Contents

Part II

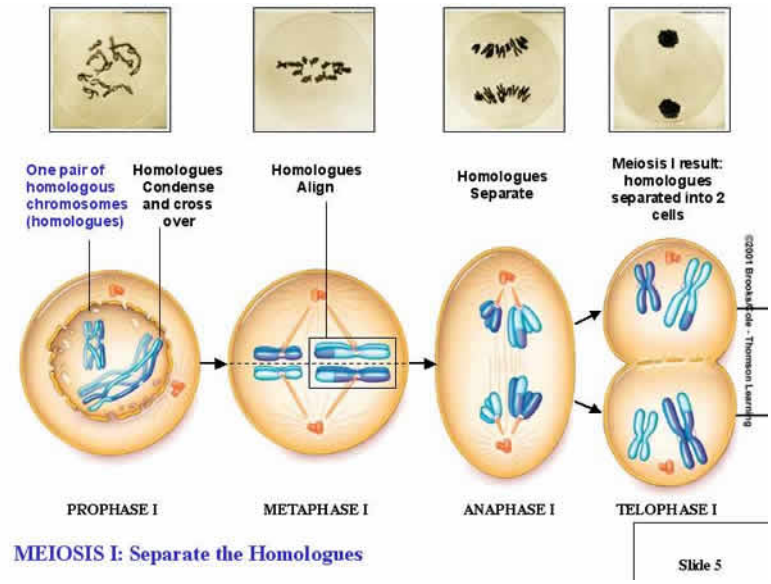
Genetic Linkage

1 Biology

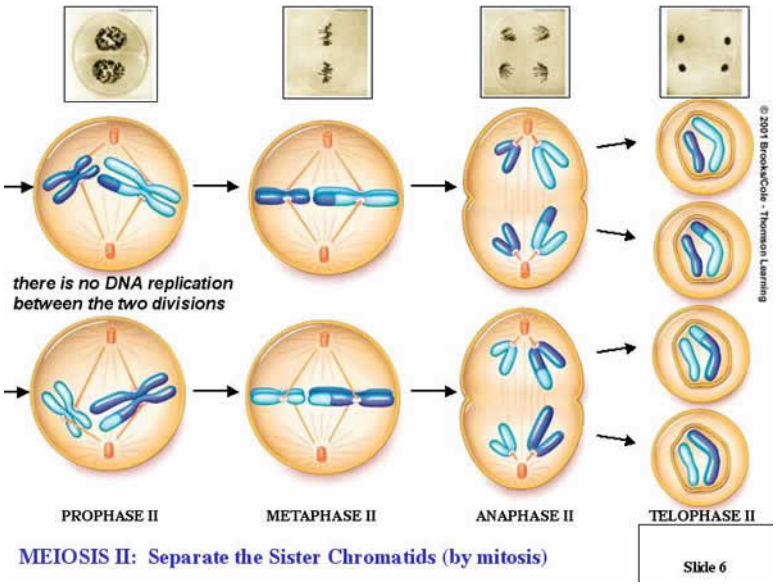
Meiosis

- Meiosis I
  - Chromosomes duplicate
  - Chromosomes condense
  - Homologs pair and *recombination* occurs
  - Homologs separate into two cells
- Meiosis II
  - Duplicate chromosomes separate into two cells

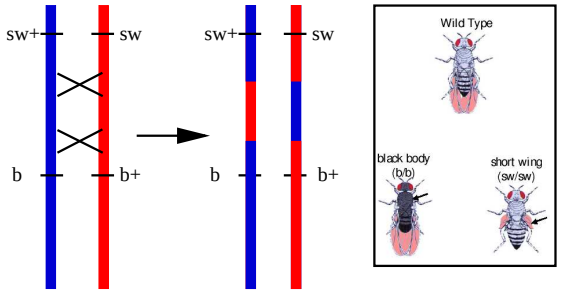
Meiosis I



Meiosis I



Genetic Linkage Example



Definition: *marker*

> A segment of known DNA along a chromosome whose inheritance can be followed, i.e. we can observe/measure it in individuals.

Definition: *haplotype*

> A haplotype (Greek *haploos* = single) is a combination of alleles at multiple loci that are transmitted together on the same chromosome. For example, $A_2B_8C_1D_5E_3$

[Reminder]

Independence If two events E and F are independent, then $P(E \cap F) = P(E)P(F)$.

Allele population frequency $p_1 = P(A_1 \text{ at locus } A)$, $p_A = P(A \text{ at an unnamed locus})$

Haplotype The linear arrangement of alleles along a region of a chromosome, a full chromosome or an entire genome.

Gamete or haplotype population frequency $p_{AB} = P(A \text{ at locus } 1 \cap B \text{ at locus } 2)$

Law of Total Probability $P(E) = P(E \mid F_1)P(F_1) + P(E \mid F_2)P(F_2)$ if $F_1 \cup F_2 = \Omega$ are *exhaustive events* and $F_1 \cap F_2 = \varnothing$ *mutually exclusive* events.

---

2 Mathematics

Following Gamete Frequencies

Let $p_{AB}(t)$ be the probability of haplotype AB in generation t. Any AB haplotype in generation $t+1$ experienced either an odd number $\{1, 3, \ldots\}$ of recombination events or an even $\{0, 2, \ldots\}$ number of recombination events between the A and B loci during meiosis. Therefore, by the law of total probability

$$p_{AB}(t+1) = (1-r)p_{AB}(t) + rp_Ap_B,$$

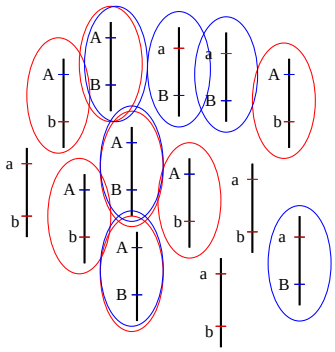
where r is the probability of an odd number of crossover events between the two loci.

To see what happens over time, subtract the haplotype frequency under the assumption of independence: p_Ap_B

$$\begin{aligned} p_{AB}(t+1) - p_Ap_B &= (1-r)p_{AB}(t) + rp_Ap_B - p_Ap_B \\ &= (1-r)p_{AB}(t) - (1-r)p_Ap_B \\ &= (1-r)\left[p_{AB}(t) - p_Ap_B\right]. \end{aligned}$$

Define $D_{AB}(t) = p_{AB}(t) - p_Ap_B$, the difference between the actual haplotype frequency and the hypothetical one achieved if the loci were independent. Call this the **linkage disequilibrium**.

Motivation for Recurrence Relation



How to make AB haplotypes from these chromosomes?

Pick a chromosome with alleles AB with probability $p_{AB}(t)$ and don't let it recombine with probability $(1-r)$

Pick a chromosome with A at the first locus with probability $p_A(t)$

Pick a chromosome B at the first locus with probability $p_B(t)$

And recombine them with probability r

Following Gamete Frequencies (cont)

$$\begin{aligned} D_{AB}(t+1) &= (1-r)D_{AB}(t) \\ &= (1-r)^2 D_{AB}(t-1) \\ &\vdots \\ &= (1-r)^t D_{AB}(0). \end{aligned}$$

As long as $1 > r > 0$, $D_{AB}(t) \to 0$.

When $D_{AB}(t) = 0$, the alleles at the two loci are independent of each other and the loci are said to be in **linkage equilibrium**.

When the two loci are on different chromosomes, can there be linkage disequilibrium? In other words, is it possible that

$$p_{AB}(t) - p_A p_B \neq 0?$$

Loci on Different Chromosomes

Yes! Consider the full complement of chromosomes passed by your grandfather and grandmother to your father. They each pass on a complete haplotype (half genome). Consider one locus on chromosome 1 and another on chromosome 2. By the Law of Segregation, the probability that you get your grandmother's locus 1 is 1/2. Similarly for locus 2. The probability that there is a genetic exchange (*not* recombination) such that locus 1 and 2 come from different grandparents is 1/2 by the Law of Independent Assortment of Chromosomes.

So, for loci on different chromosomes $r = \frac{1}{2}$ and

$$D_{AB}(t) = \left(\frac{1}{2}\right)^t D_{AB}(0),$$

where as before $D_{AB}(0) = p_{AB}(0) - p_A p_B$.

Note: Linkage disequilibrium is a misnomer because it does not always imply physical linkage, so don't let it confuse you!

Loci on different chromosomes can be in linkage disequilibrium, *but they are not physically linked.*

Speed of Approach to Equilibrium

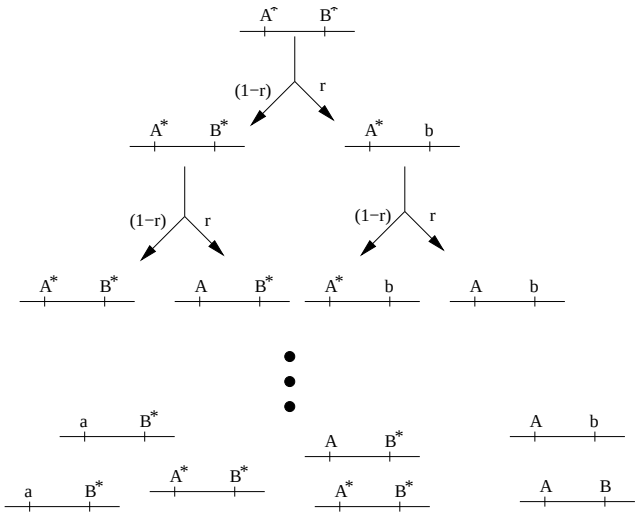
The rate of approach to equilibrium is determined by $(1 - r)$. We say there is a *geometric* rate of approach to equilibrium. This fraction is the amount by which the disequilibrium D_{AB} is decreased each generation.

Another way to think about this: Instead of using $D_{AB}(t)$, rewrite in terms of the haplotype frequencies. Then,

$$\begin{aligned} p_{AB}(t) &= p_A p_B + [p_{AB}(0) - p_A p_B]\,(1-r)^t \\ &= p_{AB}(0)(1-r)^t + \left[1 - (1-r)^t\right] p_A p_B. \end{aligned}$$

where $(1 - r)^t$ is the probability that a haplotype (if traced down the generations) has never experienced a recombination event in t generations.

Gamete Frequencies Across Generations



Consequence of Linkage Dis/Equilibrium

- Selection has no effect on neighboring loci given linkage equilibrium.
- For very tightly, physically linked loci, r is small and convergence to equilibrium is very slow.
- Founding populations can have a profound effect by determining $D_{AB}(0)$ and hence how much disequilibrium exists to persist.
- In particular, founding populations that go through bottlenecks are very important for genetics. Imagine three individuals surviving a bottleneck from a population in linkage equilibrium: AB/AB, Ab/AB, and AB/aB. This population starts with linkage disequilibrium, since $p_{AB}(0) - p_A p_B = \frac{2}{3} - \frac{5}{6}\frac{5}{6} = -\frac{1}{36}$. In fact, any small sample taken from a large population in linkage equilibrium will not likely satisfy linkage equilibrium. This bottleneck-induced linkage disequilibrium, in part, explains the importance of population isolates.

3 Testing Linkage Disequilibrium

3.1 Phase Known

Phase Known

The first case we will consider for testing linkage disequilibrium is when phase information is available. The data consist of the haplotype counts $n_{AB}, n_{Ab}, n_{aB}, n_{ab}$. Let

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

be the current level of linkage disequilibrium in the population sampled. Then, functions for the haplotype probabilities are

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

In fact, $D_{AB} = D_{Ab} = D_{aB} = D_{ab}$. For example, to show $D_{AB} = D_{Ab}$,

$$\begin{aligned} p_A &= p_{AB} + p_{Ab} \\ &= p_A p_B + D_{AB} + p_A p_b - D_{Ab} \\ &= p_A (p_B + p_b) + D_{AB} - D_{Ab} \\ &= p_A + D_{AB} - D_{Ab}. \end{aligned}$$

3.1.1 Maximum Likelihood Estimates

Likelihood

Again, the sampling distribution is multinomial because we are sampling haplotypes without replacement from a very large, homogeneous population.

Suppose we sample n individuals. Then, the likelihood of the data is

$$\begin{aligned} P(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) &= \frac{(2n)!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!} p_{AB}^{n_{AB}} p_{Ab}^{n_{Ab}} p_{aB}^{n_{aB}} p_{ab}^{n_{ab}} \\ &= \frac{(2n)!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!} (p_A p_B + D_{AB})^{n_{AB}} (p_A p_b - D_{AB})^{n_{Ab}} \\ &\quad (p_a p_B - D_{AB})^{n_{aB}} (p_a p_b + D_{AB})^{n_{ab}}. \end{aligned}$$

But since $p_a = 1 - p_A$ and $p_b = 1 - p_B$, there are 3 free parameters (p_A, p_B, D_{AB}), along with 3 degrees of freedom in the data (n_{AB}, n_{Ab}, n_{aB} since $n_{ab} = 2n - n_{AB} - n_{Ab} - n_{aB}$). Bailey's method is applicable.

Maximum Likelihood Estimates

Solve the following system of equations

$$\begin{aligned} n_{AB} &= 2n (p_A p_B + D_{AB}) \\ n_{Ab} &= 2n (p_A p_b - D_{AB}) \\ n_{aB} &= 2n (p_a p_B - D_{AB}) \\ n_{ab} &= 2n (p_a p_b + D_{AB}) \end{aligned}$$

for the MLEs

$$\begin{aligned} \hat{p}_A &= \frac{n_{AB} + n_{Ab}}{2n} = \tilde{p}_A \\ \hat{p}_B &= \frac{n_{AB} + n_{aB}}{2n} = \tilde{p}_B \\ \hat{D}_{AB} &= \frac{n_{AB}}{2n} - \frac{(n_{AB} + n_{Ab})(n_{AB} + n_{aB})}{(2n)^2} \text{ from the first equation} \\ &= \tilde{p}_{AB} - \tilde{p}_A \tilde{p}_B \end{aligned}$$

3.1.2 Bias and Variance

Bias of $\hat{D}_{AB}$

The MLE $\hat{D}_{AB}$ is biased.

$$\begin{aligned} \mathrm{E}\left(\hat{D}_{AB}\right) &= \mathrm{E}(\tilde{p}_{AB}) - \mathrm{E}(\tilde{p}_A \tilde{p}_B) \\ &= p_{AB} - \mathrm{E}(\tilde{p}_A \tilde{p}_B) \\ &= p_A p_B + D_{AB} - \mathrm{E}(\tilde{p}_A \tilde{p}_B). \end{aligned}$$

Let X_{ij} be an indicator variable that is 1 if the jth chromosome in the ith individual carries an A allele. Let Y_{ij} indicate a B allele on the jth chromosome in the ith individual. Then,

$$\begin{aligned}\tilde{p}_A &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij} \\ \tilde{p}_B &= \frac{1}{2n}\sum_{k=1}^{n}\sum_{l=1}^{2} Y_{kl}.\end{aligned}$$

E $(\tilde{p}_A\tilde{p}_B)$

$$\begin{aligned}\mathrm{E}\left(\tilde{p}_A\tilde{p}_B\right) &= \frac{1}{4n^2}\mathrm{E}\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij}\right)\left(\sum_{k=1}^{n}\sum_{l=1}^{2} Y_{kl}\right)\right] \\ &= \frac{1}{4n^2}\mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij}Y_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{2}\sum_{l\neq j,l=1}^{2} X_{ij}Y_{il}\right. \\ &\quad \left. + \sum_{i=1}^{n}\sum_{k\neq i,k=1}^{n}\sum_{j=1}^{2}\sum_{l=1}^{2} X_{ij}Y_{kl}\right] \\ &= \frac{1}{4n^2}\left[2np_{AB} + 2np_Ap_B + 4n(n-1)p_Ap_B\right] \\ &= \frac{1}{4n^2}\left[2np_Ap_B + 2nD_{AB} + 2np_Ap_B + 4n^2p_Ap_B - 4np_Ap_B\right] \\ &= p_Ap_B + \frac{1}{2n}D_{AB}.\end{aligned}$$

Bias of $\hat{D}_{AB}$

Therefore,

$$\mathrm{E}\left(\hat{D}_{AB}\right) = p_Ap_B + D_{AB} - p_Ap_B - \frac{1}{2n}D_{AB} = \frac{2n-1}{2n}D_{AB}.$$

When there is no linkage disequilibrium $D_{AB} = 0$, and the expectation becomes unbiased

$$\mathrm{E}\left(\hat{D}_{AB}\right) = 0.$$

Variance of $\hat{D}_{AB}$

Here, we assume HWE so that combinations haplotypes (or alleles) on different are just products of haplotype (or allele) frequencies. Also, Fisher's approximation applies because $\hat{D}_{AB}$ is a function of sample proportions. Applying Fisher's formula provides the variance of the MLE $\hat{D}_{AB}$.

$$\begin{aligned}\mathrm{Var}\left(\hat{D}_{AB}\right) &= \frac{1}{2n}\left[p_A(1-p_A)p_B(1-p_B)\right. \\ &\quad \left. + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2\right].\end{aligned}$$

When there is no linkage disequilibrium, $D_{AB} = 0$ and the variance is

$$\mathrm{Var}\left(\hat{D}_{AB}\right) = \frac{p_A(1-p_A)p_B(1-p_B)}{2n}.$$

Substitute in the sample allele proportions to get a number.

3.1.3 z Test & Chi-Square Test

Hypothesis Testing

We are interested in testing the null hypothesis of linkage equilibrium. In other words we wish to test

$$H_0 : D_{AB} = 0.$$

- z test uses the z statistic

$$z = \frac{\hat{D}_{AB} - \mathrm{E}\left(\hat{D}_{AB}\right)}{\sqrt{\mathrm{Var}\left(\hat{D}_{AB}\right)}} = \frac{\hat{D}_{AB}}{\sqrt{\mathrm{Var}\left(\hat{D}_{AB}\right)}} = \frac{\sqrt{2n}\hat{D}_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

- A chi-square statistic uses z^2.

$$X^2_{AB} = z^2$$

3.1.4 Exact Test

Exact Test for Linkage Disequilibrium

As we have previously mentioned, the haplotype counts $(n_{AB}, n_{Ab}, n_{aB}, n_{ab})$ are multinomially distributed

$$P\left(n_{AB}, n_{Ab}, n_{aB}, n_{ab}\right) \quad = \quad \frac{2n!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!} p_{AB}^{n_{AB}} p_{Ab}^{n_{Ab}} p_{aB}^{n_{aB}} p_{ab}^{n_{ab}}$$

And when we can assume linkage equilibrium, this probability reduces to

$$\frac{2n!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!} \left(p_A p_B\right)^{n_{AB}} \left(p_A p_b\right)^{n_{Ab}} \left(p_a p_B\right)^{n_{aB}} \left(p_a p_b\right)^{n_{ab}}$$

The allele counts are binomially distributed as before. For example,

$$P\left(n_A, n_a\right) = \frac{(2n)!}{n_A!n_a!} p_A^{n_A} p_a^{n_a}.$$

The exact test works by considering all possible haplotype counts that could have been sampled given the same allele counts.

The probability of each dataset is obtained as the following conditional probability

$$P\left(n_{AB}, n_{Ab}, n_{aB}, n_{ab} \mid n_A, n_B\right) = \frac{n_A!n_a!n_B!n_b!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!(2n)!}$$

Now, as before the procedure is as follows

- List all data sets possible with the given allele counts n_A and n_B.
- Calculate the probability of each data set.
- Rank the datasets by the probability from smallest to largest.
- Calculate the cumulative probabilities for first 2, first 3, first 4, etc...
- Report the first cumulative probability that includes the actual, observed data and report this as the p-value for the test of linkage equilibrium.

Example of Exact Test for LD

Counts		*Xho*I +	*Xho*I -	Total
*Bam*HI	+	5	6	11
	-	6	0	6
Total		11	6	17

Gamete (*Bam*HI:*Xho*I)					Cumulative	
+ +	+ -	- +	- -	Probability	Probability	
11	0	0	6	0.0001	0.0001	
10	1	1	5	0.0053	0.0054	
5	6	6	0	0.0373	0.0427	←
9	2	2	4	0.0667	0.1094	
6	5	5	1	0.2240	0.3334	
⋮	⋮	⋮	⋮	⋮	⋮	

3.1.5 Power Calculations

Power and Sample Size Calculations

One can use the chi-square statistic to perform power and sample size calculations just as before. The noncentrality parameter for the noncentral chi-square that applies under the alternative hypothesis is

$$\nu = \frac{2nD_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}.$$

As for exact tests, the allele counts n_A and n_B still follow binomial distribution even with linkage disequilibrium. The probability of the haplotype counts is

$$\begin{aligned} P\left(n_{AB}, n_{Ab}, n_{aB}, n_{ab}\right) &= \\ \frac{n!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!}\left(p_{AB}\right)^{n_{AB}}\left(p_{Ab}\right)^{n_{Ab}}\left(p_{aB}\right)^{n_{aB}}\left(p_{ab}\right)^{n_{ab}} &= \\ \frac{n!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!}\left(p_Ap_B + D_{AB}\right)^{n_{AB}}\left(p_Ap_b - D_{AB}\right)^{n_{Ab}} & \\ \times\left(p_ap_B - D_{AB}\right)^{n_{aB}}\left(p_ap_b + D_{AB}\right)^{n_{ab}} & \end{aligned}$$

for a given D_{AB}.

Assuming a hypothetical disequilibrium $D_{AB} = d$ and allele frequencies p_A and p_B. Set the allele counts $n_A = np_A$ and $n_B = np_B$ for a sample size n, and then enumerate all data sets.

For example, suppose $p_A = p_B = 0.65$ and $n = 17$, then $n_A = n_B = 17 \times 0.65 \approx 11$ and $n_a = n_b \approx 6$.

Data	H_0		H_A	
$\mathbf{n} = (n_{AB}, n_{Ab}, n_{aB}, n_{ab})$	$P\left[\mathbf{n} \mid D_{AB} = 0\right]$	Cumul.	$P\left[\mathbf{n} \mid D_{AB} = d\right]$	Cumul.
$11, 0, 0, 6$	0.0001	0.0001	0.247	0.247
$10, 1, 1, 5$	0.0053	0.0054	0.509	0.756
$5, 6, 6, 0$	0.0373	0.0427	0.199	0.955
$9, 2, 2, 4$	0.0667	0.1094	0.025	0.980
⋮	⋮	⋮	⋮	⋮

It is not quite correct because n_A and n_B will vary in sampling.

3.1.6 Multiple Alleles

Multiple Alleles

Generalizing to more than 2 alleles at just 2 loci causes little problem. Let D_{uv} be the linkage disequilibrium for allele u and v, where u is one of k alleles at locus 1 and v is one of l alleles at locus 2. Then

$$D_{uv} = p_{uv} - p_u p_v.$$

There are constraints because

$$\begin{aligned} \sum_v D_{uv} &= \sum_v p_{uv} - p_u \sum_v p_v \\ &= p_u - p_u = 0 \end{aligned}$$

and similarly

$$\sum_u D_{uv} = 0.$$

There are a total of $k + l - 1$ constraints leaving $(k-1)(l-1)$ free D_{uv} parameters.

Multiple Alleles - Data

The chi-square test-of-independence for a k by l table is equivalent to testing for linkage equilibrium.

$$H_0 : D_{uv} = 0 \text{ for all } u, v$$

		Locus 2				
		B_1	B_2	B_3	B_4	Total
Locus 1	A_1	33	54	3	8	98
	A_2	14	6	1	0	21
	A_3	87	2	54	7	150
	A_4	4	34	23	8	69
Total		138	96	81	23	338

$$X^2_{(k-1)(l-1)} = \frac{\left(33 - \frac{98\times 138}{338}\right)^2}{\frac{98\times 138}{338}} + \cdots \sim \chi^2_{(k-1)(l-1)}$$

Chi-Square Test of Independence

```
> m <- matrix(c(33,54,3,8,14,6,1,0,87,2,54,7,4,34,23,8),nrow=4,byrow=T)
> chisq.test(m)

Pearson's Chi-squared test

data:  matrix(m, nrow = 4, byrow = T)
X-squared = 147.8365, df = 9, p-value < 2.2e-16

Warning message:
In chisq.test(m) : Chi-squared approximation may be incorrect
> chisq.test(m, simulate.p.value=T)

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  matrix(m, nrow = 4, byrow = T)
X-squared = 147.8365, df = NA, p-value = 0.0004998
```

And of course exact tests can be computed using genetics software (e.g. GenePop) that should be more efficient than the simple Monte Carlo sampling of `chisq.test()`.

Normalized Linkage Disequilibria

Recall

$$\begin{aligned} p_{AB} &= p_A p_B + D_{AB} \\ p_{ab} &= p_a p_b + D_{AB}, \end{aligned}$$

or generally, for multiple alleles,

$$\begin{aligned} p_{AB} &= p_A p_B + D_{AB} \\ p_{\bar{A}\bar{B}} &= p_{\bar{A}} p_{\bar{B}} + D_{\bar{A}\bar{B}}. \end{aligned}$$

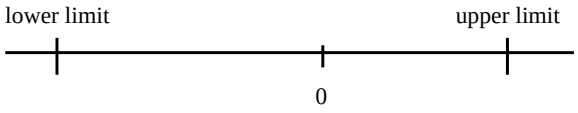
where $\bar{A}$ represents all alleles but A and similarly for B. You know that

$$\begin{aligned} 0 \leq \quad p_{AB} &\leq 1 \\ p_{AB} &\leq p_A \\ p_{AB} &\leq p_B. \end{aligned}$$

These restrictions limit the possible values of D_{AB}. In addition, the range of D_{AB} depend on the allele frequencies, making it difficult to compare multiple disequilibria.

Normalize the disequilibrium by using these limits. The above equations imply

$$\max\left(-p_A p_B, -p_{\bar{A}} p_{\bar{B}}\right) \leq D_{AB} \leq \min\left(p_{\bar{A}} p_B, p_A p_{\bar{B}}\right).$$



Set the normalized disequilibrium (somewhere in $[-1, 1]$) as

$$D'_{AB} = \begin{cases} -\frac{D_{AB}}{\max(-p_A p_B, -p_{\bar{A}} p_{\bar{B}})} & D_{AB} < 0 \\ \frac{D_{AB}}{\min(p_{\bar{A}} p_B, p_A p_{\bar{B}})} & D_{AB} > 0 \end{cases}$$

Details: The complete derivation of the limits is shown as

$$\begin{aligned} 0 &\leq p_A p_B + D_{AB} \leq \min(p_A, p_B) \\ -p_A p_B &\leq D_{AB} \leq \min(p_A p_b, p_a p_B) \\ \text{Similarly,} \quad -p_a p_b &\leq D_{AB} \leq \min(p_A p_b, p_a p_B) \end{aligned}$$

Combine the last two equations to conclude

$$-\max(p_a p_b, P_A p_B) \leq D_{AB} \leq \min(p_A p_b, p_a p_B)$$

Also, we need to check the other conditions are satisfied by the above equation.

$$\begin{aligned} 0 &\leq p_a p_B - D_{AB} \leq \min(p_a, p_B) \\ -\min(p_a, p_B) &\leq D_{AB} \leq p_a p_B \\ \text{Similarly,} \quad -\min(p_A, p_b) &\leq D_{AB} \leq p_A p_b \end{aligned}$$

It turns out these limits are more liberal than the ones we already derived because $p_i \geq p_i p_j$ for all i, j and clearly $p_a p_B$ and $p_A p_b$ are both $\geq \min(p_A p_b, p_a p_B)$. Thus the first conditions give us the most restrictive limits and the proof is complete.

Other Associations Among Loci

Suppose HWE does not apply so that there are now associations between alleles or haplotypes on different loci. The number and types of associations that can result increases very fast. Below is a complete list of associations among the four alleles at two loci in diploid individuals.

Parameter/Association	chromosome 1	chromosome 2
p_A, p_B	●	○
$D_A = p_{AA} - p_A^2, D_B = p_{BB} - p_B^2$	●	●
$D_{AB} = p_{AB} - p_A p_B$	●	○
$D_{A/B} = p_{A/B} - p_A p_B$	○	●
D_{AAB}, D_{ABB}	●	●
D_{AB}^{AB}	●	●

Trigenic and Quadrigenic Associations

Trigenic associations are dependent on digenic associations. Consider a large coefficient D_{AB}, so a strong association between A and B on the same chromosome. Then a trigenic association between A and B on the same chromosome with A on a second chromosome will also be large because already there is the A and B association on the same chromosome. To separate these affects one substracts the digenic associations from the trigenics.

$$\begin{aligned} D_{AAB} &= p_{AAB} - p_A^2 p_B - p_A D_{AB} - p_A D_{A/B} - p_B D_A \\ D_{ABB} &= p_{ABB} - p_A p_B^2 - p_B D_{AB} - p_B D_{A/B} - p_A D_B \end{aligned}$$

Similarly quadrigenic disequilibrium is defined as

$$\begin{aligned} D_{AB}^{AB} &= P_{AB}^{AB} - 2p_A D_{ABB} - 2p_B D_{AAB} - 2p_A p_B D_{AB} - 2p_A p_B D_{A/B} \\ &\quad - p_A^2 D_B - p_B^2 D_A - D_{AB}^2 - D_{A/B}^2 - D_A D_B - p_A^2 p_B^2 \end{aligned}$$

HWE implies no cross-chromosome associations (e.g. $D_A = D_B = D_{A/B} = D_{AAB} = D_{ABB} = D_{AB}^{AB} = 0$) and it was under this assumption that we derived variance for $\hat{D}_{AB}$ previously. When there are cross-chromosome associations (HWD), then the variance changes to

$$\text{Var}(\hat{D}_{AB}) = \frac{1}{2n}\left(p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2 + D_A D_B + D_{A/B}^2 + D_{AB}^{AB}\right)$$

Also, we can derive

$$\text{Var}(\hat{D}_{A/B}) = \frac{1}{2n}\left(p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D_{A/B} - D_{A/B}^2 + D_A D_B + D_{AB}^2 + D_{AB}^{AB}\right)$$

Finally, the trigenic variance is

$$\begin{aligned}\text{Var}(\hat{D}_{AAB}) &= [p_A^2(1-p_A)^2 + (1-2p_A)^2 D_A - D_A^2][p_B(1-p_B) + D_B] \\ &\quad + p_A(1-p_A)(1-2p_A)(1-2p_B)(D_{AB} + D_{A/B}) - 2D_{AAB}^2 \\ &\quad + [1 - 5p_A(1-p_A) + D_A](D_{AB} + D_{A/B})^2 + 2p_A(1-p_A)(1-2p_A)D_{ABB} \\ &\quad + [(1-2p_A)^2(1-2p_B) - 2(1-2p_B)D_A - 4(1-2p_A)(D_{AB} + D_{A/B})]D_{AAB} \\ &\quad + [(1-2p_A)^2 - 2D_A](D_{AB}^{AB} - 2D_{AB}D_{A/B})\end{aligned}$$

The variance for $\hat{D}_{ABB}$ is symmetric with the above. Variances for quadrigenic associations fill many lines, but can be found in Weir & Cockerham. (1989) "Complete characterization of the disequilibrium at two loci" in *Mathematical Evolutionary Theory.* Feldman (ed.) Princeton University Press: Princeton, pp.86–110. These formula illustrate the increasing complexity of handling multiple alleles and loci simultaneously. Adding a third locus would be deathly.

Estimating/Testing Di/Trigenic Associations

The MLEs of all disequilibria are obtained by substituting sample proportions into the formulas. Variances are obtained using Fisher's approximation. The complexity increases very fast.

Testing for associations proceeds using chi-square statistics computed with the MLE and Fisher-approximated variances

$$X_D^2 = \frac{\hat{D}^2}{\text{Var}\left(\hat{D}\right)}$$

for some disequilibrium D.

Testing Di/Trigenic Associations

Unfortunately, testing gets complicated fast because

- $X_{D_{AB}}^2$ depends on $D_{A/B}$ and $X_{D_{A/B}}^2$ depends on D_{AB}.
- $X_{D_{AAB}}^2$ and $X_{D_{ABB}}^2$ depend on D_{AB}^{AB}.

In the face of these difficulties, the recommended testing procedure is:

- Test $H_0 : D_{AB}^{AB} = 0$ first. If not rejected, set $D_{AB}^{AB} = 0$.
- Test $H_0 : D_{AAB} = D_{ABB} = 0$ next, setting $D_{AB}^{AB} = 0$ if possible.
- Test $H_0 : D_{AB} = 0$ and $H_0 : D_{A/B} = 0$ without assumptions, i.e. leaving $\hat{D}_{AB}$ in the formula.
- If either test on digenics is not rejected, retest the other. For example, if $H_0 : D_{AB} = 0$ is not rejected, retest $H_0 : D_{A/B} = 0$ this time setting $D_{AB} = 0$ in $X_{A/B}^2$.

3.2 Phase Unknown

Missing Phase Information

Often the only thing that can be observed is the genotype and the phase information is hidden.

One solution is to use the EM algorithm to estimate the haplotype frequencies with phase information. Unfortunately, this approach assumes HWE. What to do if you do not want to assume HWE?

Let $p_{A/B}$ represent the frequency of A and B on different chromosomes within individuals. The problem is we cannot distinguish this event from A and B on the same chromosome, i.e. p_{AB}. But, the sum

$$p_{AB} + p_{A/B} = 2P_{AB}^{AB} + P_{A\bar{B}}^{AB} + P_{\bar{A}B}^{AB} + \frac{1}{2}\left(P_{\bar{A}\bar{B}}^{AB} + P_{\bar{A}B}^{A\bar{B}}\right)$$

is observable because it is a function of phase-free genotype frequencies

$$p_{AB} + p_{A/B} = 2P_{AABB} + P_{AABb} + P_{AaBB} + \frac{1}{2}P_{AaBb}.$$

3.2.1 Composite Disequilibrium Δ_{AB}

Composite Digenic Disequilibrium

Define

$$\Delta_{AB} = p_{AB} + p_{A/B} - 2p_A p_B = D_{AB} + D_{A/B}.$$

which measures both linkage and cross-linkage coefficients.

Again, an MLE is obtained by substituting in sample proportions:

$$\hat{\Delta}_{AB} = \tilde{p}_{AB+A/B} - 2\tilde{p}_A\tilde{p}_B.$$

where

$$\tilde{p}_{AB+A/B} = \frac{1}{n}\left(2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}\right).$$

Then,

$$\hat{\Delta}_{AB} = \frac{1}{n}\left(2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}\right) - 2\tilde{p}_A\tilde{p}_B.$$

Testing Composite Digenic Disequilibrium

Unfortunately, the variance of $\hat{\Delta}_{AB}$ depends on higher order associations.

Trigenic disequilibrium pose no problems when phase information are not available.

$$\begin{aligned} D_{AAB} &= p_{AAB} - p_A^2 p_B - p_A D_{AB} - p_A D_{A/B} - p_B D_A \\ &= p_{AAB} - p_A^2 p_B - p_A \Delta_{AB} - p_B D_A \\ D_{ABB} &= p_{ABB} - p_A p_B^2 - p_B D_{AB} - p_B D_{A/B} - p_A D_B \\ &= p_{ABB} - p_A p_B^2 - p_B \Delta_{AB} - p_A D_B. \end{aligned}$$

Also, remembering quadrigenic disequilibrium

$$\begin{aligned} D_{AB}^{AB} &= P_{AB}^{AB} - 2p_A D_{ABB} - 2p_B D_{AAB} - 2p_A p_B D_{AB} - 2p_A p_B D_{A/B} \\ &\quad - p_A^2 D_B - p_B^2 D_A - D_{AB}^2 - D_{A/B}^2 - D_A D_B - p_A^2 p_B^2 \end{aligned}$$

motivates a phase-free quadrigenic disequilibrium

$$\begin{aligned} \Delta_{AABB} &= P_{AB}^{AB} - 2p_A D_{ABB} - 2p_B D_{AAB} - 2p_A p_B \Delta_{AB} - \Delta_{AB}^2 \\ &\quad -p_A^2 D_B - p_B^2 D_A - D_A D_B - p_A^2 p_B^2. \end{aligned}$$

There is a difference in these two equations of the term $2D_{AB}D_{A/B}$, but we *cannot* handle it if phase is unknown.

The procedure for testing is as follows

- Test $H_0 : \Delta_{AABB} = 0$. If not rejected, ignore quadrigenic disequilibrium from now on.
- Test $H_0 : D_{AAB} = D_{ABB} = 0$. If not rejected, ignore trigenic disequilibria from now on.
- Test $H_0 : \Delta_{AB} = 0$.

The great advantage of this test is that it does *not* assume Hardy-Weinberg equilibrium.

3.2.2 Other Exact Tests

Other Tests

- Genotype independence across loci, i.e. $P_{A_sA_tB_uB_v} = P_{A_sA_t}P_{B_uB_v}$?

  Exact test: condition on single-locus genotype frequencies and consider the different two-locus datasets that could arise. We need

  $$P\left(\{n_{stuv}\} \mid \{n_{rs}, n_{uv}\}\right) = \frac{\prod_{r,s} n_{rs}! \prod_{u,v} n_{uv}!}{n! \prod_{r,s,u,v} n_{rsuv}!}$$

  and the procedure is directly analagous to the exact test for linkage disequilibrium.

- LE and HWE, i.e. $P_{A_sA_tB_uB_v} = 4p_{A_s}p_{A_t}p_{B_u}p_{B_v}$?

  Exact test: consider all possible genotype arrangements given the same fixed allele frequencies. We need

  $$P\left(\{n_{stuv}\} \mid \{n_r\}, \{n_u\}\right) = \frac{n!2^{H_A}s^{H_B}\prod_r n_r! \prod_u n_u!}{(2n)!(2n)! \prod_{r,s,u,v} n_{stuv}!}$$

3.2.3 Multiple Testing

Multiple Tests Over Multiple Loci

Suppose you collect data on L different loci and you want to know about Hardy-Weinberg at these loci. Your conclusion depends on your question.

- **Does HWE apply to each locus?** Apply the test L times. If the lth test rejects at level α, conclude locus l is not in HWE.
- **Does HWE apply to the sampled population?** Apply the test L times. If the lth test rejects at level α, then conclude that the population is not at HWE, but adjust your type I error.

$$\begin{aligned}
\alpha' &= P\left(> 0 \text{ tests reject } H_0 \mid H_0 \text{ true}\right) \\
&= 1 - P\left(0 \text{ tests reject } H_0 \mid H_0 \text{ true}\right) \\
&= 1 - \left[1 - P\left(\text{ test rejects } H_0 \mid H_0 \text{ true}\right)\right]^L \\
&\quad \text{assuming independence of tests!} \\
&= 1 - (1-\alpha)^L \\
&\approx L\alpha.
\end{aligned}$$

Bonferroni Correction

If you assume that the tests are independent then $\alpha' \approx L\alpha$. To control the type I error α' on your test, note the significance level per test (**experimentwise error rate**) should be set at

$$\alpha = \frac{\alpha'}{L} \text{ for example } \frac{0.05}{L}.$$

Note, that the independence of tests assumption is probably not valid. If HWE does not apply at one locus, it is less likely to apply at a second locus because, perhaps the population is not satisfying one of the HW assumptions.

3.2.4 Test of Population Homogeneity

Tests for Homogeneity

Suppose you have two datasets collected at different times on the what you think is the same population. Neither dataset is large, but if you could combine the datasets, you would have a dataset large enough to do some real testing. Can you combine the samples?

Implicit assumptions: (1) same population, and (2) same collection procedure.

	Sample			
Category	1	2	3	Row Totals
A	n_{A1}	n_{A2}	n_{A3}	$\sum_i n_{Ai}$
B	n_{B1}	n_{B2}	n_{B3}	$\sum_i n_{Bi}$
C	n_{C1}	n_{C2}	n_{C3}	$\sum_i n_{Ci}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Col. Totals	$\sum_u n_{u1}$	$\sum_u n_{u2}$	$\sum_u n_{u3}$	n

Again, a chi-square test of independence applies.

Computing Expected Counts

The expected counts in cell u, i are given by

$$E_{ui} = \frac{\sum_j n_{uj} \sum_v n_{vi}}{n}$$

Then compute the chi-square goodness of fit statistic

$$X^2 = \sum_{u,i} \frac{(O_{ui} - E_{ui})^2}{E_{ui}},$$

which follows a χ^2_d distribution with $d = (\#\text{ rows} - 1)\,(\#\text{ columns} - 1)$ degrees of freedom.

If the hypothesis is not rejected, then conclude that the samples can be merged and proceed with your other tests.