

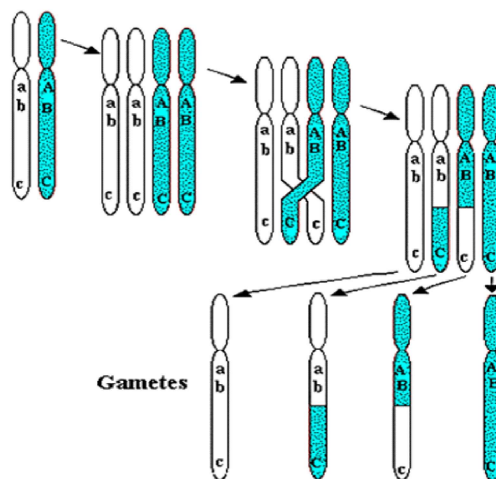
8

Estimation of the Recombination Fraction

8.1: Learning Objectives

By the end of this chapter, the reader should be able to define recombination and the recombination fraction, distinguish linked from unlinked loci, derive and apply the maximum likelihood estimator of the recombination fraction from double-backcross family data, and compute and interpret the LOD score for linkage.

8.2: Recombination and Crossing-Over



Crossing-over and recombination during meiosis: homologous chromosomes exchange segments, producing recombinant gametes.

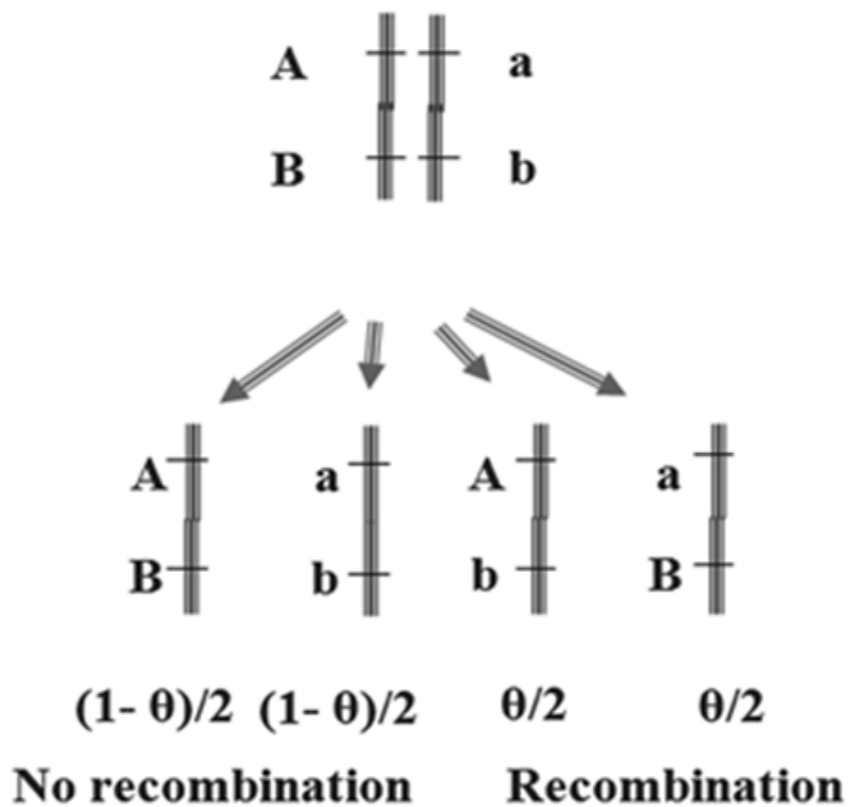
During meiosis (Chapter 1), homologous chromosomes may exchange corresponding segments through a process called **crossing-over**, which physically occurs at points called **chiasmata**. When crossing-over occurs between two loci on the same chromosome, the alleles that were originally together on one parental chromosome (“in coupling”) can end up separated into different gametes — this outcome is called a **recombinant** gamete; a gamete retaining the original parental combination is a **non-recombinant** (parental) gamete.

The **recombination fraction**, denoted θ (theta) or sometimes r , is defined as the probability that a gamete produced by a doubly heterozygous individual is recombinant at two given loci:

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

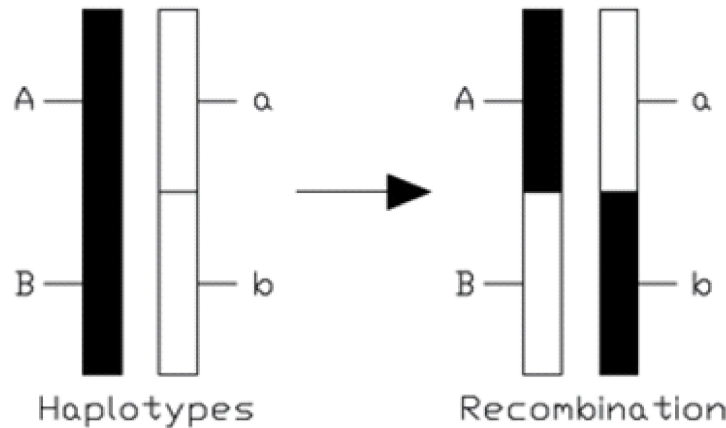
For two loci on **different chromosomes**, or very far apart on the same chromosome, segregation is effectively independent (Mendel’s Second Law, Chapter 2), and $\theta = 0.5$ (recombinant and non-recombinant gametes equally likely). For two loci **physically close together** on the same chromosome, crossing-over between them is rare, and $\theta < 0.5$; as physical distance between loci decreases toward zero, $\theta \rightarrow 0$. Two loci are said to be **linked** if $\theta < 0.5$.

Two Loci on the same chromosomes



Two loci, A/a and B/b, on the same pair of chromosomes: gametes formed without crossing-over between the loci are non-recombinant, each with probability $(1-\theta)/2$, while gametes formed with a crossover are recombinant, each with probability $\theta/2$.

8.3: The Double Backcross Design



Haplotypes before and after recombination between two linked loci, A/a and B/b .

The classical experimental (and, in humans, observational-family) design for estimating θ is the **double backcross**: a parent who is doubly heterozygous at two loci (genotype $AaBb$, with the two “high” alleles A and B in coupling on one chromosome and a , b on the homologous chromosome) is “backcrossed” to (i.e., mates with) a doubly homozygous recessive individual ($aabb$). Because the $aabb$ parent contributes only recessive alleles a and b , the offspring’s phenotype directly reveals which gamete (AB , ab , Ab , or aB) was transmitted by the doubly heterozygous parent:

- **Non-recombinant (parental) gametes:** AB and ab , each with probability $(1-\theta)/2$.
- **Recombinant gametes:** Ab and aB , each with probability $\theta/2$.

8.4: Maximum Likelihood Estimation of θ

Let n_R denote the observed number of recombinant offspring and n_{NR} the number of non-recombinant offspring, out of a total of $n = n_R + n_{NR}$ informative meioses. Since each offspring is recombinant with probability θ and non-recombinant with probability $(1-\theta)$, independently, the likelihood is binomial in form:

$$L(\theta) = \theta^{n_R} (1 - \theta)^{n_{NR}}$$

Taking logs and differentiating with respect to θ , setting the derivative to zero, gives the maximum likelihood estimator:

$$\hat{\theta} = n_R / (n_R + n_{NR}) = n_R / n$$

— simply the observed proportion of recombinant offspring, the intuitive “gene-counting” estimator. Its large-sample variance (from the information in the binomial likelihood, Chapter 10) is:

$$\text{Var}(\hat{\theta}) \approx \hat{\theta}(1 - \hat{\theta})/n$$

Theorem 8.1 (MLE of the Recombination Fraction). *For n independent, informative meioses, of which n_R are observed recombinant and $n_{NR} = n - n_R$ non-recombinant, the maximum likelihood estimator of the recombination fraction θ is $\hat{\theta} = n_R/n$, and this estimator achieves the Cramér–Rao lower bound asymptotically, with large-sample variance $\text{Var}(\hat{\theta}) = \theta(1 - \theta)/n$.*

Proof. The log-likelihood is $\ell(\theta) = n_R \log \theta + n_{NR} \log(1 - \theta)$. Differentiating with respect to θ gives the score function

$$U(\theta) = \frac{d\ell}{d\theta} = \frac{n_R}{\theta} - \frac{n_{NR}}{1 - \theta}.$$

Setting $U(\hat{\theta}) = 0$ and solving: $\frac{n_R}{\hat{\theta}} = \frac{n_{NR}}{1 - \hat{\theta}} \Rightarrow n_R(1 - \hat{\theta}) =$

n_R

n_{NR}

$n_{NR}\hat{\theta} \Rightarrow n_R = \hat{\theta}(n_R + n_{NR}) = \hat{\theta}n \Rightarrow \hat{\theta} = n_R/n$, confirming the stated MLE (and the second derivative of ℓ is everywhere negative for $\theta \in (0,1)$, confirming this is indeed a maximum, not a minimum). The Fisher information is $I(\theta) = -E\left[\frac{d^2\ell}{d\theta^2}\right] = E\left[\frac{n_R}{\theta^2} + \frac{n_{NR}}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}$ (using $E[n_R] = n\theta$, $E[n_{NR}] = n(1-\theta)$), so by the general asymptotic theory of maximum likelihood estimation (Chapter 10, Theorem 10.1), $\text{Var}(\hat{\theta}) \approx 1/I(\theta) = \theta(1-\theta)/n$. ■

Worked Example 8.2 (Recombination fraction estimation for a Nigerian consanguineous family study)

A Nigerian clinical genetics unit studies recombination between two closely linked markers flanking a rare autosomal disease locus in an extended family from a community with a tradition of consanguineous marriage (raising the informativeness of certain matings for linkage studies, Chapter 6). Genotyping 45 informative meioses across the extended pedigree reveals 6 recombinant and 39 non-recombinant gametes. By Theorem 8.1, $\hat{\theta} = 6/45 = 0.133$, with estimated standard error

$$\sqrt{\hat{\theta}(1-\hat{\theta})/n} = \sqrt{0.133 \times 0.867/45} \approx 0.0507, \text{ giving an}$$

approximate 95% confidence interval of $0.133 \pm 1.96(0.0507) = (0.034, 0.233)$ — a fairly wide interval reflecting the modest sample size typical of single extended-family linkage studies, illustrating why, historically, human linkage studies pooled data across as many informative families as possible (Section 8.5's LOD-score additivity) rather than relying on any single family's data alone.

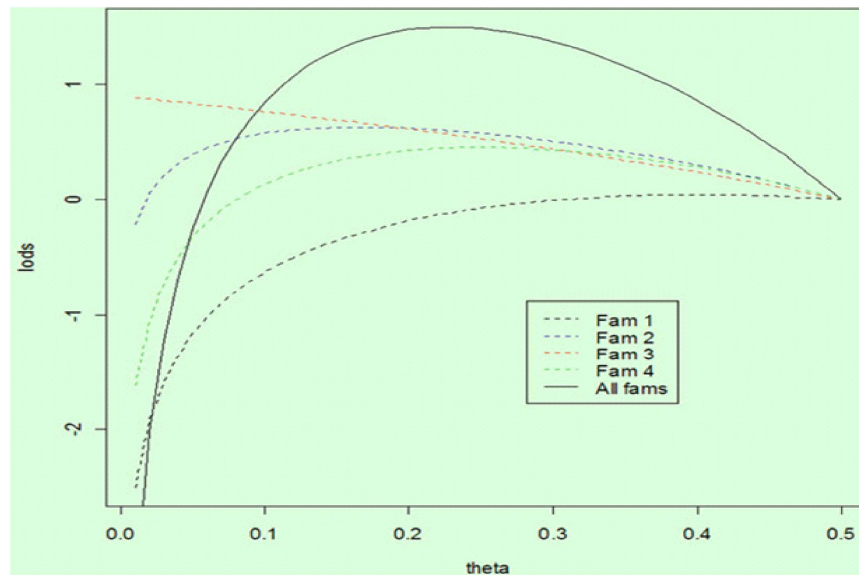
8.5: The LOD Score

Because we wish to formally test $H_0: \theta = 0.5$ (no linkage) against $H_1: \theta < 0.5$ (linkage), and because likelihoods for genetic linkage are conventionally reported on a base-10 logarithmic scale, the **LOD score** (logarithm of the odds) at a trial value of θ is defined as:

$$Z(\theta) = \log_{10} \left[\frac{L(\theta)}{L(0.5)} \right] = \log_{10} \left[\frac{\theta^{n_R} (1 - \theta)^{n_{NR}}}{0.5^{n_R + n_{NR}}} \right]$$

The **maximum LOD score**, $Z(\hat{\theta})$, is obtained by substituting the MLE $\hat{\theta}$ from Section 8.4. By long-standing convention in human genetics (Morton, 1955), a maximum LOD score $\mathbf{Z} \geq 3$ (odds of 1000:1 or more in favour of linkage) is taken as significant evidence of linkage between the two loci; $\mathbf{Z} \leq -2$ is taken as significant evidence *against* linkage (i.e., evidence that the loci assort independently); values in between are inconclusive. (This threshold reflects the low prior probability of any two randomly chosen loci being linked, so a stringent evidential bar is required — a Bayesian, not merely frequentist, rationale, developed further in Chapter 17.)

Combining evidence across families: Because LOD scores are logarithms of likelihood ratios, and independent families contribute independent likelihood factors, LOD scores are simply **additive** across families at any fixed value of θ :



LOD score curves, as a function of the trial recombination fraction θ , computed separately for four independently ascertained families and for their combined (“All fams”) analysis; combining families sharpens the peak and increases the maximum LOD score.

This additivity is one of the most practically important properties of the LOD-score framework: a single small family’s data may give a maximum LOD score well below the significance threshold of 3, yet pooling several such families — simply by adding their individual LOD curves at each θ — can push the combined maximum LOD score decisively past the threshold, exactly as illustrated above and as revisited in the multi-family linkage analysis of Chapter 17.

8.6: Worked Example 8.1

A Nigerian family-linkage study genotypes two closely spaced markers flanking a disease locus in a set of double-backcross offspring, observing $n_R = 3$ recombinants and $n_{NR} = 17$ non-recombinants ($n = 20$).

(a) Find θ .

$$\hat{\theta} = 3/20 = 0.15$$

(b) Find the standard error of θ .

$$SE(\hat{\theta}) = \sqrt{0.15(0.85)/20} = \sqrt{0.1275/20} = \sqrt{0.006375} = 0.0798$$

(c) Compute the maximum LOD score, $Z(\theta)$.

$$Z(0.15) = \log_{10}[(0.15)^3(0.85)^{17}/(0.5)^{20}]$$

$(0.15)^3 = 0.003375$. $(0.85)^{17} \approx 0.0631$ (since $\ln 0.85 = -0.1625$, $\times 17 = -2.763$, $e^{(-2.763)} \approx 0.0631$). $(0.5)^{20} \approx 9.537 \times 10^{-7}$.

Numerator = $0.003375 \times 0.0631 \approx 2.130 \times 10^{-4}$. Ratio = $2.130 \times 10^{-4} / 9.537 \times 10^{-7} \approx 223.3$.

$$Z(0.15) = \log_{10}(223.3) \approx 2.35.$$

(d) **Interpret.** A LOD score of 2.35 falls short of the conventional $Z \geq 3$ threshold for significant linkage, though it is well above zero and represents suggestive evidence of linkage (odds of roughly 223:1 in favour of $\theta = 0.15$ over independent assortment). In practice, the study would typically be extended (more families genotyped, or additional flanking markers added) before a definitive linkage claim could be made.

Theorem 8.2: (Consistency and Asymptotic Normality of the Recombination Fraction MLE)

Let $\hat{\theta}$ be the maximum likelihood estimator of the recombination fraction θ obtained from n independent backcross offspring, each classified as recombinant or non-recombinant (Section 8.4). Then $\hat{\theta}$ is a consistent estimator of the true θ , and $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a $\text{Normal}(0, I(\theta)^{-1})$ random variable, where $I(\theta)$ is the per-observation Fisher information.

Proof. Each backcross offspring's recombinant/non-recombinant status is a Bernoulli(θ) trial, and the number of recombinants R among n offspring is $\text{Binomial}(n, \theta)$, so the log-likelihood is exactly the standard binomial log-likelihood in θ . This is a regular one-parameter exponential family model satisfying the standard regularity conditions (the parameter space $(0, 0.5)$ is open, the likelihood is twice differentiable, and the Fisher information is finite and positive for θ in the interior), so the general asymptotic theory of maximum likelihood estimation established in Chapter 10 (Section 10.5) applies directly: $\hat{\theta} = R/n$ is consistent by the Law of Large Numbers applied to the sample proportion, and its asymptotic normality with variance $I(\theta)^{-1} = \theta(1-\theta)/n$ follows immediately from the general MLE asymptotic normality theorem, since the binomial Fisher information for θ is $n/[\theta(1-\theta)]$. This special case previews the general result proved formally in Chapter 10.

8.7: Chapter Summary

- Recombination fraction, θ , is the probability that a gamete from a doubly heterozygous parent is recombinant at two loci; $\theta = 0.5$ for unlinked loci, $\theta < 0.5$ for linked loci.
- Under the double backcross design, θ is estimated by simple proportion: $\hat{\theta} = n_R / n$, the maximum likelihood estimator of a binomial parameter.

The LOD score, $Z(\theta) = \log_{10}[L(\theta)/L(0.5)]$, is the standard statistic for testing and summarising evidence of linkage; $Z \geq 3$ is conventionally significant, $Z \leq -2$ excludes linkage.

8.8: Computational Corner: MLE of the Recombination Fraction and the LOD Score

R.

MLE of the recombination fraction, theta, from double-backcross data (Section 9.4)

```
n_parental <- 68; n_recombinant <- 32; n_total <-  
n_parental + n_recombinant
```

```
loglik_theta <- function(theta, r, n) r * log(theta) + (n - r)  
  * log(1 - theta)  
theta_hat <- n_recombinant / n_total  
theta_grid <- seq(0.01, 0.49, by = 0.001)  
ll <- sapply(theta_grid, loglik_theta, r = n_recombinant, n =  
  n_total)  
lod <- (ll - loglik_theta(0.5, n_recombinant, n_total)) / log(10)  
  
plot(theta_grid, lod, type = "l", xlab = expression(theta), ylab =  
  "LOD score")  
abline(v = theta_hat, col = "red", lty = 2)  
cat("theta_hat =", theta_hat, " max LOD =", max(lod), "\n")
```

Python.

```

import numpy as np

n_parental, n_recombinant = 68, 32
n_total = n_parental + n_recombinant

def loglik_theta(theta, r, n):
    return r * np.log(theta) + (n - r) * np.log(1 - theta)

theta_hat = n_recombinant / n_total
theta_grid = np.arange(0.01, 0.5, 0.001)
ll = loglik_theta(theta_grid, n_recombinant, n_total)
lod = (ll - loglik_theta(0.5, n_recombinant, n_total)) /
np.log(10)
print(theta_hat, lod.max())

```

Both scripts recover $\theta = r/n$ exactly as derived analytically in Section 9.4, and plot the LOD -score curve of Section 9.5, showing that the maximum LOD occurs precisely at θ — the computational equivalent of setting the score function to zero.

Mini-Project**Mini-Project 8: Estimating Recombination Fraction and Building a LOD Score Curve**

Simulate a double backcross dataset and reproduce the full recombination-fraction estimation and LOD score workflow of this chapter.

Step 1: Simulate double backcross offspring data at a true recombination fraction of your choosing (e.g., $\theta = 0.15$).

Step 2: Compute the maximum likelihood estimate of θ from your simulated data using Section 8.4's method.

Step 3: Compute the LOD score across a grid of candidate theta values from 0 to 0.5.

Step 4: Plot the LOD score curve and identify the theta value at which it peaks.

Step 5: Determine whether your peak LOD score would be considered significant evidence for linkage under the conventional threshold (Section 8.5).

Deliverable: *Simulation and estimation code, a LOD score curve plot, and a one-paragraph interpretation of significance.*

8.9: Practice Questions

- Q1.** A double backcross study of 50 informative offspring finds 8 recombinants. Compute θ and its standard error.
- Q2.** Using the data in Q1, compute the maximum LOD score, $Z(\theta)$, and state whether the evidence for linkage is conventionally significant.
- Q3.** Explain, intuitively, why $\theta = 0.5$ corresponds to “no linkage” while $\theta = 0$ corresponds to “complete linkage” (loci always co-transmitted).
- Q4.** A researcher reports a LOD score of -3 at $\theta = 0.5$ test point for two markers. What does this tell you about linkage between the two loci?
- Q5.** Two small families each give a maximum LOD score of 1.6 at $\theta = 0.15$. If their likelihood contributions are independent, what is their combined LOD score at $\theta = 0.15$, and does this combined value cross the conventional significance threshold?
- Q6.** Using the R or Python code of Section 8.8, verify numerically that $Z(\theta)$ computed from the pooled data of two families (obtained by summing their recombinant and non-recombinant counts) equals the sum of each family’s individually computed LOD score at the same θ .

