2 Key Statistical Issues for QTL

- general notation and data structure
- recombination model
  - two linked markers
  - flanking markers to a QTL
  - map distance and map functions
- modelling the phenotype
  - phenotype model
  - model likelihood
  - Bayesian posterior
- missing data concepts and algorithms
- model selection

interval mapping basics

- observed measurements
  - \( Y \) = phenotypic trait
  - \( X \) = markers & linkage map
  - \( i \) = individual index \( 1, \ldots, n \)
- missing data
  - missing marker data
  - \( Q \) = QT genotypes
  - alleles \( QQ, Qq, \) or \( qq \) at locus
- unknown genetic architecture
  - \( \lambda \) = QT locus (or loci)
  - \( \theta \) = genetic action
  - \( m \) = number of QTL
- \( \Pr(\hat{Q}|X,\lambda, m) \) recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for \( Q \) given \( X \)
- \( \Pr(Y|Q, \theta, m) \) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters \( \theta \) (could be non-parametric)

after Sen Churchill (2001)
2.1 general notation and data structure

- \( Y \) = phenotype values
  - as concept and realized (observed) values
- \( X \) = marker genotypes
  - type of experimental cross
  - linkage map construction
    - marker orders, positions, linkage phases
  - observed marker genotypes (possibly with error)
- \( \Pr(Y, X) \) = joint probability
  - what we “know” about \( Y \) and \( X \) for this experiment
  - usually assume linkage map is “known”

conditional data likelihood

- condition on markers and linkage map
  \[
  \Pr(Y \mid X) = \frac{\Pr(Y, X)}{\Pr(X)}
  \]
- \( \Pr(X) \) comprises information on linkage map
  - not influenced by phenotype
  - thus can “ignore” for QTL purposes
unknown QTL genotypes

• usually have sparse linkage map of markers
  – want to condition on actual QTL genotype $Q$
    \[ \text{pr}(Y|Q) \]
  – but actual QTL affecting phenotype not known

• need to consider all possibilities
  – average \( \text{pr}(Y|Q) \) over all possible genotypes $Q$
  – weight by recombination \( \text{pr}(Q|X) \)

\[
\text{pr}(Y | X) = \sum_Q \text{pr}(Y | Q) \text{pr}(Q | X)
\]

enter the (Greek) parameters

• $\theta$ = genetic effects, or gene action
  – additive, dominance, epistasis
  – may include reference values
    • grand mean ($\mu$), environmental variance ($\sigma^2$)

• $\lambda$ = location(s) of QTL
  – measured along “linear” genome
  – related to recombination and map distance

\[
L(\theta, \lambda | Y, X) = \text{pr}(Y | X, \theta, \lambda) = \sum_Q \text{pr}(Y | Q, \theta) \text{pr}(Q | X, \lambda)
\]
2.2 recombination model

• locus $\lambda$ is distance along linkage map
  – identifies flanking marker region
• flanking markers provide good approximation
  – map assumed known from earlier study
  – inaccuracy slight using only flanking markers
    • extend to next flanking markers if missing data
  – could consider more complicated relationship
    • but little change in results

$$\text{pr}(Q \mid X, \lambda) = \text{pr}(\text{geno} \mid \text{map, locus}) \approx \text{pr}(\text{geno} \mid \text{flanking markers, locus})$$

2.2.1 two linked markers

• backcross design
  – $n$ individuals, 2 markers
    – follow one gamete
• recombinants
  – $Ab, aB$
    – $n_R = n_{12} + n_{21}$
• non-recombinants
  – $ab, AB$
    – $n_{NR} = n_{11} + n_{22}$
• recombination rate
  \[
  \hat{r} = \frac{n_R}{n} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}
  \]
no linkage?

- test for no linkage: \( r = 1/2 \)
- assumption: all individuals have same rate
  - implies binomial variation
  \[
  \hat{r} = \frac{n_R}{n} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}, \quad \text{var}(\hat{r}) \approx \frac{\hat{r}(1-\hat{r})}{n}
  \]
- normal or chi-square test statistic
  \[
  Z = \frac{\hat{r} - 1/2}{\sqrt{\text{var}(\hat{r})}} \sim N(0,1) \quad \text{or} \quad Z^2 = \frac{(n_R - n/2)^2}{(n_R n_{NR}/n)} \sim \chi^2
  \]

binomial probabilities

binomial prob
\( n = 30,100 \)
\( r = 0.4,0.2 \)
\[
\text{pr}(n_R = k) = \binom{n}{k} r^k (1-r)^{n-k}
\]

normal approx
likelihood ratio and LOD test

- likelihood for linked markers
  \[ L(r) = \text{pr}(n_R \mid n, r) = Cr^n_r (1 - r)^{n_{NR}} \]
- likelihood for unlinked markers
  \[ L\left(\frac{1}{2}\right) = C\left(\frac{1}{2}\right)^n \]
- likelihood ratio and LOD
  \[ LR = 2^n \left(\hat{r}\right)^{n_R} (1 - \hat{r})^{n_{NR}}, G^2 = 2\log(LR) \sim \chi_1^2 \]
  \[ LOD = \log_{10}(LR) = \frac{G^2}{2\log(10)} = .217G^2 \]

test statistic: distribution

- \(Z^2\) and \(G^2\) are generally close to each other
  - \(Z^2\) based on properties of counts
  - \(G^2\) and LOD based on likelihood principle
  - both have approximate chi-square distribution
- (non)central chi-square distribution
  \[ r = 0.5: Z^2, G^2 \sim \chi_1^2 \]
  \[ r < 0.5: Z^2, G^2 \sim \chi_{1; ncp}^2, ncp = 4n(0.5 - r)^2 \]
backcross examples

- $n=100$ individuals, $n_R=40$ recombinants
  - $r = 0.4$, se($r$) = 0.049
  - $Z = -2.04$, $Z^2 = 4.17$, $p$-value = 0.041
  - $G^2 = 4.03$, LOD = 0.874, $p$-value = 0.045
- $n=100$ individuals, $n_R=20$ recombinants
  - $r = 0.2$, se($r$) = 0.04
  - $Z = -7.5$, $Z^2 = 56.25$, $p$-value < 0.0001
  - $G^2 = 38.55$, LOD = 8.37, $p$-value < 0.0001

- $n=30$ individuals, $n_R=12$ recombinants
  - $r = 0.4$, se($r$) = 0.089
  - $Z = -1.12$, $Z^2 = 1.25$, $p$-value = 0.26
  - $G^2 = 1.21$, LOD = 0.262, $p$-value = 0.27
- $n=30$ individuals, $n_R=6$ recombinants
  - $r = 0.2$, se($r$) = 0.073
  - $Z = -4.11$, $Z^2 = 16.87$, $p$-value < 0.0001
  - $G^2 = 11.56$, LOD = 2.51, $p$-value < 0.0001
simulations of LOD distribution

\[ n = 100, r = 0.3, 0.5 \]

1000 samples

histogram

chi-square curve

rescaled by \(2\log(10)\)

central

\[ r = 0.5 \]

non-central

\[ r = 0.3 \]

LOD and LR over possible \(r\)

\[ n = 30 \]

\[ n_R = 12 \text{ or } 6 \]

evaluate at all possible \(r\) not just “best”

LR like a density

LOD is basis for hypothesis test estimate interval
LR, LOD and $p$-values

<table>
<thead>
<tr>
<th>LR</th>
<th>LOD</th>
<th>$p$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>0.0319</td>
<td>0.1</td>
</tr>
<tr>
<td>31.6</td>
<td>1.5</td>
<td>0.0086</td>
<td>0.0316</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.0024</td>
<td>0.01</td>
</tr>
<tr>
<td>1000</td>
<td>3</td>
<td>0.0002</td>
<td>0.001</td>
</tr>
<tr>
<td>10000</td>
<td>4</td>
<td>&lt;0.0001</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

LOD-based interval estimate for $r$

Point estimate $\hat{r} = n_R / n$

Interval estimate from LOD peak down 1.5 LOD
(or 1 or 2 or …)

$n = 30, n_R = 12, 6$
LOD-based interval calculations

<table>
<thead>
<tr>
<th>n</th>
<th>n_R</th>
<th>r</th>
<th>1 LOD</th>
<th>1.5 LOD</th>
<th>2 LOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>3</td>
<td>0.1</td>
<td>0.03-0.25</td>
<td>0.02-0.29</td>
<td>0.01-0.33</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.1</td>
<td>0.05-0.17</td>
<td>0.04-0.19</td>
<td>0.04-0.21</td>
</tr>
<tr>
<td>30</td>
<td>6</td>
<td>0.2</td>
<td>0.08-0.37</td>
<td>0.06-0.42</td>
<td>0.05-0.46</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0.2</td>
<td>0.13-0.29</td>
<td>0.11-0.31</td>
<td>0.10-0.33</td>
</tr>
<tr>
<td>30</td>
<td>12</td>
<td>0.4</td>
<td>0.23-0.50</td>
<td>0.19-0.50</td>
<td>0.17-0.50</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0.4</td>
<td>0.30-0.50</td>
<td>0.28-0.50</td>
<td>0.26-0.50</td>
</tr>
</tbody>
</table>

Note skew in intervals for small recombination rates.
Note upper boundary of 0.5.
likelihood & Bayesian posterior

- recall the likelihood and likelihood ratio:
\[ L(r) = \text{pr}(n_R \mid n, r) = Cr^{n_r} (1-r)^{n_{NR}} \]
\[ LR(r) = 2^n r^{n_r} (1-r)^{n_{NR}} \]
- posterior turns likelihood into a density
  - assume \( r \) may be any value prior to seeing data
  - posterior = likelihood x prior / constant
\[ \text{pr}(r \mid n, n_R) = \frac{L(r)}{A} \text{ or } = \frac{LR(r)}{A} \]
\[ A = \text{area under likelihood or LR curve} \]
\[ \sum_r \text{pr}(r \mid n, n_R) = 1 \]

LR and Bayes posterior

imagine LR as density
area under curve = 1
\[ \text{pr}(r \mid n_R) = \frac{LR(r)}{A} \]
what is probability that \( r \) is between 0.25 and 0.5?
where is interval with highest posterior mass? (HPD region)
example: \( n=30, n_R=12,6 \)
95% HPD regions
HPD-based interval calculations

<table>
<thead>
<tr>
<th>n</th>
<th>n_R</th>
<th>r</th>
<th>HPD level 96.8%</th>
<th>HPD level 99.1%</th>
<th>HPD level 99.76%</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>3</td>
<td>0.1</td>
<td>0.03-0.25</td>
<td>0.02-0.29</td>
<td>0.01-0.33</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.1</td>
<td>0.05-0.17</td>
<td>0.05-0.19</td>
<td>0.04-0.21</td>
</tr>
<tr>
<td>30</td>
<td>6</td>
<td>0.2</td>
<td>0.08-0.37</td>
<td>0.07-0.41</td>
<td>0.05-0.45</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0.2</td>
<td>0.13-0.29</td>
<td>0.12-0.31</td>
<td>0.10-0.33</td>
</tr>
<tr>
<td>30</td>
<td>12</td>
<td>0.4</td>
<td>0.25-0.50</td>
<td>0.22-0.50</td>
<td>0.19-0.50</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0.4</td>
<td>0.31-0.50</td>
<td>0.30-0.50</td>
<td>0.28-0.50</td>
</tr>
</tbody>
</table>

Note how these almost agree with LOD-based intervals. Density height for HPD varies by \( n \) and \( r \).
who was Bayes?

• Reverend Thomas Bayes (1702-1761)
  – part-time mathematician
  – buried in Bunhill Cemetery, Moongate, London
  – famous paper in 1763 *Phil Trans Roy Soc London*
  – was Bayes the first with this idea? (Laplace)

• billiard balls on rectangular table
  – two balls tossed at random (uniform) on table
  – where is first ball if the second is to its left (right)?

where is the first ball?

\[
\begin{align*}
\text{prior} & \quad \Pr(\theta) = 1 \\
\text{likelihood} & \quad \Pr(Y|\theta) = \theta^Y(1-\theta)^{1-Y} \\
\text{posterior} & \quad \Pr(\theta|Y) = \text{?}
\end{align*}
\]

\[
\begin{align*}
\Pr(Y|\theta) &= \int_0^1 \theta^Y(1-\theta)^{1-Y} \, d\theta = \frac{1}{2} \\
\Pr(\theta|Y) &= \begin{cases} 
2\theta & \text{if } Y = 1 \\
2(1-\theta) & \text{if } Y = 0
\end{cases} \\
\text{(now throw second ball } n \text{ times)}
\end{align*}
\]
what is Bayes theorem?

• before and after observing data
  – prior: \( \text{pr}(\theta) = \text{pr}(\text{parameters}) \)
  – posterior: \( \text{pr}(\theta | Y) = \text{pr}(\text{parameters} | \text{data}) \)

• posterior = likelihood * prior / constant
  – usual likelihood of parameters given data
  – normalizing constant \( \text{pr}(Y) \) depends only on data
    • constant often drops out of calculation

\[
\text{pr}(\theta | Y) = \frac{\text{pr}(\theta, Y)}{\text{pr}(Y)} = \frac{\text{pr}(Y | \theta) \times \text{pr}(\theta)}{\text{pr}(Y)}
\]

---

Bayes rule for recombination \( r \)

likelihood
\[
\text{pr}(n_r | n, r) = L(r | n, n_r) = C r^n (1-r)^{n-2n_r}
\]

prior on recombination \( r \):
\[
\text{pr}(a \leq r \leq b) = 2(b-a)
\]

0 \( \leq \) \( a \leq b \leq 1/2 \)

Bayes rule:
\[
\text{pr}(r | n, n_r) = \frac{\text{pr}(n_r | n, r) \times \text{pr}(r)}{\text{pr}(n_r | n)}
\]

normalizing constant:
\[
\text{pr}(n_r | n) = \frac{1}{\Gamma(\frac{1}{2})} \int_0^1 \text{pr}(n_r | n, r) 2dr
\]
two markers in F2 intercross

- two meioses
  - follow both gametes
  - 16 possibilities
- ambiguity with AaBb
  - 0 or 2 recombinations
- log likelihood ratio:

\[
\log LR = \sum_i n_i \log \left( f_i(r) / f_i(0.5) \right)
\]

<table>
<thead>
<tr>
<th>genotype</th>
<th>(A_B)</th>
<th>(A_b)</th>
<th>(a_B)</th>
<th>(a_b)</th>
<th>(A_B) or (a_b)</th>
<th>(A_b)</th>
<th>(a_B)</th>
<th>(a_b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>code</td>
<td>22</td>
<td>21</td>
<td>20</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>02</td>
<td>01</td>
</tr>
<tr>
<td>frequency</td>
<td>(f(r))</td>
<td>(1 - r^2)</td>
<td>(r(1 - r))</td>
<td>(r^2)</td>
<td>(1 - r^2)</td>
<td>(r^2)</td>
<td>(r(1 - r))</td>
<td>(1 - r^2)</td>
</tr>
<tr>
<td>(f(r = 1/2))</td>
<td>1/16</td>
<td>2/16</td>
<td>1/16</td>
<td>2/16</td>
<td>4/16</td>
<td>2/16</td>
<td>1/16</td>
<td>2/16</td>
</tr>
<tr>
<td>recombinations</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\gamma = \text{probability of double recombinant}
\]
- for AaBb genotype, haplotype not known
- need to “guess” the recombinant fraction of \(n_{11}\) offspring

\[
\gamma \text{ and } r \text{ are inter-related}
\]
- no “closed” solution, need to iterate
- guess \(\gamma\), use to estimate \(r\), improve \(\gamma\), etc.

\[
\gamma = \text{pr} \left( \frac{A_{B}}{a_{B}} \bigg| AaBb \right) = \frac{r^2}{(1-r)^2 + r^2}
\]

\[
\hat{r} = \frac{1}{2n} \left[ (n_{01} + n_{10} + n_{12} + n_{21}) + 2(n_{02} + n_{20} + \gamma n_{11}) \right]
\]
EM algorithm for F2 recombination

- initial guess: \( r = 0.5, \ \gamma = 0.5 \)
- Expectation (E) step
  - substitute expected values for nuisance \( \gamma \)
  - update \( \gamma \) given current value of \( r \)
- Maximization (M) step
  - maximize likelihood for parameter \( r \)
  - update \( r \) given current value of \( \gamma \)
- iterate E-step and M-step until “convergence”
  - stop when change in log-likelihood is small
  - \( \log LR = \sum n_i \log \left( f_i(r) / f_i(0.5) \right) \)
  - usually change in \( r \) is small at this point

2.2.2 flanking markers to QTL

- most genotype information is local
  - linkage drops off with distance
  - approximate by using only flanking markers
  - exception: linkage disequilibrium
    - different chromosome regions could be correlated
    - due to selection, etc.
    - not a problem for backcross or F2 intercross
- missing marker data: use next flanking marker
backcross QTL & flanking markers

1 meiosis
8 possible genotypes
3 recombination rates
small distances & rates?
no double crossovers

\[ \rho = \frac{r_{AQ}}{r_{AB}} \]

<table>
<thead>
<tr>
<th>Marker genotype</th>
<th>QTL genotype</th>
<th>( Q_Q )</th>
<th>( Q_q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( AB/AB )</td>
<td>( \frac{1 - r_{AQ}}{1 - r_{AB}} )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( Ab/AB )</td>
<td>( \frac{1 - r_{AQ}}{1 - r_{AB}} )</td>
<td>1 - ( \rho )</td>
<td>( \rho )</td>
</tr>
<tr>
<td>( aB/AB )</td>
<td>( \frac{r_{AQ}}{r_{AB}} )</td>
<td>( \rho )</td>
<td>1 - ( \rho )</td>
</tr>
<tr>
<td>( ab/AB )</td>
<td>( \frac{r_{AQ}}{1 - r_{AB}} )</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

F2 QTL & flanking markers

2 meioses
27 possible genotypes
3 recombination rates
EM steps on \( \gamma \) and \( r_{AB} \)
small distances & rates?
no double crossovers

\[ \rho = \frac{r_{AQ}}{r_{AB}}, \gamma = \frac{r_{AB}^2}{(1 - r_{AB}^2) + r_{AB}^2} \]

<table>
<thead>
<tr>
<th>QTL</th>
<th>flanking marker genotypes</th>
<th>( AB )</th>
<th>( Ab )</th>
<th>( Ab )</th>
<th>( ab )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q/Q</td>
<td>1</td>
<td>1 - ( \rho )</td>
<td>( 1 - \rho )^2</td>
<td>( \rho )</td>
<td>( \gamma_0(1 - \rho) )</td>
</tr>
<tr>
<td>Q/q</td>
<td>0</td>
<td>( \rho )</td>
<td>( 2 \rho(1 - \rho) )</td>
<td>( 1 - \rho )</td>
<td>( \gamma_0[\rho^2 + (1 - \rho)^2] )</td>
</tr>
<tr>
<td>q/q</td>
<td>0</td>
<td>0</td>
<td>( \rho^2 )</td>
<td>0</td>
<td>( \gamma_0(1 - \rho) )</td>
</tr>
</tbody>
</table>

ch. 2 © 2003 Broman, Churchill, Yandell, Zeng
2.2.3 map distance & map functions

• How to relate genetic linkage to physical distance?
  – math assumptions = crude approximations
  – critical for map building, minor effect on QTL
• \( x = \) genetic map distance (1 Morgan = 100 cM)
  – expected number of crossovers per meiosis between two loci on a single chromatid strand (Sturtevant 1913)
• typical map functions
  – Morgan: interference \( r_{AB} = r_{AQ} + r_{QB} \)
  – Kosambi: intermediate \( r_{AB} = (r_{AQ} + r_{QB})/(1+4r_{AQ}r_{QB}) \)
  – Haldane: no interference \( r_{AB} = r_{AQ} + r_{QB} - 2r_{AQ}r_{QB} \)

2.3 modelling the phenotype

• trait = mean + genetic + environment
• \( \text{pr}(\text{trait } Y \mid \text{genotype } Q, \text{effects } \theta) \)
  \( \text{pr}(Y \mid Q, \theta) = \text{normal}(G_Q, \sigma^2) \)
  \( Y = \mu + G_Q + E \)
2.3.1 phenotype model

- how is phenotype related to genotype?
- typical assumptions
  - normal environmental variation
    - residuals $e$ (not $Y$) have bell-shaped histogram
  - genetic value $G_Q$ is composite of a few QTL
    - other polygenic effects same across all individuals
  - genetic effect uncorrelated with environment

\[
Y = \mu + G_Q + e, e \sim N(0, \sigma^2)
\]

\[
E(Y \mid Q, \theta) = \mu + G_Q, \quad \operatorname{var}(Y \mid Q, \theta) = \sigma^2
\]

\[
\theta = (\mu, G_Q, \sigma^2) \text{ effects}
\]
F2 intercross phenotype model

- here assume only one QTL
- genotypes QQ, Qq, qq
- genotypic values $G_{QQ}$, $G_{Qq}$, $G_{qq}$
- decompose as additive, dominance effects

<table>
<thead>
<tr>
<th>genotype: $Q=$</th>
<th>QQ</th>
<th>Qq</th>
<th>qq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mather-Jinx: $G_Q =$</td>
<td>$\mu + \alpha$</td>
<td>$\mu + \delta$</td>
<td>$\mu - \alpha$</td>
</tr>
<tr>
<td>Fisher-Cockerham: $G_Q =$</td>
<td>$\mu + \alpha - \frac{\delta}{2}$</td>
<td>$\mu + \frac{\delta}{2}$</td>
<td>$\mu - \alpha - \frac{\delta}{2}$</td>
</tr>
</tbody>
</table>

2.3.2 model likelihood

- why study the likelihood?
  - uncover hidden aspects of QTL
  - loci $\lambda$, effects $\theta$, given data $(Y,X)$
- what is evidence to support a QTL?
- where are the QTL?
- how precise can estimate the loci & effects?
- what genetic architecture is supported?
building the model likelihood

- likelihood links phenotype & recombination
  - through unknown QTL genotypes $Q$
  - mixture over all possible genotypes
- contribution from individual $i$
  \[
  \Pr(Y_i \mid X_i, \theta, \lambda) = \sum_Q \Pr(Y_i \mid Q, \theta) \Pr(Q \mid X_i, \lambda)
  \]
- product over all individuals
  \[
  L(\theta, \lambda \mid Y, X) = \prod_i \sum_Q \Pr(Y_i \mid Q, \theta) \Pr(Q \mid X_i, \lambda)
  \]

and if there are no QTL?

- $Y = \mu + e$, or $L(\mu \mid Y)$
- no relationship with markers & map $X$
- for normal data, maximum likelihood yields
  \[
  L(\mu \mid Y) = N(Y \mid \mu, \sigma^2)
  \]
  \[
  \hat{\mu} = \bar{Y} = \sum Y_i / n
  \]
  \[
  \hat{\sigma}^2 = s^2 = \sum_i (Y_i - \bar{Y})^2 / n
  \]
maximum likelihood & LOD

• likelihood peaks at some \((\theta, \lambda)\)
  – use “hat” (\(^\wedge\)) to signify value at maximum

• LOD profiles likelihood peak
  – find \(\theta\) to maximize likelihood for each \(\lambda\)
  – profile (scan) loci \(\lambda\) over entire genome
  \[
  L(\theta, \lambda | Y, X) = \text{pr}(Y | X, \theta, \lambda) = \text{prod} \, \text{pr}(Y_i | X_i, \theta, \lambda)
  \]

\[LOD(\lambda | Y, X) = \log_{10}\left(\frac{\max_\theta L(\theta, \lambda | Y, X)}{\max_\mu L(\mu | Y)}\right)\]

2.3.3 Bayesian posterior

• treat unknowns as random
  – build “uncertainty” into model framework
  – genetic architecture: gene action \(\theta\), QTL locus \(\lambda\)

• interpret weighted likelihood as a density
  – weights based on prior “beliefs”

\[
\text{pr}(\theta, \lambda | Y, X) = \frac{\text{pr}(Y | X, \theta, \lambda) \text{pr}(\theta, \lambda | X)}{\text{pr}(Y | X)}
\]

\[
\text{pr}(\theta, \lambda | X) = \text{pr}(\theta) \text{pr}(\lambda | X)
\]
choice of Bayesian priors

- elicited priors
  - higher weight for more probable parameter values
  - based on prior empirical knowledge
  - use previous study to inform current study
    - weather prediction, previous QTL studies on related organisms

- conjugate priors
  - convenient mathematical form
  - essential before computers, helpful now to simply computation
  - large variances on priors reduces their influence on posterior

- non-informative priors
  - may have “no” information on unknown parameters
  - prior with all parameter values equally likely
    - may not sum to 1 (improper), which can complicate use

- always check sensitivity of posterior to choice of prior

incorporate missing genotypes $Q$

- augment data with missing genotypes $Q$
  - use recombination model to state uncertainty
  - avoid likelihood mixture by augmentation

- marginal posterior on unknowns of interest
  - average over fully augmented posterior

\[
\begin{align*}
\pr(\theta, \lambda, Q | Y, X) &= \frac{\pr(Y | Q, \theta) \pr(Q | X, \lambda) \pr(\theta, \lambda | X)}{\pr(Y | X)} \\
\pr(\theta, \lambda | Y, X) &= \sum_Q \pr(\theta, \lambda, Q | Y, X)
\end{align*}
\]
Bayesian parameter estimates

- estimates are posterior means or modes
  - mean = weighted average of all possible values
  - mode = maximum
- can get joint or marginal estimates

\[
\hat{\theta}_{\text{mean}} = \sum_{\theta, \lambda} \theta \Pr(\theta, \lambda | Y, X) \\
\hat{\theta}_{\text{mode}} = \arg\max_{\theta} \left( \sum_{\lambda} \Pr(\theta, \lambda | Y, X) \right)
\]

2.4 missing data concepts

- missing QTL genotype \( Q \)--see section 2.3
- missing marker data \( X \)
  - errors in genotyping
  - difficulty reading signal (on gel)
  - marker not fully informative
- distinguish full data \( X \) from observed \( X_{\text{obs}} \)
  - weighted average over all possible marker values

\[
\Pr(Q | X_{\text{obs}}, \lambda) = \sum_{X} \Pr(Q | X, \lambda) \Pr(X | X_{\text{obs}})
\]