Chapter 7 Linkage

Introduction

Recombination fraction

Genetic map construction

Locating genes affecting traits

Distances Between Genes

- The recombination fraction C_{AB} between loci *A* and
  B is the probability that the gamete transmitted by an
  individual is recombinant

- Linkage parameter $\lambda_{AB} = 1 - 2C_{AB}, 0 \le C_{AB} \le \frac{1}{2}$

Physical distance d_{AB} (kb)

Genetic map distance X_{AB} (cM)

Haldanc map function (1919)

$$C_{AB} = \frac{1}{2}(1 - e^{-2X_{AB}})$$

$$X_{AB} = -\frac{1}{2}\ln(1 - 2C_{AB})$$

Kosambi map function (1944)

$$X_{AB} = \frac{1}{4}\ln\left(\frac{1+2C_{AB}}{1-2C_{AB}}\right)$$

Estimation of Recombination

- Backcross method

Table7.1

Two-locus counts for a backcross. Each locus has two alleles, with one dominant to the other.

		Offspring				
	Genotype	AB/ab	Ab/ab	aB/ab	ab/ab	
	Phenotype	AB	Ab	aB	ab	Total
Observed		*a*	*b*	*c*	*d*	*n*
Expected		$\frac{n}{4}$	$\frac{n}{4}$	$\frac{n}{4}$	$\frac{n}{4}$	*n*

$$X_A^2 = \frac{(a + b - n/2)^2}{n/2} + \frac{(c + d - n/2)^2}{n/2} = \frac{(a + b - c - d)^2}{n}$$

$$X_B^2 = \frac{(a + c - b - d)^2}{n}$$

$$X_{AB}^2 = \frac{(a + d - b - c)^2}{n}$$

$$X^2 = \frac{(a - n/4)^2}{n/4} + \frac{(b - n/4)^2}{n/4} + \frac{(c - n/4)^2}{n/4} + \frac{(d - n/4)^2}{n/4}$$

$$= X_A^2 + X_B^2 + X_{AB}^2$$

Example: The maize data reported by Goodman et al (1980)

$\frac{E12,U4}{E12,U4}$	$\frac{E12,U8}{E12,U4}$	$\frac{E15,U4}{E12,U4}$	$\frac{E15,U8}{E12,U4}$
53	13	9	55

$$X^2_{Mdh5} = \frac{(66 - 64)^2}{130} = 0.03$$

$$X^2_{Got3} = \frac{(62 - 68)^2}{130} = 0.28$$

$$X^2_{Mdh5,Got3} = \frac{(108 - 22)^2}{130} = 56.89$$

$$X^2 = \frac{20.5^2 + 19.5^2 + 23.5^2 + 22.5^2}{32.5} = 57.02$$

Table 7.2
Two-locus counts for a backcross for linked loci.

		Offspring				
	Genotype	AB/ab	Ab/ab	aB/ab	ab/ab	
	Phenotype	AB	Ab	aB	ab	Total
Observed		a	b	c	d	n
Expected		$\frac{n}{2}(1-c_{AB})$	$\frac{n}{2}c_{AB}$	$\frac{n}{2}c_{AB}$	$\frac{n}{2}(1-c_{AB})$	n

$$\hat{c} = \min(\tilde{p}, 1-\tilde{p})$$

$$L(c_{AB}) \propto (1\text{-}c_{AB})^{a+d}(c_{AB})^{b+c}$$

$$S_{cAB} = -\frac{a+d}{1\text{-}c_{AB}} + \frac{b+c}{c_{AB}}$$

$$I_{cAB} = \frac{n}{c_{AB}(1-c_{AB})}$$

$$\hat{c}_{AB} = \frac{b+c}{n}$$

$$Var(\hat{c}_{AB}) = \frac{c_{AB}(1\text{-}c_{AB})}{n}$$

$$\hat{c} = \min\left(\frac{b+c}{n}, 0.5\right)$$

- F_2 population method: (homework)
- Direct method: Three-generation method

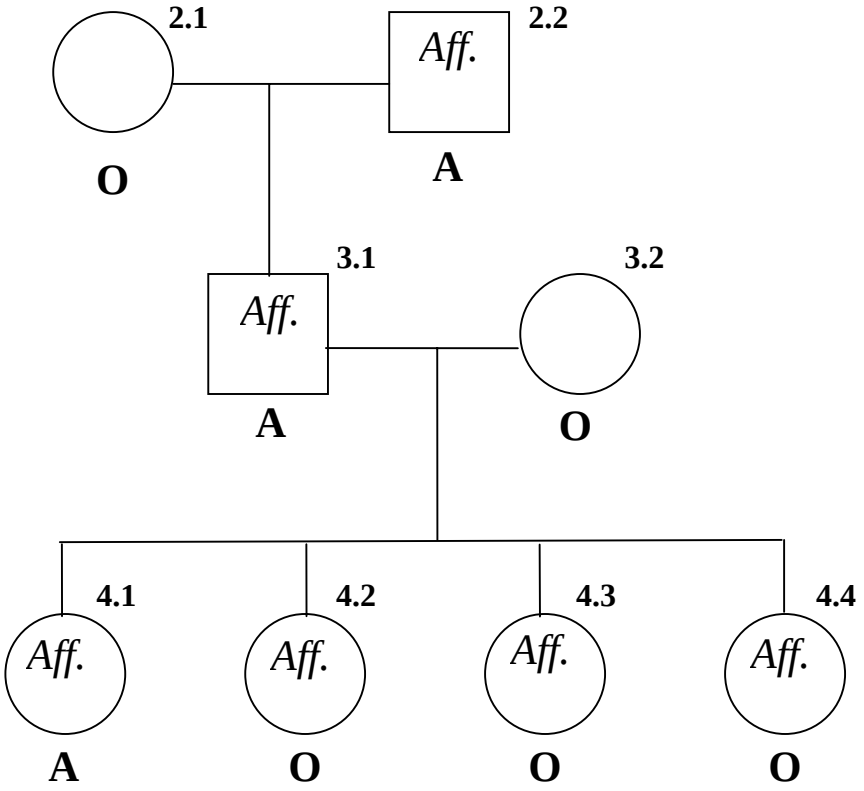


Figure 7.1 Three generation pedigree

- Lod scores with phase known (pg 238)

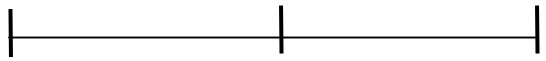
$$Z(c) = \log[\frac{L(c)}{L(0.5)}]$$

$$L(c) \propto (1-c)c^3$$

$$Z(c) = \log[16(1-c)c^3]$$

- Lod scores with phase unknown (pg 240)

Genetic Map Construction



A **B** **C**

- By Haldane map function,

$$C_{AC} = C_{AB}(1-C_{BC}) + (1-C_{AB})C_{BC}$$

- Sum of adjacent recom. fraction (sar)
- If the order of the loci *A*, *B* , *C* was mistakenly believed to be *B* , *A*, *C* , then the sar under the wrong order would be calculated as $C_{AB} + C_{AC}$. From the relation derived under the true order,

$$C_{AB} + C_{AC} \geq C_{AB} + C_{BC}$$

The sar under the wrong order is not less than that under the correct order.

- Two problems with using the sar as a criterion for

determining the correct order for a set of loci (pg 241).

- Seriation
- Simulated annealing
- Branch and bound
- Error detection

Linkage Disequilibrium Mapping

Pedigree-based and population-based approaches.

Random sample

- Genotype data : pg 126
- Haplotype data :

Conditional data

- Genotype data : Suppose a disease locus has alleles *N*, *D* and a marker locus has two alleles M_1, M_2. In the whole population, the allele frequencies are $p_N\ p_D$ at the disease locus and m_1, m_2 at the marker locus. Allowing for the linkage disequilibrium D between the loci, the four gametes have frequencies

$$NM_1\text{:}\ y_1 = p_N m_1 + D$$
$$NM_2\text{:}\ y_2 = p_N m_2 - D$$
$$DM_1\text{:}\ y_3 = p_D m_1 - D$$
$$DM_2\text{:}\ y_4 = p_D m_2 + D$$

For a random-mating population, the conditional frequencies a of marker allele M_1 in each of the three disease genotypic categories are

$$a_1 = \Pr(M_1 \mid NN) = \frac{y_1^2 + y_1 y_2}{p_N^2} = m_1 + \frac{D}{p_N}$$

$$a_3 = \Pr(M_1 \mid DD) = \frac{y_3^2 + y_3 y_4}{p_D^2} = m_1 - \frac{D}{p_D}$$

$$a_2 = \Pr(M_1 \mid ND) = \frac{y_1 y_4 + y_2 y_3 + 2 y_1 y_3}{2 p_N p_D}$$

$$= m_1 + \frac{D}{2 p_N} - \frac{D}{2 p_D} = \frac{1}{2}(a_1 + a_3)$$

$$L(a_1, a_3)_1 \propto (a_1)^{x_1} (1 - a_1)^{2n_1 - x_1}$$

$$L(a_1, a_3)_2 \propto (a_1 a_3)^{n_{21}} (a_1 + a_3 - 2a_1 a_3)^{n_{22}} (1 - a_1 - a_3 + a_1 a_3)^{n_{23}}$$

$$L(a_1, a_3)_3 \propto (a_3)^{x_3} (1 - a_3)^{2n_3 - x_3}$$

$$Var(\hat{a}_1) = \delta_1 \beta_1 \beta_{13} / \gamma$$

$$Var(\hat{a}_3) = \delta_3 \beta_3 \beta_{13} / \gamma$$

$$Cov(\hat{a}_1, \hat{a}_3) = -n_2 \delta_1 \beta_1 \beta_{13} / \gamma$$

$$\beta_1 = a_1(1 - a_1) \qquad \beta_3 = a_3(1 - a_3)$$

$$\beta_{13} = a_1 + a_3 - 2a_1 a_3$$

$$\delta_1 = \beta_{13}(2n_3 + n_2) - n_2 \beta_1$$

$$\delta_3 = \beta_{13}(2n_1 + n_2) - n_2 \beta_3$$

$$\gamma = \delta_1 \delta_3 - n_2^2 \beta_1 \beta_3$$

$$\hat{m}_1 = p_N \hat{a}_1 + p_D \hat{a}_3$$

$$\hat{D} = p_N p_D (\hat{a}_1 - \hat{a}_3)$$

$$Var(\hat{m}_1) = p_N^2 Var(\hat{a}_1) + p_D^2 Var(\hat{a}_3) + 2 p_N p_D Cov(\hat{a}_1, \hat{a}_3)$$

$$Var(\hat{D}) = p_N^2 p_D^2 [Var(\hat{a}_1) + Var(\hat{a}_3) - Cov(\hat{a}_1, \hat{a}_3)]$$

Yule's coefficient of association

- Haplotype data :

$$a_1 = \Pr(M_1 \mid N) = \frac{\Pr(NM_1)}{\Pr(N)} = m_1 + \frac{D}{p_N}$$

$$a_3 = \Pr(M_1 \mid D) = \frac{\Pr(DM_1)}{\Pr(D)} = m_1 - \frac{D}{p_D}$$

$$\hat{m}_1 = p_N \frac{n_{N1}}{n_N} + p_D \frac{n_{D1}}{n_D} = p_N \hat{a}_1 + p_D \hat{a}_3$$

$$\hat{D} = p_N p_D \left(\frac{n_{N1}}{n_N} - \frac{n_{D1}}{n_D}\right) = p_N p_D (\hat{a}_1 - \hat{a}_3)$$

- Heterozygote data :

Other tests:

Case-Control test: The comparison of marker allele frequencies in samples of disease and normal genotypes or haplotypes is called case controlled test.

- **Mean Haplotype-sharing tests：**

  Affected sib-pair test ：It considers the marker alleles received by two afftected children. If there is no linkage between disease and marker loci, the distribution of ibd alleles should have nothing to do with disease status.

Table 7.7 Observed and expected frequencies for four haplotypes when only heterozygotes at a disease locus are available.

	M_1		M_2		
	Exp.	Obs.	Exp.	Obs.	Total
N	$\frac{1}{2}a_1$	$\frac{1}{2n_2}(n_{21}+n_{22}^c)$	$\frac{1}{2}(1-a_1)$	$\frac{1}{2n_2}(n_{23}+n_{22}^r)$	$\frac{1}{2}$
D	$\frac{1}{2}a_3$	$\frac{1}{2n_2}(n_{21}+n_{22}^r)$	$\frac{1}{2}(1-a_3)$	$\frac{1}{2n_2}(n_{23}+n_{22}^c)$	$\frac{1}{2}$
Total	a_2	$\frac{1}{2n_2}(2n_{21}+n_{22})$	$1-a_2$	$\frac{1}{2n_2}(2n_{23}+n_{22})$	1

- **Transmission disequilibrium test:**

Table 7.8 Genotypes and transmitted gametes for heterozygous markers.

Genotype	Frequency	Probability of				
		D	$T:M_1D$ $NT:M_1$	$T:M_1D$ $NT:M_2$	$T:M_2D$ $NT:M_1$	$T:M_2D$ $NT:M_2$
M_1D/M_1D	$\Pr(M_1D)^2$	1	1	0	0	0
M_1D/M_1N	$2\Pr(M_1D)\Pr(M_1N)$	1/2	1/2	0	0	0
M_1N/M_1N	$\Pr(M_1N)^2$	0	0	0	0	0
M_1D/M_2D	$2\Pr(M_1D)\Pr(M_2D)$	1	0	1/2	1/2	0
M_1D/M_2N	$2\Pr(M_1D)\Pr(M_2N)$	1/2	0	$(1-c)/2$	$c/2$	0
M_1N/M_2D	$2\Pr(M_1N)\Pr(M_2D)$	1/2	0	$c/2$	$(1-c)/2$	0
M_1N/M_2N	$2\Pr(M_1N)\Pr(M_2N)$	0	0	0	0	0
M_2D/M_2D	$\Pr(M_2D)^2$	1	0	0	0	1
M_2D/M_2N	$2\Pr(M_2D)\Pr(M_2N)$	1/2	0	0	0	1/2
M_2N/M_2N	$\Pr(M_2N)^2$	0	0	0	0	0

$$\Pr(M_1D) = pm + D$$
$$\Pr(T:M_1, NT:M_1 \mid T:D) = m^2 + mD/p$$
$$\Pr(T:M_1, NT:M_2 \mid T:D) = m(1-m) + (1-c-m)D/p$$
$$\Pr(T:M_2, NT:M_1 \mid T:D) = m(1-m) + (c-m)D/p$$
$$\Pr(T:M_2, NT:M_2 \mid T:D) = (1-m)^2 - (1-m)D/p$$

$$\Pr(T:M_1, NT:M_2 \mid M_1M_2, T:D) = \frac{1}{2} + \frac{(1-2c)D}{2[2m(1-m)p + D(1-2m)]}$$
$$\Pr(T:M_2, NT:M_2 \mid M_1M_2, T:D) = \frac{1}{2} - \frac{(1-2c)D}{2[2m(1-m)p + D(1-2m)]}$$

- **Population admixture**

$$p_{DM_I} = p_{D_I} p_{M_I}$$
$$p_{DM_{II}} = p_{D_{II}} p_{M_{II}}$$
$$p_{DM} = \alpha p_{DM_I} + (1-\alpha) p_{DM_{II}}$$
$$= [\alpha p_{D_I} + (1-\alpha) p_{D_{II}}][\alpha p_{M_I} + (1-\alpha) p_{M_{II}}]$$
$$+ \alpha(1-\alpha)(p_{D_I} - p_{D_{II}})(p_{M_I} - p_{m_{II}})$$
$$= p_D p_M + \alpha(1-\alpha)(p_{D_I} - p_{D_{II}})(p_{M_I} - p_{m_{II}})$$

- Estimating marker-disease distances
  - Old disease : When the disease mutation has been present in the population for a long time, the amount of linkage disequilibuium even to nearby markers is expected to be zero.

$$r^2 = \frac{\hat{D}^2}{\hat{p}_N \hat{p}_D \hat{m}_1 \hat{m}_2} = \frac{X^2}{n}$$

$$E(r^2) = \frac{1}{1+4\mathrm{NC}}$$

  - Young disease : For young disease, chromosome carrying the disease allele are also likely to carry the marker allele that was present on the chromosome(s) on which the disease mutation occurred.

$$p_{D_i} = \alpha \Pr(\textit{no recom.}) + \left[1 - \alpha \Pr(\textit{no recom.})\right] p_{N_i}$$

Homework:

1. Page 236: The F_2 individuals will fall into four recognizable classes when there is dominance at both loci. (a) verify Table 7.4 (b) show that the likelihood of $C_{AB} \propto (2+\theta)^a (1-\theta)^{b+c} (\theta)^d$ (c) verify $S(\theta)$, $I(\theta)$, $\hat{c}_{AB}$ and $Var(\hat{c}_{AB})$.
2. Page 240: If the parents are double heterozygotes of unknown phase, verify Pr(type I child), Pr(two type I children), Z(c) of two type II children, Z(c) for two independent families, one with two type I children and one with one of each type of child.