Composite Interval Mapping

extend from one to multiple QTL
condition on additional marker loci
use conditional probabilities for multilocus genotypes
explicit models for two or three linked QTL
gets complicated: multidimensional

goal: test for QTL in interval
with statistic independent of effects
of other QTL along chromosome

improve precision and efficiency
of mapping multiple QTL

multiple regression analysis for linked QTL
use multiple markers as surrogates
extract position information from coefficients

composite interval mapping

interval mapping scan for single QTL
while using other markers as surrogates

simplify mapping for multiple QTL
from multiple dimensional search problem
to one dimensional scan

each interval test independent
(assume no interference!)
consider presence of only a single QTL in interval
combine interval mapping with multiple regression
issues of marker selection

multiple regression analysis

linear structure of gene located on chromosomes
linear model structure of regression
assumptions:
– no crossover interference (Haldane)
– no epistasis

partial regression coefficient of trait on marker
expected to depend only on QTL in current interval
independent of other QTL beyond flanking markers
Zeng (1994); Jansen Stam (1994)

regress trait values y on t markers observed in B_1 populations

$$y_j = \mu + \sum_{k=1}^{t} b_k x_{jk} + e_j$$

$x_{jk} = 0, 1$ for kth marker on jth individual
b_k is regression coefficient for kth marker

Backcross Design
covariance among markers

variance of kth marker	$\sigma_k^2 = Var(x_{jk}) = 1/4$
covariance between markers	$\sigma_{ik} = (1 - 2r_{ik})/4$

variance of marker k conditional on i

$$\begin{aligned}\sigma_{k\cdot i}^2 &= \sigma_k^2 - \sigma_{ik}^2/\sigma_i^2 \\ &= [1 - (1-2r_{ik})^2]/4 = r_{ik}(1-r_{ik})\end{aligned}$$

covariance between two markers conditional on third

$$\begin{aligned}\sigma_{ik\cdot l} &= \sigma_{ik} - \sigma_{il}\sigma_{kl}/\sigma_l^2 \\ &= [(1-2r_{ik}) - (1-2r_{il})(1-2r_{kl})]/4 \\ &= \begin{cases} 0 & \text{for order } ilk \text{ or } kli \\ r_{kl}(1-r_{kl})(1-2r_{ik}) & \text{for order } ikl \text{ or } lki \\ r_{il}(1-r_{il})(1-2r_{ik}) & \text{for order } lik \text{ or } kil \end{cases}\end{aligned}$$

assumption: no interference leads to

$(1-2r_{ik}) = (1-2r_{il})(1-2r_{kl})$ for order ilk or kli

zero covariance between flanking markers or QTL
conditional on intermediate marker

Backcross Design
partial regression coefficient $b_{yk.s_k}$

covariance between y and kth marker is

$$\sigma_{yk} = \sum_{g=1}^{m} (1 - 2r_{gk}) b_g^*/4$$

unconditional regression coefficient

$$b_k = \sigma_{yk}/\sigma_k^2 = \sum_{g=1}^{m} (1 - 2r_{gk}) b_g^*$$

conditional regression coefficient

$$b_{yk \cdot s_k} = \sum_{g \in S(k-1,k)} \frac{r_{k-1,g}(1 - r_{k-1,g})(1 - 2r_{gk})}{r_{k-1,k}(1 - r_{k-1,k})} b_g^*$$

$$+ \sum_{g \in S(k,k+1)} \frac{r_{g,k+1}(1 - r_{g,k+1})(1 - 2r_{kg})}{r_{k,k+1}(1 - r_{k,k+1})} b_g^*$$

s_k = set of all markers except M_k first sum: all QTL between markers $k-1$ and k
second sum: all QTL between markers k and $k+1$
$S(i,l)$ = set of QTL g between M_i and M_l

foundation for composite interval mapping

zero covariance with QTL beyond flanking markers conditional on intermediate markers

regression coefficient depends only on QTL located between markers $k-1$ and $k+1$
greatly simplifies testing

test presence of QTL within a marker interval without bias from other QTL outside

properties of multiple regression relevant to QTL mapping

Property 1 *In the multiple regression analysis, assuming additivity of QTL effects between loci (i.e., ignoring epistasis), the expected partial regression coefficient of the trait on a marker depends only on those QTL which are located on the interval bracketed by the two neighboring markers, and is unaffected by the effects of QTL located on other intervals*. This property essentially says that a conditional (interval) test can be constructed based on the partial regression coefficient and such a test would test the linkage effect of only those QTL which are located within the defined interval.

properties of multiple regression relevant to QTL mapping

Property 2 *Conditioning on unlinked markers in the multiple regression analysis will reduce the sampling variance of the test statistic by controlling some residual genetic variation and thus will increase the power of QTL mapping*. This means that even unlinked markers contain useful information which can be used to increase the statistical power of the test and the efficiency of the genetic mapping. This useful information has not been utilized in the current QTL mapping methods.

properties of multiple regression relevant to QTL mapping

Property 3 *Conditioning on linked markers in the multiple regression analysis will reduce the chance of interference of possible multiple linked QTL on hypothesis testing and parameter estimation, but with a possible increase of sampling variance*. The first part of the sentence restates Property 1, and the second part of the sentence says that an interval test may entail a loss in the statistical power of the test because the test is a conditional test. This summarizes the advantage and disadvantage of the interval test: that is, there is a trade-off between precision and efficiency of mapping by using an interval test. Effective balance on these two issues will be the major consideration in practical mapping of QTL.

properties of multiple regression relevant to QTL mapping

Property 4 *Two sample partial regression coefficients of the trait value on two markers in a multiple regression analysis are generally uncorrelated unless the two markers are adjacent markers*. This is related to the correlation between two test statistics in two intervals for an interval test. It has been shown that, for an interval test, a test statistic on an interval is generally asymptotically uncorrelated to the test statistic on another interval unless two intervals are adjacent intervals. Even when the two intervals are adjacent intervals, the correlation between two test statistics in two intervals is usually very small. This property is related to the issue of determining an appropriate critical value of a test statistic under a null hypothesis for an overall test covering a whole genome.

composite interval mapping (CIM) model

extension of interval mapping
selected markers included as cofactors
to control genetic variation of linked or unlinked QTL beyond interval of interest

model 1 QTL on interval between M_i and M_{i+1}

$$y_j = \mu + b^* x_j^* + \sum_k b_k x_{jk} + e_j$$

x_j^* = putative QTL genotype
x_{jk} = marker genotypes for genetic background
(how to select markers as cofactors?)

hypotheses at position λ on genome
$H_0 : b^* = 0$ and $H_1 : b^* \neq 0$
can be performed at any genome position
systematic strategy to search for QTL
test statistic is almost independent for each interval
roughly test QTL cluster in that interval only

CIM likelihood analysis

likelihood function $L(b^*, \mathbf{B}, \sigma^2)$

$$= \prod_{j=1}^{n} \left[p_{1j} \phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - b^*}{\sigma} \right) + p_{0j} \phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma} \right) \right]$$

other stuff: $\mathbf{X}_j\mathbf{B} = \mu + \sum_k b_k x_{jk}$
maximum likelihood estimates of parameters found by same method as interval mapping

posterior probability that $x_j^* = 1$

$$P_j = \frac{p_{1j} \phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - b^*}{\sigma} \right)}{p_{1j} \phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - b^*}{\sigma} \right) + p_{0j} \phi(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma})}$$

vector notation: $\mathbf{Y} = \{y_j\}_{n\times 1}$, $\mathbf{P} = \{P_j\}_{n\times 1}$
vector or matrix transposition denoted by "T"
sum of probabilities $P_\cdot = \sum_{j=1}^{n} P_j$

QTL effect b^*

differentiate log-lilelihood with respect to b^*

$$\frac{\partial \ln L}{\partial b^*} = \sum_{j=1}^{n} P_j \frac{[y_j - \mathbf{X}_j\mathbf{B} - b^*]}{\sigma^2}$$

setting this derivative to zero provides

$$\sum_{j=1}^{n} P_j(y_j - \mathbf{X}_j\mathbf{B} - b^*) = 0$$

solution for QTL effect (Zeng 1994)

$$\begin{aligned} \hat{b}^* &= \sum_{j=1}^{n} (y_j - \mathbf{X}_j\hat{\mathbf{B}})P_j/P_\cdot \\ &= (\mathbf{Y} - \mathbf{X}\mathbf{B})^\mathsf{T}\mathbf{P}/P_\cdot \end{aligned}$$

$\mathbf{B}$: vector of mean μ and cofactors b_k

differentiate log-likelihood with respect to $\mathbf{B}$

$$\frac{\partial \ln L}{\partial \mathbf{B}} = \sum_{j=1}^{n} \frac{P_j\mathbf{X}_j^\mathsf{T}(y_j - \mathbf{X}_j\mathbf{B} - b^*) + (1-P_j)\mathbf{X}_j^\mathsf{T}(y_j - \mathbf{X}_j\mathbf{B})}{\sigma^2}$$

set to zero using matrix notation (normal equations)

$$\mathbf{X}^\mathsf{T}(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \mathbf{X}^\mathsf{T}\mathbf{P}b^*$$

$$\hat{\mathbf{B}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{Y} - \mathbf{P}\hat{b}^*)$$

variance σ^2

$$\frac{\partial \ln L}{\partial \sigma^2} = \sum_{j=1}^{n} \frac{P_j(y_j - \mathbf{X}_j\mathbf{B} - b^*)^2 + (1-P_j)(y_j - \mathbf{X}_j\mathbf{B})^2}{2\sigma^4} - \frac{n}{2\sigma^2}$$

set to zero and solve for variance

$$\hat{\sigma}^2 = [(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\mathsf{T}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) - \hat{b}^{*2}c]/n$$

Hypothesis Test

hypotheses at position λ on genome
$H_0 : b^* = 0$ and $H_1 : b^* \neq 0$.

likelihood function under null hypothesis

$$L(b^* = 0, \mathbf{B}, \sigma^2) = \prod_{j=1}^{n} \phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma}\right)$$

has maximum likelihood estimates

$$\hat{\hat{\mathbf{B}}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$$

$$\hat{\hat{\sigma}}^2 = (\mathbf{Y} - \mathbf{X}\hat{\hat{\mathbf{B}}})^\mathsf{T}(\mathbf{Y} - \mathbf{X}\hat{\hat{\mathbf{B}}})/n$$

likelihood ratio (LR) test statistic

$$\mathsf{LR} = -2\ln\frac{L(b^* = 0, \hat{\hat{\mathbf{B}}}, \hat{\hat{\sigma}}^2)}{L(\hat{b}^*, \hat{\mathbf{B}}, \hat{\sigma}^2)}$$

$$\mathsf{LOD} = -\log_{10}\frac{L(b^* = 0, \hat{\hat{\mathbf{B}}}, \hat{\hat{\sigma}}^2)}{L(\hat{b}^*, \hat{\mathbf{B}}, \hat{\sigma}^2)}$$

Analysis in an F_2 population

mixture model includes additive & dominance effects

$$y_j = \mu + a^*x_j^* + d^*z_j^* + \sum_k a_k x_{jk} + \sum_k d_k z_{jk} + e_j$$

where

a^* = additive effect of the putative QTL

d^* = dominance effect of the putative QTL

$$x_j^* = \begin{cases} 2 & \text{if the QTL genotype is } QQ \\ 1 & \text{if the QTL genotype is } Qq \\ 0 & \text{if the QTL genotype is } qq \end{cases}$$

$$z_j^* = \begin{cases} 1 & \text{if the QTL genotype is } Qq \\ 0 & \text{if the QTL genotype is } QQ \text{ or } qq \end{cases}$$

F_2 probability of recombination

distribution of QTL genotype x_j^* (trinomial)

$$p_{kj} = Prob(x_j^* = k | M_i, M_{i+1}, \theta) \quad k = 0, 1, 2.$$

$\theta = r_{M_1Q}/r_{M_1M_2}$
r_{M_1Q} is recombination between M_1 and Q
$r_{M_1M_2}$ is recombination between M_1 and M_2
double recombination is ignored

	QTL genotype		
Marker genotype	QQ (2)	Qq (1)	qq (0)
$M_1M_1M_2M_2$	1	0	0
$M_1M_1M_2m_2$	$1-\theta$	θ	0
$M_1M_1m_2m_2$	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2
$M_1m_1M_2M_2$	θ	$1-\theta$	0
$M_1m_1M_2m_2$	$\eta\theta(1-\theta)$	$1-2\eta\theta(1-\theta)$	$\eta\theta(1-\theta)$
$M_1m_1m_2m_2$	0	$1-\theta$	θ
$m_1m_1M_2M_2$	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$
$m_1m_1M_2m_2$	0	θ	$1-\theta$
$m_1m_1m_2m_2$	0	0	1

$$\eta = \frac{r^2_{M_1M_2}}{(1-r_{M_1M_2})^2+r^2_{M_1M_2}}$$

F_2 likelihood function

$$L(a^*, d^*, \mathbf{B}, \sigma^2) = \prod_{j=1}^{n} \left[p_{2j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - 2a^*}{\sigma}\right) \right.$$

$$\left. + p_{1j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - a^* - d^*}{\sigma}\right) + p_{0j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma}\right) \right]$$

maximum likelihood estimates

$$\hat{a}^* = (\mathbf{Y} - \mathbf{XB})^\mathsf{T}\mathbf{P}_2/(2 \times 1^\mathsf{T}\mathbf{P}_2)$$

$$\hat{d}^* = (\mathbf{Y} - \mathbf{XB})^\mathsf{T}\mathbf{P}_1/(1^\mathsf{T}\mathbf{P}_1) - \hat{a}^*$$

$$\hat{\mathbf{B}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{Y} - (2\mathbf{P}_2 + \mathbf{P}_1)\hat{a}^* - \mathbf{P}_1\hat{d}^*)$$

$$\hat{\sigma}^2 = \frac{1}{n}[(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\mathsf{T}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) - 4(1^\mathsf{T}\mathbf{P}_2)\hat{a}^{*2} - (1^\mathsf{T}\mathbf{P}_1)(\hat{a}^* + \hat{d}^*)^2]$$

$$P_{2j} = \frac{p_{2j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - 2a^*}{\sigma}\right)}{p_{2j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - 2a^*}{\sigma}\right) + p_{1j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - a^* - d^*}{\sigma}\right) + p_{0j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma}\right)}$$

$$P_{1j} = \frac{p_{1j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - a^* - d^*}{\sigma}\right)}{p_{2j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - 2a^*}{\sigma}\right) + p_{1j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - a^* - d^*}{\sigma}\right) + p_{0j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma}\right)}$$

F_2 test statistic

$$\text{LOD}(\theta) = -\log_{10}\frac{L(a^* = 0, d^* = 0, \hat{\hat{\mathbf{B}}}, \hat{\hat{\sigma}}^2)}{L(\hat{a}^*, \hat{d}^*, \hat{\mathbf{B}}, \hat{\sigma}^2)}$$

under the hypotheses

$$H_0: \ a^* = 0, d^* = 0 \ \text{ and } \ H_1: \ a^* \neq 0, d^* \neq 0$$

maximum likelihood analysis
similar to IM but more complicated
depends on genetic model and experimental design
general formulae available (Kao Zeng 1997)

simulation example from Zeng (1994)

4 "chromosomes" $\times$ 16 markers, 10 cM intervals
simulated for 300 offspring from backcross population
trait affected by 10 QTLs (see figure 7.3)
QTLs account for 70% of phenotypic variance

trait = sum of effects of QTLs plus environment
environment is normally distributed
with variance rescaled so $h^2 = .7$

three models (I, II, III)

I: all other markers used as cofactors (CIM)
II: only unlinked markers used as cofactors
III: interval mapping (IM) analysis

three models (I, II, III)

I: all other markers used as cofactors (CIM)
II: only unlinked markers used as cofactors
III: interval mapping (IM) analysis

test statistic I (composite interval mapping)
roughly independent between intervals
correctly identified 6 largest QTL with high precision

II & III lack interval test property
affected by linked QTL, biased mapping
II has higher power to identify association
than I, III (because of Property 2)

Marker selection

Which markers should be added?
no simple solution:
depends on number and positions of QTL

too few markers:
–bias in estimates of position, effect
–high residual genetic variation
too many markers:
–reduce the power of analysis to detect QTL

tuning parameters for CIM

n_p = number of markers as cofactors
W_s = width of testing window
adjusting n_p and W_s changes model

generally want $n_p < 2\sqrt{n} << n$ (Jansen Stam 1994)
or determine n_p automatically by F-to-enter or
F-to-drop criterion in stepwise regression analysis
want $W_s >$10,15cM: depends on sample size

QTL Cartographer
http://statgen.ncsu.edu/qtlcart/cartographer.html

three step procedure:
1: (cofactor step) select n_p markers significantly associated with trait using (forward or backward) stepwise regression
2: (mapping step) for each interval, pick 2 markers as testing window, at least W_s cM beyond testing interval (one for each direction)
3: (fitting step) fit model using subset of n_p markers outside the testing window as cofactors

rule of thumb

–decide n_p based on stepwise regression ($SRmapqtl$): use F-to-enter (forward) or F-to-drop (backward) with level $\alpha = 0.01$
–start with large w_s, then gradually decrease, but do not let peaks drop much