## Multiple Interval Mapping

Kao Zeng (1997); Kao Zeng Teasdale (1998)
idea suggested by Lander and Botstein (1989)
multiple QTL oriented method combining
simultaneous multiple QTL mapping analysis
analysis of genetic architecture of quantitative traits

search algorithm for number, positions, effects and interaction of significant QTL

other multiple QTL analyses
Bayesian interval mapping
via Markov chain Monte Carlo (MCMC) for $m$ fixed
(Satagopan et al. 1996; Uimari and Hoeschele 1997)
reversible jump MCMC for random number $m$ of QTL
Bayesian: (Satagopan Yandell 1996; Sillanpää Arjas 1997, 1998; Stephens Fisch 1997)
Frequentist: (Heath 1997)

reversible jump among different sized models
(Green 1995; Richardson Green 1997).



## components of MIM

1. **evaluation procedure**
   analyze likelihood of data given genetic model
   (number, positions and epistasis of QTL)

2. **search strategy**
   search and select "best" genetic model
   (among those sampled) in the parameter space

3. **estimation procedure**
   estimate parameters of genetic architecture of quantitative traits
   – number, positions, effects and epistasis of QTL
   – QTL genetic variances and covariances

4. **prediction procedure**
   predict genotypic values of individuals
   based on selected genetic model
   *e.g.* for marker assisted selection



## MIM model

consider $m$ putative QTL in backcross

$$y_j = \mu + \sum_{k=1}^{m} a_k^* x_{jk}^* + \sum_{kl \in \{\text{pairs}\}} b_{kl}^* (x_{jk}^* x_{jl}^*) + e_j$$

$j$ indexes individuals of the sample: $j = 1, 2, \cdots, n$
$m$ = number of putative QTL
$y_j$ = phenotypic value of individual $j$
$\mu$ = model mean
$a_k^*$ = marginal effect of putative QTL $k$
$x_{jk}^*$ = genotype indicator of putative QTL $k$
(unobserved but inferred from flanking marker data)
$b_{jkl}^*$ = epistatic effect between putative QTL $k$ and $l$
$\{\text{pairs}\}$ = subset of QTL pairs with epistatis
(danger of over-parameterizing model)
$m_{pairs}$ = number of pair-wise epistatic effects
$e_j$ = residual effect assumed $N(0, \sigma^2)$



## MIM likelihood analysis

likelihood of data given model is mixture of normals
genotypes of individual observed only at markers
model contains missing data

$$L(\mathbf{B}, \mu, \sigma^2) = \prod_{j=1}^{n} \left[ \sum_{g=1}^{2^m} p_{jg} \phi\left(y_j | \mu + \mathbf{D}_{jg}\mathbf{B}, \sigma^2\right) \right]$$

term in bracket = weighted sum of normal densities
one for each of $2^m$ possible multiple-QTL genotypes

| | |
|---|---|
| $p_{jg}$ | = probability of genotype $g$ conditional on markers for individual $j$ |
| $\mathbf{B}$ | = vector of QTL parameters $a^*$, $b^*$ |
| $\mathbf{D}_{jg}$ | = vector of genetic model design specifying configuration of $x^*$ with $a^*$ and $b^*$ for $j$th QTL genotype |
| $\mu + \mathbf{D}_{jg}\mathbf{B}$ | = mean for $g$th QTL genotype |
| $\phi(y\|\mu, \sigma^2)$ | = normal density for $y$ with mean $\mu$ and variance $\sigma^2$ |

same idea as IM and CIM, but more machinery

## procedure to obtain maximum likelihood estimates using an EM algorithm

described by Kao and Zeng (1997)
EM iterative procedure: Expectation-Maximization
Expectation step $t+1$

$$P_{jg}^{[t+1]} = \frac{p_{jg}\phi(y_j|\mu^{[t]} + \mathbf{D}_{jg}\mathbf{B}^{[t]}, \sigma^{2[t]})}{\sum_{g=1}^{2^m} p_{jg}\phi(y_j|\mu^{[t]} + \mathbf{D}_{jg}\mathbf{B}^{[t]}, \sigma^{2[t]})}$$

Maximization step $t+1$

$$\mu^{[t+1]} = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \sum_{g=1}^{2^m}\sum_{k=1}^{m+m_{pairs}} P_{jg}^{[t+1]} D_{jgk} B_k^{[t+1]}\right)$$

$$B_k^{[t+1]} = \frac{\sum_{j,g} P_{jg}^{[t+1]} D_{jgk}[(y_j - \mu^{[t]}) - \sum_{l=1}^{k-1} D_{jgl} B_l^{[t+1]} - \sum_{l=k+1}^{m+m_{pairs}} D_{jgl} B_l^{[t]}]}{\sum_{j,g} P_{jg}^{[t+1]} D_{jgr}^2}$$

$$\sigma^{2[t+1]} = \frac{1}{n}\left[\sum_j (y_j - \mu^{[t+1]})^2 - 2\sum_j (y_j - \mu^{[t+1]})\left[\sum_{g,k} P_{jg}^{[t+1]} D_{jgk} B_k^{[t+1]}\right] + \sum_{j,g,k,l} P_{jg}^{[t+1]} D_{jgk} D_{jgl} B_k^{[t+1]} B_l^{[t+1]}\right]$$

$B_k = k$th model parameter in $\mathbf{B}$
$D_{jgk} = k$th design element of $\mathbf{D}_{jg}$



## general matrix form

Kao Zeng (1997); Kao Zeng Teasdale (1999);
Zeng Kao Basten (2000)

$$\mu = \frac{1}{n}\mathbf{1}^{\mathsf{T}}[\mathbf{Y} - \mathbf{PDB}]$$

$$\mathbf{B}^{(t+1)} = diag(\mathbf{V})^{-1}[\mathbf{D}^{\mathsf{T}}\mathbf{P}^{\mathsf{T}}(\mathbf{Y}-\mu) - nondiag(\mathbf{V})\mathbf{B}^{(t)}]$$

$$\sigma^2 = \frac{1}{n}[(\mathbf{Y}-\mu)^{\mathsf{T}}(\mathbf{Y}-\mu) - 2(\mathbf{Y}-\mu)^{\mathsf{T}}\mathbf{PDB} + \mathbf{B}^{\mathsf{T}}\mathbf{VB}]$$

with $\mathbf{V} = \{V_{kl}\}_{k,l=1}^{m+m_{pairs}}$

$$V_{kl} = \sum_{j,g} P_{jg} D_{jgk} D_{jgl}$$

$diag(\mathbf{V})$ = diagonal matrix with $V_{11}, V_{22}, \cdots$
$nondiag(\mathbf{V})$ = matrix $\mathbf{V}$ with 0 down diagonal



## posterior genotype probability $P_{jg}$

probability of QTL genotype given markers

$$Prob(\text{genotype}|\text{Markers}) = Pr(g|M) = p_{jg}$$

conditional density of phenotype given genotype

$$Prob(\text{pheno}|\text{geno}) = Pr(y|g) = \phi\left(y_j|\mu + \mathbf{D}_{jg}\mathbf{B}, \sigma^2\right)$$

probability conditional on markers and phenotype

$$Prob(\text{genotype}|\text{Markers,phenotype}) =$$

$$\begin{aligned} P_{jg} &= Pr(g|M,y) \\ &= \frac{Pr(g|M)Pr(y|g)}{\sum_g Pr(g|M)Pr(y|M)} \end{aligned}$$

usual mixture situation
$P_{jg}$ plays same role as $P_j$ earlier



## dealing with many QTL

$m$ QTL $\longrightarrow 2^m$ possible mixture components
may be prohibitive for efficient numerical analysis
but most genotypes have negligible probabilities $P_{jg}$
can we skip these evaluations?

## practical implementation of MIM algorithm

select subset of "significant" mixture components
for each individual for evaluation:
set any $p_{jg} > \delta$ (=0.001) to zero (drop them)
sum of "significant" $p_{jg} > .95$ (adjust $\delta$ if needed)
normalize "significant" probs: $\sum_g p_{jg} = 1$

number of "significant" mixture components $\sim$ 10-100
(depends on marker density, $m$, position of QTL)
negligible loss of accuracy of likelihood evaluation
compared to no selection

### conditional likelihood ratio test

test for each QTL effect $B_k$
likelihood ratio test conditional on other QTL effects

$$LOD = \log_{10} \frac{L(\text{all } B_l \neq 0)}{L(B_k = 0, \text{ all other } B_l \neq 0)}$$

can proceed as above if we have
–positions of $m$ putative QTL
–selected $m + m_{pairs}$ QTL effects

how do we search for multiple QTL?
how do we decide how many QTL to include?
how do we select best genetic model?
(number, positions, gene action, epistasis)
criterion: fit data well in some sense



### Model selection
### premodel selection

evaluation of MIM model is computationally intensive
important to select good premodel for MIM analysis

1. select subset of significant markers
stepwise regression: backward, forward, combined
stopping rule based on $F$-to-drop or $F$-to-enter

2. use selected marker to perform CIM
to scan the genome for candidate positions

3. identify candidate epistatic pairs
treat marker pairs as one unit in stepwise reg

4. compare 2 & 3 to reach consensus premodel

5. test each premodel parameter under MIM
drop non-significant terms in stepwise fashion



### model selection under MIM

1. premodel selection of QTL, main & epistatic effects

2. scan genome to search for 1 additional QTL
test marginal effect & retain if significant

3. search for 1 additional epistatic effect
among those pairwise terms not yet included
test & retain if significant; repeat

4. re-evaluate significance of all model effects
drop nonsignificant effects unless part of epistasis
repeat in stepwise fashion

5. optimize estimates of each model QTL position
scan region between neighbor QTL to find MLE of position
conditional on current estimates of positions and effects of all other QTL and epistasis)
repeat sequentially until negligible change

6. return to step 2 and repeat until no more significant QTL can be added and position estimates are optimized



### epistasis with no main effect?

epistatis between selected QTL and all others
stepwise search of largest epistatic effect
scan genome and test with all selected QTL
intense numerical calculation

### challenges of search for multiple QTL

high, unknown dimension–complicated, difficult
search on whole genome, not just markers

numerous peaks & valleys in likelihood “landscape”
danger of selecting a local peak far from maximum

appropriate criteria for model selection?
appropriate strategies to search for epistatic QTL?

open questions:
global (genome wide) search for multiple QTL
genetic architecture: multiple components

## MIM stopping rules

when to stop search algorithm?
criterion for comparing different models?

**multiple regression analysis with model selection**

stopping rules usually based on
– minimizing the final prediction error (FPE) criterion
– information criteria (IC)
(Stuart and Ord 1991; Miller 1990)

## QTL mapping as model selection

Broman (1997)
predictors (QTL genotypes) not observed
model selection with markers informative but insufficient to find QTL positions



## QTL mapping as model selection

likelihood ratio or F statistic
test fitted genetic effect: model selection
adjustment on level to account for multiple tests
(Lander Botstein 1987; Haley Knott 1992; Zeng 1994; Jansen Stem 1994)

final prediction error (FPE) criterion

$$S_k = (n+k)RSS_k/(n-k)$$

$RSS(k) =$ residual sum of squares
$k =$ number of parameters fitted in mode
$L_k =$ likelihood of data given $k$-parameter model

Information Criteria
general form in regression analysis

$$\begin{aligned} \text{IC} &= -2(\log L_k - kc_n/2) \\ &\approx \log[RSS_k/n] + kc_n/n \end{aligned}$$



## Various Information Criteria

$$\begin{aligned} \text{IC} &= -2(\log L_k - kc_n/2) \\ &\approx \log[RSS_k/n] + kc_n/n \end{aligned}$$

$c_n = 2$, AIC (Akaike 1969)
$c_n = \log(n)$, BIC (Bayesian; Schwarz 1978)
$c_n = 2\log(\log n)$ (Hannan Quinn 1979)

$S_k \sim$ AIC as $n \to \infty$ (Shibata 1981, 1984)
$S_k$ "optimal": minimizes prediction error as $n \to \infty$
(Breiman Freedman 1983)

Schwarz & Hannan-Quinn consistent
$Prob$(select true model) $\to 1$ as $n \to \infty$
other measures typically include too many terms

asymptotics do not show behavior for finite $n$



## Information Criteria and $F$-to-enter

stepwise selection procedure (Miller 1990, p.208)
IC leads to $F$-to-enter statistic for regression analysis
minimum $F$-to-enter (provided $c_n/n$ is small)

$$\begin{aligned} \frac{RSS_k - RSS_{k+1}}{RSS_{k+1}/(n-k-1)} &\le (n-k-1)(e^{c_n/n}-1) \\ &\approx 2c_n\left(1 - \frac{k+1}{n}\right) \end{aligned}$$

since $LR = n\log(RSS_k/RSS_{k+1})$
similar result holds for LR or LOD

$$\begin{aligned} LR_k &= -2\log\frac{L_k}{L_{k+1}} \le n\log(c_n/n+1) \\ &\approx c_n\left(1 - \frac{k+1}{n}\right) \end{aligned}$$

choice of $c_n$ defines criterion
$c_n = 2$, AIC $\to$ LOD threshold of 0.43

**Information Criteria and $F$-to-enter**

$$LR_k \approx c_n \left(1 - \frac{k+1}{n}\right)$$

$c_n = \delta \log n$, $2 < \delta < 3$ (Broman 1997)
suppose $n$ between 100 $\sim$ 500
LOD threshold = 2 $\sim$ 2.7 for $\delta = 2$
LOD threshold = 3 $\sim$ 4 for $\delta = 3$
similar level to current practice in IM
(Lander Botstein 1989; Zeng 1994)

but: argument is arbitrary
does not relate to
–genetic length of linkage map
–number of markers & linkage group
–density of markers

open questions on stopping rules!



**estimate genotypic values of individual $j$**

indirect estimate of QTL genotype from markers
weighted mean of all possible genotypic values
$\hat{P}_{jg}$ conditional on markers, phenotype

$$\hat{y}_j = \hat{\mu} + \sum_g \hat{P}_{jg} \left[\sum_k D_{jgk} \hat{B}_k\right]$$

first sum: all $2^m$ possible QTL genotypes $g$
(or subset of "significant" QTL genotypes)
second sum: $m$ main effects + $m_{pairs}$ epistatis
$\hat{}$ = maximum likelihood estimate (MLE)

**predict genotypic values based on markers only**

marker assisted selection, cross prediction

$$\hat{y}_j = \hat{\mu} + \sum_g p_{jg} \left[\sum_k D_{gk} \hat{B}_k\right]$$

only have $p_{jg}$ since $\hat{P}_{jg}$ depends on unobserved phenotype $y_j$



**genetic variances and covariances**
**partition of estimated phenotypic variance**

use maximum likelihood estimates of effects

$$\hat{\mathbf{B}} = \hat{\mathbf{V}}^{-1}\mathbf{D}^{\mathsf{T}}\hat{\mathbf{P}}^{\mathsf{T}}(\mathbf{Y} - \hat{\mu})$$

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}[(\mathbf{Y} - \hat{\mu})^{\mathsf{T}}(\mathbf{Y} - \hat{\mu}) - \hat{\mathbf{B}}^{\mathsf{T}}\hat{\mathbf{V}}\hat{\mathbf{B}}] \\
&= \frac{1}{n}\left[\sum_{j=1}^{n}(y_j - \hat{\mu})^2 - \sum_{j,g,k,l} \hat{P}_{jg} D_{jgk} D_{jgl} \hat{B}_k \hat{B}_l\right] \\
&= \frac{1}{n}\left[\sum_{j=1}^{n}(y_j - \bar{y}_\cdot)^2 - \sum_{j,g,k,l} \hat{P}_{jg}(D_{jgk} - \bar{D}_{\cdot\cdot k})(D_{jgl} - \bar{D}_{\cdot\cdot l}) \hat{B}_k \hat{B}_l\right]
\end{aligned}$$

$\bar{y}_\cdot = \sum_{i=1}^{n} y_i/n$
$\bar{D}_{\cdot\cdot k} = \sum_{j=1}^{n} \sum_{g=1}^{2^m} \hat{P}_{jg} D_{jgk}/n$

phenotype = genotype + environment

$$\begin{aligned}
\hat{\sigma}^2 &= \hat{\sigma}_p^2 - \hat{\sigma}_g^2 \\
&= \text{phenotypic variance – genotypic variance}
\end{aligned}$$

$\hat{\sigma}_g^2/\hat{\sigma}_p^2$ = explained variation ($R^2$) of MIM model



**partition of genotypic variance**

$$\begin{aligned}
\hat{\sigma}_g^2 &= \frac{1}{n}\left[\sum_{j,g,k} \hat{P}_{jg}(D_{jgk} - \bar{D}_{\cdot\cdot k})^2 \hat{B}_k^2\right] \\
&\quad + \frac{1}{n}\left[\sum_{j,g,k \neq l} \hat{P}_{jg}(D_{jgk} - \bar{D}_{\cdot\cdot k})(D_{jgl} - \bar{D}_{\cdot\cdot l}) \hat{B}_k \hat{B}_l\right] \\
&= \sum_k \hat{\sigma}_{B_k}^2 + \frac{1}{2}\sum_{k \neq l} \hat{\sigma}_{B_k, B_l}
\end{aligned}$$

$\hat{\sigma}_{B_k}^2$ estimates genetic variance due to QTL effect $B_k$
$\hat{\sigma}_{B_k, B_l}$ estimates genetic covariance between $B_k$, $B_l$

**variance component for single QTL effect**

$$\hat{\sigma}_k^2 = \hat{\sigma}_{B_k}^2 + \frac{1}{2}\sum_{l \neq k} \hat{\sigma}_{B_k, B_l}$$

$\hat{\sigma}_k^2$ estimates variance of $k$th QTL effect
adjusted for linkage disequilibrium ($\sigma_{B_k, B_l} \neq 0$)
$\hat{\sigma}_k^2$ not guaranteed to be positive

**Genetic architecture of a morphological shape difference**

We show as an example the mapping results of an experiment in *Drosophila* (Zeng et al. 1998). Two *Drosophila* species, *D. simulans* and *D. mauritiana*, were crossed to make $F_1$ hybrids. Because $F_1$ males are sterile, females of $F_1$ population were backcrossed to each of the parental lines. Two independent samples (of size 200 and 300) were drawn from each backcross population and genotyped and phenotyped at two different times. Therefore the total sample size of the experiment is about 1000. We refer to the two samples from backcross to *D. simulans* as BS-S1 and BS-S2, and those to *D. mauritiana* as BM- S1 and BM-S2. The trait is the morphology of the posterior lobe of the male genital arch analyzed as the first principal component in an elliptical Fourier analysis (Liu et al. 1995). The results of MIM analysis are shown in Figures **??** and **??** and Tables **??** and **??**. The final



model selected contains 19 QTL (based on the joint analysis of the samples in two backcrosses) distributed on the three *Drosophila* major chromosomes, X, II and III. Figure **??**b depicts the likelihood profile (LOD score) for each QTL that spans from one QTL to its neighbors. The threshold used in analysis is also shown in the figure. As a comparison, Figure **??**a shows the mapping result based on CIM. Forty five markers were genotyped and their map positions are indicated by filled triangles in the figure.

Table **??** shows the estimates of positions and effects of these 19 QTL as percentage of the observed difference of the trait means between the respective parental populations and the $F_1$ hybrid. The sum of the 19 QTL effects explain 99% of the observed differences. Also, because the estimates of substitution effects are estimates of $a + d$ in BM and $a - d$ in BS, where $a$ is the additive effect of a QTL and $d$ is the dominance effect, $a$ and $d$ can be jointly



estimated and are expressed as percentage of half the observed difference between two parental populations in Table **??**. Again the additive effects of these 19 QTL explain 99% of the observed difference. There are substantial dominance effects, but overall the dominance effects are marginal compared to the additive effects. Six QTL pairs show significant epistatic effects in BM (Table **??**), and none in BS according to the threshold adopted for the study. Together, these 19 QTL explain 93.2% of the total variance in BS and 91.6% (plus epistatic variances in Table **??**) in BM. These are the coefficients of determination ($R^2$) of the MIM model in the respective populations, an estimate of heritability of the trait. With these estimates, the genetic architecture of the trait difference between *D. simulans* and *D. mauritiana* becomes clear.

Because this experiment contains two independent samples for each backcross, we asked whether the mapping results obtained from one sample can be