does this make sense?

D_{AB}=P(AB)−P(A)P(B)
Suppose D_{AB} = P(a)P(B).
Then P(AB)−P(A)P(B)= P(a)P(B)
P(AB)=P(A)P(B)+ P(a)P(B) = P(B)[P(A) + P(a)]=P(B)

This means that the haplotype frequency of AB is equal to the allele frequency of B. That is, the allele A only occurs when the allele B occurs at the other locus. This is an extreme case of allelic dependence, so it makes sense that the coefficient D is at its maximum.

Estimating D

Suppose we have a situation where we can consider single chromosomes to have been sampled from a population of interest. The data might be arranged in a table such as:

		First Locus		
		B	b	
Second Locus	A	n_{AB}	n_{Ab}	n_A
	a	n_{aB}	n_{ab}	n_a
		n_B	n_b	N

We would like to estimate D_{AB} from the data. The maximum likelihood estimate of D_{AB} is

$$\hat{D}_{AB} = \hat{p}_{AB} - \hat{p}_A \hat{p}_B$$

Estimating D

…where the population frequencies are just estimated by the sample frequencies. The MLE turns out to be slightly biased. If N gametes have been sampled, then

$$E(\hat{D}_{AB}) = \frac{N-1}{N} D_{AB}$$

The variance of this estimate depends on both the true allele frequencies and the true level of linkage disequilibrium:

$$Var(\hat{D}_{AB}) = \frac{1}{N}[p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2]$$

Testing for LD with D

Since D_{AB}=0 corresponds to the status of no linkage disequilibrium, it is sometimes of interest to test the null hypothesis H_0: D_{AB}=0. One way to do this is to use a chi-square statistic. It is constructed by squaring the asymptotically normal statistic z:

$$z = \frac{\hat{D}_{AB} - E(\hat{D}_{AB})}{\sqrt{Var(\hat{D}_{AB})}}$$

after setting D_{AB}=0 in both the expectation and the variance. The test statistic is compared against a Chi-Square distribution with one degree of freedom.

R^2

Define a random variable X_A to be 1 if the allele at the first locus is A and 0 if the allele is a.

Define a random variable X_B to be 1 if the allele at the second locus is B and 0 if the allele is b.

Then the correlation between these random variables is:

$$r_{AB} = \frac{\mathrm{Cov}(X_A, X_B)}{\sqrt{\mathrm{Var}(X_A)\mathrm{Var}(X_B)}} = \frac{D_{AB}}{\sqrt{f_A f_a f_B f_b}}.$$

It is usually more common to consider the square:

$$r^2 = \frac{D^2}{f_A f_a f_B f_b}.$$

R^2

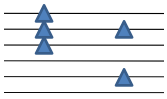
- Has the same value however the alleles are labeled
- Relationship between R^2 and the power of association studies
- Tests for LD: A natural test statistic to consider is the contingency table test. Compute a test statistic using the Observed haplotype fequencies and the Expected frequency if there were no LD:

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

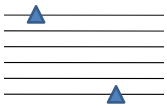
Under the null hypothesis of No LD, the statistic is asymptotically χ^2 distributed with 1 df. It turns out that

$$X^2 = n\hat{r}^2.$$

Is LD high or low?



$R^2=|D'|=0$



$R^2 \approx 0$ $|D'|=1$

"perfect LD"

$R^2=|D'|=1$

D' is problematic to interpret with rare alleles.

"Perfect LD"

haplotype	frequency
AB	98%
Ab	0%
aB	0%
ab	2%

D=P(AB)-P(A)P(B)=.98-(.98)(.98)= 0.0196
Maximum value for these allele frequencies:
min(.98 · .02, .02 · .98)= 0.0196
D'=1

$R^2 = (0.0196)^2/[(.98)(.98)(.02)(.02)]=1$

(This is sometimes called "perfect LD": two alleles at two loci have the same allele frequency and don't ever "separate".)

when D' and R^2 "disagree"

haplotype	frequency
AB	90%
Ab	5%
aB	5%
ab	0%

D=P(AB)-P(A)P(B)=.9-(.95)(.95)= -0.0025
Maximum value for these allele frequencies:
max(-.95 · .95, -0.05 · .05)=max(-.9025, -.0025)=-.0025
D'=1

(When D'=1 then at least one haplotype doesn't exist in the data.)

$R^2 = (-0.0025)^2/[(.95)(.95)(.05)(.05)]=0.0028$

How are allelic associations generated?

- Genetic drift: In a finite population, the gene pool of one generation can be regarded as a random sample of the gene pool of the previous generation. As such, allele and haplotypes frequencies are subject to sampling variation – random chance. The smaller the population is, the larger the effects of genetic drift are.
- Mutation: If a new mutation appears in a population, alleles at loci linked with the mutant allele will maintain an association for many generations. The association lasts longer when linkage is greater (that is, the recombination fraction is much smaller than ½ – very close to 0).

How are allelic associations generated?

•Founder effects: Applies to a population that has grown rapidly from a small group of ancestors. For example, the 5,000,000 Finns mostly descended from about 1000 people who lived about 2000 years ago. Such a population is prone to allelic disequilibrium.

•Selection: When an individual's genotype influences his/her reproductive fitness. For example, if two alleles interact to decrease reproductive fitness, the alleles will tend to be negatively associated.

•Stratification: Some populations consist of two or more subgroups that, for cultural or other reasons, have evolved more or less separately. Two loci that are in linkage equilibrium for each subpopulation may be in linkage disequilibrium for the larger population.

Exercise 5

population stratification

Consider a population with three subpopulations. Consider two biallelic loci, the first with alleles A and a; the second with alleles B and b.

N	A allele frequency	B allele frequency	AB haplotype frequency
1000	0.3	0.5	0.15
2000	0.2	0.4	0.08
10000	0.05	0.1	0.005

Exercise 5

Do any of three subpopulations show allelic association?

No.
0.15=0.3x0.5
0.08=0.2x0.4
0.005=0.05x0.1

Exercise 5

Does the larger population show allelic association?

A allele frequency:
[1000(0.3)+2000(0.2)+10000(0.05)]/13000=0.0923
B allele frequency:
[1000(0.5)+2000(0.4)+10000(0.1)]/13000=0.1770

The equilibrium frequency of AB is
(0.0923)(0.1770)=0.0163
AB haplotype frequency:
[1000(0.15)+2000(0.08)+10000(0.005)]/13000=0.0277
Thus they are positively associated in the larger population.

How are allelic associations maintained?

- selection
- non-random mating (e.g., pop stratification)
- Linkage:

Consider again two loci A and B, with alleles

$A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ with frequencies

$p_1, p_2, \ldots, p_m$ and $q_1, q_2, \ldots, q_n$ in the population. The haplotypes are $A_1B_1, A_1B_2, \ldots, A_mB_n$, with frequencies

$$h_{11}^0, h_{12}^0, \ldots, h_{mn}^0,$$

in generation 0. What is the frequency of haplotype A_iB_j, in the next generation? In the following calculation we assume random mating in the population.

$$h_{ij}^1 = P(haplotype^1 = A_iB_j)$$
$$= P(haplotype^1 = A_iB_j \mid \text{no recombination})P(\text{no recombination})$$
$$+P(haplotype^1 = A_iB_j \mid \text{recombination})P(\text{recombination})$$
$$= P(haplotype^1 = A_iB_j \mid \text{no recombination})(1-c)$$
$$+P(haplotype^1 = A_iB_j \mid \text{recombination})c$$
$$= h_{ij}^0(1-c) + p_iq_jc$$

From this we can deduce that the difference in haplotype frequency between the generations is:

$$h_{ij}^1 - h_{ij}^0 = c(p_iq_j - h_{ij}^0)$$

When will this difference be 0? That is, when are haplotype frequencies stable?

When c=0 or no LD.

We can also use this expression to characterize the difference between the true haplotype frequency and what the haplotype frequency would be under equilibrium:

$$h_{ij}^1 - p_i q_j = (1-c)(h_{ij}^0 - p_i q_j)$$

Extending this to the k*th* generation, we get:

$$h_{ij}^k - p_i q_j = (1-c)^k (h_{ij}^0 - p_i q_j)$$

Alternatively we can write:

$$D_{ij}^1 = (1-c) D_{ij}^0$$

$$D_{ij}^k = (1-c)^k D_{ij}^0$$

The figure on the next slide shows the decline of linkage disequilibrium in a large, randomly-mating population for several different values of c (called theta in the figure).

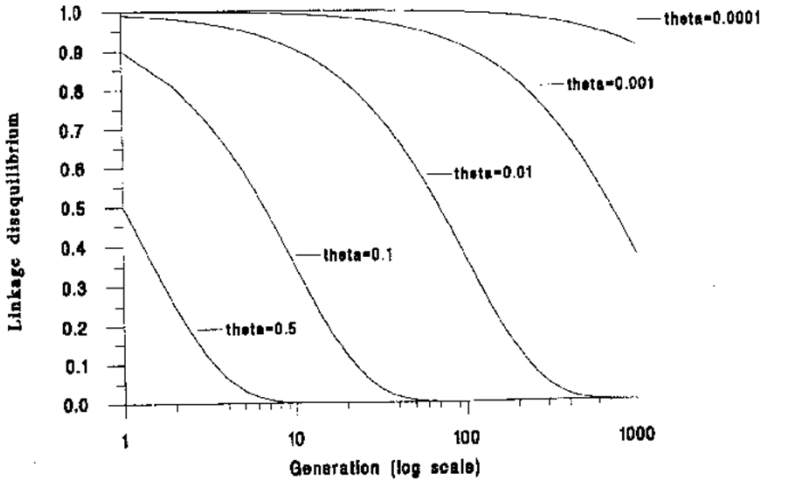


Figure 4.1 Decay of linkage disequilibrium by generation.