Chapter 3: Genes on chromosomes

3.1.1 MEIOSIS INDICATORS AT MULTIPLE LOCI :

• *Segregation* of DNA at a locus j is fully specified by *meiosis indicators*

$$\begin{aligned} S_{i,j} &= 0 \text{ if DNA copied at locus} j \text{ is parent's maternal DNA} \\ &= 1 \text{ if DNA copied at locus} j \text{ is parent's paternal DNA} \end{aligned}$$

where $i = 1, ..., m$ indexes the meioses.

• Mendel's First Law: $S_{i,j}$ are independent over i with

$$\Pr(S_{i,j} = 0) = \Pr(S_{i,j} = 1) = \frac{1}{2}.$$

• $S_{i,j}$ are independent for loci j on different chromosome pairs
$S_{i,j}$ are dependent among loci j on the same chromosome pair

• The vector $S_{\bullet,j} = \{S_{i,j}, i = 1, ..., m\}$ is known as the inheritance vector at that locus.
ibd at locus j is a function of the inheritance vector $S_{\bullet,j}$.

3.1.2 RECOMBINATION BETWEEN TWO LOCI:

- Recall there is recombination if the genes at the two loci j and l come from different parental chromosomes (different grandparents).

- For two given loci (l and j) the recombination frequency ρ between them is

$$\rho \;=\; \Pr(S_{i,l} \neq S_{i,j}) \quad \text{for each } i, \quad 0 \leq \rho \leq \frac{1}{2}.$$

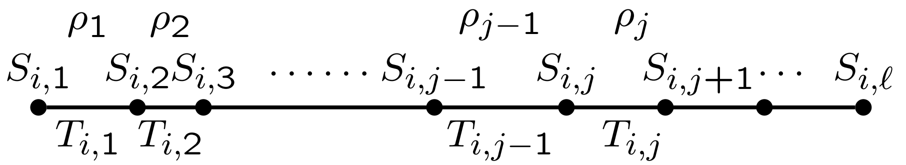
- The pairwise distribution of $(S_{i,j}, S_{i,l})$ is determined by ρ.

- For loci that are close together on a chromosome, ρ is close to 0. For independently segregating loci, $\rho = \frac{1}{2}$.

$S_{i,l}$	$S_{i,j}$		
	0	1	
0	$(1-\rho)/2$	$\rho/2$	1/2
1	$\rho/2$	$(1-\rho)/2$	1/2
	1/2	1/2	1

- In practice, recombination frequencies vary among meioses, a major factor in this variation being the sex of the parent. Computationally, this can be incorporated.

3.1.3 MULTILOCUS RECOMBINATION: NO INTERFERENCE:



- $S_{i,j}$ specifies inheritance at locus j in meosis i.
  ρ_j is probability of recombination between locus j and locus $j+1$.
  $T_{i,j}$ is indicator of recombination between locus j and locus $j+1$ in meiosis i.

- $\Pr(T_{i,j} = 1) = \Pr(S_{i,j} \neq S_{i,j+1}) = \rho_j$.

- Assume all the $T_{i,j}$ are independent.
  Then $S_{i,j}$ are Markov over j
  Given $S_{i,j}$, $S_{i,j-1}$ is independent of $S_{i,j+1}$.

- Recombination probabilities are not additive.($\rho_{1:3} \neq \rho_1 + \rho_2$)
For example: $\rho_{1:3} = \Pr(S_{i,1} \neq S_{i,3}) = \rho_1(1-\rho_2) + \rho_2(1-\rho_1)$.
For example: if $\rho_1 = \rho_2 = 0.1$ then $\rho_{1:3} = 0.18$ (see lab 3).

3.1.4 GENETIC DISTANCE AND THE CROSSOVER PROCESS:

- Between two loci, the genetic distance d in Morgans is the expected number of crossovers between the loci on a given gamete.

- Regardless of the crossover process (i.e. regardless of dependence in number and locations of crossovers), genetic distance is always additive, since expectations are additive.

- Usually we measure genetic distance in centiMorgans, because a Morgan can be a whole chromosome. 100cM = 1 Morgan.

- Genetic distance has little to do with physical distance;
  but 1cM $\approx 10^6$bp is a very useful overall rule.

- In a given meiosis, between two loci, recall there is recombination, if, in the offspring gamete, there is an odd number of crossovers between the loci.

- The recombination probability, $\rho(d)$, as a function of d is the *map function*.

3.1.5 THE HALDANE MAP FUNCTION:

- In the model of Haldane (1919), crossovers are assumed to occur as a Poisson process, rate 1 per Morgan (by defn.). The number of crossovers $C(d)$ in genetic distance d is Poisson with mean d.

- This is a model of no genetic interference. The numbers of crossovers in disjoint intervals are independent, and, conditionally on the number occurring, their locations are uniformly and independently distributed.

- Under Haldane's model, $\rho(d)$ is the probability that a Poisson random variable with mean d is odd:

$$\begin{aligned}\rho(d) &= \sum_{k \text{ odd}} e^{-d}\frac{d^k}{k!} = \frac{1}{2}e^{-d}\sum_{k=0}^{\infty}\left(\frac{d^k}{k!} - \frac{(-d)^k}{k!}\right) \\ &= \frac{1}{2}e^{-d}(e^d - e^{-d}) = \frac{1}{2}(1 - \exp(-2d)).\end{aligned}$$

- Under this model, $\rho(d)$ is an increasing function of d, $\rho(d) \to \frac{1}{2}$ as $d \to \infty$, and $\rho(d) \approx d$ as $d \to 0$.

- If $\rho = 0.1$, $d = 0.1115$ (11.15 cM). If $\rho = 0.18$, $d = 0.223$.

3.1.6 INTERFERENCE and OTHER MAP FUNCTIONS:

• In fact, interference exists, mainly in crossovers inhibiting the nearby presence of others, and the requirement for reliable meiosis that there is at least one chiasma on every chromosome pair.

• Pairwise interference can be characterized by coincidence function:

$$c_i(j,j') \;=\; \frac{\Pr(T_{i,j} = T_{i,j'} = 1)}{\Pr(T_{i,j} = 1)\,\Pr(T_{i,j'} = 1)} \left\{ \begin{array}{l} < 1 \text{ positive interference} \\ = 1 \text{ no interference} \\ > 1 \text{ negative interference} \end{array} \right.$$

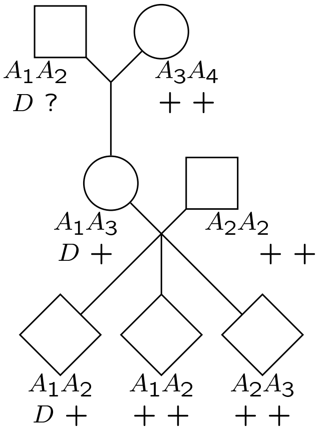
• There are lots of models – the model determines the map function. The reverse is not true. An important positive interference map function is the Kosambi (1944) map function: many published maps of 1970s-1990s are in Kosambi cM.

• Inference and estimation is always in terms of ρ. A map function simply allows representation on a linear map. The important thing is to use the correct map function when transforming between published genetic distances and ρ.

• Almost all multilocus computations assume no interference.

3.2.1 COUNTING RECOMBINANTS:

• *Linkage analysis* is concerned with estimating ρ and with testing the null hypothesis $H_0 : \rho = \frac{1}{2}$ against the alternative $H_1 : \rho < \frac{1}{2}$. Estimates and tests are based on likelihoods and likelihood ratios.



• At a DNA marker locus, two grandparents have types A_1A_2 and A_3A_4; their daughter has type A_1A_3.

• She marries someone of type A_2A_2 and their three children are of types A_1A_2, A_1A_2 and A_2A_3.

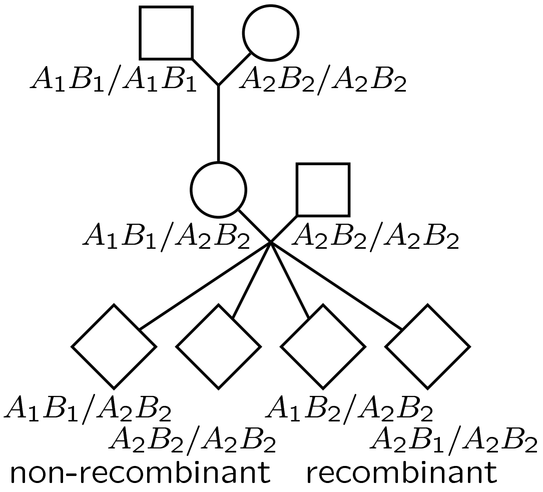
• Granddad, the daughter, and the first child all carry some trait allele D. Other individuals carry only normal $+$ alleles.

• From grandparents to Mom we can phase her as A_1D/A_3+. To her three kids, kids 1 and 3 are non-recombinant ($X_1 = X_3 = 0$) and kid 2 is recombinant ($X_2 = 1$). So $n = 3$, the number of recombinants $T \sim B(3,\rho)$, and T takes the value $t = 1$.

3.2.2 BACKCROSS DESIGN (Phase known):

- Where each offspring can be classified recombinant or non-recombinant, as above, the number of recombinants in n observed offspring is $T \sim B(n, \rho)$.

- Such data arise in a *backcross experiment* using two inbred lines:



Line 1: alleles A_1 and B_1 (genotype A_1B_1/A_1B_1),
Line 2: alleles A_2 and B_2 (genotype A_2B_2/A_2B_2).
Hybrid (F1): all have genotype A_1B_1/A_2B_2.
Backcross to line 2: all get A_2B_2 from the line-2 parent; combination A_1B_1, A_2B_2 (non-recombinant), A_1B_2 or A_2B_1 (recombinant) from the F1 parent observable.

(Note: Have A-locus, (alleles $A_1,\ A_2$) and B-locus.)

3.2.3 BACKCROSS ANALYSIS:

- Suppose n offspring of such matings are scored, and t are recombinant. To test for linkage, compare the likelihood to its value in the absence of linkage ($\rho = \frac{1}{2}$): the log-likelihood difference is

$$\begin{aligned} \mathrm{lod}(\rho) &= \ell(\rho) - \ell(\frac{1}{2}) \\ &= t\log(\rho) \ + \ (n-t)\log(1-\rho) \ + \ n\log(2). \end{aligned}$$

- With base-10 logs, this is known as the **lod score**.

The MLE of ρ is $\widehat{\rho} = t/n$, provided $2t \leq n$ (since $\rho \leq \frac{1}{2}$). To test $\rho = \frac{1}{2}$ against $\rho < \frac{1}{2}$, the maximized lod score is:

$$\mathrm{lod}(\widehat{\rho}) \ = \ t\log t \ + \ (n-t)\log(n-t) - n\log(n/2)$$

provided $2t \leq n$, and 0 otherwise. This is a decreasing function of t, and we reject the null hypothesis $\rho = \frac{1}{2}$ if $t < t_0$ with critical value t_0 chosen to give a specified size of the test (type I error).

3.2.4 TYPE-1 ERROR and CRITICAL VALUES:

- When n is large, T is approximately $N(n\rho, n\rho(1-\rho))$. If $\rho = \frac{1}{2}$. $T \sim N(\frac{n}{2}, \frac{n}{4})$, is good approximation.

- Then $\frac{2}{\sqrt{n}}(T - \frac{n}{2}) \sim N(0,1)$. For a test size (type-1 error) α, reject H_0 in favor of $H_1 : \rho < \frac{1}{2}$ if $\frac{2}{\sqrt{n}}(T - \frac{n}{2}) \leq \Phi^{-1}(\alpha)$ where Φ is the standard Normal cdf.

- For example, for $\alpha = 0.025$, $\Phi^{-1}(\alpha) = -1.96 \approx -2$, so reject H_0 if $T \leq \frac{n}{2} - \sqrt{n} = k^*$.

offspring sampled n	critical value k^*	recombinant proportion k^*/n	lod score $\text{lod}_{10}(k^*/n)$	recombinants for lod score 3
25	≈ 7	≈ 0.3	1.088	≤ 3
100	≈ 40	≈ 0.4	0.874	≤ 31
625	≈ 287	≈ 0.46	0.905	≤ 267
1024	≈ 480	≈ 0.48	0.869	≤ 452

- Table of critical values for a test size $\alpha = 0.025$ and base-10 lod scores for binomial samples.

Added comments:

- The (base 10) lod score is around 1 for a number of recombinants at the critical value for a test of size $\alpha = 0.025$ of $H_0 : \rho = \frac{1}{2}$.

- Traditionally, a base-10 lod score of 3 is required to infer linkage. This is a more stringent test, the idea being that if two arbitrary locations in the genome are chosen the prior probability of linkage is small.

- Also given in the table is the upper bound on the number of recombinants that will provide a lod score of 3.

- The type-1 error at this lod-score-3 critical value k^{**} is $\Phi((2k^{**} - n)/\sqrt{n})$ which is order 10^{-4}.

3.2.5 TESTING USING natural (base-e) LOD SCORES:

- Here and on next page we use **base-e** logs !!

$$\begin{aligned} \ell(\rho) &= t\log(\rho) + (n-t)\log(1-\rho), \qquad \widehat{\rho} = t/n \\ \ell(\widehat{\rho}) &= t\log t + (n-t)\log(n-t) - n\log n \\ \ell(1/2) &= t\log(1/2) + (n-t)\log(1-1/2) = n\log(1/2) \\ \mathrm{lod}(\rho) &= \ell(\rho) - \ell(1/2) \end{aligned}$$

- Example 1: $H_0: \ \rho = 0.1$
$2(\ell(\widehat{\rho}) - \ell(0.1)) \sim \chi^2_1$ if H_0 is true.

- Example 2: But we want to test $H_0: \ \rho = 0.5$
Then $2(\ell(\widehat{\rho}) - \ell(0.5)) = 2\mathrm{lod}(\widehat{\rho})$
If H_0 is true, then half the time $\widehat{\rho} = 0.5$, and $\ell(\widehat{\rho}) = \ell(0.5)$.
So $2(\ell(\widehat{\rho}) - \ell(0.5))$ is $(1/2) \times 0 + (1/2) \times \chi^2_1$ if there is no linkage.

- Note this is the analogue of doing a one-sided test in testing based on the number of recombinants.

3.2.6 Testing equality of recombination frequencies:

- Suppose we see t_m recombinants in n_m male meioses and t_f recombinants in n_f female meioses. Then we can test $H_0: \rho_m = \rho_f$.

- Unconstrained case (general hypothesis):

$$\begin{aligned} \ell(\rho_m, \rho_f) &= t_m\log(\rho_m) + (n_m - t_m)\log(1-\rho_m) \\ &\quad + t_f\log(\rho_f) + (n_f - t_f)\log(1-\rho_f) \end{aligned}$$

maximized by $\widehat{\rho_m} = t_m/n_m, \ \widehat{\rho_f} = t_f/n_f$.

- Under H_0: if $\rho_m = \rho_f = \rho$,

$$\ell(\rho,\rho) = (t_m + t_f)\log(\rho) + (n_m + n_f - t_m - t_f)\log(1-\rho)$$

maximized by $\widehat{\rho} = (t_m + t_f)/(n_m + n_f)$.

- If H_0 is true, $2(\ell(\widehat{\rho_m}, \widehat{\rho_f}) - \ell(\widehat{\rho}, \widehat{\rho}))$ is χ^2_1.
(2 parameters in general, 1 under H_0)
Remember to use **base-e** logs.