Chapter 11
Mapping Segregation Distortion Loci

A basic assumption in QTL mapping is that genomic loci (QTL and markers) follow the Mendelian segregation ratio. The Mendelian ratio depends on the population under investigation. For example, in a BC population, the Mendelian ratio is $1 : 1$ for the two genotypes (A_1A_1, A_1A_2). In an $\mathtt{F}_2$ population, the Mendelian ratio is $1 : 2 : 1$ for the three genotypes (A_1A_1, A_1A_2, and A_2A_2). If the segregation ratio of a locus deviates from the Mendelian segregation ratio, we say that the locus is a non-Mendelian locus or segregation distortion locus (SDL). In fact, a marker whose segregation deviates from the Mendelian ratio is not necessarily an SDL. It is most likely that a true SDL sits nearby the marker and the observed segregation distortion of the marker is caused by the SDL because of linkage. Sometimes we may see markers in several regions of the genome that show segregation distortion. This may be caused by several SDL across the genome. The SDL themselves may be caused by viability selection. In other words, different genotypes of the SDL may have different viabilities. Genotypes that are favored by the viability selection are overrepresented, while genotypes that are against by the viability selection are underrepresented. Therefore, an SDL may also be called viability locus (VL). Viability selection may happen in the gametic level or zygotic level or both. But it is hard to tell the difference between gametic selection and zygotic selection unless we can directly observe the gametes. Like quantitative trait loci, segregation distortion loci can be mapped using marker information. Evolutionary biologists may be more interested in SDL, while agricultural scientists may be more interested in QTL. In a single experiment of genetic mapping, we may simultaneously investigate both SDL and QTL.

The earliest work in SDL mapping was Fu and Ritland (1994). For the first time, the authors proposed the viability selection hypothesis and tried to map SDL using marker information under the maximum likelihood framework. Mitchell-Olds (1995) also developed a similar ML method to map SDL in an $\mathtt{F}_2$ population. A more systematic treatment of SDL mapping was made by Lorieux et al. (1995a,b) using genome-wide markers. Vogl and Xu (2000) took a Bayesian approach to mapping viability selection loci. Luo and Xu (2003) developed an EM algorithm to estimate the segregation ratio under the ML framework. Luo et al. (2005) eventually

S. Xu, *Principles of Statistical Genomics*, DOI 10.1007/978-0-387-70807-2_11,

developed a quantitative genetics model to estimate the genetic effects of viability loci using a four-way cross design. Some of the methods have been applied to SDL mapping in rice (Wang et al. 2005a).

This chapter will introduce methods for mapping SDL under two different models. One is called the probabilistic model (Luo and Xu 2003), and the other is called the liability model (Luo et al. 2005). Under some special situations, both models will generate the same result, but in most situations, the liability model is more efficient. In the last section, we will combine QTL mapping and SDL mapping together and jointly map QTL and SDL (Xu and Hu 2009).

11.1 Probabilistic Model

Consider an SDL with an arbitrary segregation ratio in an $\mathtt{F}_2$ family derived from the cross of two inbred lines. Let M and N be the left and right flanking markers bracketing the SDL (denoted by G for short). The interval of the genome carrying the three loci is denoted by a segment MGN. The three genotypes of the SDL are denoted by G_1G_1, G_1G_2, and G_2G_2, respectively. Similar notation also applies to the genotypes of the flanking markers. The interval defined by markers M and N is divided into two segments. Let r_1 and r_2 be the recombination fractions for segment MG and segment GN, respectively. The joint distribution of the marker genotypes conditional on the SDL genotype can be derived using the Markovian property under the assumption of no segregation interference between consecutive loci. Let us order the three genotypes, G_1G_1, G_1G_2, and G_2G_2, as genotypes 1, 2, and 3, respectively. If individual j takes the κth genotype for the SDL, we denote the event by $G_j = \kappa,\ \forall\kappa = 1,2,3$. The joint probability of the two markers conditional on the genotype of the SDL is

$$\begin{aligned}\Pr(M_j = \xi, N_j = \zeta | G_j = \kappa)\\ &= \Pr(M_j = \xi | G_j = \kappa)\Pr(N_j = \zeta | G_j = \kappa) \qquad (11.1)\end{aligned}$$

for all $\kappa, \xi, \zeta = 1,2,3$, where $\Pr(M_j = \xi | G_j = \kappa) = T_1(\kappa,\xi)$ and $\Pr(N_j = \zeta | G_j = \kappa) = T_2(\kappa,\zeta)$. We use $T_i(\kappa,\xi)$ to denote the κth row and the ξth column of the following transition matrix

$$T_i = \begin{bmatrix} (1-r_i)^2 & 2r_i(1-r_i) & r_i^2 \\ r_i(1-r_i) & (1-r_i)^2 + r_i^2 & r_i(1-r_i) \\ r_i^2 & 2r_i(1-r_i) & (1-r_i)^2 \end{bmatrix}, \forall i = 1,2 \qquad (11.2)$$

For example,

$$\begin{aligned}\Pr(M_j = 1, N_j = 2 | G_j = 3)\\ &= \Pr(M_j = 1 | G_j = 3)\Pr(N_j = 2 | G_j = 3) \qquad (11.3)\\ &= T_1(3,1)\ T_2(3,2) = 2r_1^2 r_2(1-r_2) \qquad (11.4)\end{aligned}$$

Let $\omega_\kappa = \Pr(G = \kappa), \forall\kappa = 1, 2, 3$, be the probability that a randomly sampled individual from the $\mathtt{F}_2$ family takes the κth genotype. Let $\omega = \{\omega_1, \omega_2, \omega_3\}$ be the array of the genotype frequencies, and it is the vector of parameters for estimation and test. Under Mendelian segregation, the three genotype frequencies are denoted by $\phi = \{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\}$. Therefore, the null hypothesis is that the $\mathtt{F}_2$ population is a Mendelian population, i.e., $\omega = \phi$. We use a generic notation p for probability, so that $p(G_j = \kappa)$ represents $\Pr(G_j = \kappa)$ and $p(M_j, N_j|G_j = \kappa)$ stands for $\Pr(M_j, N_j|G_j = \kappa)$. Given the parameters ω, the data (flanking marker genotypes), and the multinomial probability model, we are ready to construct the log likelihood function, which is

$$\begin{aligned} L(\omega) &= \sum_{j=1}^{n} \ln\left[\sum\nolimits_{\kappa=1}^{3} p(G_j = \kappa)p(M_j, N_j|G_j = \kappa)\right] \\ &= \sum_{j=1}^{n} \ln\left[\sum\nolimits_{\kappa=1}^{3} \omega_\kappa T_1(\kappa, M_j)T_2(\kappa, N_j)\right] \end{aligned} \tag{11.5}$$

where the parameters have a restriction $\sum_{\kappa=1}^{3} \omega_\kappa = 1$. Note that without any other information, $p(G_j = \kappa) = \omega_\kappa, \ \forall j = 1, \ldots, n$. Under the assumption of Mendelian segregation, $\omega = \phi$, i.e., $\omega_1 = \omega_3 = \frac{1}{2}\omega_2 = \frac{1}{4}$. However, we treat ω as unknown parameters. We postulate that deviation of ω from the Mendelian ratio will cause a marker linked to locus G to show distorted segregation. This likelihood function has been used by Luo et al. (2005) for mapping SDL.

11.1.1 The EM Algorithm

The MLE of the parameters can be solved via the EM algorithm (Dempster et al. 1977). We need to rewrite the likelihood function in a form of complete-data. Let us define a delta function as

$$\delta(G_j, \kappa) = \begin{cases} 1 & \text{if } G_j = \kappa \\ 0 & \text{if } G_j \neq \kappa \end{cases} \tag{11.6}$$

If the genotypes of the SDL are known for all individuals, i.e., given $\delta(G_j, \kappa)$ for all $j = 1, \ldots, n$ and $\kappa = 1, 2, 3$, the complete-data log likelihood is

$$L(\omega, \delta) = \sum_{j=1}^{n} \ln[p(M_j, N_j|G_j)p(G_j)] \tag{11.7}$$

where

$$\begin{aligned} p(M_j, N_j|G_j) &= \prod_{\kappa=1}^{3} p(M_j, N_j|G_j = \kappa)^{\delta(G_j,\kappa)} \\ &= \prod_{\kappa=1}^{3} [T_1(\kappa, M_j)T_2(\kappa, N_j)]^{\delta(G_j,\kappa)} \end{aligned} \tag{11.8}$$

and

$$p(G_j) = \prod_{\kappa=1}^{3} \omega_\kappa^{\delta(G_j,\kappa)} \tag{11.9}$$

Therefore, the complete-data log likelihood function can be rewritten as

$$L(\omega, \delta) = \sum_{j=1}^{n}\sum_{\kappa=1}^{3} \delta(G_j, \kappa)\{\ln[T_1(\kappa, M_j)] + \ln[T_2(\kappa, N_j)] + \ln(\omega_\kappa)\} \tag{11.10}$$

This log likelihood function involves missing value $\delta(G_j, \kappa)$ and thus cannot be used directly. We need to take expectation of this function with respect to $\delta(G_j, \kappa)$. In addition, we introduce a Lagrange multiplier to make sure that the parameters are estimated within their restriction, i.e., $\sum_{\kappa=1}^{3} \omega_\kappa = 1$. Therefore, the actual log likelihood function that is maximized in the EM algorithm is

$$\begin{aligned} E[L(\omega, \delta)] =& \sum_{j=1}^{n}\sum_{\kappa=1}^{3} E[\delta(G_j, \kappa)]\{\ln[T_1(\kappa, M_j)] + \ln[T_2(\kappa, N_j)] + \ln(\omega_\kappa)\} \\ &+ \lambda\left(1 - \sum\nolimits_{\kappa=1}^{3} \omega_\kappa\right) \end{aligned} \tag{11.11}$$

where λ is a Lagrange multiplier and is treated as a parameter for estimation. Before we maximize the above expected complete-data log likelihood function, we need to calculate $E[\delta(G_j, \kappa)]$, which is called the posterior expectation of the missing genotype and is calculated using Bayes' theorem,

$$E[\delta(G_j, \kappa)] = \frac{\omega_\kappa T_1(\kappa, M_j)T_2(\kappa, N_j)}{\sum_{\kappa'}^{3} \omega_{\kappa'} T_1(\kappa', M_j)T_2(\kappa', N_j)} \tag{11.12}$$

Note that this posterior expectation requires the value of parameter ω, which happens to be what we want to estimate. Therefore, iterations are required. Once an initial value of ω is provided, we can find the posterior expectation of the missing genotype, which further allows us to maximize the expected complete-data log likelihood function, (11.11). To maximize $E[L(\omega, \delta)]$, we take the partial derivatives of $E[L(\omega, \delta)]$ with respect to the parameters and equate the partial derivative to zero and solve for the parameters. The partial derivatives are

$$\frac{\partial}{\partial \omega_\kappa} E[L(\theta,\delta)] = \sum_{j=1}^{n} E[\delta(G_j,\kappa)]\frac{1}{\omega_\kappa} - \lambda, \forall \kappa = 1,2,3 \tag{11.13}$$

and

$$\frac{\partial}{\partial \lambda} E[L(\omega,\delta)] = 1 - \sum_{\kappa=1}^{3} \omega_\kappa \tag{11.14}$$

Setting (11.13) to zero, we get

$$\omega_\kappa = \frac{1}{\lambda}\sum_{j=1}^{n} E[\delta(G_j,\kappa)], \forall \kappa = 1,2,3 \tag{11.15}$$

Equation (11.14) is just the restriction, which allows us to solve for λ. Note that $\sum_{\kappa=1}^{3} E[\delta(G_j,\kappa)] = 1$, i.e., the sum of the three conditional probabilities is unity. This leads to

$$\sum_{\kappa=1}^{3} \omega_\kappa = \frac{1}{\lambda}\sum_{\kappa=1}^{3}\sum_{j=1}^{n} E[\delta(G_j,\kappa)] = \frac{1}{\lambda}\sum_{j=1}^{n}\sum_{\kappa=1}^{3} E[\delta(G_j,\kappa)] = \frac{n}{\lambda} = 1 \tag{11.16}$$

As a result, we get $\lambda = n$. Substituting $\lambda = n$ into (11.15) leads to

$$\omega_\kappa = \frac{1}{n}\sum_{j=1}^{n} E[\delta(G_j,\kappa)], \forall \kappa = 1,2,3 \tag{11.17}$$

The Lagrange multiplier λ is a nuisance parameter that allows us to find solution of ω_k in a convenient way. The EM algorithm is summarized as:

1. Initialize $\omega = \omega^{(0)}$.
2. Calculate $E[\delta(G_j,\kappa)]$ using (11.12) (the E-step).
3. Update ω using (11.17) (the M-step).
4. Repeat the E-step and the M-step until a certain criterion of convergence is satisfied.

11.1.2 Hypothesis Test

The null hypothesis is H_0: $\omega = \phi$. The alternative hypothesis is H_A: $\omega \neq \phi$. The likelihood ratio test statistic is used to test the null hypothesis. The likelihood ratio test statistic is

$$\mathrm{LRT} = -2[L_0(\phi) - L_1(\hat{\omega})] \tag{11.18}$$

where

$$L_1(\hat{\omega}) = \sum_{j=1}^{n} \ln\left[\sum\nolimits_{\kappa=1}^{3} \hat{\omega}_\kappa T_1(\kappa, M_j) T_2(\kappa, N_j)\right] \tag{11.19}$$

is the observed log likelihood function evaluated at $\omega = \hat{\omega}$ and

$$L_0(\phi) = \sum_{j=1}^{n} \ln\left[\sum\nolimits_{\kappa=1}^{3} \phi_\kappa T_1(\kappa, M_j) T_2(\kappa, N_j)\right] \tag{11.20}$$

is the log likelihood function evaluated at $\omega = \phi$. Under the null hypothesis, LRT will follow a chi-square distribution with 2 degrees of freedom. The reason for the 2 degrees of freedom is that we only have two (not three) independent parameters to estimate.

11.1.3 Variance Matrix of the Estimated Parameters

There are three parameters in vector $\omega = \{\omega_1, \omega_2, \omega_3\}$, but only two are independent because of the restriction $\sum_{\kappa=1}^{3} \omega_\kappa = 1$. Therefore, we only need to find the variance matrix of two components. Let us express $\omega_3 = 1 - \omega_1 - \omega_2$ so that

$$\text{var}(\hat{\omega}_3) = \text{var}(\hat{\omega}_1) + \text{var}(\hat{\omega}_2) + 2\text{cov}(\hat{\omega}_1, \hat{\omega}_2) \tag{11.21}$$

$$\text{var}(\hat{\omega}_1, \hat{\omega}_3) = \text{cov}(\hat{\omega}_1, 1 - \hat{\omega}_1 - \hat{\omega}_2) = -\text{var}(\hat{\omega}_1) - \text{cov}(\hat{\omega}_1, \hat{\omega}_2) \tag{11.22}$$

and

$$\text{var}(\hat{\omega}_2, \hat{\omega}_3) = \text{cov}(\hat{\omega}_2, 1 - \hat{\omega}_1 - \hat{\omega}_2) = -\text{var}(\hat{\omega}_2) - \text{cov}(\hat{\omega}_1, \hat{\omega}_2) \tag{11.23}$$

We now redefine $\omega = \{\hat{\omega}_1, \hat{\omega}_2\}$ as a vector with two components only. Therefore, we only need to derive the variance–covariance matrix for vector $\omega = \{\hat{\omega}_1, \hat{\omega}_2\}$ because the variance for $\hat{\omega}_3$ and the covariances involving $\hat{\omega}_3$ are all functions of $\text{var}(\omega)$. Let

$$L_j(\omega, \delta) = \sum_{\kappa=1}^{3} \delta(G_j, \kappa)\{\ln[T_1(\kappa, M_j)] + \ln[T_2(\kappa, N_j)] + \ln(\omega_\kappa)\} \tag{11.24}$$

be the complete-data log likelihood function for individual j so that $L(\omega, \delta) = \sum_{j=1}^{n} L_j(\omega, \delta)$. Note that whenever ω_3 occurs, it is replaced by $\omega_3 = 1 - \omega_1 - \omega_2$. The Louis (1982) information matrix is

$$I(\hat{\omega}) = -E[H(\hat{\omega}, \delta)] - \text{var}[S(\hat{\omega}, \delta)] \tag{11.25}$$

where

$$E[H(\hat{\omega},\delta)] = \begin{bmatrix} \sum_{j=1}^{n} E\left(\frac{\partial^2 L_j(\omega,\delta)}{\partial\omega_1^2}\right) & \sum_{j=1}^{n} E\left(\frac{\partial^2 L_j(\omega,\delta)}{\partial\omega_1\partial\omega_2}\right) \\ \sum_{j=1}^{n} E\left(\frac{\partial^2 L_j(\omega,\delta)}{\partial\omega_1\partial\omega_2}\right) & \sum_{j=1}^{n} E\left(\frac{\partial^2 L_j(\omega,\delta)}{\partial\omega_2^2}\right) \end{bmatrix} \tag{11.26}$$

is the expectation of the Hessian matrix of the complete-data log likelihood function and

$$\mathrm{var}[S(\hat{\omega},\delta)] = \begin{bmatrix} \sum_{j=1}^{n} \mathrm{var}\left(\frac{\partial L_j(\omega,\delta)}{\partial\omega_1}\right) & \sum_{j=1}^{n} \mathrm{cov}\left(\frac{\partial L_j(\omega,\delta)}{\partial\omega_1}, \frac{\partial L_j(\omega,\delta)}{\partial\omega_2}\right) \\ \sum_{j=1}^{n} \mathrm{cov}\left(\frac{\partial L_j(\omega,\delta)}{\partial\omega_1}, \frac{\partial L_j(\omega,\delta)}{\partial\omega_2}\right) & \sum_{j=1}^{n} \mathrm{var}\left(\frac{\partial L_j(\omega,\delta)}{\partial\omega_2}\right) \end{bmatrix} \tag{11.27}$$

is the variance–covariance matrix of the score vector of the complete-data log likelihood function. Both the expectation and the variance are taken with respect to the missing value $\delta(G_j,\kappa)$ using the posterior distribution of the genotype of the SDL. The inverse of the information matrix is used as an approximation of the variance matrix of $\hat{\omega} = \{\hat{\omega}_1,\hat{\omega}_2\}$ as shown below:

$$\mathrm{var}(\hat{\omega}) = \begin{bmatrix} \mathrm{var}(\hat{\omega}_1) & \mathrm{cov}(\hat{\omega}_1,\hat{\omega}_2) \\ \mathrm{cov}(\hat{\omega}_1,\hat{\omega}_2) & \mathrm{var}(\hat{\omega}_2) \end{bmatrix} \tag{11.28}$$

i.e., $\mathtt{var}(\hat{\omega}) \approx I^{-1}(\hat{\omega})$.

We now evaluate each element of $E[H(\hat{\omega},\delta)]$ and $\mathrm{var}[S(\hat{\omega},\delta)]$. For the expected Hessian matrix, we have

$$\begin{aligned} E\left(\frac{\partial^2 L_j(\omega,\delta)}{\partial\omega_1^2}\right) &= -E[\delta(G_j,1)]\frac{1}{\omega_1^2} - E[\delta(G_j,3)]\frac{1}{(1-\omega_1-\omega_2)^2} \\ E\left(\frac{\partial^2 L_j(\omega,\delta)}{\partial\omega_2^2}\right) &= -E[\delta(G_j,2)]\frac{1}{\omega_2^2} - E[\delta(G_j,3)]\frac{1}{(1-\omega_1-\omega_2)^2} \\ E\left(\frac{\partial^2 L_j(\omega,\delta)}{\partial\omega_1\partial\omega_2}\right) &= -E[\delta(G_j,3)]\frac{1}{(1-\omega_1-\omega_2)^2} \end{aligned} \tag{11.29}$$

For the variance matrix of the score vector, we have

$$\begin{aligned} \mathrm{var}\left(\frac{\partial L_j(\omega,\delta)}{\partial\omega_1}\right) = {} & \frac{1}{\omega_1^2}\mathrm{var}[\delta(G_j,1)] + \frac{1}{\omega_3^2}\mathrm{var}[\delta(G_j,3)] \\ & - \frac{2}{\omega_1\omega_3}\mathrm{cov}[\delta(G_j,1),\delta(G_j,3)] \end{aligned}$$

$$\begin{aligned}
\operatorname{var}\left(\frac{\partial L_j(\omega,\delta)}{\partial \omega_2}\right) &= \frac{1}{\omega_2^2}\operatorname{var}[\delta(G_j,2)] + \frac{1}{\omega_3^2}\operatorname{var}[\delta(G_j,3)] \\
&\quad - \frac{2}{\omega_2\omega_3}\operatorname{cov}[\delta(G_j,2),\delta(G_j,3)] \\
\operatorname{cov}\left(\frac{\partial L_j(\omega,\delta)}{\partial \omega_1}, \frac{\partial L_j(\omega,\delta)}{\partial \omega_2}\right) &= \frac{1}{\omega_1\omega_2}\operatorname{cov}[\delta(G_j,1),\delta(G_j,2)] \\
&\quad - \frac{1}{\omega_2\omega_3}\operatorname{cov}[\delta(G_j,2),\delta(G_j,3)] \\
&\quad - \frac{1}{\omega_1\omega_3}\operatorname{cov}[\delta(G_j,1),\delta(G_j,3)] \\
&\quad + \frac{1}{\omega_3^2}\operatorname{var}[\delta(G_j,3)] \qquad (11.30)
\end{aligned}$$

Note again that $\omega_3 = 1-\omega_1-\omega_2$ for notational simplicity. The variance–covariance matrix of the score vector requires the variance–covariance matrix of vector $\delta_j = \{\delta(G_j,1), \delta(G_j,2), \delta(G_j,3)\}$. Let $\pi_{j\kappa} = E[\delta(G_j,\kappa)], \forall\kappa = 1,2,3$ be the short notation for the posterior expectation of $\delta(G_j,\kappa)$. The variance–covariance matrix of vector δ_j is

$$\operatorname{var}(\delta_j) = \begin{bmatrix} \pi_{j1}(1-\pi_{j1}) & -\pi_{j1}\pi_{j2} & -\pi_{j1}\pi_{j3} \\ -\pi_{j1}\pi_{j2} & \pi_{j2}(1-\pi_{j2}) & -\pi_{j2}\pi_{j3} \\ -\pi_{j1}\pi_{j3} & -\pi_{j2}\pi_{j3} & \pi_{j3}(1-\pi_{j3}) \end{bmatrix} \qquad (11.31)$$

Elements of the score vector and the Hessian matrix for individual j are the first and second partial derivatives of $L_j(\omega,\delta)$ with respect to ω. These are given as follows:

$$\begin{aligned}
\frac{\partial L_j(\omega,\delta)}{\partial \omega_1} &= \delta(G_j,1)\frac{1}{\omega_1} - \delta(G_j,3)\frac{1}{1-\omega_1-\omega_2} \\
\frac{\partial L_j(\omega,\delta)}{\partial \omega_2} &= \delta(G_j,2)\frac{1}{\omega_2} - \delta(G_j,3)\frac{1}{1-\omega_1-\omega_2} \qquad (11.32)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 L_j(\omega,\delta)}{\partial \omega_1^2} &= -\delta(G_j,1)\frac{1}{\omega_1^2} - \delta(G_j,3)\frac{1}{(1-\omega_1-\omega_2)^2} \\
\frac{\partial^2 L_j(\omega,\delta)}{\partial \omega_2^2} &= -\delta(G_j,2)\frac{1}{\omega_2^2} - \delta(G_j,3)\frac{1}{(1-\omega_1-\omega_2)^2} \\
\frac{\partial^2 L_j(\omega,\delta)}{\partial \omega_1 \partial \omega_2} &= -\delta(G_j,3)\frac{1}{(1-\omega_1-\omega_2)^2} \qquad (11.33)
\end{aligned}$$

11.1.4 Selection Coefficient and Dominance

In viability selection, we often use selection coefficient and the degree of dominance to express the intensity of selection. There is a unique relationship between segregation distortion and the selection intensity. Let w_{11}, w_{12}, and w_{22} be the relative fitness of the three genotypes (G_1G_1,G_1G_2, and G_2G_2) of the SDL, respectively. Let s and h be the selection coefficient and degree of dominance. The relative fitness can be expressed as (Hartl and Clark 1997)

$$\begin{aligned} w_{11} &= 1 \\ w_{12} &= 1 - sh \\ w_{22} &= 1 - s \end{aligned} \tag{11.34}$$

In an $\mathtt{F}_2$ population, the average fitness is

$$\bar{w} = \frac{1}{4}w_{11} + \frac{1}{2}w_{12} + \frac{1}{4}w_{22} = \frac{1}{4} + \frac{1}{2}(1 - sh) + \frac{1}{4}(1 - s) \tag{11.35}$$

The segregation ratio after the viability selection is

$$\begin{aligned} \omega_1 &= \frac{\frac{1}{4}w_{11}}{\bar{w}} = \frac{1}{1 + 2(1 - sh) + (1 - s)} \\ \omega_2 &= \frac{\frac{1}{2}w_{11}}{\bar{w}} = \frac{2(1 - sh)}{1 + 2(1 - sh) + (1 - s)} \\ \omega_3 &= \frac{\frac{1}{4}w_{11}}{\bar{w}} = \frac{1 - s}{1 + 2(1 - sh) + (1 - s)} \end{aligned} \tag{11.36}$$

This equation system represents the relationship between the segregation ratio and the intensity of viability selection. The inverse relationship is given by Luo et al. (2005)

$$\begin{aligned} s &= \frac{\omega_1 - \omega_3}{\omega_1} \\ h &= \frac{\omega_1 - \frac{1}{2}\omega_2}{\omega_1 - \omega_3} \end{aligned} \tag{11.37}$$

which is used to obtain the MLE of s and h given the MLE of $\omega = \{\omega_1, \omega_2, \omega_3\}$.

11.2 Liability Model

Systematic environmental effects may mask the effects of viability loci and cause low power of detection. It is impossible to remove the systematic error from the analysis using the probabilistic model described above. However, the liability model proposed here provides an extremely convenient way to remove such systematic errors. Let y_j be an underlying liability for individual j in the $\texttt{F}_2$ population. We use the following linear model to describe y_j:

$$y_j = X_j\beta + Z_j\gamma + \varepsilon_j \tag{11.38}$$

where β is a vector of nongenetic effects (systematic error effects), X_j is a design matrix for the systematic errors, $Z_j = \{Z_{j1}, Z_{j2}\}$ represents the genotypes of the SDL and has been defined earlier in QTL mapping, $\gamma = \{a, d\}$ are the genetic effects of QTL as defined earlier, and $\varepsilon_j \sim N(0, 1)$ is the residual error for the liability. We can see that the liability is simply a regular quantitative trait, except that it is not observable. Because the liability is a hypothetical variable, the residual variance cannot be estimated, and thus, we set the variance to unity. We assume that viability selection acts on the liability under the truncation selection scheme, i.e., individual j will survive if $y_j \geq 0$; otherwise, it will be eliminated from the population. Since all individuals observed in the $\texttt{F}_2$ population are survivors, $y_j \geq 0$ applies to all individuals. The probability that $y_j \geq 0$ is

$$\Pr(y_j \geq 0) = \Phi(X_j\beta + Z_j\gamma) \tag{11.39}$$

where $\Phi(.)$ is the standardized cumulative normal function. This probability may be considered as the relative fitness. Recall that

$$Z_{j1} = \begin{cases} +1 & \text{for } G_1G_1 \\ 0 & \text{for } G_1G_2 \\ -1 & \text{for } G_2G_2 \end{cases} \tag{11.40}$$

and

$$Z_{j2} = \begin{cases} 0 & \text{for } G_1G_1 \\ 1 & \text{for } G_1G_2 \\ 0 & \text{for } G_2G_2 \end{cases} \tag{11.41}$$

are the indicator variables for the QTL genotype. Therefore, given each of the three genotypes, we have

$$\begin{aligned} \Pr(y_j \geq 0|G_1G_1) &= w_j(11) = \Phi(X_j\beta + a) \\ \Pr(y_j \geq 0|G_1G_2) &= w_j(12) = \Phi(X_j\beta + d) \\ \Pr(y_j \geq 0|G_2G_2) &= w_j(22) = \Phi(X_j\beta - a) \end{aligned} \tag{11.42}$$

Let us define the expected relative fitness for individual j by

$$\begin{aligned} \bar{w}_j &= \frac{1}{4}w_j(11) + \frac{1}{2}w_j(12) + \frac{1}{4}w_j(22) \\ &= \frac{1}{4}\Phi(X_j\beta + a) + \frac{1}{2}\Phi(X_j\beta + d) + \frac{1}{4}\Phi(X_j\beta - a) \end{aligned} \tag{11.43}$$

The normalized fitness for individual j is

$$\omega_j(1) = \frac{\frac{1}{4}w_j(11)}{\bar{w}_j} = \frac{\Phi(X_j\beta + a)}{\Phi(X_j\beta + a) + 2\Phi(X_j\beta + d) + \Phi(X_j\beta - a)}$$

$$\omega_j(2) = \frac{\frac{1}{2}w_j(12)}{\bar{w}_j} = \frac{2\Phi(X_j\beta + d)}{\Phi(X_j\beta + a) + 2\Phi(X_j\beta + d) + \Phi(X_j\beta - a)}$$

$$\omega_j(3) = \frac{\frac{1}{4}w_j(22)}{\bar{w}_j} = \frac{\Phi(X_j\beta - a)}{\Phi(X_j\beta + a) + 2\Phi(X_j\beta + d) + \Phi(X_j\beta - a)} \tag{11.44}$$

Under the liability model, the parameter vector is $\theta = \{\beta, \gamma\}$. We have formulated the problem of mapping SDL into that of mapping QTL. The log likelihood function is

$$L(\theta) = \sum_{j=1}^{n} \ln\left[\sum\nolimits_{\kappa=1}^{3} \omega_j(\kappa)T_1(\kappa, M_j)T_2(\kappa, N_j)\right] \tag{11.45}$$

11.2.1 EM Algorithm

Due to the complexity of the likelihood function, there has been no simple algorithm for the MLE of the parameters. Therefore, Luo et al. (2005) used the simplex algorithm (Nelder and Mead 1965) to search for the MLE of parameters. An EM algorithm does exist except that the maximization step is much more complicated than that under the probabilistic model. Let us look at the log likelihood function used in the complete-data situation, i.e., $\delta(G_j, \kappa)$ is treated as known:

$$L(\theta, \delta) = \sum_{j=1}^{n} L_j(\theta, \delta) \tag{11.46}$$

where

$$\begin{aligned} L_j(\theta, \delta) = &+ \delta(G_j, 1)[\ln(T_1(1, M_j)) + \ln(T_2(1, N_j)) + \ln\Phi(X_j\beta + a)] \\ &+ \delta(G_j, 2)[\ln(T_1(2, M_j)) + \ln(T_2(2, N_j)) + \ln 2 + \ln\Phi(X_j\beta + d)] \\ &+ \delta(G_j, 3)[\ln(T_1(3, M_j)) + \ln(T_2(3, N_j)) + \ln\Phi(X_j\beta - a)] \\ &- \ln[\Phi(X_j\beta + a) + 2\Phi(X_j\beta + d) + \Phi(X_j\beta - a)] \end{aligned} \tag{11.47}$$

The first partial derivatives are

$$S(\theta,\delta)=\sum_{j=1}^{n}S_j(\theta,\delta)=\begin{bmatrix}\sum\limits_{j=1}^{n}\frac{\partial L_j(\theta,\delta)}{\partial\beta}\\ \sum\limits_{j=1}^{n}\frac{\partial L_j(\theta,\delta)}{\partial a}\\ \sum\limits_{j=1}^{n}\frac{\partial L_j(\theta,\delta)}{\partial d}\end{bmatrix} \tag{11.48}$$

where

$$\begin{aligned}\frac{\partial L_j(\theta,\delta)}{\partial\beta}&=+\,\delta(G_j,1)\frac{X_j^T\phi(X_j\beta+a)}{\Phi(X_j\beta+a)}+\delta(G_j,2)\frac{X_j^T\phi(X_j\beta+d)}{\Phi(X_j\beta+d)}\\&\quad+\delta(G_j,3)\frac{X_j^T\phi(X_j\beta-a)}{\Phi(X_j\beta-a)}\\&\quad-\frac{X_j^T\phi(X_j\beta+a)+2X_j^T\phi(X_j\beta+d)+X_j^T\phi(X_j\beta-a)}{\Phi(X_j\beta+a)+2\Phi(X_j\beta+d)+\Phi(X_j\beta-a)}\\ \frac{\partial L_j(\theta,\delta)}{\partial a}&=+\,\delta(G_j,1)\frac{\phi(X_j\beta+a)}{\Phi(X_j\beta+a)}-\delta(G_j,3)\frac{\phi(X_j\beta-a)}{\Phi(X_j\beta-a)}\\&\quad-\frac{\phi(X_j\beta+a)-\phi(X_j\beta-a)}{\Phi(X_j\beta+a)+2\Phi(X_j\beta+d)+\Phi(X_j\beta-a)}\\ \frac{\partial L_j(\theta,\delta)}{\partial d}&=+\,\delta(G_j,2)\frac{\phi(X_j\beta+d)}{\Phi(X_j\beta+d)}\\&\quad-\frac{2\phi(X_j\beta+d)}{\Phi(X_j\beta+a)+2\Phi(X_j\beta+d)+\Phi(X_j\beta-a)}\end{aligned} \tag{11.49}$$

The Fisher information matrix is

$$I(\theta)=\sum_{j=1}^{n}\begin{bmatrix}E\left[\frac{\partial L_j(\theta,\delta)}{\partial\beta}\frac{\partial L_j(\theta,\delta)}{\partial\beta^T}\right] & E\left[\frac{\partial L_j(\theta,\delta)}{\partial\beta}\frac{\partial L_j(\theta,\delta)}{\partial a}\right] & E\left[\frac{\partial L_j(\theta,\delta)}{\partial\beta}\frac{\partial L_j(\theta,\delta)}{\partial d}\right]\\ E\left[\frac{\partial L_j(\theta,\delta)}{\partial a}\frac{\partial L_j(\theta,\delta)}{\partial\beta^T}\right] & E\left[\frac{\partial L_j(\theta,\delta)}{\partial a}\frac{\partial L_j(\theta,\delta)}{\partial a}\right] & E\left[\frac{\partial L_j(\theta,\delta)}{\partial a}\frac{\partial L_j(\theta,\delta)}{\partial d}\right]\\ E\left[\frac{\partial L_j(\theta,\delta)}{\partial d}\frac{\partial L_j(\theta,\delta)}{\partial\beta^T}\right] & E\left[\frac{\partial L_j(\theta,\delta)}{\partial d}\frac{\partial L_j(\theta,\delta)}{\partial a}\right] & E\left[\frac{\partial L_j(\theta,\delta)}{\partial d}\frac{\partial L_j(\theta,\delta)}{\partial d}\right]\end{bmatrix} \tag{11.50}$$

Let $S(\theta)=E[S(\theta,\delta)]$ be the expectation of the first partial derivative. We have the following iteration equation, which is the maximization step of the EM algorithm:

$$\theta^{(t+1)}=\theta^{(t)}+I^{-1}(\theta^{(t)})S(\theta^{(t)}) \tag{11.51}$$

The expectation step is to calculate the expectation of δ_j using

$$E[\delta(G_j,\kappa)] = \frac{\omega_j(\kappa)T_1(\kappa,M_j)T_2(\kappa,N_j)}{\sum_{\kappa'}^{3}\omega_j(\kappa')T_1(\kappa',M_j)T_2(\kappa',N_j)} \tag{11.52}$$

Before we proceed to the next section, let us look at the details of the Fisher information matrix. In a slightly more compact notation, it is rewritten as

$$I(\theta) = \sum_{j=1}^{n} E\left[S_j(\theta,\delta)S_j^T(\theta,\delta)\right] \tag{11.53}$$

where $S_j(\theta,\delta)$ can be expressed as a linear function of vector δ_j, i.e.,

$$S_j(\theta,\delta) = A_j^T\delta_j + C_j \tag{11.54}$$

where A_j is a $3 \times (p+2)$ matrix and C_j is a $(p+2) \times 1$ vector. The expressions of A_j and C_j can be found from (11.49). The dimension of vector β is p. Since var(δ_j) and $E(\delta_j)$ are known (given before), we can write

$$\begin{aligned} I(\theta) &= \sum_{j=1}^{n} E\left(A_j^T\delta_j\delta_j^T A_j + A_j^T\delta_j C_j^T + C_j\delta_j^T A_j + C_j C_j^T\right) \\ &= \sum_{j=1}^{n} A_j^T E(\delta_j\delta_j^T)A_j + A_j^T E(\delta_j)C_j^T + C_j E(\delta_j^T)A_j + C_j C_j^T \end{aligned} \tag{11.55}$$

where

$$E(\delta_j\delta_j^T) = \mathrm{var}(\delta_j) + E(\delta_j)E(\delta_j^T) \tag{11.56}$$

Definition of var(δ_j) can be found in (11.31).

11.2.2 Variance Matrix of Estimated Parameters

The variance–covariance matrix of the estimated parameters can be approximated by $\mathrm{var}(\hat{\theta}) \approx I^{-1}(\hat{\theta})$. However, a better approximation is to adjust the Fisher information matrix by the variance–covariance matrix of the score vector, i.e.,

$$\mathrm{var}(\hat{\theta}) \approx \left[I(\hat{\theta}) - \sum_{j=1}^{n}\mathrm{var}[S_j(\theta,\delta)]\right]^{-1} \tag{11.57}$$

where

$$\mathrm{var}[S_j(\theta,\delta)] = \mathrm{var}(A_j\delta_j) = A_j^T \mathrm{var}(\delta_j)A_j \tag{11.58}$$

This adjustment gives the Louis (1982) information matrix.

11.2.3 Hypothesis Test

The null hypothesis is that there is no segregation distortion. This has been formulated as $H_0 : a = d = 0$. The log likelihood function evaluated at $\theta = \hat{\theta}$ is

$$L_1(\hat{\theta}) = \sum_{j=1}^{n} \ln\left[\sum\nolimits_{\kappa=1}^{3} \omega_j(\kappa)T_1(\kappa, M_j)T_2(\kappa, N_j)\right] \tag{11.59}$$

The log likelihood function evaluated under the null model is

$$L_0(\phi) = \sum_{j=1}^{n} \ln\left[\sum\nolimits_{\kappa=1}^{3} \phi_\kappa T_1(\kappa, M_j)T_2(\kappa, N_j)\right] \tag{11.60}$$

This is because under $H_0 : a = d = 0$, we have $\omega_j(\kappa) = \phi_\kappa, \forall\kappa = 1, 2, 3$. Given L_0 and L_1, the usual likelihood ratio test statistic LRT is used to test the null hypothesis, where $\mathtt{LRT} = -2(L_0 - L_1)$.

The liability model has two advantages over the probabilistic model: (1) Cofactors can be removed from the analysis by fitting a β vector in the model and (2) the Wald (1943) test statistic may be used to test the null hypothesis.

11.3 Mapping QTL Under Segregation Distortion

Segregation distortion has long been treated as an error in the area of QTL mapping. Its impact on the result of QTL mapping is generally considered detrimental. Therefore, QTL mappers usually delete markers with segregation distortion before conducting QTL mapping. However, a recent study (Xu 2008) shows that segregation distortion can help QTL mapping in some circumstances. Rather than deleting markers with segregation distortion, we can take advantage of these markers in QTL mapping. This section will combine QTL mapping and SDL mapping to map QTL and SDL jointly. The method was recently published by Xu and Hu (2009).

11.3.1 Joint Likelihood Function

Consider that a QTL itself is also an SDL, i.e., the QTL is not necessarily a Mendelian locus. We now go back to the probabilistic model for the SDL. The parameter for SDL is $\omega = \{\omega_1, \omega_2, \omega_3\}$. Let y_j be the phenotypic value of a quantitative trait (not the liability) measured from individual j. The probability density of y_j conditional on $G_j = \kappa$ for individual j is normal with mean $\mu_j = X_j\beta + H_\kappa\gamma$ and variance σ^2, i.e.,

$$p(y_j|G_j = \kappa) = f_\kappa(y_j) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y_j - X_j\beta - H_\kappa\gamma)^2\right] \quad (11.61)$$

The conditional probability for the (flanking) markers is

$$p(M_j, N_j|G_j = \kappa) = T_1(\kappa, N_j)T_2(\kappa, N_j) \quad (11.62)$$

The probability that $G_j = \kappa$ is

$$p(G_j = \kappa) = \omega_\kappa \quad (11.63)$$

The joint likelihood function can be obtained by combining the three probabilities,

$$L(\theta) = \sum_{j=1}^{n}\ln\left[\sum\nolimits_{\kappa=1}^{3} p(G_j = \kappa)p(y_j|G_j = \kappa)p(M_j, N_j|G_j = \kappa)\right] \quad (11.64)$$

which is rewritten as

$$L(\theta) = \sum_{j=1}^{n}\ln\left\{\sum\nolimits_{\kappa=1}^{3} \omega_\kappa f_\kappa(y_j)T_1(\kappa, M_j)T_2(\kappa, N_j)\right\} \quad (11.65)$$

where the parameter vector is $\theta = \{\beta, \gamma, \omega\}$.

11.3.2 EM Algorithm

Derivation of the EM algorithm is given by Xu and Hu (2009). Here we only provide the final result. The expectation step of the EM algorithm requires computing the expectation of δ_j conditional on the data and θ. Because δ_j is a multivariate Bernoulli variable, the expectation is simply the probability of $\delta(G_j, \kappa) = 1$, i.e.,

$$\begin{aligned} E[\delta(G_j, \kappa)] &= \frac{p(G_j = \kappa)p(y_j|G_j = \kappa)p(M_j, N_j|G_j = \kappa)}{\sum_{\kappa'=1}^{3} p(G_j = \kappa')p(y_j|G_j = \kappa')p(M_j, N_j|G_j = \kappa')} \\ &= \frac{\omega_\kappa f_\kappa(y_j)T_1(\kappa, M_j)T_2(\kappa, N_j)}{\sum_{\kappa'=1}^{3} \omega_{\kappa'} f_\kappa(y_j)T_1(\kappa', M_j)T_2(\kappa', N_j)} \end{aligned} \quad (11.66)$$

The maximization step of the EM algorithm involves the following equations:

$$
\begin{aligned}
\beta &= \left[\sum_{j=1}^{n} X_j^T X_j^T\right]^{-1} \left[\sum_{j=1}^{n}\sum_{\kappa=1}^{3} E[\delta(G_j,\kappa)](y_j - H_\kappa\gamma)\right] \\
\gamma &= \left[\sum_{j=1}^{n}\sum_{\kappa=1}^{3} E[\delta(G_j,\kappa)]\left(H_\kappa^T H_\kappa\right)\right]^{-1} \left[\sum_{j=1}^{n}(y_j - X_j\beta)\right] \\
\sigma^2 &= \frac{1}{n}\sum_{j=1}^{n}\sum_{\kappa=1}^{3} E[\delta(G_j,\kappa)]\left(y_j - X_j\beta - H_\kappa\gamma\right)^2 \\
\omega_\kappa &= \frac{1}{n}\sum_{j=1}^{n} E[\delta(G_j,\kappa)],\ \forall\kappa = 1,2,3
\end{aligned}
\tag{11.67}
$$

11.3.3 Variance–Covariance Matrix of Estimated Parameters

Let us define the complete-data log likelihood function for individual j as

$$
\begin{aligned}
L_j(\theta,\delta) =& -\frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)(y_j - X_j\beta - H_\kappa\gamma)^2 \\
&+ \sum_{\kappa=1}^{3}\delta(G_j,\kappa)\{\ln[T_1(\kappa,M_j)] + \ln[T_2(\kappa,N_j)]\} \\
&+ \sum_{\kappa=1}^{3}\delta(G_j,\kappa)\ln\omega_\kappa
\end{aligned}
\tag{11.68}
$$

where $\omega_3 = 1 - \omega_1 - \omega_2$ so that ω_3 is excluded from the parameter vector. Elements of the score vector for individual j are

$$
\begin{aligned}
\frac{\partial L_j(\theta,\delta)}{\partial\beta} &= \frac{1}{\sigma^2}\sum_{k=1}^{3}\delta(G_j,\kappa)X_j^T(y_j - X_j\beta - H_\kappa\gamma) \\
\frac{\partial L_j(\theta,\delta)}{\partial\gamma} &= \frac{1}{\sigma^2}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)H_\kappa^T(y_j - X_j\beta - H_\kappa\gamma) \\
\frac{\partial L_j(\theta,\delta)}{\partial\sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)(y_j - X_j\beta - H_\kappa\gamma)^2
\end{aligned}
$$

$$\begin{aligned}\frac{\partial L_j(\theta,\delta)}{\partial \omega_1} &= \delta(G_j,1)\frac{1}{\omega_1} - \delta(G_j,3)\frac{1}{1-\omega_1-\omega_2}\\ \frac{\partial L_j(\theta,\delta)}{\partial \omega_2} &= \delta(G_j,2)\frac{1}{\omega_2} - \delta(G_j,3)\frac{1}{1-\omega_1-\omega_2}\end{aligned} \tag{11.69}$$

Elements of the Hessian matrix are

$$\begin{aligned}\frac{\partial^2 L_j(\theta,\delta)}{\partial\beta\partial\beta^T} &= -\frac{1}{\sigma^2}X_j^T X_j\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\beta\partial\gamma^T} &= -\frac{1}{\sigma^2}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)X_j^T H_\kappa\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\beta\partial\sigma^2} &= -\frac{1}{\sigma^4}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)X_j^T(y_j - X_j\beta - H_\kappa\gamma)\end{aligned} \tag{11.70}$$

$$\begin{aligned}\frac{\partial^2 L_j(\theta,\delta)}{\partial\gamma\partial\gamma^T} &= -\frac{1}{\sigma^2}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)H_\kappa^T H_\kappa\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\gamma\partial\beta^T} &= -\frac{1}{\sigma^2}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)H_\kappa^T X_j\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\gamma\partial\sigma^2} &= -\frac{1}{\sigma^4}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)H_\kappa^T(y_j - X_j\beta - H_\kappa\gamma)\end{aligned} \tag{11.71}$$

$$\begin{aligned}\frac{\partial^2 L_j(\theta,\delta)}{\partial\sigma^2\partial\sigma^2} &= +\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)(y_j - X_j\beta - H_\kappa\gamma)^2\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\sigma^2\partial\beta^T} &= -\frac{1}{\sigma^4}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)(y_j - X_j\beta - H_\kappa\gamma)X_j\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\sigma^2\partial\gamma^T} &= -\frac{1}{\sigma^4}\sum_{\kappa=1}^{3}\delta(G_j,\kappa)(y_j - X_j\beta - H_\kappa\gamma)H_\kappa\end{aligned} \tag{11.72}$$

$$\begin{aligned}\frac{\partial^2 L_j(\theta,\delta)}{\partial\omega_1\partial\omega_1} &= -\delta(G_j,\kappa)\frac{1}{\omega_1^2} + \delta(G_j,3)\frac{1}{\omega_3^2}\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\omega_1\partial\omega_2} &= +\delta(G_j,3)\frac{1}{\omega_3^2}\\ \frac{\partial^2 L_j(\theta,\delta)}{\partial\omega_2\partial\omega_2} &= -\delta(G_j,\kappa)\frac{1}{\omega_2^2} + \delta(G_j,3)\frac{1}{\omega_3^2}\end{aligned} \tag{11.73}$$

The score vector and the Hessian matrix provide the original material from which $-E[H(\theta,\delta)]$ and $\mathtt{var}[S(\theta,\delta)]$ are calculated (see Xu and Hu 2009). The Louis

(1982) information matrix is

$$I(\theta) = -E[H(\theta, \delta)] - \mathrm{var}[S(\theta, \delta)] \tag{11.74}$$

from which, we can get the variance matrix of the estimated parameters using $\mathrm{var}(\hat{\theta}) \approx I^{-1}(\hat{\theta})$.

11.3.4 Hypothesis Tests

Hypothesis 1

There are several different hypotheses we can test. The first null hypothesis is $H_0 : \gamma = 0$, i.e., there is no QTL for the quantitative trait. To test this hypothesis, we need the full-model likelihood value as shown below:

$$L_1(\hat{\theta}) = \sum_{j=1}^{n} \ln \left\{ \sum_{\kappa=1}^{3} \omega_\kappa f_\kappa(y_j) T_1(\kappa, M_j) T_2(\kappa, N_j) \right\} \tag{11.75}$$

where the parameters in the right-hand side of the equation are replaced by the MLE. The reduced-model likelihood value is calculated using

$$\begin{aligned} L_0(\hat{\hat{\theta}}) = & \sum_{j=1}^{n} \ln \left\{ \sum_{\kappa=1}^{3} \omega_\kappa T_1(\kappa, M_j) T_2(\kappa, N_j) \right\} \\ & - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - X_j \beta)^2 - \frac{n}{2} \ln(\sigma^2) \end{aligned} \tag{11.76}$$

where $\gamma = 0$ is enforced and $\hat{\hat{\theta}}$ is the estimated parameter vector under the reduced model. The usual likelihood ratio test statistic is then constructed using the two likelihood values.

Hypothesis 2

The second hypothesis is $H_0 : \omega = \phi$, i.e., the population is Mendelian. The log likelihood functions under the full model are

$$L_1(\hat{\theta}) = \sum_{j=1}^{n} \ln \left\{ \sum_{\kappa=1}^{3} \omega_\kappa f_\kappa(y_j) T_1(\kappa, M_j) T_2(\kappa, N_j) \right\} \tag{11.77}$$

This is the same as that given in (11.75). The likelihood value under the reduced model is

$$L_1(\hat{\hat{\theta}}) = \sum_{j=1}^{n} \ln \left\{ \sum_{\kappa=1}^{3} \phi_\kappa f_\kappa(y_j) T_1(\kappa, M_j) T_2(\kappa, N_j) \right\} \tag{11.78}$$

where $\omega = \phi$ is enforced and $\hat{\hat{\theta}}$ is the estimated parameter vector under the restricted model. The usual likelihood ratio test statistic is then constructed using the two likelihood values.

Hypothesis 3

The third hypothesis is $H_0 : \gamma = 0 \ \& \ \omega = \phi$, i.e., Mendelian population with no QTL effect for the quantitative trait. The full model remains the same as that given in (11.75) and (11.77). The reduced model is

$$\begin{aligned} L_0(\hat{\hat{\theta}}) = &\sum_{j=1}^{n} \ln \left\{ \sum_{\kappa=1}^{3} \phi_\kappa T_1(\kappa, M_j) T_2(\kappa, N_j) \right\} \\ &- \frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - X_j \beta)^2 - \frac{n}{2} \ln(\sigma^2) \end{aligned} \tag{11.79}$$

where $\hat{\hat{\theta}}$ is the estimated parameter vector under the restricted model. This hypothesis will be rejected if $\gamma \neq 0$ or $\omega \neq \phi$ or both inequalities hold. This hypothesis is particularly interesting for QTL mapping under selective genotyping. If the $\mathtt{F}_2$ population is a Mendelian population, i.e., there is no segregation distortion, individuals are only genotyped based on the extremity of the phenotype values. Selective genotyping will lead to $\omega \neq \phi$, even if the original $\mathtt{F}_2$ population is Mendelian.

11.3.5 Example

The mouse data introduced in Sect. 8.1 of Chap. 8 is used again for the joint QTL and SDL analysis. The mouse genome is scanned for QTL of the 10th-week body weight, the segregation distortion locus (SDL), and both QTL and SDL with a 1-cM increment for all the 19 chromosomes (excluding the sex chromosome) of the genome. The LOD scores are depicted in Fig. 11.1. Let LOD = 3 be the criterion of significance for gene detection. Two QTL appear to be significant, and both are on chromosome 2. Three SDL are significant with one on chromosome 6 ($\mathtt{LOD} \approx 42.5$), one on chromosome 14 ($\mathtt{LOD} \approx 5.5$), and one on chromosome 18 ($\mathtt{LOD} \approx 3.5$). The joint test has the highest LOD score across the entire genome.

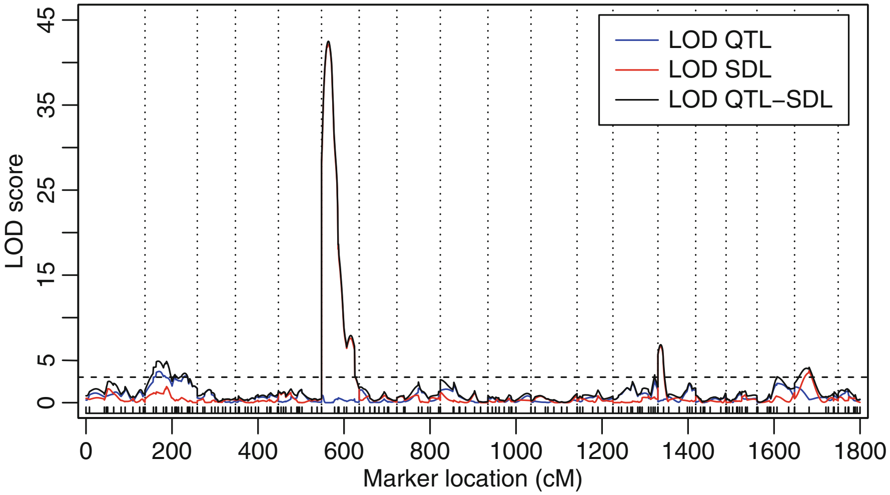

Fig. 11.1 The LOD test statistics profiles for the mouse genome (excluding the sex chromosome). The three LOD score profiles represent (1) the LOD test for QTL of the 10th-week body weight (*blue*), (2) the LOD score for SDL (segregation distortion locus, *red*), and (3) the LOD score for both the QTL and SDL (*black*). The dashed horizontal line indicates the LOD = 3 criterion. The 19 chromosomes are separated by the *vertical reference dotted lines*

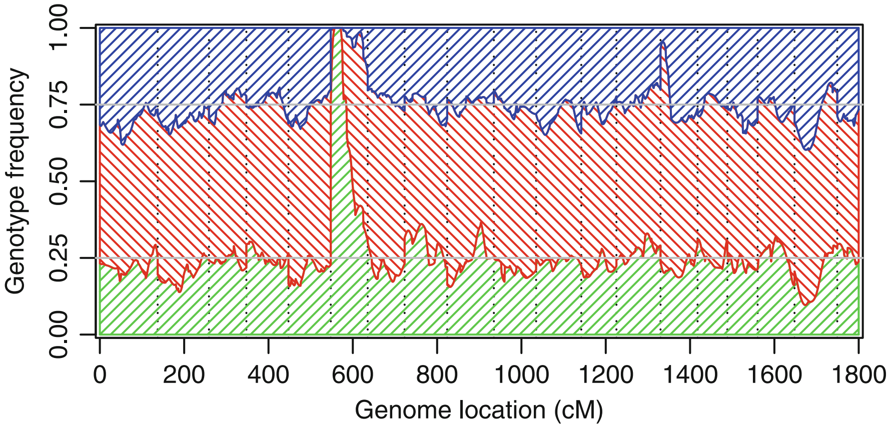

Fig. 11.2 Estimated genotypic frequencies for the mouse genome. Frequencies of the three genotypes are represented by areas with different patterns (A_1A_1 at *top*, A_1A_2 in the *middle*, and A_2A_2 at the *bottom*). The chromosomes are separated by the reference lines on the horizontal axis. The two reference lines on the vertical axis (0.25 and 0.75) divide the area into three parts based on the Mendelian segregation ratio (0.25, 0.5, and 0.25)

The estimated frequencies of the three genotypes (A_1A_1, A_1A_2, and A_2A_2) are shown in Fig. 11.2. The large SDL on chromosome 6 was extremely strong, and it wiped out all heterozygotes and homozygotes of the other type. The allele of this locus was fixed for the A_2 allele.