Typos and Comments

These comments on Ben Liu's (1998) *Statistical Genomics* are provided by Brian Yandell with assistance from David Butruille. Barry Pittendrigh, Diane Austin, Pablo Quijada and others in the Fall 1998 Statistical Genomics class have contributed as well. See
`www.stat.wisc.edu/~yandell/statgen/`
for further details.

ppp.ll:
page ppp, line ll (from bottom if negative)
ppp.rr.ll:
page ppp, paragraph rr, line ll (from bottom if negative)

1: Introduction

2, Table 1.2:
One might quibble with the use of "Classical Genomics" for the contents of this book. It is primarily "Statistical Genetics" it seems, with some attention to genomics issues. Also, some would argue that Sequence Comparison should be listed under "DNA Sequence Analysis", although it surely involves considerable "Genome Informatics".

3: Introduction to Genomics

52, Table 3.2:
There is a similar table in Kearsey and Pooni (1996, p. 177). It gives references on total genome size to Weaver and Hedrick (1989) and Arumaganathan and Earle (1991).

4: Statistics in Genomics

90, Table 4.2:
mean of binomial should be np
90.2.11:

$$mgf(X) = \sum_{x=0}^{\infty} e^{tx} \cdots$$

94.1:

$$L(\theta) = L(\theta|x) = PR\{X = x|\theta\}$$

94.2.6–:

$$L(\theta) = L(n,p) = \cdots = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

The logarithm is

$$\log\{l(\theta)\} = \log(\begin{array}{c} n \\ x \end{array}) + x\log(p) + (n-x)\log(1-p)$$

(and elsewhere there is confusion of θ and p)
95, Section 4.3.3:
Good to point out here that variance is the inverse of information: the higher the information, the lower the variance.
101, Equation (4.36), and 103.8:
Strictly speaking, the non-centrality parameter is *not* the expected value of the statistic, as the notation $E(G)$ implies. Rather, one substitutes the "true" parameter θ for the estimator $\hat{\theta}$ in the statistic to find the non-centrality parameter. Liu uses c for the non-centrality parameter on p. 103, but various authors use δ or δ^2. The mean and variance of a non-central chi-square with parameter c and degrees of freedom df are

$$E(\chi^2_{c,df}) = df + c \text{ and } V(\chi^2_{c,df}) = 2df + 4c \; .$$

For large df, the non-central chi-square is approximately normal. See for instance Yandell (1997) or Johnson, Kotz and Balakrishnan (1994).
111.8:
Replace 0.5 by $\frac{1}{2}$
119, Equation (4.67):

$$\hat{\sigma}^2_{\hat{\theta}} = 1/nI(\hat{\theta})$$

120-121:
There is some confusion here as Liu considers θ_i as random variables, while they were parameters estimated by $\hat{\theta}_i$ on the previous page. Double sums should have $\sum_{j=1}^{i-1}$.
122-123:
The random variables are the confidence limits T_1 and T_2. The parameter θ should not have a "hat" on it for (4.76) through (4.79).
125–128:
The rough 95% confidence interval guidelines of 2 units for log likelihoods and 1 for LOD scores are just that. More exactly, 2 units of log likelihood corresponds to 4 units of a χ^2_1 variate, which would yield a

95.45% confidence interval. Similarly, 1 unit of LOD score is $2\log(10) \approx 4.605$ units of a χ^2_1 variate, which would yield a 96.81% confidence interval. Notice that the 1 LOD intervals are larger than the 2 log likelihood intervals, since they have higher confidence of covering the true θ. The 'exact' 95% would be 1.921 units of log likelihood or 0.8341 units of LOD score. However keep the following in mind: (i) the χ^2 is only an approximation to the distribution; (ii) the degrees of freedom are not always 1 (depends on the number of parameter restrictions in the hypothesis – here $\theta = 0.5$).

126, Figure 4.6:
"likelihood function of Equation (4.83)"

134.2.:
Here and through later sections of the book, $L(\theta)$ now refers to the log likelihood, rather than the likelihood as developed in this chapter.

5: Single-Locus Models

140.2.:
The item n is used in two contexts here. Change all but the last n in this paragraph to N, the number of individuals, leaving n as the number of detectable genotypic classes.

141, Table 5.2:
Notice that the smallest p-value across the 10 plants is .24 (or .26 in Table 5.3). One would expect to have one plant with p-value less than .1 under the null hypothesis of 3:1 ratio. However, it is believed that Mendel stopped counting when he found more or less the ratio he expected. That is, these larger-than-expected p-values support the idea that Mendel shaded his data to support his theory.

143.4.:
Just a clarification. $(.5)^5$ is the probability that all 5 individuals are AA, and similarly for aa, in a backcross or doubled haploid cross. That is $2(.5)^5$ is the probability that a particular marker is *not* polymorphic for 5 individuals. When one considers codominant markers in F2, then a marker will be monomorphic if all individuals are the same, either AA (with prob. $(.25)^5$) or aa (ditto) or Aa (with prob. $(.5)^5$); add these three up to get .0332.

144, Figure 5.2:
The lower curve is for a test based on the null hypothesis of a 1:1 ratio when in fact the true ratio is 3:1.

145.1.:
Once again, the "parametric value" G_E is the non-centrality parameter discussed on p. 101. It is derived by substituting the expectation of the observed value. The non-centrality for a null hypothesized ratio of 1:1 when the true ratio is 3:1 is given in Equation (5.8). A similar form for the reverse situation (guess 3:1, true 1:1) is

$$2\{0.5n \log\left[\frac{0.5n}{0.75n}\right] + 0.5n \log\left[\frac{0.5n}{0.25n}\right]\}$$

which comes out to be $0.2877n$. The following table may help. Here the value of $n = 15$ is used as in the book.

null	true	non-centrality	power
1:1	3:1	$0.2616n = 3.924$	0.51
3:1	1:1	$0.2877n = 4.316$	0.55

These power values are plotted for n from 5 to 100 in Figure 5.2.

145, Equation (5.9):
This follows from a little algebra. Notice that the expected number of the first class is x : 1 ratio is $nx/(x+1)$, and for the second class is $n/(x+1)$. These add up to $n = a + b$. Set up the chi-square statistic as sum of (obs – exp)^2/exp with just two terms here, and simplify the algebra. Same exact calculations for the y : 1 ratio.

146, Equation (5.12):
The numerator inside the square brackets should be $x^{0.5} - y^{0.5}$. The way to develop this is to plug in the solutions for a and b from (5.10) into equation (5.9). Now set this equal to the α critical value of the chi-square distribution and solve for n.

148.7:
"and the covariance between" (drop the word "estimated")

148.13:
The variance simplifies as well to

$$\text{var}(\hat{p}_i) = p_i(1 - p_i)/2N$$

148.2.–149.1.:
The discussion on the bottom half of 148 through the top half of 149 concerning dominant markers is somewhat confusing. Consider first the case with only two alleles. Then $p_i + p_j = 1$ and $n_{i^\circ} = N - n_{jj}$. A little algebra with Equation (5.17) shows that $\hat{p}_i + \hat{p}_j = 1$, even though the form is more complicated. Further, the Equation (5.18) and the equation above it for a recessive allele also have this property. With only two alleles, there is no iteration needed via EM, and both of these give the same estimate.

Now on to more than two alleles, but where one allele (A_i) is dominant to the recessive allele A_j. In this case, $p_i + p_j < 1$ and one cannot observe the genotype A_{ij} distinct from A_{ii}. However, one can still observe all other genotypes distinctly (e.g. A_{ik}, $k \neq i, j$ and others). In this case, the EM method becomes important. The initial values might be taken from (5.18) and from the equation above it for the recessive allele. But the conditional probabilities in (5.16) are not quite right:

$$\begin{aligned} P[A_i|p_{i^\circ}] &= \frac{p_i^2+p_ip_j}{p_i^2+2p_ip_j} \\ P[A_j|p_{i^\circ}] &= \frac{p_ip_j}{p_i^2+2p_ip_j} \end{aligned}$$

The fallacy is that the conditioning is with respect to these two alleles, not all the alleles. Similarly the denominators in (5.17) must be changed from $[1-p_j^2]$ to $p_i^2 + 2p_ip_j$. I tried out the EM, and it converges in one step.

149.3.1:
"with a certain codominant allele"

150.3.:
The formula in Equation (5.23) and then rest of this page seem to be wrong.

151–154:
The discussion of allele frequencies which decay geometrically (or exponentially) is thin. Here $p_i = (1-\lambda)\lambda^{i-1}$ for $i = 1, \cdots$ has heterozygosity

$$H = 1 - \sum p_i^2 \approx 1 - \frac{(1-\lambda)^2}{1-\lambda^2} \ .$$

[The approximation ($\approx$) comes in since there are a finite number of alleles l. The sum of a geometric series $\sum_{i=0}^{\infty} x^i$ is $1/(1-x)$; here $x = \lambda^2$.] The effective number of alleles (see related discussion of effective number of alleles ch. 16, p. 483, and 19, p. 560–561) for a gene is $N_e = 1/(1-H)$, which in this case is

$$N_e = \frac{1-\lambda^2}{(1-\lambda)^2} = \frac{1+\lambda}{1-\lambda} \ .$$

This can be found, for instance, in Crow and Kimura (1970). The following relation between λ and N_e

λ	1/3	2/3=4/6	9/11	19/21	39/41
N_e	2	5	10	20	40

explains the choices of λ in tables.

160, Figure 5.6:
The power curves rise so quickly for small sample size because there is a high probability of picking up a new allele with each new individual.

6: Two-Locus: Controlled Crosses

165 & 167, Table 6.2 & Table 6.3:
Locus A column, cross 1, the test statistic should be .03 instead of .06

169.2:
which has $c-1$ degrees of freedom, will test $\cdots$

171, Equation (6.8):
Replace $\log L_\theta(x)$ by $L(\theta)$. Also, here and elsewhere Liu drops the extra set of brackets around the $[\cdots]^2$ term. The expectation is of this squared term, rather than the square of the expectation.

172, Equation (6.10):

$$+ f_5 \log(1 - 2\theta + 2\theta^2) + \cdots$$

This log likelihood and others ignore the fractions (e.g. $f_1 \log(0.25)$). This is OK, since they drop out in (6.11) when taking the derivative in terms of θ.

173.1.4:
Replace $\log L_\theta(x)$ by $L(\theta)$

173.2.1:
"Setting Equation (6.11)" $\cdots$

174.2:
Second term should be "20log" rather than "40log".

172, Table 6.8:
It would help to lay this out as well as a square table with AA, Aa, aa along the rows and BB, Bb, bb along the columns. Cell entries could be the expected frequency. Then show how this collapses to Table 6.10, p. 174, by summing over the first two rows and the first two columns.

185.2.5:
In the second line of SD for F2 population, the sample size is 200 rather than 400.

185.2.11:
The square bracket term should each be inverted.

$$\sqrt{\frac{1}{N_{F2}+N_{BC}}\left[\frac{2(3-4\hat\theta+2\hat\theta^2)}{\hat\theta(2-\hat\theta)(3-2\hat\theta+\hat\theta^2)} + \frac{1}{\hat\theta(1-\hat\theta)}\right]^{-1}}$$

186–187, Figures 6.3 & 6.4:
The y-axis is wrong (descending) in both of the leftmost (F2) panels. Confidence interval is based on maximum log likelihood + 2 (–101.03) instead of log likelihood – 2 as indicated in the F2 panel of Figure 6.3.

189, Equation (6.30):
Drop the N inside the square brackets.

189–190:

There may be some errors in the calculations for the G_{EO} non-centrality parameters.

190.2.:
The Power and Sample Size section is confusing, particularly the notation $[\chi^2_{..}]^{-1}$. First, there is a critical value $x = \chi_{\alpha;df}$ set by the hypothesis test to insure the test has size α. The investigator wants to ensure the test has some selected power γ. That implies that the non-centrality parameter G_E has to be "so big". A picture showing non-central chi-squares could be used to illustrate that as G_E increases, the probability $Pr[\chi_{G_E,df} \geq x]$ increases from α (when $G_E = 0$) to 1 (for large G_E). This material is first introduced on p. 131, sec. 4.7.

190.2.7–8:
G_E is the expected log likelihood. G_{EO} is G_E/N, the expected log likelihood divided by the sample size.

195.2.8–10:
The simplification has a mistake. f_4 should be replaced by $2f_4$ everywhere. This changes the solution, but later he considers $f_4 = 0$ so this drops out. Note that then $N = f_1 + f_2 + f_3$ (line 17 of paragraph).

196–197, Figure 6.8 & Tables 6.26, 6.27:
It appears that c and C are reversed in Figure 6.8, based on Table 6.26. It would help if all the C's (and c's) were changed to D's (and d's), as in Table 6.27, to prevent confusion with the next section, Table 6.28, where a new codominant marker C is created from TDLM's. Table 6.27 should say "dominant marker D".

198, Equation (6.36):
This is $I(r_2|r_1)$, the conditional information on r_2 given r_1.

201:
The additive distortion parameters a and b have slightly different meanings in the BC and F2 breeding systems. There is no direct relation between them. To examine this, collapse to just the A locus. In the BC case, one has Aa:aa and A:a being $(.5 + 2a) : (.5 - 2a)$, while in the F2 case, one has AA:Aa:aa as $(.25 + 3a^*) : .5 : (.25 - 3a^*)$, or on a gamete bases A:a is $(.5 + 1.5a^*) : (.5 - 1.5a^*)$. Note that there is no obvious mechanistic interpretation here, and one cannot take the gamete ratios for the BC and recover the genotype ratios for the F2.

201, Figure 6.10:
Does it make biological sense to have a and b be functions of θ (e.g. $a = b = .1\theta$ or $.2\theta$)?

7: Two-Locus: Natural Populations

217.2.–2:

$$Pr[\chi^2_{df=1} \geq G_C] > \alpha$$
and
$$Pr[\chi^2_{df=1} \geq G_R] > \alpha$$

219, Equation 7.3:
Here again is the likelihood, using $L(\theta)$ notation. Should not have a "hat" in $L_{BC\overline{PUK}}(\theta)$.

219.3.2:
Reference for Nordheim *et al.* (1984) is missing from references. This paragraph on the Bayes procedure could be better developed for readability.

219, Equation 7.4:
Should have parentheses and no "hat" (appears twice): $L_{BC\overline{PUK}}(\theta)d\theta$

221, Table 7.5:
Mixtures of recombinants (every other term) should have expected frequencies $0.25(1 - 2\theta + 2\theta^2)$. This also changes Equation (7.5) to

$$L_{MP}(\theta) = (f_1 + f_3 + f_5 + f_7 + f_9)[\log(\theta) + \log(1 - \theta)] + (f_2 + f_4 + f_6 + f_8)\log(1 - 2\theta + 2\theta^2)$$

and changes the remainder of this subsection. The Solution is much more complicated, requiring the solution of a cubic polynomial in θ.

222, Table 7.6:
Outcrossed progeny for AaBB has expected frequency $0.5v[(1 - u)(1 - \theta) + u\theta]$. Selfed progeny for AaBb has expected frequency $0.25(1 - 2\theta + 2\theta^2)$ as in Table 6.7, p. 172. It would help to have a table similar to Table 7.4, such as

	Maternal Gamete	
Pollen Gamete	AB	Ab
	$0.5(1-\theta)$	0.5θ
AB	AABB	AABb
uv	$0.5uv(1-\theta)$	$0.5uv\theta$
Ab	AABB	AABb
$u(1-v)$	$0.5u(1-v)(1-\theta)$	$0.5u(1-v)\theta$
aB	AABB	AABb
$(1-u)v$	$0.5(1-u)v(1-\theta)$	$0.5(1-u)v\theta$
ab	AABB	AABb
$(1-u)(1-v)$	$0.5(1-u)(1-v)(1-\theta)$	$0.5(1-u)(1-v)\theta$

	Maternal Gamete	
Pollen Gamete	aB	ab
	0.5θ	$0.5(1-\theta)$
AB	AaBB	AaBb
uv	$0.5uv\theta$	$0.5uv(1-\theta)$
Ab	AaBB	AaBb
$u(1-v)$	$0.5u(1-v)\theta$	$0.5u(1-v)(1-\theta)$
aB	AaBB	AaBb
$(1-u)v$	$0.5(1-u)v\theta$	$0.5(1-u)v(1-\theta)$
ab	AaBB	AaBb
$(1-u)(1-v)$	$0.5(1-u)(1-v)\theta$	$0.5(1-u)(1-v)(1-\theta)$

This material is used in Table 7.6 and Table 7.11.

222.4.2:
"A anb B are linked with a recombination of θ"

223, Table 7.8:
"frequency of u"
224–226:
In a linkage analysis experiment as conducted nowadays, one usually plans to take information on many loci for any given plant (or other type of organism) and use these plants or there progenies to make QTL analysis. In such case the population parameter t is not as critical as knowing which plant is actually the result of a self or the result of an outcross (using information about all the marker loci, inference should be possible in that regard). This is very important because the selfed plant will be suffering from inbreeding depression and it would be a serious confounding factor in the QTL analysis if not taken into account.
225, Table 7.10:
Average Information is correct, but estimates are not.

Maternal	$\hat{t}$
AA	$n_{AA}/[(1-u)(n_{AA}+n_{Aa})]$
Aa	$(n_{aa}-n_{AA})/[(n_{AA}+n_{aa})(1-2u)]$
aa	$n_{Aa}/[u(n_{aa}+n_{Aa})]$

225, Figure 7.2:
This is (absolute) average information from Table 7.10, but vertical axis is mislabeled. In left and right figures, the maximum information for homozygotes (AA and aa) is $1/.09 \approx 11$, while in the middle panel, the maximum is $1/.25 = 4$. It does not make sense to consider relative information here, since homozygotes and heterozygotes (Aa) can have zero information.
226, Table 7.11, & 227.1.8–9:
"$p_i = p_{oi}$ is the expected frequency of genotype i for the outcrossed progeny in Table 7.6." Notice that conditional probabilities for Maternal Plant in outcross complement those of Pollen Plant: if genotype is AABb and Pollen gamete is AB, then Maternal gamete must be Ab. I would propose redefining $p_{Mi\overline{AB}}$ to refer only to selfing. Thus the conditional probabilities for recombinants among selfed progeny given the genotype (numbered as in Table 7.6 and 7.11) would be $p_{M3Ab} = p_{M7aB} = 1.0$, $p_{M2Ab} = p_{M4aB} = p_{M6Ab} = p_{M8aB} = \theta$, $p_{M5Ab} = p_{M5aB} = 0.5\theta$ and all other $p_{MiAb} = p_{MiaB} = 0.0$. With this redefinition, Equation (7.15), page 227, seems correct.
227.2.–2:
Numerical Recipes is a classic, but there might be more recent sources. Online software can be found at NetLib (`netlib.bell-labs.com/netlib`) or StatLib (`lib.stat.cmu.edu`).
227:
Put 'hat's over the initial estimates, as $\hat{\theta}^0$ and similarly for u and v at bottom of page.
228, Equation (7.16):
"The maximum likelihood estimates of u, v and θ given the current estimates of probabilities $p_{Pi\overline{AB}}$ and $p_{Mi\overline{AB}}$ are, respectively," $\cdots$
227–229:
The idea in the iterations (say for Method II using EM) is to use the current estimates of u, v and θ to compute the expectations (E step), and then use the probabilities from the E step in the M step to update the estimates. It would help to note that after each iteration, the '0' values are replaced by the '1' values, e.g. replace $\hat{\theta}^0$ by $\hat{\theta}^1$, etc. This could be made explicit, for instance, in the last sentence of step [4] after Equation (7.16).

8: Two-Locus: Linkage Disequilibrium

246, Table 8.1:
The notation f_i is used elsewhere for the (observed) frequencies. That is, $E(f_1) = p_{AB}^2$ and so on.
247, Equation (8.12):
The notation switch is confusing. The 'hat's seem to indicate that one is using the initial frequencies f_i to estimate gamete frequencies, but this is not used consistently. The *'s seem to refer to the next generation. Why not use superscripts as is done with D_{AB}^0?
252–254:
It would help to see the connection between Weir's approach and this approach to TDD. Otherwise, this appears quite nice and clean.

9: Locus Ordering

276–278:
There is potential for confusion with the recombination parameters θ_1 and θ_2 and with the observed counts f_i, $i = 1, \cdots, 8$. The text talks of θ_1 and θ_2 as the recombination rates between the first and second ordered markers and between the second and third ordered markers, respectively, using the example of the order ABC. Thus $\theta_1 = \theta_{ab}$ and $\theta_2 = \theta_{bc}$ for this ordering. However, with a different ordering, they correspond to different pairs of markers. The consequence is that the expected frequencies in Table 9.2 are correct *only* if the ordering is ABC. This might explain errors in Table 9.5 noted below. The author

does make this explicit, and does define the log likelihood in Equation (9.1) in terms of L_{ABC}. However, it would be nice to see the θ_1 and θ_2 dropped, and to see the log likelihoods for the other two orders.

277, Equation (9.4):
Drop 2 from numerator in first equation.

277, Equation (9.5):
Last term in numerator should be $\hat{\theta}_{ab}^2\hat{\theta}_{bc}^2 Var(\hat{\theta}_{ac})$. This whole expression is derived from a Taylor expansion: $V(f(y)) \approx [f'(y)]^2 V(y)$ (cf. Yandell (1997, sec. 15.1). It is a little more complicated because there are three random variates, the $\hat{\theta}$s.

278, Equation (9.7):

$$L_{ABC}(\theta_1, \theta_2, 0) = \cdots$$

This is defined and valid only if $f_3 + f_6 = 0$: no double crossovers in the data.

278, Table 9.3:
Switch the 5 and 11 in the 2nd and 3rd row of the last column.

278, Table 9.4:
Second column, 2nd and 3rd rows, respectively, should be $f_2 + f_6$ and $f_3 + f_7$.

279, Table 9.5:
Drop the row for $C = 0$, since this is impossible for any order (all would have at least one double crossover). Entry for $C = 1$ and BAC order should be –85.20.

280.4.–3:

$$\frac{Pr[\text{BAC}]}{Pr[\text{ABC}]} = \exp[-85.20 - (-93.60)] = 4447$$

282, Equation (9.10):
$R = \{a_1, a_2, \cdots, a_l\}$ is not defined. The θs should be replaced with $\hat{\theta}$s. In other words, $n_{ij}\hat{\theta}_{ij}$ is the number of recombinants between loci i and j, and $n_{ij}(1-\hat{\theta}_{ij})$ is the number of non-recombinants.

283, Equation (9.12):

$$PARF = \cdots$$

283, Equation (9.16):

$$SALOD^{EN} = \sum_{i=1}^{l-1} z_{a_i a_j} EN_{a_i a_j}$$

283, Equation (9.17):

$$EN = [\log_{10}(2) + \hat{\theta}\log_{10}(\hat{\theta}) + (1-\hat{\theta})\log_{10}(1-\hat{\theta})]^{-1}$$

284–289:
The author missed an opportunity to refer to the DNA solution of the travelling salesman problem (Adelman 1994). Further, this could readily replace the abstract discussion of TSP. The basic idea is that there are 7 cities connected by certain bridges and roads. What order can a travelling salesman go to visit all seven cities exactly once, without returning to any. Adelman, a computer scientist, solves this with 20-mers of DNA and PCR technology. (He actually spent a semester in a biology lab to learn techniques.)

287.4.11:
There is some error in this equation, as c is undefined.

289, Table 9.7:
Either the first frequency should be 5 or the estimate in the paragraph below should have 7×0.5. In any event, the sum of terms is not 152: it is 154.5 or 155.5.

290.=2.–3:
"is G', the likelihood ratio $\cdots$"

290.2.–1 & 290.4.5:
"orders BAC, ABC and ACB is 1 : $2.25 \times 10^{-}4$: $3.88 \times 10^{-}13$" (that is, BAC:ABC is 4447:1 or $1{:}2.25 \times 10^{-}4$)

292, Table 9.8:
"2-R = $\cdots$ the previous adjacent locus"

10: Multi-Locus Models

325–327:
Evans *et al.* (1993) showed that most commonly used map functions are not "multi-locus feasible", extending the work of Karlin.

References

LM Adelman (1994) "Molecular computation of solutions to combinatorial problems", *Science 266*, 1021–1023.

K Arumaganathan and ED Earle (1991) "Nuclear DNA content of some important plant species", *Plant Mol Biol Rep 9(3)* 208–18.

SN Evans, MS McPeek and TP Speed (1993) "Crossover models and map functions", *Theoretical Population Biology 43*, 80–90.

NL Johnson, S Kotz and N Balakrishnan (1994) *Continuous Univariate Distributions*, v 1, 2nd ed. Wiley-Interscience.

MJ Kearsey and HS Pooni (1996) *The Genetical Analysis of Quantitative Traits*. Chapman & Hall.

EV Nordheim, DM O'Malley and S-C Chow (1984) "On the performance of a likelihood ratio test for genetic linkage", *Biometrics 40*, 785–790.

RF Weaver and PW Hedrick (1989) *Genetics*. WC Brown.

BS Yandell (1997) *Practical Data Analysis for Designed Experiments*, Chapman & Hall / CRC Press.