Module 17: Advanced QTL Mapping
Zhao-Bang Zeng, Brian S. Yandell
Presentation Schedule

Monday

Unit	pages	title
1 ZBZ	2-24	Multiple interval mapping (MIM)
2 ZBZ	25-36	Score statistic to aid MIM model selection
lunch		
A BSY	75-93	Overview: multiple QTL approaches
B BSY	94-103	R/qtl software demo

Tuesday

Unit	pages	title
C BSY	114-136	Bayesian interval mapping (BIM)
D BSY	103-113	R/qtlbim software demo
lunch		
E BSY	137-147	Data examples in detail
	148-157	Multiple traits and co-mapping
3 ZBZ	37-57	Multiple trait analysis
Wednesday		
4 ZBZ	58-74	Gene expression QTL (eQTL) analysis
F BSY	158-170	eQTL Tools
	171-204	Causal networking

Slide 3

Multiple Interval Mapping

Zhao-Bang Zeng

Department of Statistics

North Carolina State University

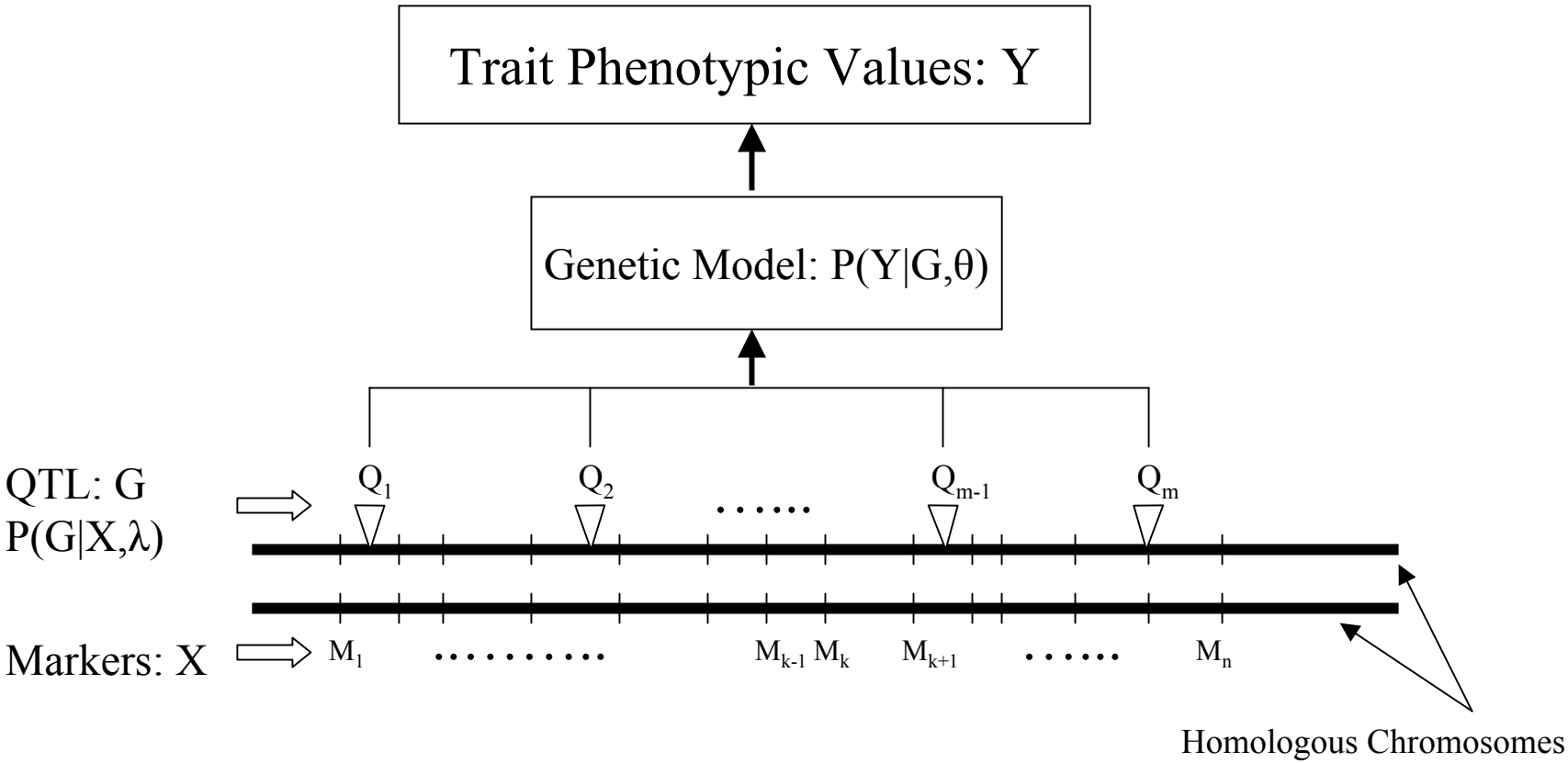
Slide 4

Purpose of Multiple Interval Mapping

- Simultaneously search and map multiple QTL: more powerful and accurate, though computationally intensive.
- Study QTL epistasis: Epistasis is a context-dependent phenomenon. It is important to study the epistasis in its totality (as much as possible) jointly with QTL main effects.
- Comprehensive estimation of genetic architecture of quantitative traits: number, positions, effects and interactions of QTL; Partition of genetic variance due to each QTL effect; ...
- Produce efficient prediction function for marker-assisted selection, pharmocogenetics jointly with co-factors, ...

It combines QTL mapping analysis with the study of genetic architecture of quantitative traits and marker-assisted selection.

QTL Mapping



Likelihood of Data: $P(Y,X) = P(Y|X)\,P(X) = \sum_G P(Y|G,\theta)\,P(G|X,\lambda)\,P(X|\delta)$

Slide 5

Statistical Framework

- Data:

  Y = Quantitative trait phenotype

  X = Molecular marker genotypes

- Joint probability of data:

$$P(Y, X) = P(Y|X)P(X)$$

$P(X)$: Marker analysis, including segregation and linkage analysis, linkage phase inference, and linkage map construction.

$P(Y|X)$: QTL analysis which is to analyze the relationship between markers and trait through QTL, and through this analysis to infer the genetic structure (or architecture) of quantitative traits, such as the number, genomic position, allelic frequencies, effects and interaction of QTL.

Slide 6

Marker Likelihood Analysis

The likelihood of marker data can be symbolically denoted as $P(X|\gamma, \phi, \omega)$ with parameters

γ = recombination frequencies between markers

ϕ = marker linkage phases

ω = marker linkage order

Through this likelihood analysis, we can infer γ, ϕ, ω.

Slide 7

QTL Analysis

QTL analysis contains two parts:

$$P(Y|X) = \sum_G P(Y|G)P(G|X)$$

$P(G|X)$ involves the segregation analysis of QTL given marker genotypes and is a function of QTL position (λ).

$P(Y|G)$ is a link function between QTL genotypes (G) and trait phenotypes (Y), and can be modeled as a function of QTL effect parameters (θ), such as the additive, dominance and epistatic effects of QTL and any other parameters that link QTL genotypes to trait phenotypes.

Together, λ and θ represent the genetic architecture parameters of quantitative traits.

Thus, $P(Y|X, \theta, \lambda) = \sum_G P(Y|G, \theta)P(G|X, \lambda)$

Slide 8

MIM model

For m putative QTL, the multiple interval mapping model (for a backcross population) is defined by

$$y_i = \mu + \sum_{r=1}^{m} \alpha_r x_{ir}^* + \sum_{r \neq s \in (1,\cdots,m)}^{t} \beta_{rs}(x_{ir}^* x_{is}^*) + e_i$$

where

- y_i is the phenotypic value of individual i;
- i indexes individuals of the sample: $i = 1, 2, \cdots, n$;
- μ is the mean of the model;
- α_r is the marginal effect of putative QTL r,
- x_{ir}^* is a coded variable denoting the genotype of putative QTL r (defined by $1/2$ or $-1/2$ for the two genotypes), which is unobserved but can be inferred from maker data in sense of

Slide 9

probability;

- β_{irs} is the epistatic effect between putative QTL r and s;
- $r \neq s \in (1, \cdots, m)$ denotes a subset of QTL pairs that each shows a significant epistatic effect to avoid the over-parameterization that could result when using all pairs;
- m is the number of putative QTL chosen based on either their significant marginal effects or significant epistatic effects;
- t is the number of significant pairwise epistatic effects;
- e_i is a residual effect of the model assumed to be normally distributed with mean zero and variance σ^2.

Slide 10

Likelihood

The likelihood function of the data given the model is a mixture of normal distributions

$$L(\mathbf{E}, \mu, \sigma^2) = \prod_{i=1}^{n} \sum_{j=1}^{2^m} P(G_j|X_i) P(Y_i|G_j)$$

$$= \prod_{i=1}^{n} \left[\sum_{j=1}^{2^m} p_{ij} \phi\left(y_i | \mu + \mathbf{G}_j \mathbf{E}, \sigma^2\right) \right]$$

- p_{ij} is the probability of each multilocus genotype conditional on marker data;
- $\mathbf{E}$ is a vector of QTL parameters (α's and β's);
- $\mathbf{G}_j$ is a vector specifying the configuration of x^*'s associated with each α and β for the jth QTL genotype;
- $\phi(y|\mu, \sigma^2)$ denotes a normal density function

Slide 11

Calculate $P(G|X)$

Example: For a backcross population, let a QTL (g) be located between two markers (x_1 and x_2). Let the recombination frequency between x_1 and g be r_1, that between g and x_2 be r_2, and between x_1 and x_2 be r_{12}. The conditional probability $P(g|x_1, x_2)$ is:

Marker genotype		QTL genotype, g	
x_1, x_2	Freq	1	0
11	$\frac{1-r_{12}}{2}$	$\frac{(1-r_1)(1-r_2)}{1-r_{12}} \approx 1$	$\frac{r_1 r_2}{1-r_{12}} \approx 0$
10	$\frac{r_{12}}{2}$	$\frac{(1-r_1)r_2}{r_{12}} \approx 1-\lambda$	$\frac{r_1(1-r_2)}{r_{12}} \approx \lambda$
01	$\frac{r_{12}}{2}$	$\frac{r_1(1-r_2)}{r_{12}} \approx \lambda$	$\frac{(1-r_1)r_2}{r_{12}} \approx 1-\lambda$
00	$\frac{1-r_{12}}{2}$	$\frac{r_1 r_2}{1-r_{12}} \approx 0$	$\frac{(1-r_1)(1-r_2)}{1-r_{12}} \approx 1$

where $\lambda = r_1/r_{12}$

Slide 12

For multiple QTL in multiple marker intervals

$$P(G|X, \lambda) = \prod_{r=1}^{m} P(G_r|X, \lambda_r)$$

For more complicated cases, a hidden Markov model is generally used for the calculation.

Slide 13

EM algorithm

Take derivative of $logL$ with respect to each model parameter (μ, E_r, σ^2) and equate the derivatives to zero, we can obtain a series of functions with the following algorithm (called EM algorithm).

EM is an iterative procedure involving an E-step (Expectation) and M-step (Maximization) in each iteration. In the $[t+1]$th iteration,

E-step:

$$\pi_{ij}^{[t+1]} = \frac{p_{ij}\phi(y_i|\mu^{[t]} + \mathbf{D}_j\mathbf{E}^{[t]}, \sigma^{2[t]})}{\sum_{j=1}^{2^m} p_{ij}\phi(y_i|\mu^{[t]} + \mathbf{D}_j\mathbf{E}^{[t]}, \sigma^{2[t]})}$$

Slide 14

M-step:

$$E_r^{[t+1]} = \frac{\sum_i \sum_j \pi_{ij}^{[t+1]} G_{jr}[(y_i - \mu^{[t]}) - \sum_{s=1}^{r-1} G_{js}E_s^{[t+1]} - \sum_{s=r+1}^{m+t} G_{js}E_s^{[t]}]}{\sum_i \sum_j \pi_{ij}^{[t+1]} D_{jr}^2}$$

$$\mu^{[t+1]} = \frac{1}{n}\sum_i \left(y_i - \sum_j \sum_r \pi_{ij}^{[t+1]} G_{jr} E_r^{[t+1]} \right)$$

$$\sigma^{2[t+1]} = \frac{1}{n}\left[\sum_i (y_i - \mu^{[t+1]})^2 - 2\sum_i (y_i - \mu^{[t+1]}) \sum_j \sum_r \pi_{ij}^{[t+1]} G_{jr} E_r^{[t+1]} \right.$$
$$\left. + \sum_r \sum_s \sum_i \sum_j \pi_{ij}^{[t+1]} G_{jr} G_{js} E_r^{[t+1]} E_s^{[t+1]} \right]$$

Slide 15

Conditional QTL genotype probability π_{ij}

Probability of QTL genotype conditional on marker genotype:

$$P(G_j|X_i) = p_{ij}$$

Conditional density of trait phenotype given QTL genotype:

$$P(Y_i|G_j) = \phi(y_i|\mu + \mathbf{G}_j\mathbf{E}, \sigma^2)$$

Probability of QTL genotype conditional on marker genotype and trait phenotype:

$$P(G_j|X_i, Y_i) = \pi_{ij} = \frac{P(G_j|X_i)P(Y_i|G_j)}{\sum_G P(G_j|X_i)P(Y_i|G_j)}$$

$$= \frac{p_{ij}\phi(y_i|\mu + \mathbf{G}_j\mathbf{E}, \sigma^2)}{\sum_j p_{ij}\phi(y_i|\mu + \mathbf{G}_j\mathbf{E}, \sigma^2)}$$

Slide 16

Advantage and disadvantage of EM algorithm

- Advantages: Generally stable i.e. likelihood increases steadily and do not diverge; easy to implement; easy to generalize to more complicated situation (such as multiple QTL with epistasis)
- Disadvantages: Very slow to converge to optimum

Slide 17

GEM-NR algorithm

Speed up the converging process while preserving the property of stability. Combine EM (generalized EM) with a quadratic algorithm in the maximization step (Newton-Ralphson algorithm).

In the E-step, still calculate the conditional probability ($\pi's$) of QTL genotypes given current estimates of model parameters.

The M-step is obtained by

$$\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)} \left. \frac{\partial Q(\theta|\theta^{(t)})^{-1}}{\partial\theta \cdot \partial\theta'} \right|_{\theta^{(t)}} \left. \frac{\partial Q(\theta|\theta^{(t)})}{\partial\theta} \right|_{\theta^{(t)}}$$

where $Q(\theta|\theta^{(t)})$ is the expected complete-data loglikelihood conditional on the current estimates of parameter values.

Slide 18

Comments on the algorithms

- GEM-NR algorithm (Generalized EM with Newton-Ralphson algorithm):

$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) =$$

$$(\theta^{(t+1)} - \theta^{(t)})' \left. \frac{\partial Q(\theta|\theta^{(t)})}{\partial\theta} \right|_{\theta^{(t)}} - \frac{1}{2}(\theta^{(t+1)} - \theta^{(t)})' \left. \frac{\partial Q(\theta|\theta^{(t)})}{\partial\theta \cdot \partial\theta'} \right|_{\theta^{(t)}} (\theta^{(t+1)} - \theta^{(t)})$$

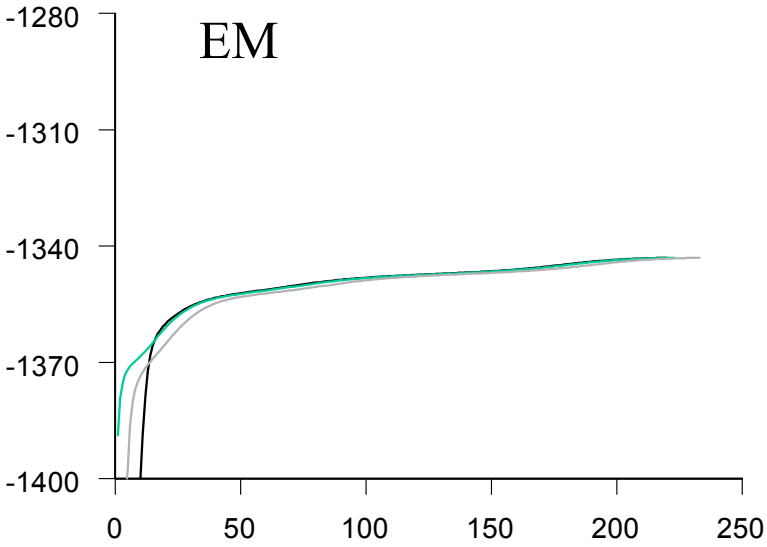
- EM algorithm (Expectation-Maximization algorithm):

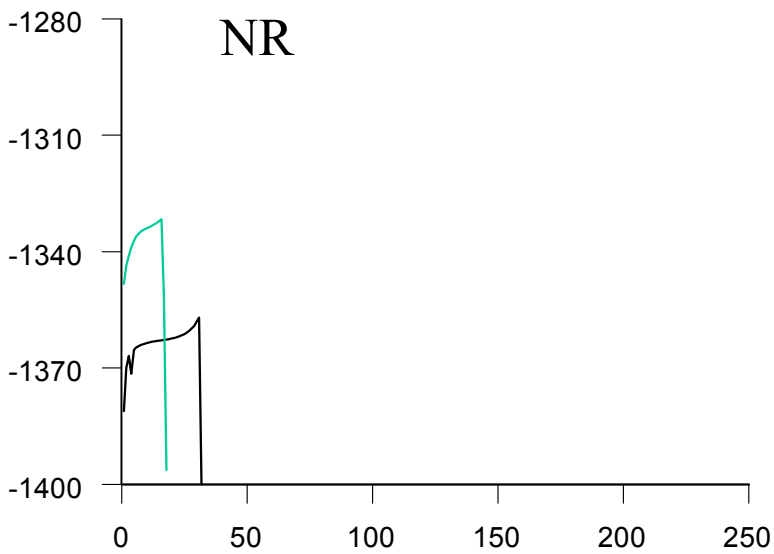
$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) = (\theta^{(t+1)} - \theta^{(t)})' \left. \frac{\partial Q(\theta|\theta^{(t)})}{\partial\theta} \right|_{\theta^{(t)}}$$

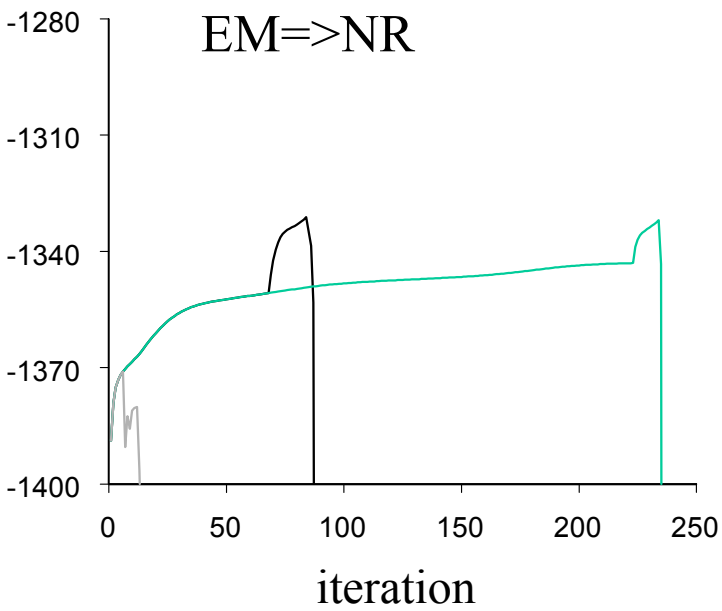
- NR algorithm (Newton-Ralphson algorithm):

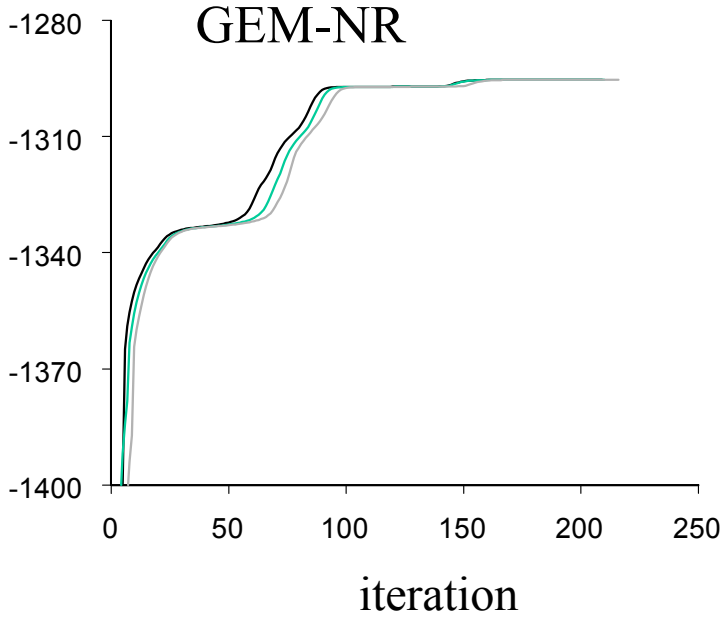
$$l(\theta^{(t+1)}) - l(\theta^{(t)}) =$$

$$(\theta^{(t+1)} - \theta^{(t)})' \left. \frac{\partial l(\theta)}{\partial\theta} \right|_{\theta^{(t)}} - \frac{1}{2}(\theta^{(t+1)} - \theta^{(t)})' \left. \frac{\partial l(\theta)}{\partial\theta \cdot \partial\theta'} \right|_{\theta^{(t)}} (\theta^{(t+1)} - \theta^{(t)})$$


EM
-1280
-1310
-1340
-1370
-1400
0
50
100
150
200
250


NR
-1280
-1310
-1340
-1370
-1400
0
50
100
150
200
250


EM=>NR
-1280
-1310
-1340
-1370
-1400
0
50
100
150
200
250
iteration


GEM-NR
-1280
-1310
-1340
-1370
-1400
0
50
100
150
200
250
iteration

Dealing with many QTL

- m $QTL \longrightarrow 2^m$ possible mixture components.
- This can be prohibitive for efficient numerical analysis.
- But most genotypes have negligible probabilities.
- Can we skip these evaluations?

Slide 19

Practical implementation of MIM algorithm:

Select a subset of "significant" mixture components for each individual for evaluation: (1) set any $p_{ij} < \delta$ $(= 0.005)$ to zero (drop them); (2) Sum of "significant" $p_{ij} > 0.95$ (adjust δ if needed); (3) normalize "significant" probs: $\sum_j p_{ij} = 1$.

Number of "significant" components $\sim$ 10-100, depending on marker density, number and position of QTL. It ha negligible loss of accuracy of likelihood evaluation as compared to no selection.

Conditional likelihood ratio test

Test for each QTL effect E_r conditional on other QTL effects:

$$LOD = \log_{10} \frac{L(all\ E_s \neq 0)}{L(E_r = 0,\ all\ other\ E_s \neq 0)}$$

Slide 20

It can proceed as above if we have positions of m putative QTL and selected $m + t$ 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.

Slide 21

Interactive model selection procedure in Window QTL Cartographer

1. Initial model: select *New Model* to use an automatical stepwise selection procedure, CIM or stepwise marker selection for initial model selection.
2. Search for new QTL: select *Refine Model => Search for New QTL => Search for QTL* to scan the genome for new QTL and determine whether to accept QTL based on the selected criterion.
3. Search for QTL epistasis: select *Refine Model => Search for New QTL => Search for Epistasis* to search for epistatic effects among identified QTL based on the selected criterion (better to use a lower criterion such as AIC).
4. Re-evaluate QTL effects: select *Refine Model => Testing for Existing QTL* to re-evaluate the significance of each QTL effect

Slide 22

in the model based on the selected criterion. This procedure can remove non-significant effects from the model.

5. Optimize QTL positions: select *Refine Model => Optimizing QTL Position* to optimize QTL position estimates in the current model. QTL position is optimized one by one in a sequential order.
6. Procedure 2 to 5 can be repeated if needed.
7. Selection criterion: Quite a few provided (such as BIC and AIC).
8. Display graphic result: select *Refine Model => MIM Model Summary => Graphic Result File* to calculate and display the likelihood profile for each QTL.
9. Show MIM estimates: select *Refine Model => MIM Model Summary => Model Summary File* to show the MIM output result file. Information includes position, likelihood ratio and

Slide 23

effect of each QTL, epistatic effects of QTL, partition of the variance explained by QTL (due to main and interaction effects), estimate of genotypic values of individuals based on the model.

Slide 24

Model selection criteria

- Akaike information criterion (AIC): minimize $-2(\log L_k - k)$.
- C_p method: minimize adjusted R^2.
- Bayes information criterion (BIC): minimize $-2(\log L_k - kc(n)/2)$ with $c(n) = \log(n)$, or $c(n) = 2\log(\log n)$, or other penalty function.
- Final prediction error (FPE) method: minimize prediction error.
- Delete-one cross-validation, Delete-d cross-validation, and generalized cross-validation: different ways to implement FPE.
- Bootstrap model selection: use bootstrap resampling to implement FPE.
- Minimizing posterior predictive loss: similar to FPE in concept.

Slide 25

Residual permutation test

Suppose that we want to test the hypotheses: H_0: k QTL vs. H_1 $(k+1)$ QTL. Assume that the k QTL of H_0 is contained in the model of $(k+1)$ QTL of H_1.

Let $\hat{y}_{i|H_0}$ and $\hat{y}_{i|H_1}$ be the estimated genotypic value of individual i under H_0 and H_1, respectively.

Residual permutation sample: Let $\hat{e}_i = y_i - \hat{y}_{i|H_0}$ and $\hat{\epsilon}_i = \hat{e}_i - \bar{e}$ with $\bar{e} = \sum_i \hat{e}_i/n$. To generate a residual permutation sample $\{(X_i^*, Y_i^*)\}$, we first generate a random sample of residuals $\{\epsilon_i^*\}$ from $\{\hat{\epsilon}_j, j = 1, \cdots, n\}$ without replacement, and define $X_i^* = X_i$ and $Y_i^* = \hat{y}_{i|H_0} + \epsilon_i^*$.

Slide 26

Residual permutation test is performed as follows:

1. Draw a residual permutation sample $\{(X_i^*, Y_i^*)\}$.
2. Search for the best position in the genome (other than the positions of the k QTL) for the hypothetical $k+1$ QTL and perform the likelihood ratio test for the hypotheses.
3. Repeat step 1 and 2 for a predetermined number of times to obtain an empirical bootstrap distribution of the test statistic, T^*.
4. Reject H_0 if the test statistic in the original data exceeds $\hat{T}_\alpha$, where $\hat{T}_\alpha$ is the $(1-\alpha)$th quantile of the bootstrap distribution of T^*.

Slide 27

Estimating the variance explained by QTL

Variance explained by QTL effect E_r can be estimated as

$$\hat{\sigma}^2_{E_r} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2^m}\hat{\pi}_{ij}(G_{jr} - \bar{G}_r)^2\hat{E}_r^2$$

Covariance explained by QTL effect E_r and E_s is

$$\hat{\sigma}_{E_r,E_s} = \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{2^m}\hat{\pi}_{ij}(G_{jr} - \bar{G}_r)(G_{js} - \bar{G}_s)\hat{E}_r\hat{E}_s$$

Thus the total genetic variance explained by QTL is

$$\hat{\sigma}^2_g = \sum_r \hat{\sigma}^2_{E_r} + \sum_{r\neq s}\hat{\sigma}_{E_r,E_s}$$

Slide 28

It is convenient and may also be informative to combine the variance due to each QTL effect with half of the covariances between this QTL effect and other effects and report it as the variance component associated with this QTL effect

$$\hat{\sigma}^2_r = \hat{\sigma}^2_{E_r} + \frac{1}{2}\sum_{s\neq r}\hat{\sigma}_{E_r,E_s}$$

Slide 29

Estimation of genotypic values

The genotypic value of an individual can be estimated as:

$$\hat{y}_i = \hat{\mu} + \sum_{j=1}^{2^m} \sum_{r=1}^{m+t} \hat{\pi}_{ij} G_{jr} \hat{E}_r$$

To predict the genotypic values of quantitative traits based on marker information only (e.g. in cross-prediction; early selection), we need to use

$$\hat{y}_i = \hat{\mu} + \sum_{j=1}^{2^m} \sum_{r=1}^{m+t} \hat{p}_{ij} G_{jr} \hat{E}_r$$

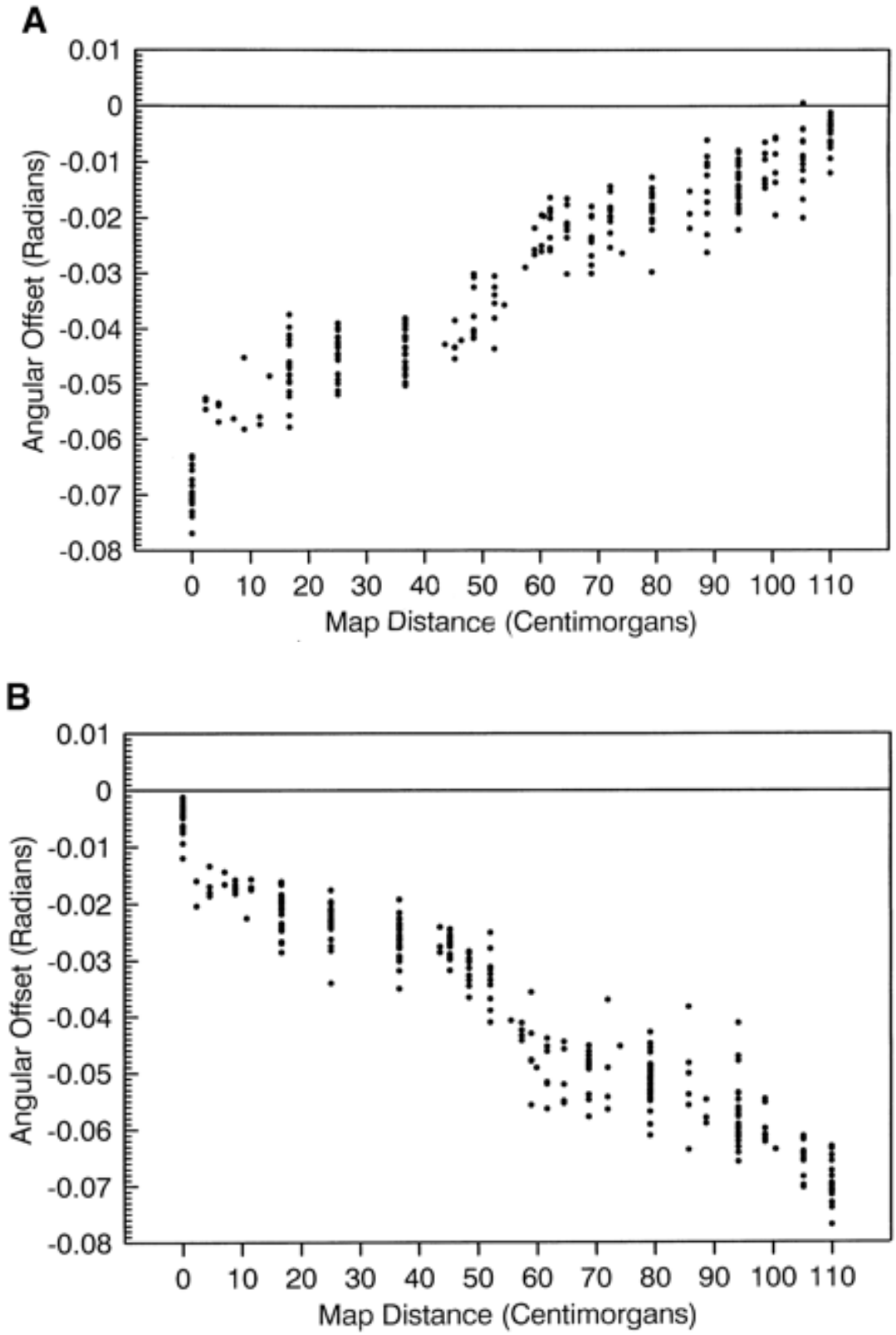
as $\hat{\pi}_{ij}$ is a function of phenotype y_i which is unavailable in early selection.

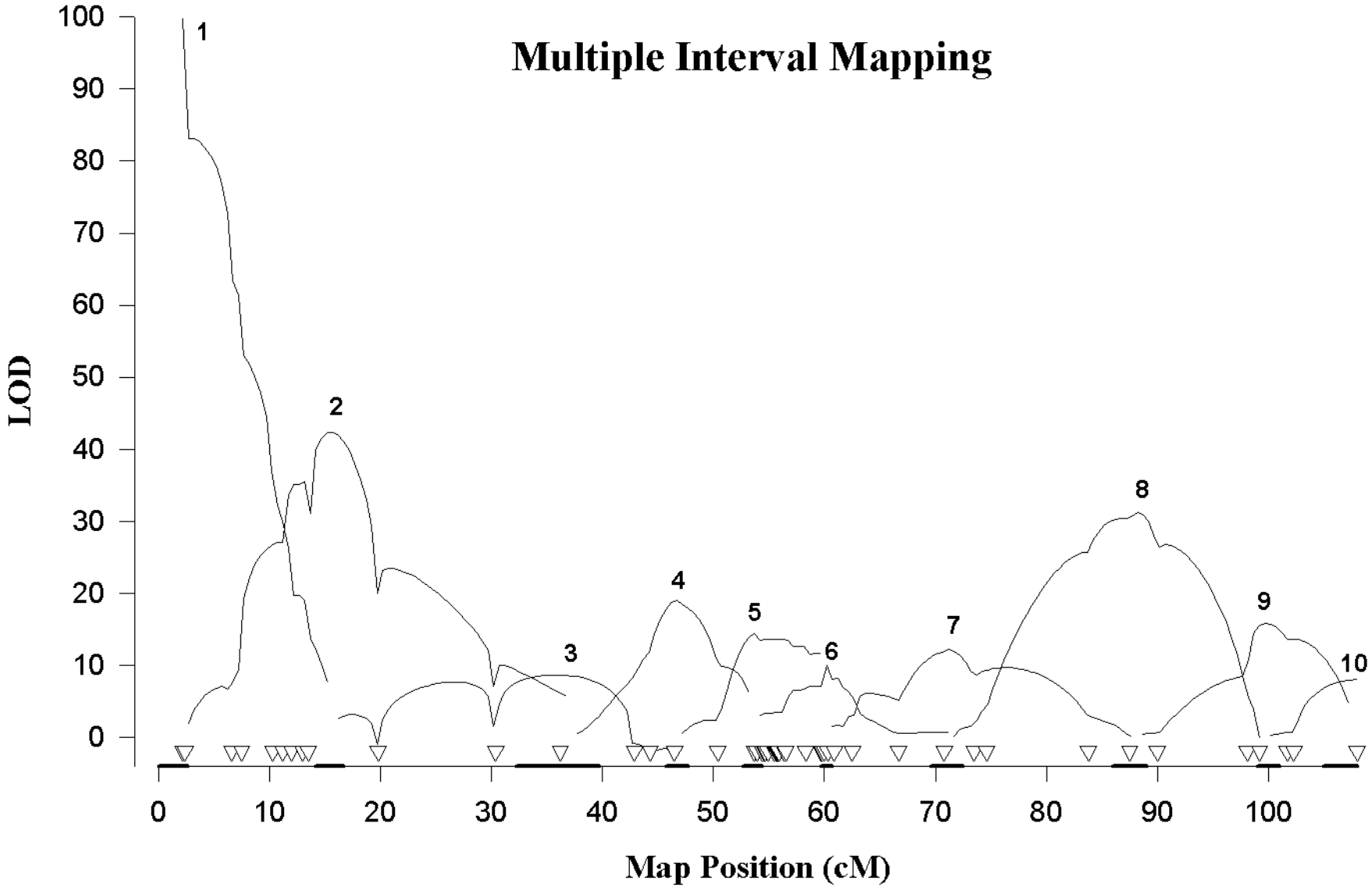
These can be used for marker-assisted selection.

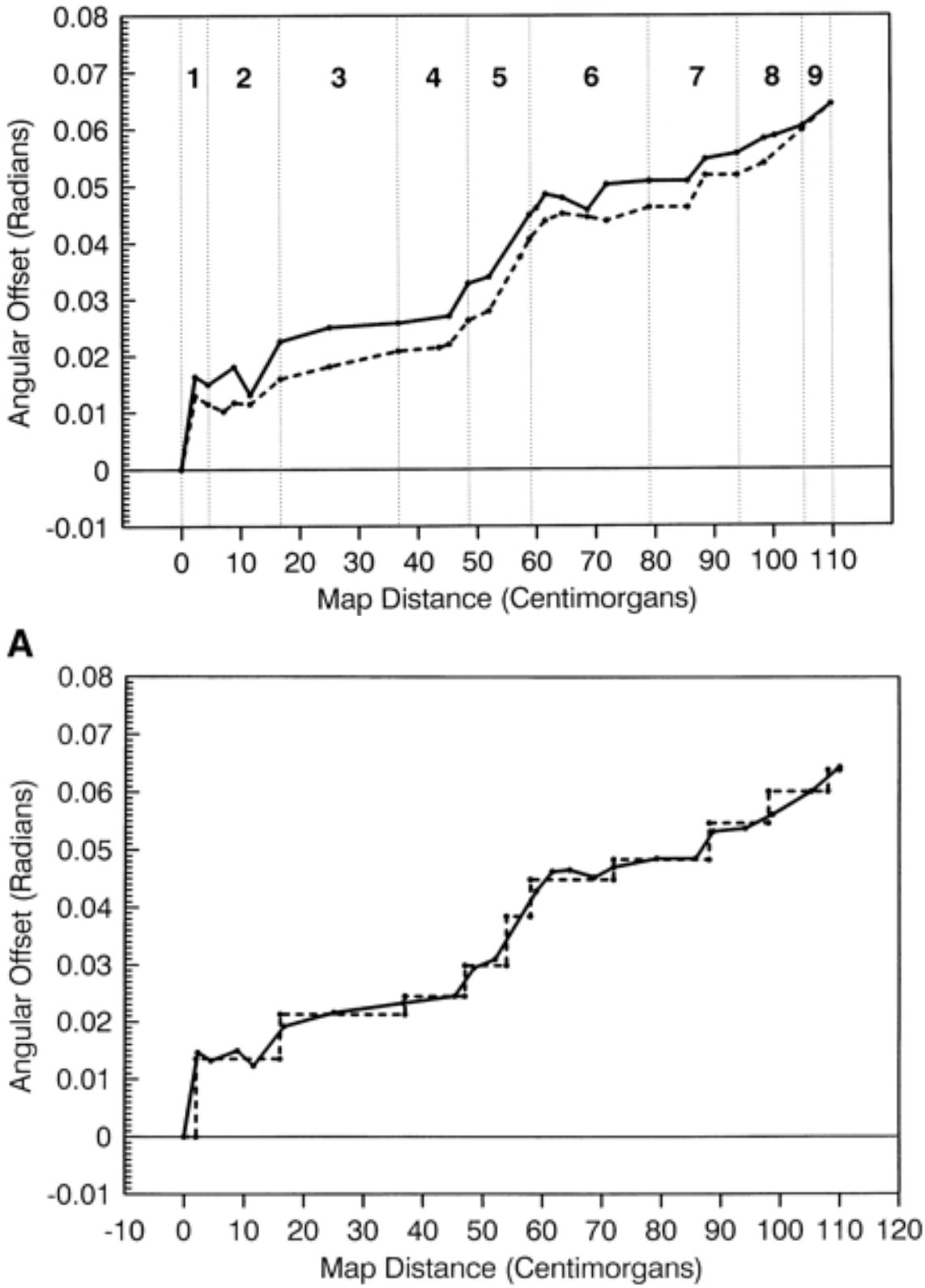
Slide 30

MIM example: Genetic architecture of wing size of *Drosophila melanogaster* on chromosome 2 (Ken Weber):

- Population: 701 recombinant inbred lines originating from a cross between high and low selected lines on wing size. Only QTL on chromosome 2 are segregating in the population, and other chromosomes are identical for all RIL.
- Trait: wing size measured in radian in an allometric analysis.
- 10 QTL are identified by MIM analysis. There is a good agreement between the sum of estimated additive effects of QTL and the observed parental genotype difference.
- There are some significant additive by additive interaction effects between QTL. The interaction pattern is complex.
- Together, 10 additive and 14 additive by additive QTL effects explain 95% of the total variance in the population.

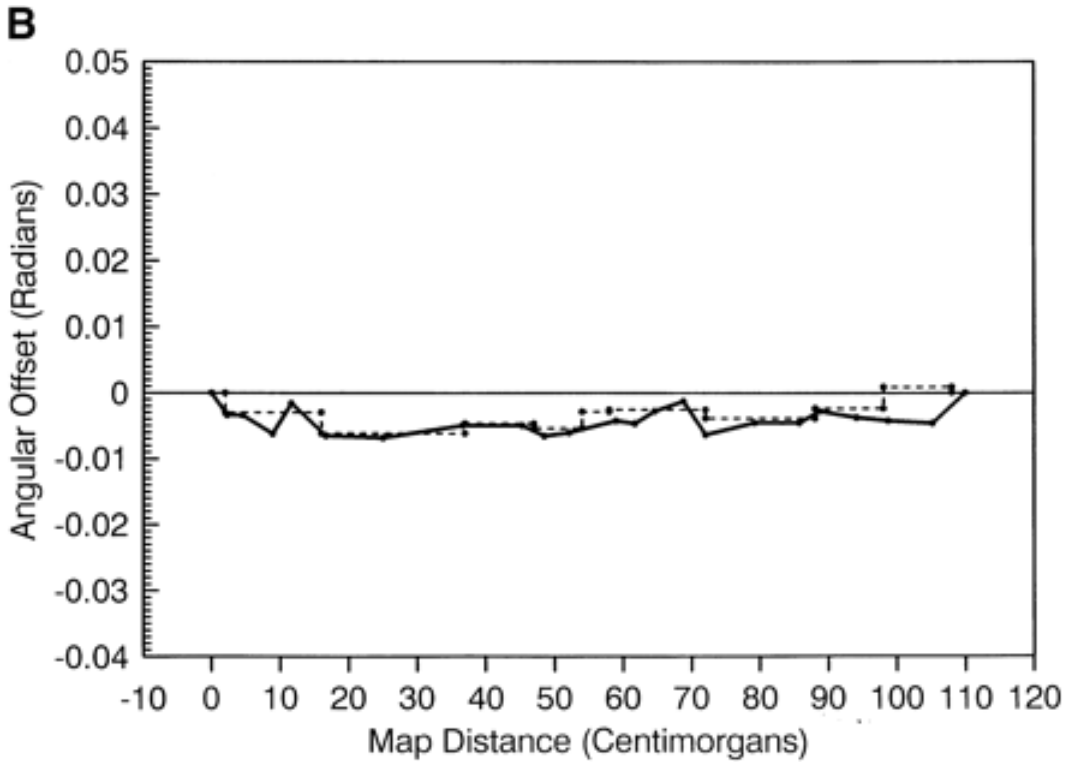

A
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
Angular Offset (Radians)
0 10 20 30 40 50 60 70 80 90 100 110
Map Distance (Centimorgans)
B
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
Angular Offset (Radians)
0 10 20 30 40 50 60 70 80 90 100 110
Map Distance (Centimorgans)


Multiple Interval Mapping
LOD
Map Position (cM)
0
10
20
30
40
50
60
70
80
90
100
1
2
3
4
5
6
7
8
9
10


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
Angular Offset (Radians)
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100
110
Map Distance (Centimorgans)

A

0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
Angular Offset (Radians)
-10
0
10
20
30
40
50
60
70
80
90
100
110
120
Map Distance (Centimorgans)


B
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
Angular Offset (Radians)
-10
0
10
20
30
40
50
60
70
80
90
100
110
120
Map Distance (Centimorgans)

Slide 31

Statistics of several models in the backward stepwise elimination process for selecting significant QTL epistatic effects

Model	QTL #	Epis #	2ln(Likelihood)	R^2
M_{45}	10	45	7094	0.952
M_{14}	10	14	7088	0.954
M_{13}	10	13	7075	0.954
M_5	10	5	7058	0.953
M_4	10	4	7041	0.953
M_3	10	3	7008	0.952
M_0	10	0	6997	0.947

Slide 32

Estimates of QTL positions and additive effects

QTL	Posi (cM)	Effect ($\times 10^{-2}$)	Effect (%)
1	2	1.35	20.9
2	16	0.78	12.2
3	37	0.32	4.9
4	47	0.54	8.4
5	54	0.86	13.4
6	58	0.64	9.9
7	72	0.35	5.4
8	88	0.63	9.8
9	98	0.55	8.5
10	108	0.37	5.8
Total			99.1

Slide 33

Estimates of QTL additive by additive interaction effects

QTL pair	Effect ($\times 10^{-2}$)	QTL pair	Effect ($\times 10^{-2}$)
(1&2)	1.49	(2&8)	-0.22
(1&3)	-1.10	(3&4)	-1.08
(1&5)	-0.52	(3&7)	0.13
(1&9)	-0.17	(4&6)	0.33
(2&4)	1.48	(8&9)	0.73
(2&5)	0.27	(8&10)	-0.80
(2&6)	-0.36	(9&10)	0.89

Slide 34

Estimated variances and covariances of the QTL additive effects in percentage of the total phenotypic variance

QTL	1	2	3	4	5	6	7	8	9	10	Sum
1	16.6	7.0	0.9	0.2	-1.5	-1.1	-0.7	-1.6	-2.2	-1.7	15.6
2	7.0	5.6	1.1	1.1	0.6	0.3	-0.1	-0.6	-1.2	-1.0	12.8
3	0.9	1.1	0.9	1.2	1.4	1.0	0.3	0.1	-0.2	-0.2	6.4
4	0.2	1.1	1.2	2.7	3.5	2.4	0.8	0.6	-0.2	-0.3	11.8
5	-1.5	0.6	1.4	3.5	6.8	4.8	1.6	1.1	-0.2	-0.5	17.4
6	-1.1	0.3	1.0	2.4	4.8	3.7	1.2	0.9	-0.1	-0.3	12.9
7	-0.7	-0.1	0.3	0.8	1.6	1.2	1.1	1.2	0.5	0.2	6.2
8	-1.6	-0.6	0.1	0.6	1.1	0.9	1.2	3.6	2.2	1.2	8.7
9	-2.2	-1.2	-0.2	-0.2	-0.2	-0.1	0.5	2.2	2.8	1.6	3.0
10	-1.7	-1.0	-0.2	-0.3	-0.5	-0.3	0.2	1.2	1.6	1.3	0.3
Total											95.1

Slide 35

Estimated variances and covariances of the QTL additive by additive interaction effects in percentage of the total phenotypic variance

QTLs	1&2	1&3	1&5	1&9	2&4	2&5	2&6	2&8	3&4	3&7	4&6	8&9	8&10	9&10	Sum
1&2	2.4	-1.2	-0.4	0.0	-0.9	-0.2	0.2	0.1	0.2	0.0	0.0	-0.1	0.2	-0.1	-0.6
1&3	-1.2	2.7	0.8	0.0	-1.6	-0.2	0.3	0.0	-0.4	0.1	0.1	0.2	-0.3	0.1	0.8
1&5	-0.4	0.8	0.6	0.0	-1.1	-0.2	0.3	0.0	0.2	0.0	-0.1	0.1	-0.2	0.0	0.3
1&9	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
2&4	-0.9	-1.6	-1.1	0.0	4.7	0.7	-0.9	-0.2	-1.0	0.0	-0.2	-0.2	0.3	-0.1	-0.4
2&5	-0.2	-0.2	-0.2	0.0	0.7	0.2	-0.2	0.0	-0.2	0.0	0.0	-0.1	0.1	0.0	-0.1
2&6	0.2	0.3	0.3	0.0	-0.9	-0.2	0.3	0.0	0.2	0.0	-0.1	0.1	-0.1	0.0	0.1
2&8	0.1	0.0	0.0	0.0	-0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.1	-0.1	0.0	0.0
3&4	0.2	-0.4	0.2	0.0	-1.0	-0.2	0.2	0.0	1.2	-0.1	0.0	0.0	0.0	0.0	0.1
3&7	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0
4&6	0.0	0.1	-0.1	0.0	-0.2	0.0	-0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
8&9	-0.1	0.2	0.1	0.0	-0.2	-0.1	0.1	0.1	0.0	0.0	0.0	0.6	-0.6	-0.1	-0.1
8&10	0.2	-0.3	-0.2	0.0	0.3	0.1	-0.1	-0.1	0.0	0.0	0.0	-0.6	1.0	-0.4	0.1
9&10	-0.1	0.1	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.4	0.5	0.0
Total															0.3

Slide 36

.

Slide 1

Use of score statistic to aid model selection for mapping multiple QTL

Zhao-Bang Zeng

Departments of Statistics & Genetics

Bioinformatics Research Center

North Carolina State University

Email: zeng@stat.ncsu.edu

Slide 2

Multiple Interval Mapping (MIM)

We developed MIM in 1999 (Kao et al. 1999; Zeng et al. 1999) and implemented it in Windows QTL Cartographer (Wang et al. 2001-2007).

MIM is a model selection-based method for identifying multiple QTL (mostly sequentially), and based on the identification to estimate the parameters of genetic architecture including epistasis.

A few model selection criteria, such as BIC, were implemented. *However, it is still not quite clear what appropriate criteria should be used for model selection.* The criteria should take into account a number of experimental factors, such as genome size, genetic map density, informativeness of markers, and proportion of missing data.

Slide 3

MIM Model and Likelihood

Model (for m putative QTL in a backcross population):

$$y_i = \mu + \sum_{k=1}^{m} a_k x_{ik} + \sum_{k \neq l \in \{1,\ldots,m\}} \delta_{kl}\gamma_{kl} x_{ik} x_{il} + \varepsilon_i.$$

where x_{ik} is unobserved QTL genotype with known conditional probability from genetic markers and $\varepsilon_i \sim N(0,1)$. Likelihood:

$$L(\theta; v) = \prod_{i=1}^{n} \sum_{j=1}^{2^m} P(G_j|X_i) P(Y_i|G_j) = \prod_{i=1}^{n} \left[\sum_{j=1}^{2^m} p_{ij} \phi\left(y_i | \mu + \mathbf{G}_j \mathbf{E}, \sigma^2\right) \right]$$

$$l(\theta; v) = \sum_{i=1}^{n} l_i(\theta; v) = \sum_{i=1}^{n} ln \left\{ \sum_{j=1}^{2^m} p_{ij} \phi(y_i | \mu_j, \sigma^2) \right\}$$

We have worked out an efficient algorithm that combines generalized EM and Newton-Raphson method (GEM-NR) to maximize the likelihood for complex genetic models.

Slide 4

Model Search and Model Selection

There are several ways to do model selection with MIM. One way is to perform sequential search, adding one parameter at a time.

- Starting with no QTL, scan the genome and compute test statistics for adding one QTL at a time. The putative QTL position with the maximum test statistic is added to the model **if the statistic exceeds a specified threshold**. This process is repeated until no more QTL is found.
- Then we search for adding parameters for epistasis between pairs of QTL identified.
- The model can be refined iteratively by dropping insignificant parameters and searching for new parameters if necessary.

The key ingredient in the model selection is choosing an appropriate test statistic and corresponding threshold value.

Slide 5

Score Statistic

Zou *et al.* (2004 Genetics 168:2307-2316) proposed using score statistic to test QTL effect and a resampling procedure for determining the appropriate threshold.

Suppose we have identified the $m-1$ QTL with parameters η and want to test for adding the m^{th} QTL with parameter β.

Let $U(d)$ denote the score function for β, at genomic position d, evaluated at $\beta = 0$ and $\widehat{\eta}$.

$$\widehat{U_i}(d) = U_{\beta,i}(0,\widehat{\eta};d) - \left(\frac{\partial^2 l(0,\widehat{\eta};d)}{\partial\beta\partial\eta}\right)\left(\frac{\partial^2 l(0,\widehat{\eta};d)}{\partial\eta^2}\right)^{-1} U_{\eta,i}(0,\widehat{\eta};d)$$

$$U_{\beta,i}(\beta,\eta;d) = \frac{\partial l_i(\beta,\eta:d)}{\partial\beta}$$

$$U_{\eta,i}(\beta,\eta;d) = \frac{\partial l_i(\beta,\eta;d)}{\partial\eta} = (\frac{\partial l_i}{\partial\theta_1},\cdots,\frac{\partial l_i}{\partial\theta_{m-1}},\frac{\partial l_i}{\partial\mu},\frac{\partial l_i}{\sigma^2})'$$

$$\widehat{U}(d) = \sum_{i=1}^{n}\widehat{U}_i(d)$$

Slide 6

The score statistic for H_0: $\beta = 0$ against H_1: $\beta \neq 0$ at location d is

$$W(d) = \widehat{U}'(d)\widehat{V}^{-1}(d)\widehat{U}(d)$$

where $\widehat{V}(d) = \sum_{i=1}^{n}\widehat{U}_i(d)\widehat{U}_i'(d)$.

Slide 7

Resampling with Score Statistic

An efficient way to simulate conditional null distribution

1. Generate G_i, $i = 1, 2, \cdots, n$ from $N(0, 1)$.
2. Calculate $U^*(d) = \sum_{i=1}^{n} \widehat{U}_i(d) G_i$, $W^*(d) = U^{*'}(d) \widehat{V}^{-1} U^*(d)$, and $S^* = max_d W^*(d)$.
3. Repeat step 1 and 2 for N times to find ${S_k}^*$ for $k = 1, \cdots, N$.
4. Compute the $100(1-\alpha)^{th}$ percentile of $\{{S_k}^* : k = 1, \cdots, \mathrm{N}\}$ to determine the threshold value.
5. Accept the position being tested as identifying a new QTL if the observed score statistic for the position exceeds the threshold value.

Simulation

Case I:

- 9 chromosomes each with 12 markers in 10cM interval
- 8 QTL with equal effect on different chromosomes
- Heritability = 0.4
- 1000 times score statistics bootstraps
- 1000 Replications

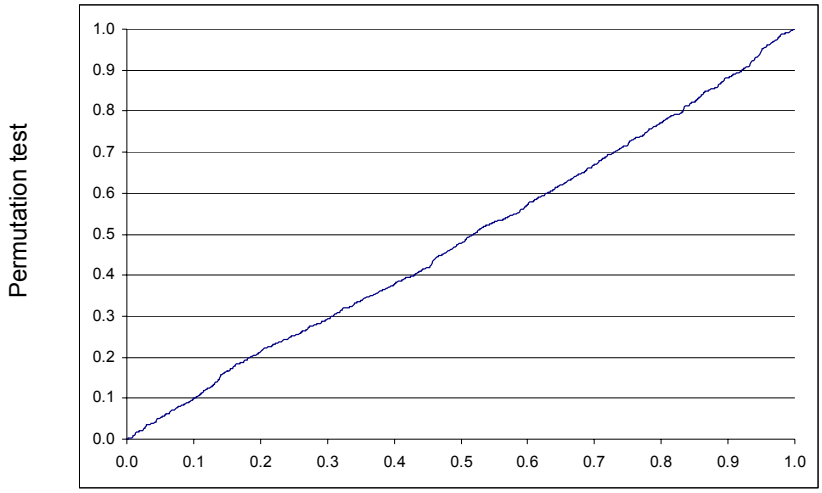
Simulation

Case II:

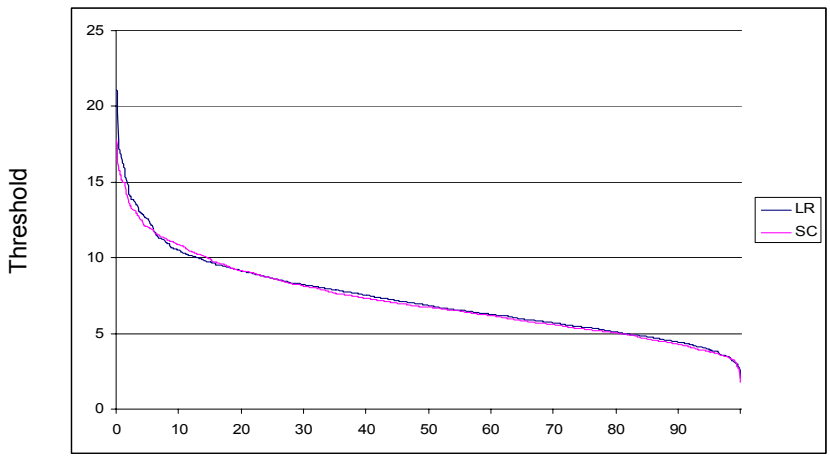
- 9 chromosomes each with 12 markers in 10cM interval
- 8 QTL with different effects and some linkage
- Heritability = 0.4
- 1000 times score statistics bootstraps
- 1000 Replications

Chromosome	1	1	2	3	3	6	7	9
Positions (cM)	27.4	90.3	49.0	32.5	88.9	9.3	70.7	63.2
QTL effect	0.503	0.670	0.798	0.590	-0.503	0.710	0.255	0.399

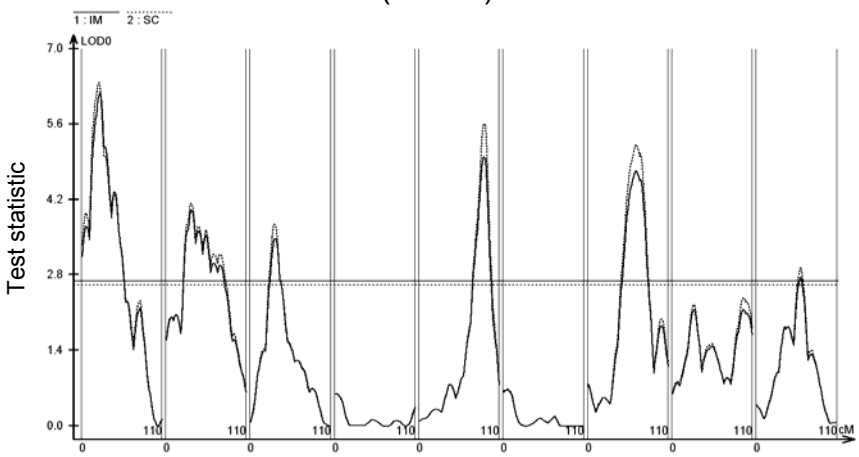
Q-Q plot of score resampling test statistic vs. likelihood ratio with permutation at null (no QTL) (Case I)



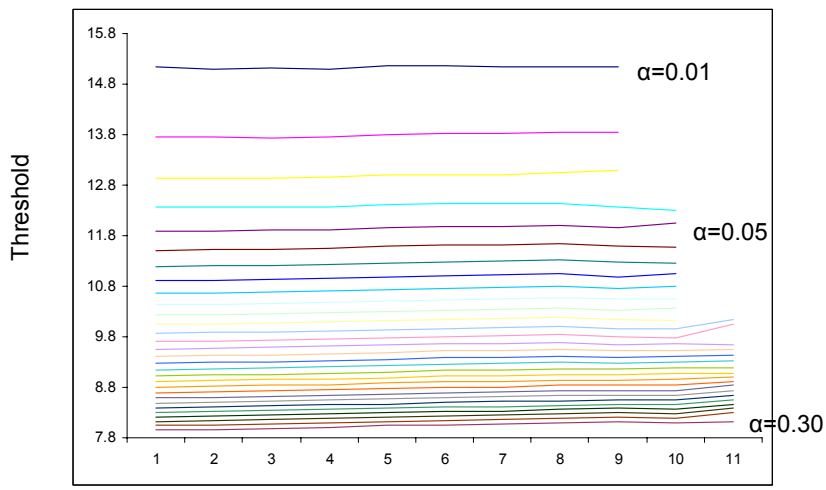
Comparison of thresholds of score statistic and Likelihood ratio with permutation (Case I)



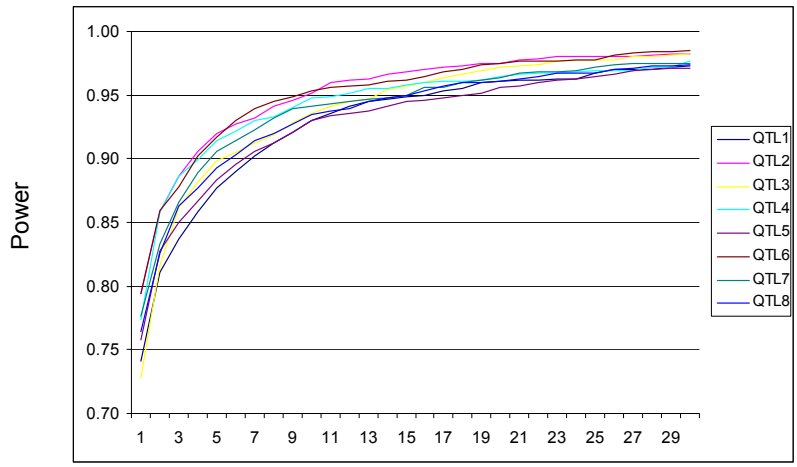
Score statistic vs likelihood ratio (Case I)



Change of threshold in indifferent steps for model selection with score resampling test statistic (case I)

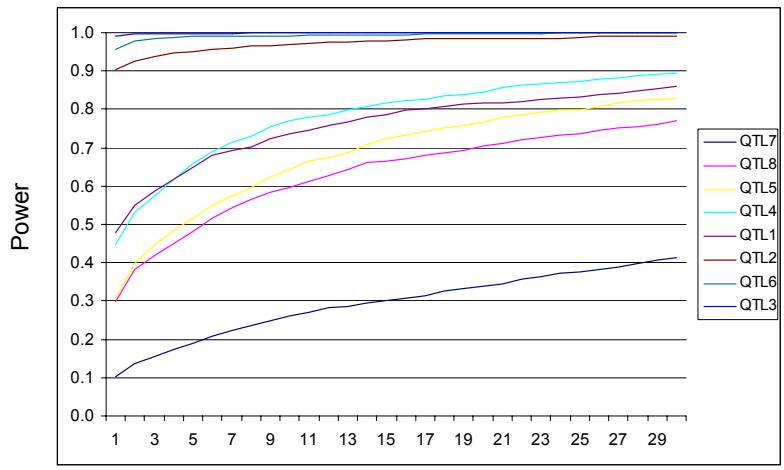


Statistical power of detecting QTL for Case I



Genome-wide significance level (α level)

Statistical power of detecting QTL for Case II



Genome-wide significance level (α level)

False discovery rate (FDR) of QTL identification

by using different genome-wide significance level and different LOD support interval for defining true or false identification (case I and II)

α level		5%	10%	15%	20%	25%	30%
LOD-1.0	I	0.090	0.094	0.100	0.105	0.110	0.117
	II	0.090	0.107	0.112	0.119	0.127	0.134
LOD-1.5	I	0.042	0.046	**0.051**	0.056	0.061	0.069
	II	0.045	0.052	**0.057**	0.063	0.071	0.077
LOD-2.0	I	0.023	0.028	0.033	0.038	0.042	0.049
	II	0.023	0.029	0.034	0.038	0.043	0.047

Confidence of LOD support interval

(% of true QTL inside the interval, case II)

α level	5%			10%			15%		
LOD	1.0	1.5	2.0	1.0	1.5	2.0	1.0	1.5	2.0
QTL-1	0.879	0.952	0.985	0.887	0.954	0.984	0.889	**0.955**	0.979
QTL-2	0.911	0.960	0.981	0.917	0.962	0.981	0.921	**0.967**	0.982
QTL-3	0.938	0.969	0.981	0.939	0.967	0.976	0.930	**0.959**	0.970
QTL-4	0.926	0.978	0.993	0.922	0.976	0.989	0.916	**0.973**	0.985
QTL-5	0.929	0.970	0.987	0.923	0.970	0.985	0.925	**0.966**	0.980
QTL-6	0.938	0.971	0.984	0.935	0.970	0.983	0.934	**0.968**	0.980
QTL-7	0.709	0.842	0.923	0.673	0.828	0.918	0.686	**0.825**	0.920
QTL-8	0.800	0.911	0.953	0.806	0.905	0.956	0.818	**0.907**	0.956
Average	**0.879**	**0.944**	**0.973**	**0.875**	**0.942**	**0.972**	**0.877**	**0.940**	**0.969**

Confidence of LOD support interval (% of true QTL inside the interval, case II)

α level	20%			25%			30%		
LOD	**1.0**	**1.5**	**2.0**	**1.0**	**1.5**	**2.0**	**1.0**	**1.5**	**2.0**
QTL-1	0.888	0.954	0.978	0.887	0.954	0.979	0.884	0.952	0.975
QTL-2	0.924	0.968	0.983	0.920	0.960	0.974	0.922	0.958	0.970
QTL-3	0.928	0.954	0.969	0.921	0.945	0.961	0.915	0.939	0.957
QTL-4	0.914	0.970	0.983	0.911	0.968	0.983	0.909	0.963	0.980
QTL-5	0.921	0.964	0.981	0.916	0.962	0.979	0.911	0.957	0.976
QTL-6	0.923	0.961	0.972	0.911	0.951	0.967	0.897	0.939	0.958
QTL-7	0.689	0.829	0.924	0.687	0.827	0.922	0.692	0.838	0.926
QTL-8	0.818	0.904	0.958	0.819	0.906	0.956	0.814	0.904	0.951
Average	**0.876**	**0.938**	**0.969**	**0.872**	**0.934**	**0.965**	**0.868**	**0.931**	**0.962**

LOD support interval width (cM) (case II)

α level	5%			10%			15%		
LOD	**1.0**	**1.5**	**2.0**	**1.0**	**1.5**	**2.0**	**1.0**	**1.5**	**2.0**
QTL-1	18.9	26.3	34.8	19.7	27.5	36.5	20.4	**28.4**	37.7
QTL-2	13.1	17.4	21.9	13.3	17.8	22.3	13.4	**17.8**	22.3
QTL-3	11.7	15.1	18.6	11.7	15.2	18.8	11.8	**15.3**	19.0
QTL-4	17.4	24.0	32.3	17.8	24.5	33.3	17.7	**24.3**	32.6
QTL-5	17.4	23.7	33.0	17.4	24.1	33.2	17.8	**24.8**	34.2
QTL-6	12.5	15.9	19.3	12.5	15.8	19.3	12.6	**16.0**	19.6
QTL-7	22.9	33.7	47.8	23.7	35.6	51.6	24.7	**37.2**	54.8
QTL-8	22.4	33.5	46.7	23.1	34.8	49.2	24.0	**36.7**	52.1
Average	**17.0**	**23.7**	**31.8**	**17.4**	**24.4**	**33.0**	**17.8**	**25.1**	**34.0**

LOD support interval width (cM) (case II)

α level	20%			25%			30%		
LOD	**1.0**	**1.5**	**2.0**	**1.0**	**1.5**	**2.0**	**1.0**	**1.5**	**2.0**
QTL-1	20.9	29.2	38.7	21.0	29.4	39.3	21.5	29.9	40.0
QTL-2	13.4	17.9	22.4	13.7	18.1	22.7	13.7	18.2	22.9
QTL-3	11.9	15.4	19.1	11.9	15.6	19.3	12.0	15.7	19.4
QTL-4	17.7	24.4	32.2	17.5	24.2	31.7	17.6	24.3	31.8
QTL-5	18.0	24.9	33.9	18.0	25.0	33.4	18.1	25.3	33.3
QTL-6	12.7	16.2	20.1	12.7	16.5	20.8	12.8	16.9	21.5
QTL-7	25.8	39.5	59.7	26.3	41.3	62.7	27.1	43.6	65.4
QTL-8	24.4	37.4	54.0	25.0	38.7	55.8	25.1	39.3	56.9
Average	**18.1**	**25.6**	**35.0**	**18.3**	**26.1**	**35.7**	**18.5**	**26.6**	**36.4**

QTL parameter estimation (averaged over detected replicates for each QTL, Case II)

QTL	Position	Mean Estimate	SD	Effect	Mean Estimate	SD
1	27.4	27.4	7.47	0.503	0.557	0.112
2	90.3	89.6	4.59	0.670	0.716	0.141
3	49.0	49.1	3.76	0.798	0.816	0.139
4	32.5	31.5	6.22	0.590	0.629	0.141
5	88.9	89.3	6.43	-0.503	-0.572	0.126
6	9.3	9.9	4.04	0.710	0.732	0.128
7	70.7	68.8	16.76	0.255	0.463	0.093
8	63.2	62.7	12.91	0.399	0.485	0.093

General Conclusion

- For mapping multiple QTL, sequential search via MIM offers a promising solution.
- It is appropriate to use model- and data-based criterion to aid for the model selection, such as score statistic and its resampling threshold. It takes the data and model complexity into account and computationally is very efficient. However, the threshold does not seem to change in different cycle of model search process, and may need to be calculated only once to save more computational time.

General Conclusion

- It is too conservative to use genome-wide significance level 0.05 for model selection. With significance level 0.1 to 0.2, FDR is still controlled approximately at 0.05. If higher FDR, such as 0.1, can be tolerated, we can even use significance level 0.2 to 0.4 for model selection to increase statistical power to detect relatively weaker QTL.
- It is more appropriate to use 1.5-LOD, rather than 1-LOD, support interval to report likely region of QTL. For moderate and strong QTL, 1.5-LOD support interval can give approximately 95% confidence interval for QTL position.

Slide 37

Multiple Trait Analysis: Genetic Basis of Trait Correlation

Zhao-Bang Zeng

Department of Statistics

North Carolina State University

Email: zeng@stat.ncsu.edu

URL: http://statgen.ncsu.edu/zeng/zeng.html

Slide 38

Why multiple traits/environments?

- Most QTL mapping experiments record observations on multiple traits.
- Different attributes of a general biological character: For example
  - Size of an area (*e.g.* wing size): size *vs.* shape.
  - Body weight: different parts of the weight.
  - Fitness components: viability *vs.* fecundity.
- Same trait at different developmental stages: Developmental aspects of the character.
- Same trait at different character states: growth in tropic area *vs.* growth in temporal area, grain yield in NC *vs.* Iowa. This is the issue of genotype by environment interaction, and can be analyzed by multiple trait analysis.

Slide 39

Why analyzing multiple traits/environments?

By taking into account the correlated structure of multiple traits, the joint analysis on multiple traits for mapping QTL can

- Improve the statistical power to detect QTL,
- Improve the resolution to estimate QTL positions and effects,
- Provide formal procedures to test a number of biologically interesting hypotheses concerning the nature of genetic correlations between different traits, such as
  - Joint mapping QTL (testing and estimating QTL affecting multiple traits)
  - Testing pleiotropy of QTL
  - Testing QTL $\times$ environment interaction
  - Testing whether significant effects at a genome region on multiple traits is due to pleiotropy of the same QTL or close

Slide 40

linkage of different QTL: pleiotropy *vs.* close linkage

- Provide a comprehensive estimation about the genetic architecture of quantitative traits including the structure of genetic correlations between traits.

Slide 41

Two types of data structures

We consider two types of data structures for multiple trait analysis.

- Design I: Multiple traits are measured on the same individuals. Data matrices may look like the following (for m traits, t markers, n individuals)

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t1} & x_{t2} & \cdots & x_{tn} \end{bmatrix}$$

Slide 42

- Design II: Multiple traits or trait states are measured on different individuals. Data matrices may look like the following (with one set of traits measured in population one with n_1 individuals and t_1 markers, and another set measured in another population with n_2 individuals and t_2 markers)

$$\mathbf{Y}_1 = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n_1} \\ y_{21} & y_{22} & \cdots & y_{2n_1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m_1 1} & y_{m_1 2} & \cdots & y_{m_1 n_1} \end{bmatrix} \quad \mathbf{X}_1 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n_1} \\ x_{21} & x_{22} & \cdots & x_{2n_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t_1 1} & x_{t_1 2} & \cdots & x_{t_1 n_1} \end{bmatrix}$$

$$\mathbf{Y}_2 = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n_2} \\ y_{21} & y_{22} & \cdots & y_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m_2 1} & y_{m_2 2} & \cdots & y_{m_2 n_2} \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n_2} \\ x_{21} & x_{22} & \cdots & x_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t_2 1} & x_{t_2 2} & \cdots & x_{t_2 n_2} \end{bmatrix}$$

Slide 43

This can represent several situations: for example

- The same traits measured in two backcrosses (B_1 and B_2) on different individuals. In this case a test on QTL×backcross interaction is a test about dominance (and some epistasis as well) of QTL.
- The same trait measured in two sexes: test QTL×sex interaction.
- Different groups of individuals are planted in two or multiple geographic locations.

Slide 44

Model and likelihood

For m putative QTL of T traits in S environments/populations, the multiple interval mapping model (for a backcross population) is defined by

$$y_{sti} = \mu_{st} + \sum_{r=1}^{m} \alpha_{str} x^*_{sir} + e_{sti}$$

where

- y_{sti} is the phenotypic value of trait t for individual i in environment/population s;
- i indexes individuals of the sample: $i = 1, 2, \cdots, n_s$;
- t indexes traits: $t = 1, 2, \cdots, T$;
- s indexes environments/populations: $s = 1, 2, \cdots, S$;
- μ_{st} is the mean of the model;
- α_{str} is the effect of putative QTL r on trait t in population s,

Slide 45

- x^*_{sir} is a coded variable denoting the genotype of putative QTL r (defined by $1/2$ or $-1/2$ for the two genotypes) for individual i in population s, which is unobserved but can be inferred from maker data in sense of probability;
- e_{sti} is a residual effect of the model assumed to be multivariate normal distributed with mean vector $\mathbf{0}$ and variance matrix $\mathbf{V}_s$.

Slide 46

Likelihood

The likelihood function of the data given the model is a mixture of normal distributions

$$L = \prod_{s=1}^{S} \prod_{i=1}^{n_s} \left[\sum_{j=1}^{2^m} p_{sij} \phi \left(\mathbf{y}_{si} | \mu_s + \mathbf{E}_s \mathbf{D}_{sj}, \mathbf{V}_s \right) \right]$$

- p_{sij} is the probability of each multilocus genotype conditional on marker data;
- $\mathbf{E}_s$ is a matrix of QTL parameters (α's) for population s;
- $\mathbf{D}_{sj}$ is a vector specifying the configuration of x^*'s associated with each α for the jth QTL genotype;
- $\phi(\mathbf{y}|\mu, \mathbf{V})$ denotes a multivariate normal density function for $\mathbf{y}$ with mean vector μ and variance matrix $\mathbf{V}$.

Slide 47

EM algorithm

E-step:

$$\pi_{sij}^{[t+1]} = \frac{p_{sij}\phi(\mathbf{y}_{si}|\mu_s^{[t]} + \mathbf{E}_s^{[t]}\mathbf{D}_{sj}, \mathbf{V}_s^{[t]})}{\sum_{j=1}^{2^m} p_{sij}\phi(\mathbf{y}_{si}|\mu_s^{[t]} + \mathbf{E}_s^{[t]}\mathbf{D}_{sj}, \mathbf{V}_s^{[t]})}$$

M-step:

$$E_{str}^{[t+1]} = \frac{\sum_s \sum_i \sum_j \pi_{sij}^{[t+1]} D_{srj}[(y_{sti} - \mu_{st}^{[t]}) - \sum_{u=1}^{r-1} D_{suj}E_{stu}^{[t+1]} - \sum_{u=r+1}^{m} D_{suj}E_{stu}^{[t]}]}{\sum_s \sum_i \sum_j \pi_{sij}^{[t+1]} D_{srj}^2}$$

$$\mu_{st}^{[t+1]} = \frac{1}{n_s}\sum_i \left(y_{sti} - \sum_j \sum_r \pi_{sij}^{[t+1]} D_{srj} E_{str}^{[t+1]} \right)$$

$$v_{suw}^{[t+1]} = \frac{1}{n_s}\left[\sum_i \sum_j \pi_{sij}^{[t+1]} (y_{sui} - \mu_{su}^{[t+1]} - \sum_r D_{srj}E_{sur}^{[t+1]}) \right.$$

$$\left. (y_{swi} - \mu_{sw}^{[t+1]} - \sum_r D_{srj}E_{swr}^{[t+1]}) \right]$$

Slide 48

Model selection

Model selection can proceed as in MIM. In this case, when a QTL is selected, its effects are fitted and estimated for all traits in all environments or populations, irregardless whether the QTL effect is significant for a particular trait in a particular environment.

1. Initial model: Use multivariate backward stepwise regression on markers to select an initial model.
2. Optimize the estimates of QTL positions based on the currently selected model.
3. Scan the genome to determine the best position for adding a new QTL.
4. Repeat (2) and (3) for a few times to select a few competing models.
5. If epistasis is considered, select significant epistatic terms.

Slide 49

6. Select the final model based on some information criterion.

Slide 50

Hypotheses testing

Given a selected genetic model, we can still test a number of hypotheses. Testing these hypotheses can help us to understand and interprete the genetic architecture of quantitative traits in those environments/populations.

- Joint mapping of QTL: Here we can test the hypothesis

$$H_0 : \alpha_r = \mathbf{0} \quad vs. \quad H_1 : \alpha_r \neq \mathbf{0}$$

with

$$LR = -2 \ln \frac{L(\alpha_r = \mathbf{0}, \hat{\mu}, \hat{\alpha}_{-r}, \hat{\mathbf{V}})}{L(\hat{\mu}, \hat{\alpha}, \hat{\mathbf{V}})}$$

Slide 51

- QTL × environment interaction:

$$H_0 : \ \alpha_{1r} = \cdots = \alpha_{Sr} \text{ or } \alpha_{Tr}$$

$$vs.\ H_1 : \ \alpha_{1r} \neq \cdots \neq \alpha_{Sr} \text{ or } \alpha_{Tr}$$

with

$$LR = -2 \ln \frac{L(\alpha_{1r} = \cdots = \alpha_{Sr}, \hat{\mu}, \hat{\alpha}_{-r}, \hat{\mathbf{V}})}{L(\hat{\mu}, \hat{\alpha}, \hat{\mathbf{V}})}$$

Rejection of the null hypothesis means that the QTL effects depend on environment, being different in different environments.

Note: For design II (S populations) where the joint likelihood is just a product of the separate likelihoods assuming independent samples, it is for testing this null hypothesis that the joint analysis is required.

Slide 52

How does the joint analysis improve the power and resolution of mapping QTL?

Joint Analysis vs. Separate Analysis

If we let the likelihood ratios under separate tests be

$$LR_{S1} \simeq n\beta_1^2 \quad \text{where } \beta_1 = a_1^*/(\sqrt{2}\sigma_{y_1 x^* \cdot \mathbf{X}})$$

$$LR_{S2} \simeq n\beta_2^2 \quad \text{where } \beta_2 = a_2^*/(\sqrt{2}\sigma_{y_2 x^* \cdot \mathbf{X}})$$

The likelihood ratio under the joint test can be approximated as

$$LR_J \simeq n\frac{\beta_1^2 + \beta_2^2 - 2\rho\beta_1\beta_2}{1 - \rho^2}$$

Slide 53

Some observations:

1. $LR_J \geq \text{maximum}[LR_{S1},\ LR_{S2}]$.
2. If $\rho = 0$, $LR_J \simeq LR_{S1} + LR_{S2}$.
3. If $\beta_2 = 0$, $LR_J \simeq \frac{LR_{S1}}{1-\rho^2} \geq LR_{S1}$.
4. If $\rho\beta_1\beta_2 < 0$, *i.e.* ρ and $\beta_1\beta_2$ are in different signs, $LR_J > LR_{S1} + LR_{S2}$. In this case, $\text{power}(J) > \text{maximum}[\text{power}(S1),\ \text{power}(S2)]$.

This shows that LR_J can be significantly higher than $\max[LR_{S1}, LR_{S2}]$. Of course, the threshold for LR_J is higher than that for LR_{S1} and LR_{S2} as well. However, in general we find that the power of LR_J is significantly higher than that of LR_{S1} and LR_{S2}.

Slide 54

A simulation example on joint vs. separate analysis:

We show a simulation study with three QTL having effects on three quantitative traits, distributed on a chromosome with 150 cM length in an F_2 population with 150 sample size. We assume markers are observed at every 10cM position. Genetic parameters of QTL and estimates of QTL positions and effects (averaged over 100 replicates with standard deviations in bracket) are shown in Table 3, and phenotypic parameters and estimates in Table 4.

TABLE 3: Parameters of QTL positions and effects used in simulations and their mean estimates (standard deviations), over 100 replicates, by the joint mapping on three traits (J-123) and on two traits at a time (J-12, J-13 and J-23) and by the separate mapping on each trait (S-1, S-2 and S-3)

QTL	Position	Additive effect			Dominance effect		
	(cM)	Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
				Parameters			
1	21.0	1.00	1.00	0.30	0.43	0.43	0.13
2	84.0	-0.30	-1.00	-1.00	-0.09	-0.30	-0.30
3	142.0	-1.00	0.30	1.00	0.19	0.06	0.19
				Estimates by J-123			
1	21.0 (4.3)	1.00 (0.44)	1.00 (0.42)	0.35 (0.36)	0.43 (0.44)	0.42 (0.28)	0.13 (0.25)
2	84.9 (4.7)	-0.24 (0.51)	-1.00 (0.38)	-1.01 (0.40)	-0.01 (0.51)	-0.27 (0.25)	-0.25 (0.23)
3	142.4 (3.9)	-1.03 (0.38)	0.27 (0.27)	1.02 (0.31)	0.17 (0.38)	0.05 (0.29)	0.07 (0.22)
				Estimates by J-12			
1	20.9 (5.2)	1.03 (0.43)	1.02 (0.41)		0.44 (0.32)	0.42 (0.29)	
2	83.7 (9.6)	-0.25 (0.56)	-0.97 (0.43)		0.03 (0.36)	-0.24 (0.30)	
3	141.3 (6.8)	-1.08 (0.36)	0.26 (0.34)		0.17 (0.30)	0.02 (0.30)	
				Estimates by J-13			
1	22.4 (7.8)	1.05 (0.48)		0.30 (0.39)	0.47 (0.32)		0.14 (0.25)
2	85.6 (6.0)	-0.26 (0.52)		-1.03 (0.40)	-0.01 (0.33)		-0.25 (0.24)
3	143.1 (2.8)	-1.00 (0.38)		1.03 (0.30)	0.18 (0.29)		0.08 (0.23)
				Estimates by J-23			
1	21.8 (6.5)		1.05 (0.41)	0.34 (0.40)		0.41 (0.30)	0.12 (0.26)
2	84.9 (4.4)		-1.01 (0.38)	-1.04 (0.37)		-0.27 (0.25)	-0.25 (0.23)
3	142.4 (4.5)		0.28 (0.30)	1.04 (0.35)		0.03 (0.29)	0.07 (0.23)
				Estimates by S-1, S-2 and S-3			
1	21.9,21.6,25.8 (7.7,5.7,13.6)	1.08 (0.42)	1.12 (0.34)	0.38 (0.49)	0.48 (0.33)	0.41 (0.30)	0.11 (0.33)
2	84.9,84.3,86.0 (17.1,8.7,6.7)	-0.29 (0.69)	-1.03 (0.38)	-1.06 (0.37)	0.06 (0.42)	-0.26 (0.29)	-0.24 (0.26)
3	141,136,143 (6.2,11.7,4.1)	-1.13 (0.31)	0.33 (0.45)	1.06 (0.30)	0.17 (0.30)	-0.01 (0.37)	0.08 (0.24)

Slide 55

Seven mapping analyses were performed and compared: J-123 (joint mapping on three traits); J-12, J-13, J-23 (joint mapping on two traits); S-1, S-2, S-3 (separate mapping on each trait). Conclusions:

- Standard deviations of estimates of QTL positions are generally smaller with joint analyses. This shows that joint analysis can significantly improve the resolution on the estimation of QTL position.
- There is some improvement on the estimation of QTL effects, but the improvement does not seem to be very significant.
- The joint analyses also improve significantly on the power to detect QTL (Table 5).

Slide 56

Table 5: Observed statistical power (proportion of significant replicates over all replicates) of seven methods of QTL mapping from 100 replicates of simulations

QTL	J-123	J-12	J-13	J-23	S-1	S-2	S-3
1	0.80	0.78	0.51	0.64	0.46	0.64	0.04
2	0.79	0.37	0.36	0.84	0.00	0.39	0.41
3	0.89	0.51	0.84	0.64	0.42	0.00	0.64

Slide 57

A practical data analysis with three populations and four trait states (locations):

- Three related backcross populations with a trait evaluated in four locations ($S = 3$, $T = 4$, $n_s = 128$).
- QTL analysis was performed with four locations and three populations combined in model identification and estimation.
- An initial model was selected based on a combined forward-and-backward stepwise regression on markers. Then, an extensive search and position-optimization analysis was performed repeatedly under multiple interval mapping.
- The final model contains 19 QTL that have significant effects on the trait in different locations and different populations.
- The following figure plots the LOD score profile for the 19 QTL. It depicts the statistical evidence and strength of mapping for each QTL.

Slide 58

- The LOD scores for each QTL in each population (with four locations combined) and estimates of QTL effects in each location and population provide statistical evidence and information about the significance of QTL effects in each location and population.
- Estimated genetic variances, covariances and genetic correlations provide further information about the genetic architecture of QTL. These estimates can be further partitioned into components due to each individual QTL.
- Genetic correlations between different locations provide an overall structure of QTL by location interaction. High genetic correlation reflects lack of QTL by location interaction. Low or negative genetic correlation reflects some or strong QTL by location interaction. It is clear that QTL x Location interaction is prevalent for many QTL in all the three populations.

Slide 59

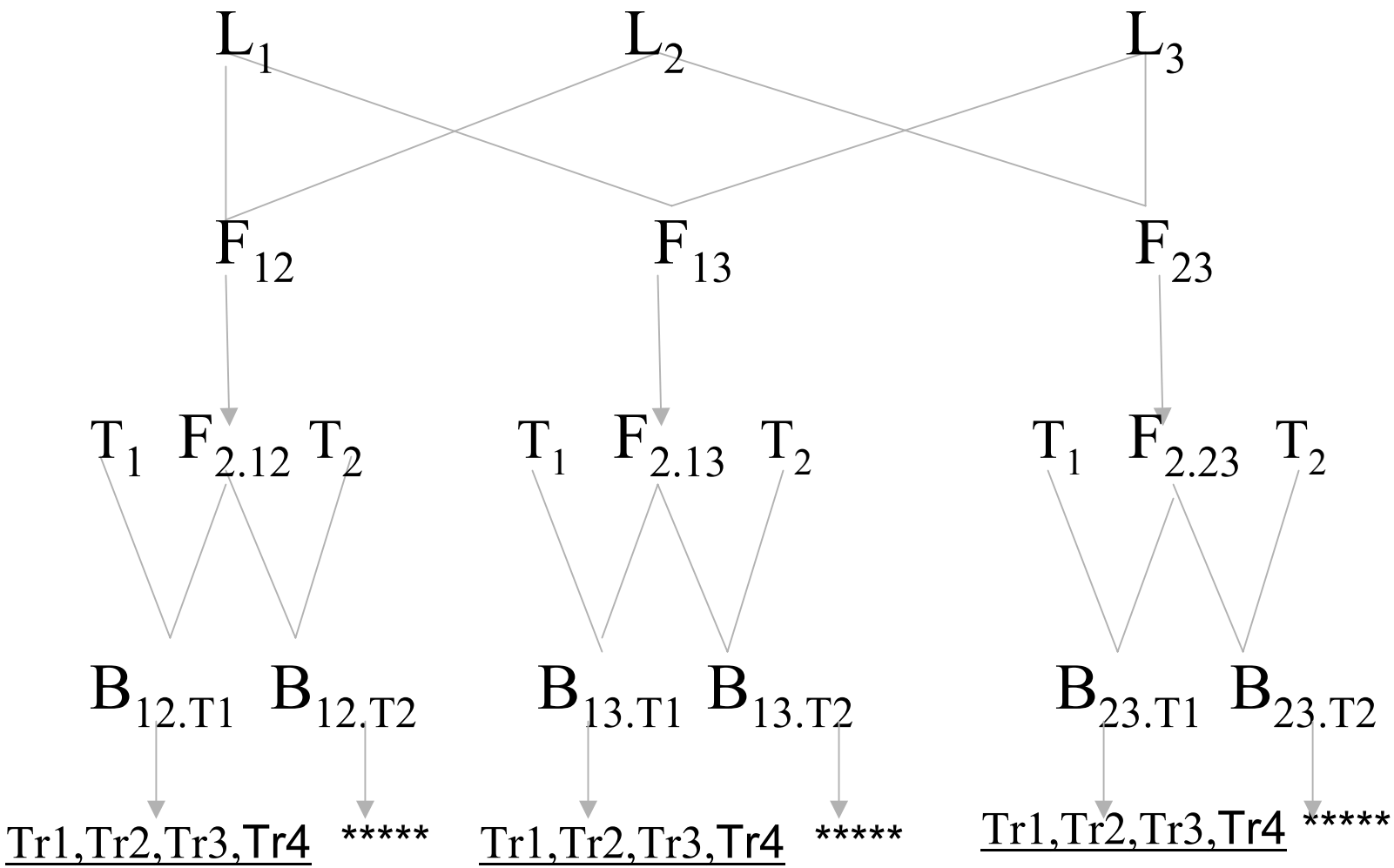
For example, in population cross 3, QTL effects in location C are in an opposite direction to those in other locations for QTL 1, 7, 8, 11, 14, 15, and 17. Together, these produce the negative genetic correlations between location C and location A, B and D. In population 1, genetic correlation between location B and D is also negative, due to mostly QTL 13, 14, 17 and 18. The genetic correlation between location A and B is also low, due to mostly QTL 4, 17 and 18. The low genetic correlation between location C and D is mostly due to QTL 4, 11 and 14. In population 2, genetic correlations between location C and location B and D are also relatively lower, indicating some QTLxLocation interaction (due to mostly QTL 11 and 19).

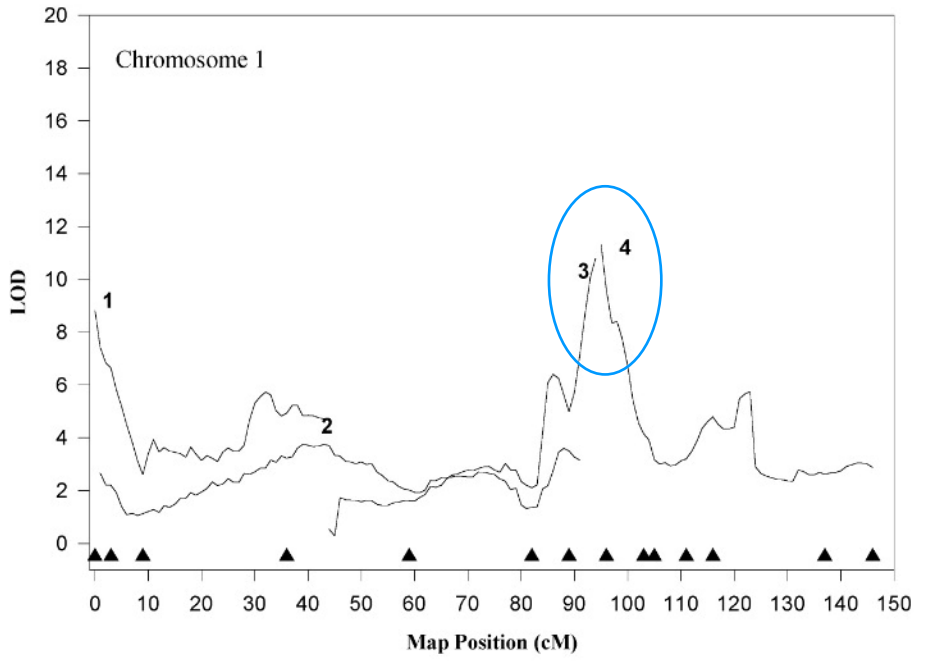
Slide 60

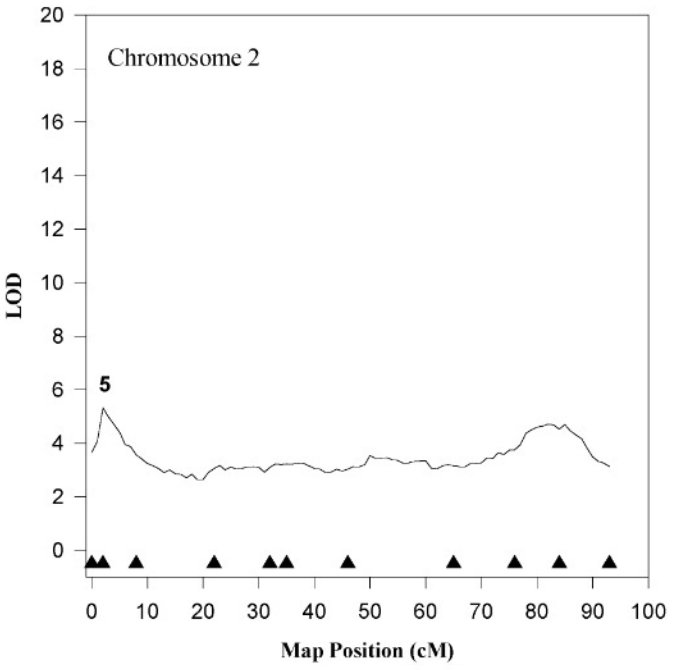
This integrated multiple QTL oriented approach for multiple traits, environments and populations provides a comprehensive method to study the details of genetic archiectrure of quantitative traits in a variety of populations. It also facilitates the comparison and test of genetic basis for multiple traits in multiple environments and populations.

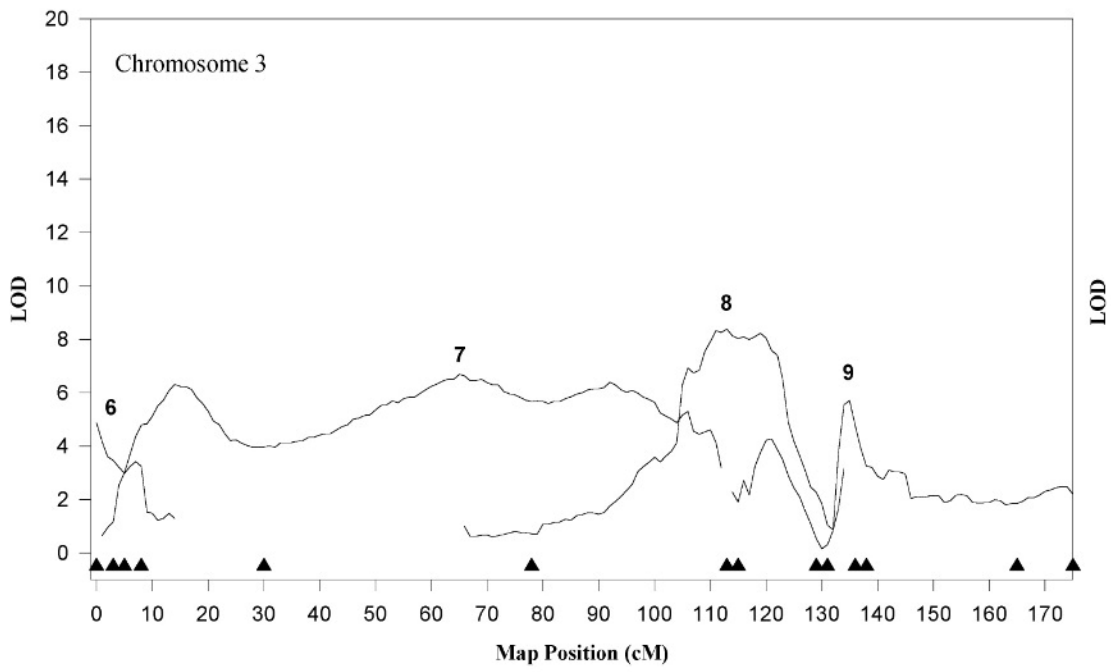
"Total is more than the sum!"

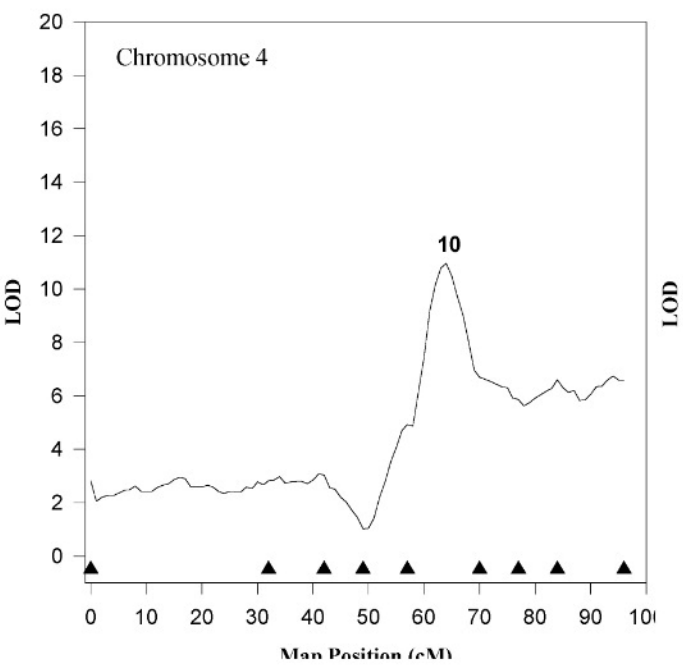
Experiment Design

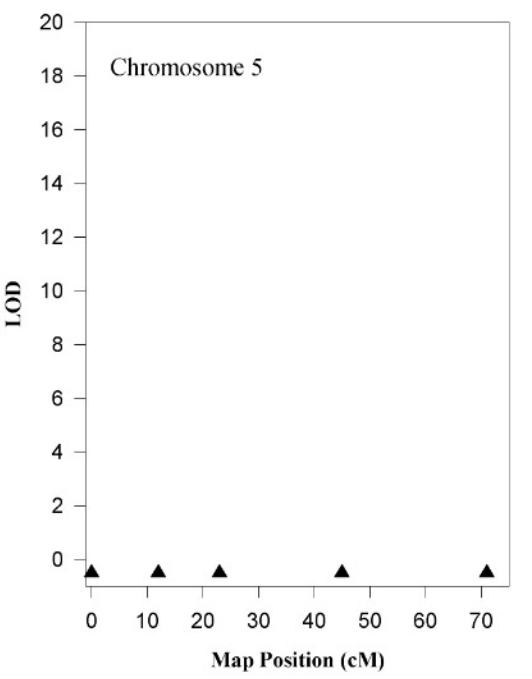


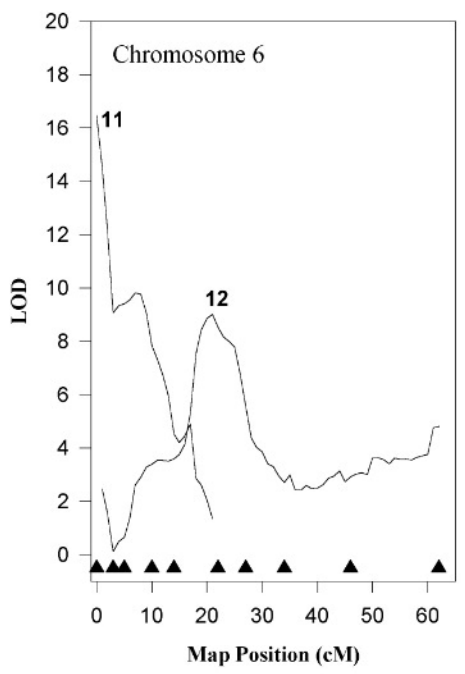

Chromosome 1
LOD
Map Position (cM)
1
2
3
4

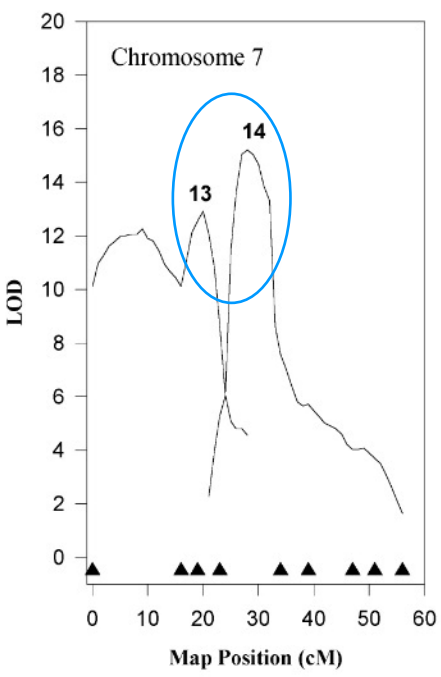

Chromosome 2
LOD
Map Position (cM)
5

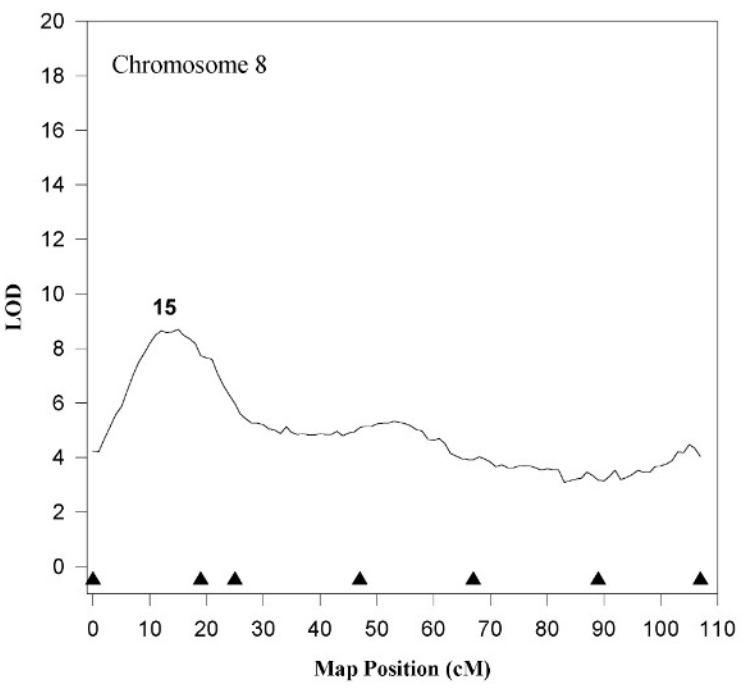

Chromosome 3
LOD
Map Position (cM)
6
7
8
9

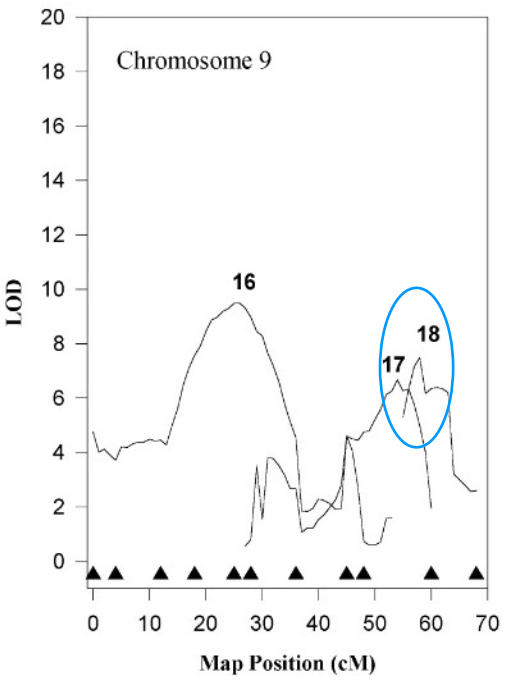

Chromosome 4
LOD
10

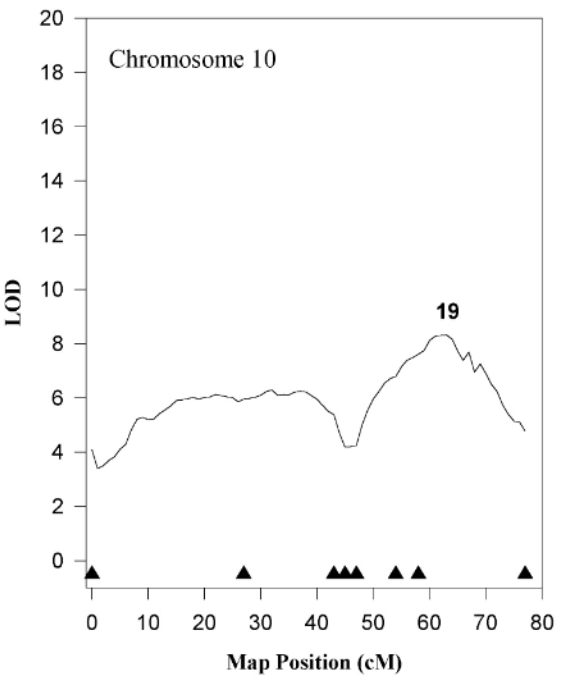

Chromosome 5
LOD
Map Position (cM)


Chromosome 6
LOD
11
12
Map Position (cM)


Chromosome 7
LOD
13
14
Map Position (cM)


Chromosome 8
LOD
15
Map Position (cM)


Chromosome 9
LOD
16
17
18
Map Position (cM)


Chromosome 10
LOD
19
Map Position (cM)

LOD score 19 putative QTL

	Posi	LOD in population			
QTL	Ch:cM	1	2	3	Sum
1	1:0	.73	2.37	5.95	9.04
2	1:43	.76	.62	3.93	5.30
3	1:92	4.12	4.46	2.71	11.29
4	1:95	4.74	6.01	1.81	12.56
5	2:2	2.27	2.86	9.37	14.49
6	3:0	.98	.83	2.39	4.19
7	3:65	.03	.44	4.77	5.24
8	3:113	1.82	.67	8.36	10.85
9	3:135	1.16	.10	6.06	7.32
10	4:64	11.06	.84	1.50	13.41
11	6:0	2.59	5.24	9.29	17.11
12	6:21	3.94	.01	.01	3.96
13	7:20	9.83	.80	5.16	15.80
14	7:28	10.38	1.52	6.67	18.57
15	8:15	8.44	1.38	1.54	11.36
16	9:26	2.32	4.98	2.26	9.55
17	9:54	4.63	2.99	1.76	9.37
18	9:58	4.51	2.67	1.80	8.98
19	10:63	1.00	2.47	2.87	6.35

Estimates of QTL effects in three populations and four locations (A,B,C,D)

	Population 1				Population 2				Population 3			
QTL	A	B	C	D	A	B	C	D	A	B	C	D
1	3.7	-1.4	1.5	0.1	-4.6	-5.0	1.2	-2.4	5.3	6.0	-2.6	2.6
2	-2.2	-1.3	-3.2	-2.6	5.2	1.7	1.0	1.1	-5.7	-5.0	0.8	-3.8
3	6.7	5.2	1.4	8.7	-9.6	-3.2	-1.5	-12.7	-9.6	-10.0	-7.1	-7.4
4	-3.0	-0.9	4.1	-8.0	17.5	6.8	1.0	18.2	7.5	4.3	4.8	4.8
5	1.5	-3.7	-0.5	0.9	5.1	6.8	0.7	-0.8	-4.0	1.4	1.2	-0.6
6	0.7	-2.1	-2.7	-0.5	1.7	-1.1	-2.9	-1.0	0.1	1.6	-4.3	-2.0
7	-1.6	0.0	-0.7	-0.1	-2.6	-3.3	0.0	-2.6	-6.1	-0.7	5.7	-4.1
8	-6.3	-1.1	-6.1	-2.9	-4.8	-5.0	-3.9	-0.9	16.0	7.2	-6.8	12.5
9	-7.6	-2.3	-0.3	-3.0	2.3	-1.0	-1.0	-1.6	-10.2	-5.7	6.3	-5.2
10	17.0	3.9	7.2	3.3	0.4	-1.3	1.8	2.8	-2.9	1.5	3.8	1.9
11	-1.0	7.5	-1.3	0.8	0.7	-5.7	4.9	-3.7	13.0	-1.7	-6.0	0.7
12	0.7	-3.9	4.0	-2.9	3.6	1.9	-1.7	1.1	4.3	2.3	-1.9	-1.6
13	9.3	12.1	1.2	-6.9	-5.7	-3.8	0.7	0.9	-3.0	4.3	-4.4	3.9
14	-9.3	-13.5	-6.7	6.9	9.6	9.8	2.2	0.6	-1.7	-7.1	12.0	-2.7
15	4.5	4.2	5.5	-1.9	4.1	-1.6	0.3	-2.0	5.2	1.7	1.2	-2.3
16	-3.0	0.3	0.7	2.3	13.8	6.9	4.3	3.7	3.9	2.2	0.6	2.1
17	-16.4	9.9	-1.9	-7.9	-20.0	-1.3	-3.4	-10.5	4.0	2.9	-4.1	-4.0
18	20.3	-9.3	-0.2	8.5	17.3	1.7	4.2	10.1	1.9	-5.0	2.9	4.7
19	-4.4	-0.9	1.6	-1.0	1.8	-3.1	3.7	-2.3	-6.3	-1.5	-0.2	-0.8

Phenotypic mean, variance (V_p), proportion of variance explained by the QTL model (V_g/V_p) for each location in each population

Env	mean	V_p	V_g/V_p
Population 1			
A	130.99	244.97	0.567
B	139.46	128.82	0.358
C	128.45	83.23	0.464
D	142.03	65.02	0.479
Population 2			
A	124.54	295.34	0.555
B	130.49	152.52	0.407
C	125.57	86.40	0.283
D	139.05	97.86	0.448
Population 3			
A	148.49	183.03	0.580
B	147.40	102.98	0.407
C	135.22	90.16	0.434
D	148.99	62.39	0.562

Estimated r_p, r_e, r_g, and (CV_g/CV_p)

Env pair	r_p	r_e	r_g	CV_g/CV_p
Population 1				
A,B	0.33	0.40	0.25	0.348
A,C	0.43	0.28	0.60	0.696
A,D	0.52	0.59	0.42	0.436
B,C	0.24	0.13	0.40	0.691
B,D	0.24	0.53	-0.16	-0.278
C,D	0.37	0.56	0.12	0.158
Population 2				
A,B	0.43	0.26	0.56	0.665
A,C	0.46	0.42	0.66	0.525
A,D	0.46	0.26	0.56	0.681
B,C	0.34	0.34	0.33	0.333
B,D	0.44	0.38	0.51	0.503
C,D	0.46	0.54	0.29	0.230
Population 3				
A,B	0.25	0.13	0.33	0.720
A,C	0.08	0.49	-0.27	-1.339
A,D	0.31	0.02	0.49	0.964
B,C	0.12	0.31	-0.20	-0.870
B,D	0.29	0.07	0.55	0.888
C,D	0.13	0.38	-0.15	-0.634

Partition of genetic correlation between location A and C in population 3

QTL:C\A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum
1	-.053	.021	.026	-.019	.007	.000	.012	-.023	.024	.005	-.016	-.009	.004	.004	-.007	-.008	-.006	-.003	.006	-.036
2	.006	-.017	-.010	.008	-.003	.000	-.003	.006	-.005	-.002	.012	.003	-.002	-.002	.004	.002	.002	.001	-.002	-.003
3	-.040	.056	.258	-.160	.011	.000	.036	-.097	.067	.026	-.099	-.033	.018	.015	-.024	-.020	-.025	-.010	.018	-.003
4	.025	-.036	-.139	.138	-.011	.000	-.028	.083	-.051	-.020	.065	.023	-.012	-.010	.023	.014	.018	.008	-.018	.072
5	.004	-.006	-.005	.005	-.018	.000	-.004	.011	-.008	-.002	.008	.005	-.003	-.002	.003	.000	.001	.000	-.001	-.012
6	-.016	.017	.031	-.027	.016	-.001	.035	-.081	.050	.013	-.052	-.024	.011	.009	-.014	-.004	-.010	-.006	.028	-.026
7	.023	-.020	-.045	.041	-.013	.001	-.131	.177	-.100	-.019	.086	.026	-.009	-.008	.021	.015	.012	.004	-.007	.053
8	-.020	.020	.056	-.055	.016	-.001	.080	-.412	.186	.020	-.087	-.034	.013	.011	-.026	-.007	-.025	-.010	.015	-.259
9	.030	-.025	-.056	.049	-.017	.001	-.066	.274	-.243	-.019	.088	.034	-.013	-.010	.026	.021	.027	.011	-.012	.098
10	.014	-.022	-.046	.041	-.010	.000	-.026	.062	-.040	-.042	.049	.023	-.012	-.009	.021	.008	.014	.006	-.018	.015
11	-.015	.040	.062	-.047	.012	.000	.042	-.095	.066	.017	-.296	-.039	.021	.017	-.019	-.018	-.007	-.002	.015	-.243
12	-.008	.010	.020	-.016	.007	.000	.012	-.035	.024	.008	-.037	-.032	.006	.005	-.008	-.005	-.005	-.002	.007	-.050
13	-.011	.022	.037	-.028	.013	.000	.014	-.044	.031	.013	-.068	-.019	.049	.017	-.010	-.010	-.016	-.005	.006	-.009
14	.056	-.081	-.149	.108	-.039	.001	-.064	.189	-.119	-.053	.272	.088	-.085	-.075	.042	.036	.048	.019	-.065	.129
15	.003	-.006	-.007	.008	-.003	.000	-.005	.014	-.009	-.004	.009	.004	-.001	-.001	.023	.002	.002	.001	-.004	.025
16	.003	-.002	-.004	.003	.000	.000	-.002	.002	-.005	-.001	.006	.002	-.001	-.001	.001	.009	.003	.001	-.001	.012
17	-.013	.012	.035	-.028	.002	.000	.013	-.060	.045	.011	-.015	-.011	.011	.007	-.008	-.019	-.063	-.021	.010	-.094
18	.010	-.008	-.021	.018	.000	.000	-.006	.037	-.027	-.007	.006	.009	-.006	-.004	.005	.009	.031	.020	-.010	.057
19	.000	.000	.001	-.001	.000	.000	.000	-.001	.001	.000	-.001	.000	.000	.000	.000	.000	.000	.000	.004	.002
Sum	-.003	-.025	.041	.038	-.029	.000	-.092	.007	-.115	-.054	-.071	.015	-.011	-.036	.053	.025	.000	.010	-.028	-.273

Total genetic correlation: -.273

Due to pleiotropy (sum of diagonal elements): -.883

Due to linkage (sum of off-diagonal elements): .609

Partition of genetic correlation between location B and D in population 3

QTL:D\B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum
1	.104	-.032	-.048	.019	.004	.005	-.003	.018	-.023	.005	-.004	.009	.010	-.028	.004	.008	.008	-.015	-.003	.038
2	-.055	.124	.090	-.037	-.009	-.007	.003	-.025	.025	-.010	.013	-.013	-.024	.054	-.010	-.007	-.010	.015	.004	.121
3	-.081	.088	.484	-.164	-.007	-.015	.008	-.078	.067	-.024	.023	-.032	-.048	.116	-.015	-.019	-.032	.049	.008	.327
4	.048	-.055	-.246	.134	.006	.011	-.006	.063	-.048	.018	-.015	.021	.030	-.069	.013	.013	.022	-.035	-.007	-.102
5	-.004	.004	.004	-.002	-.005	-.001	.000	-.004	.004	-.001	.001	-.002	-.003	.005	-.001	.000	.000	.000	.000	-.005
6	-.014	.012	.026	-.013	-.005	-.021	.003	-.030	.023	-.006	.006	-.011	-.013	.032	-.004	-.002	-.006	.013	.005	-.003
7	-.032	.022	.058	-.029	-.006	-.015	.019	-.097	.068	-.012	.014	-.017	-.016	.043	-.009	-.010	-.011	.012	.002	-.014
8	.072	-.057	-.185	.100	.018	.040	-.031	.585	-.329	.033	-.036	.057	.058	-.148	.028	.012	.057	-.086	-.011	.177
9	-.048	.031	.084	-.040	-.008	-.016	.011	-.174	.193	-.014	.017	-.026	-.027	.062	-.012	-.016	-.028	.041	.004	.033
10	.014	-.017	-.041	.020	.003	.005	-.003	.023	-.019	.018	-.005	.011	.014	-.034	.006	.004	.009	-.013	-.003	-.008
11	.003	-.007	-.013	.005	.001	.002	-.001	.008	-.007	.002	-.008	.004	.006	-.015	.001	.002	.001	-.001	-.001	-.016
12	-.013	.012	.029	-.013	-.003	-.006	.002	-.022	.019	-.006	.007	-.024	-.012	.032	-.004	-.004	-.005	.010	.002	.002
13	.020	-.029	-.059	.024	.007	.009	-.003	.030	-.026	.010	-.014	.015	.108	-.112	.005	.008	.017	-.023	-.002	-.013
14	-.024	.027	.059	-.024	-.005	-.009	.003	-.032	.025	-.010	.014	-.018	-.047	.123	-.005	-.008	-.013	.020	.006	.081
15	-.012	.017	.026	-.015	-.003	-.004	.002	-.020	.017	-.006	.004	-.007	-.007	.018	-.025	-.003	-.004	.006	.003	-.013
16	.017	-.009	-.026	.012	.000	.001	-.002	.007	-.017	.003	-.005	.005	.009	-.019	.002	.029	.012	-.014	-.002	.004
17	-.025	.017	.059	-.027	-.001	-.006	.003	-.044	.041	-.009	.003	-.010	-.026	.048	-.004	-.017	-.076	.091	.004	.021
18	.033	-.019	-.062	.030	.000	.009	-.002	.046	-.042	.010	-.002	.013	.024	-.050	.005	.013	.064	-.153	-.007	-.089
19	-.003	.003	.006	-.004	.000	-.002	.000	-.003	.002	-.002	.001	-.002	-.001	.009	-.001	-.001	-.002	.004	.008	.010
Sum	-.001	.134	.245	-.024	-.013	-.020	.006	.252	-.026	-.001	.014	-.026	.034	.068	-.027	.002	.003	-.079	.010	.551

Total genetic correlation: .551

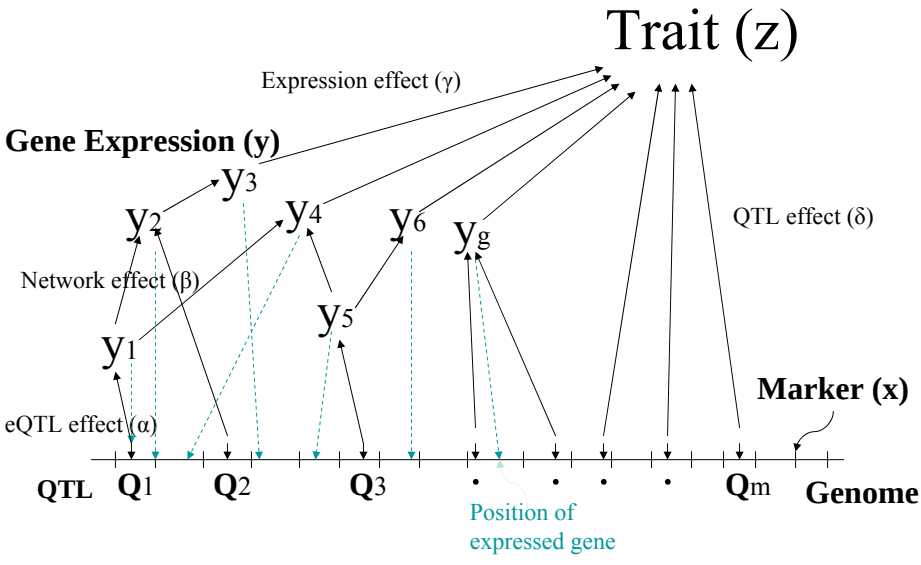
Due to pleiotropy (sum of diagonal elements): 1.617

Due to linkage (sum of off-diagonal elements):-1.066

Multiple Interval Mapping for eQTL Analysis

Zhao-Bang Zeng
Bioinformatics Research Center
Departments of Statistics & Genetics
North Carolina State University
zeng@stat.ncsu.edu

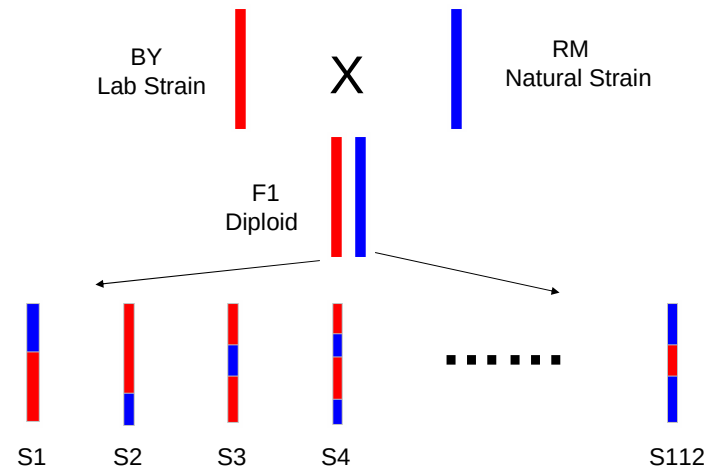
Genetic Effect Network



Goals and Issues of eQTL Mapping Analysis

- Identify and map genomic regions that significantly affect expression levels of different genes
  - Statistical methods and power to map eQTL
  - Justification of mapping procedures and results (e.g. FDR)
  - Epistasis of eQTL
  - Multiple trait analysis
- Through the mapping to identify *cis*- and *trans*-regulation of eQTL
- Identify gene expression co-regulation patterns (eQTL hot-spots)
  - Why are they co-regulated? Is there any functional relationship among those co-regulated genes?
- Prioritize candidate genes (from eQTL to genes)
  - By using regulative and functional relationship between candidate genes in eQTL regions and genes whose expressions being regulated, we might be able to prioritize and suggest candidate causal genes for some eQTL.
- Toward to network and pathway analysis

***An eQTL study on a yeast hy*brid segregant population**



Brem and Kruglyak (2005) PNAS 102:1572-1577

Experimental Design and Data

- **Sample:** BY (lab strain), RM (natural strain) and **112** F1 segregants.
- **Markers: 3312** using yeast oligoarrays
- **Gene expression:** Samples were labeled and hybridized to cDNA microarrays, containing **6215** open reading frames (ORF).
- **Reference design:** Each two-color experiment involved one sample and one reference, with the same BY RNA reference being used for all experiments.
- **Dye swap:** Two hybridizations were carried out for each sample, one with the sample labeled with Cy3 and the reference with Cy5, and one with the fluors reversed; for each gene, the two log ratios were averaged.

Yeast experiment data structure

	Markers		Expressions	
Ind	1 2 3 4 5 6 7 8 9 10 11 …… 3312		1 2 3 4 5 6 7 8 9 10 11 …… 6215	
YU	1 1 1 1 1 1 1 1 1 1 1 …….	1	X X X X X X X X X X X …….	X
RM	0 0 0 0 0 0 0 0 0 0 0 …….	0	X X X X X X X X X X X …….	X
S1	1 1 1 1 0 0 0 0 0 1 1 …….	0	X X X X X X X X X X X …….	X
S2	0 0 0 1 1 1 1 1 1 0 0 …….	1	X X X X X X X X X X X …….	X
S3	0 0 0 0 0 0 1 1 1 1 0 …….	0	X X X X X X X X X X X …….	X
….				
….				
S112	1 1 1 0 0 0 0 0 0 1 1 …….	0	X X X X X X X X X X X …….	X

Data: For I = 1, 2, ….., 112+2

Xij for j = 1, 2, …., 3312 Yik for k = 1, 2, ……, 6215

The Simplest Analysis – t-test

- For every j (marker) and k (expression trait):

  $$y_{ik} = \alpha + \beta x_{ij} + e_i \quad \text{for } i = 1, 2, \cdots, n$$

  test $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$

- Genome-wide search (multiple test issue): We want to test

  H_0 : there is no QTL in the genome

  H_1 : there is QTL in the genome

  Want to know the distribution and threshold of $max(T(j), j \in \text{genome})$ at the null.

  Solution: permutation test

- Declare those j's with $T(j) >$ threshold linked to QTL

Interval Mapping – smooth the likelihood profile and define a confidence region

- Model:

  $$y_{ik} = \alpha + \beta x_{il} + e_i$$

  l: a location in the genome

  x_{il} : missing data, but with known $Prob(x_{il} = 1 \text{ or } 0 | \text{markers})$

- Analyze the mixture model and calculate

  $$LOD(l) = \log_{10} \frac{L(\beta \neq 0)_l}{L(\beta = 0)_l}$$

- The genome-wide threshold for $LOD(l)$ can be determined by a permutation test

- 1.5-LOD support interval: The genome region covered by dropping LOD by 1.5 from the peak gives a confidence interval of QTL position

Search for multiple eQTL and epistasis

- Exhaustive 2D (or multi-D) search for 2 or more eQTL
  - There is empirical evidence that sequential search is more powerful than exhaustive 2D search (Storey et al 2005)
- Sequential search (Storey et al version)
- Sequential search (Zou and Zeng version)

Sequential search for eQTL pairs and epistasis

(Storey et al 2005 PLoS Biology)

Storey et al (2005) proposed a two-stage sequential search to detect interactive eQTL:

- For each eTrait, perform one-dimensional genome scan to identify the best position for a putative eQTL.
- Conditional on the 1st putative eQTL position, perform another one-dimensional genome scan to detect the 2nd putative eQTL with a model of both main and interaction effects.
- Final selection on the (interactive) eQTL is based on FDR.
  - "A significant expression trait is called a false discovery if any of the loci selected for that trait is a false positive. That is, a true discovery is an expression trait where all selected loci are truly linked."

Storey et al (2005)

- "Based on the joint linkage probabilities, we estimate that 2,300 traits (approximately 37%) are jointly linked to two loci, although we cannot identify all of these with high confidence. Of these 2,300 traits, **170** can be identified at a FDR of 10%."
- "In total,58 traits demonstrate a *cis* linkage"

Multiple interval mapping for eQTL analysis

(Zou and Zeng 2006)

- Model: $$y_{ik} = \alpha + \sum_{t} \beta_t x_{il_t} + \sum_{s<t} \gamma_{st} x_{il_s} x_{il_t} + e_i$$
- Sequential search for each eQTL conditional on the significance in the previous cycle for each eTrait.
- For each eTrait:
  - In cycle 1, if the max test statistic > threshold, the first eQTL is identified and continue the next step; otherwise stop the search.
  - In cycle t+1, if the conditional max test statistic > threshold, one more eQTL is added and continue the search; otherwise stop.
  - After the search for the main effects, epistatic effects of eQTL are tested based on the threshold and then added to the model.
  - Obtain 1.5-LOD support interval for each identified eQTL

MIM for eQTL analysis

- Threshold is first determined by a permutation test with a controlled type I error rate for the genome scan (e.g. 95 percentile of test statistic in a genome scan under the null).
- Then the threshold is evaluated or adjusted based on the calculation of False Discovery Rate (FDR) in the sequential genome scans for the whole detected eQTL for all the expression traits.

Table 1: Sequential genome scan using MIM

Cycle	#Scanned[1]	#Retained[2]	#Claimed[3]
1	6195	3367	3354
2	3367	1617	1242
3	1617	578	422
4	578	197	122
5	197	66	37
6	66	10	5

1. # of eTraits in each cycle
2. # of eTraits in the initial genome scans using the 10% genome-wide type I error rate
3. # of eTraits in the final result using the 5% genome-wide type I error rate

With the 5% genome-wide type I error in each genome scan, the False Discovery Rate (FDR) for all the detected eQTL is estimated about 8%.

Two procedure differences between Storey et al and Zou-Zeng

- Storey et al (2005) selected the 2nd eQTL for each trait from the genome based on both main and epistatic effects. This test has two degree of freedom. **[1] Both main and epistatic effects**
- They attempted to find the 2nd eQTL for each trait no matter how significant the 1st eQTL is. **[2] Unrestricted search**

Number of two QTL genetic models declared with 10% FDR

Methods	Number of eTraits
Storey et al: [1] and [2]	174
Main effect only: relax [1]	746
Restricted search: relax [2]	662

A simulation study

- Simulate 620 (10%) traits from the pool of inferred models with two eQTL among 6215 traits.
- To study the effect of epistasis on searching strategies, we simulate 5%, 50% and 95% proportions of genetic models with epistasis.

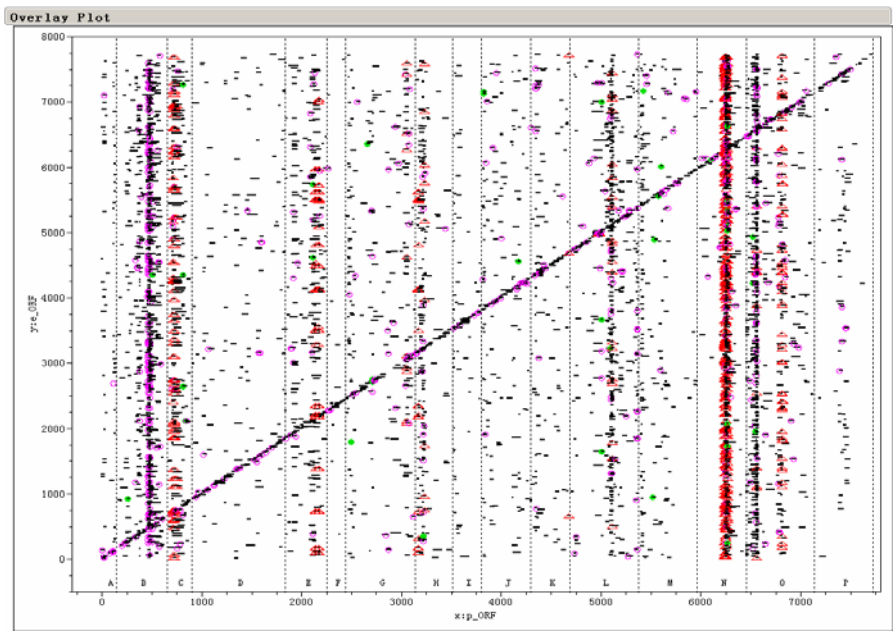
Simulation result: Mean number (SD) of 2 QTL model with 10% FDR

Interaction proportion	Storey et al method	Main effect only	Restricted search
5%	99(37)	238(71)	168(72)
50%	161(40)	194(39)	236(67)
95%	226(62)	190(58)	306(64)

The role of threshold in MIM-eQTL

- In the later cycles of genome scans, the search is restricted within the parameter space where the chance to detect strong association is high as we focus on those traits that have shown significant QTL in the previous cycles;
- It serves as a stopping rule to decide how many QTL we can find for each trait.

Example: Re-analysis of Brem & Kruglyak (2005).



Bayesian candidate gene prioritization analysis
(Wei Zou and Z-B Zeng 2006)

Table 4.1: A singe gene as the most probably gene for large number of traits

gene	max[1]	all[2]	annotation[3]
YNL067W	92	294	Protein component of the large (60S) ribosomal subunit
YNL069C	116	367	N-terminally acetylated protein component of the large (60S) ribosomal subunit
YBR154C	146	436	RNA polymerase subunit ABC27, common to RNA polymerases I, II, and III; contacts DNA and affects transactivation
YNL096C	224	325	Protein component of the small (40S) ribosomal subunit
YOL077C	284	452	Nucleolar protein, constituent of 66S pre-ribosomal particles

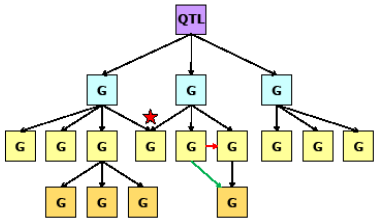
1. # of overlapping eQTL that the gene is identified as the most probable underlying gene by our Bayesian gene prioritization analysis.
2. Total # of overlapping eQTL that include the gene as a candidate gene.
3. From http://www.yeastgenome.org/

Acknowledgement

- NC State University Bioinformatics Research Center
  - Chris Basten
  - Wei Zou
  - Jessica Maia
- Funding
  - NIH GM45344
  - USDA Plant Genome
- NC State University Forest Biotechnology Group (for *Eucalyptus* data)
  - Ronald R. Sederoff
  - Matias Kirst
- U. Washington (for yeast data)
  - Leonid Kruglyak

Identify network structure from eQTL hotspots
Duarte and Zeng (2009)

- A method for finding a "framework network" that decomposes the network into a series of hierarchical layers through a series of statistical tests of conditional independence or dependence.
- Model: $P(\mathbf{G}|T) = \prod_{S \in \mathbf{S}} P(S|T) \prod_{G \in \mathbf{G} \setminus \mathbf{S}} P(G|\mathbf{S})$
- Example:



17

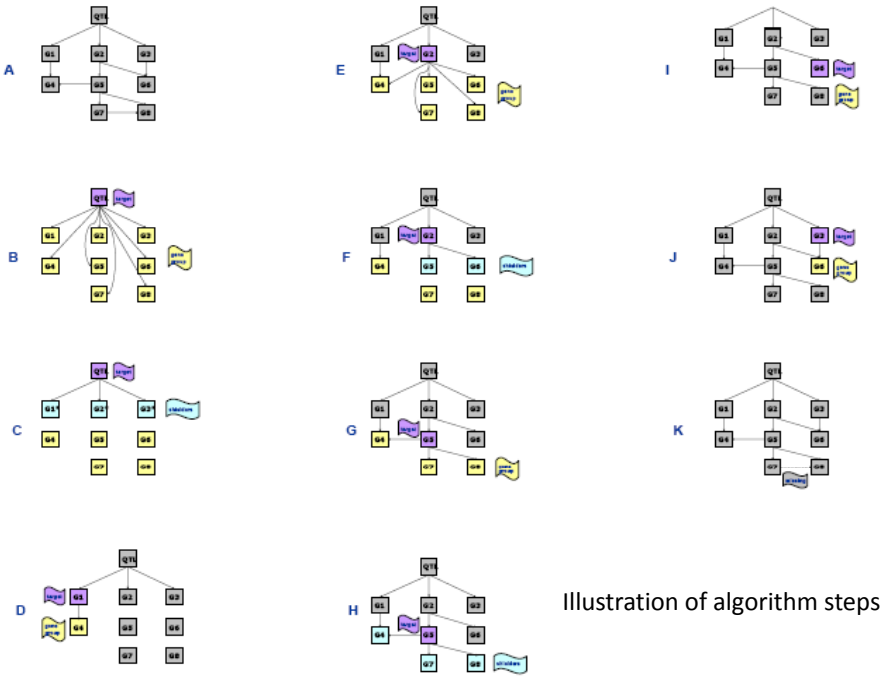


Illustration of algorithm steps

An example of discovered network

shielder	shielded
AFR1	HHF2 TOS4 YDR396W YER156C
AGA1	FAR1 PCL2 PRM6
FIG2	FIG1 PRM1 PRM2
FUS1	ASG7 BUD14 CHS1 EST3 FUS3 GFA1 HAL1 HYM1 INP52 KAR5 KNH1 MCM3 NIS1 NTA1 PDS1 PRM4 PRM5 RGD2 SHU1 SMY1 SNL1 SPP1 SST2 STE14 TEC1 UME6 YBP2 YDR124W YDR249C YDR282C YFR026C YIL080W YJR039W YJR054W YOR343C
YHR127W	KAR1 SRP102 YIL158W

Table 4.3: Discovered Network 6: eQTL at Chromosome 8 at 98,513 bp.

- Strong evidence that the eQTL is a deletion in GPA1 Gene.
- GPA1 is the alpha subunit of the G protein coupled to mating factor receptor and is involved in the mating pheromone signal transduction pathway.
- FUS1, AFR1, AGA1, and FIG2 are all associated with the GO biological process "response to pheromone during conjugation with cellular fusion".
- STE12 is a transcription factor that is activated by a MAPK kinase signaling cascade involved in pheromone.
- FUS1 is the first downstream target of STE12.
- 61.1% of genes shielded by FUS1 are targets of the STE12 transcription factor.
- Itself a target of STE12, it may modulate the transcriptional effects of STE12 on other targets.

19

A model for inferring gene pathway from gene knockout experiment (Aylor and Zeng, 2007)

Classical genetic model and interpretation:

$$A^+B^+ : y = \mu + \varepsilon$$

$$A^-B^+ : y = \mu + \beta_A + \varepsilon$$

$$A^+B^- : y = \mu + \beta_B + \varepsilon$$

$$A^-B^- : y = \begin{cases} \mu + \beta_A + \varepsilon & \text{if } A \text{ is epistatic to } B \\ \mu + \beta_B + \varepsilon & \text{if } B \text{ is epistatic to } A \end{cases}$$

Mostly applied for lethal phenotype or viability

20

A quantitative genetic model

$$y = \mu + \beta_A x_A + \beta_B x_B + \beta_I x_A x_B + \varepsilon$$

$$x_A = \begin{cases} 0 \text{ for } A^+ \\ 1 \text{ for } A^- \end{cases} \qquad x_B = \begin{cases} 0 \text{ for } B^+ \\ 1 \text{ for } B^- \end{cases}$$

$$\text{Model 1} : y = \mu + \beta_A + \varepsilon$$
$$\text{Model 2} : y = \mu + \beta_B + \varepsilon$$
$$\text{Model 3} : y = \mu + \beta_I + \varepsilon$$
$$\text{Model 4} : y = \mu + \beta_A + \beta_B + \varepsilon$$
$$\text{Model 5} : y = \mu + \beta_A + \beta_I + \varepsilon$$
$$\text{Model 6} : y = \mu + \beta_B + \beta_I + \varepsilon$$
$$\text{Model 7} : y = \mu + \beta_A + \beta_B + \beta_I + \varepsilon$$
$$\text{Model 8} : y = \mu + \varepsilon$$

21

Gene pathway interpretation

- We considered all combinations of gene order and action within simple ON/OFF models and then predicted the hypothetical effect of deleting genes on each of them.
- There are four points of variation to model for each gene pair relationship.
  - The first is the identity of the upstream gene, i.e. the gene order.
  - Secondly, the upstream gene will turn the downstream gene either on (enhance) or off (repress).
  - Thirdly, the downstream gene can enhance or repress the expression of a target gene for which expression is observed.
  - Lastly, we consider that the upstream gene itself will be enhanced or repressed by some initiating factor such as a developmental cue or environmental perturbation.

22

Linking quantitative model to pathway interpretation: An example

Genotype	Upstream Gene	Gene Action	Target Gene Expression	Regression Model
A^+B^+	ON	$A^{ON} \dashv B^{OFF} \rightarrow X^{OFF}$	μ	$\mu + \beta_A + \beta_I$
A^-B^+		$X \dashv B^{ON} \rightarrow X^{ON}$	$\mu + \beta_A$	
A^+B^-		$A^{ON} \dashv X \rightarrow X^{OFF}$	μ	
A^-B^-		$X \dashv X \rightarrow X^{OFF}$	μ	
A^+B^+	OFF	$A^{OFF} \dashv B^{ON} \rightarrow X^{ON}$	μ	$\mu + \beta_B$
A^-B^+		$X \dashv B^{ON} \rightarrow X^{ON}$	μ	
A^+B^-		$A^{OFF} \dashv X \rightarrow X^{OFF}$	$\mu + \beta_B$	
A^-B^-		$X \dashv X \rightarrow X^{OFF}$	$\mu + \beta_B$	

23

Another example

Genotype	Upstream Gene	Gene Action	Target Gene Expression	Regression Model
A^+B^+	ON	$A^{ON} \rightarrow B^{ON} \rightarrow X^{ON}$	μ	$\mu + \beta_A + \beta_B + \beta_I$
A^-B^+		$X \rightarrow B^{OFF} \rightarrow X^{OFF}$	$\mu + \beta_A$	
A^+B^-		$A^{ON} \rightarrow X \rightarrow X^{OFF}$	$\mu + \beta_B$	
A^-B^-		$X \rightarrow X \rightarrow X^{OFF}$	$\mu + \beta_B$	
A^+B^+	OFF	$A^{OFF} \rightarrow B^{OFF} \rightarrow X^{OFF}$	μ	μ
A^-B^+		$X \rightarrow B^{OFF} \rightarrow X^{OFF}$	μ	
A^+B^-		$A^{OFF} \rightarrow X \rightarrow X^{OFF}$	μ	
A^-B^-		$X \rightarrow X \rightarrow X^{OFF}$	μ	

24

Match quantitative models with pathway interpretations

a. Hierarchical Relationships

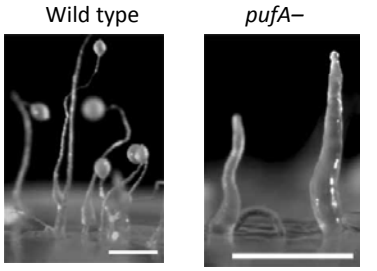
Upstream Gene	A upstream of B		B upstream of A	
	ON	OFF	ON	OFF
Repressor	$\mu+\beta_A+\beta_I$ [5]	$\mu+\beta_B$ [2]	$\mu+\beta_B+\beta_I$ [6]	$\mu+\beta_A$ [1]
Enhancer	$\mu+\beta_A+\beta_B+\beta_I$ [7]	μ [8]	$\mu+\beta_A+\beta_B+\beta_I$ [7]	μ [8]

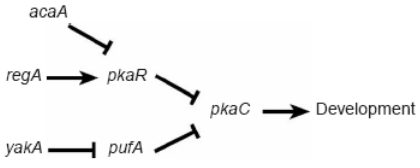
b. Non-hierarchical Relationships

State of A/B	ON/ON	ON/OFF	OFF/ON	OFF/OFF
Enhancer/Enhancer	$\mu+\beta_I$ [3]			
Enhancer/Repressor Or Repressor/Enhancer	$\mu+\beta_A+\beta_B$ [4]	$\mu+\beta_A$ [1]	$\mu+\beta_B$ [2]	μ [8]
Repressor/Repressor	$\mu+\beta_I$ [3]			

An application: Study of gene pathway in *Dictyostelium* (Van Driessche et al. 2007)

- Upon removal of nutrients, *D. discoideum* executes a developmental program in which single cells aggregate and form multicellular organisms.
- PKA pathway is known and important for the process. The pathway gene single and double knockout strains were created.
- Whole genome gene expression profiles were assayed and used to infer the pathway.

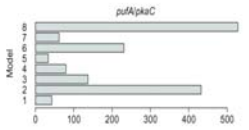




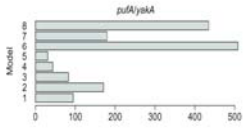
The protein kinase A (PKA) pathway

26

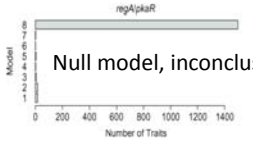
Aylor and Zeng (2007)



pufA ⊣ *pkaC* → Development



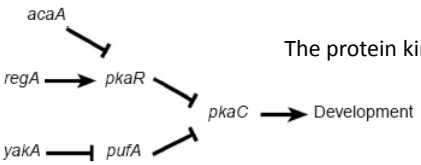
yakA ⊣ *pufA* ⊣ Development



regA → *pkaR* ⊣ Development Or others

Null model, inconclusive

Recall:



The protein kinase A (PKA) pathway

27

Concluding remarks

- The genetics of complex traits is very complex, and has been a black box.
- Through QTL mapping, a number of genetic components can be identified.
- With the aid of other omics information in a systems oriented study, some genetic pathways and networks that contribute to complex trait variation can also be illuminated.

28

Seattle Summer Institute 2010
Advanced QTL
Brian S. Yandell, UW-Madison

www.stat.wisc.edu/~yandell/statgen

- overview: multiple QTL approaches
- Bayesian QTL mapping & model selection
- data examples in detail
- software demos: R/qtl and R/qtlbim
- mapping multiple traits

Real knowledge is to know the extent of one's ignorance.
Confucius (on a bench in Seattle)

Overview of Multiple QTL

1. what is the goal of multiple QTL study?
2. gene action and epistasis
3. Bayesian vs. classical QTL
4. QTL model selection
5. QTL software options

1. what is the goal of QTL study?

- uncover underlying biochemistry
  - identify how networks function, break down
  - find useful candidates for (medical) intervention
  - epistasis may play key role
  - statistical goal: maximize number of correctly identified QTL
- basic science/evolution
  - how is the genome organized?
  - identify units of natural selection
  - additive effects may be most important (Wright/Fisher debate)
  - statistical goal: maximize number of correctly identified QTL
- select "elite" individuals
  - predict phenotype (breeding value) using suite of characteristics (phenotypes) translated into a few QTL
  - statistical goal: mimimize prediction error

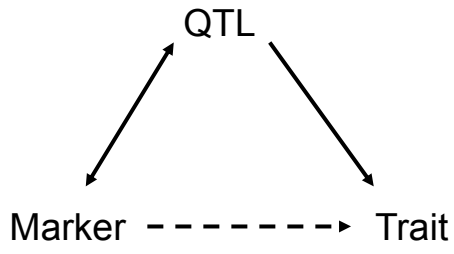
cross two inbred lines

→ linkage disequilibrium

→ associations

→ linked segregating QTL

(after Gary Churchill)



problems of single QTL approach

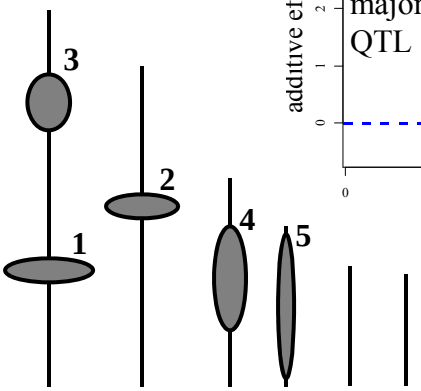
- wrong model: biased view
  - fool yourself: bad guess at locations, effects
  - detect ghost QTL between linked loci
  - miss epistasis completely
- low power
- bad science
  - use best tools for the job
  - maximize scarce research resources
  - leverage already big investment in experiment

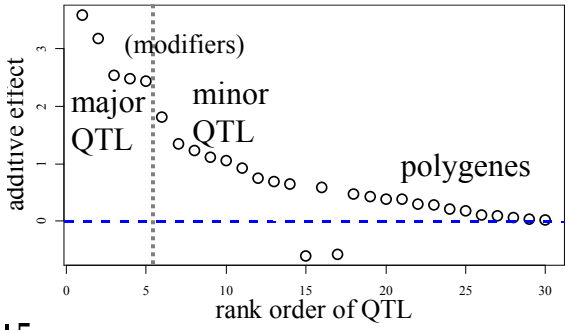
advantages of multiple QTL approach

- improve statistical power, precision
  - increase number of QTL detected
  - better estimates of loci: less bias, smaller intervals
- improve inference of complex genetic architecture
  - patterns and individual elements of epistasis
  - appropriate estimates of means, variances, covariances
    - asymptotically unbiased, efficient
  - assess relative contributions of different QTL
- improve estimates of genotypic values
  - less bias (more accurate) and smaller variance (more precise)
  - mean squared error = MSE = $(\text{bias})^2$ + variance

Pareto diagram of QTL effects

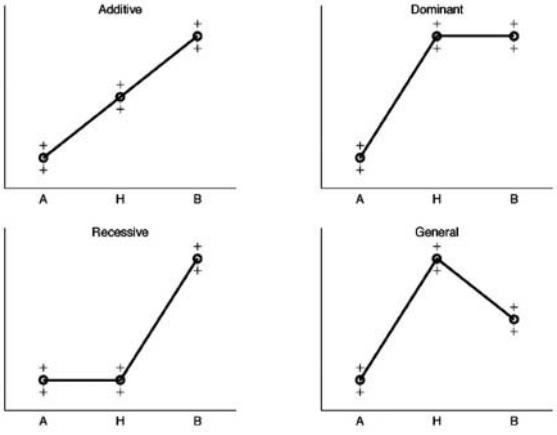
major QTL on linkage map





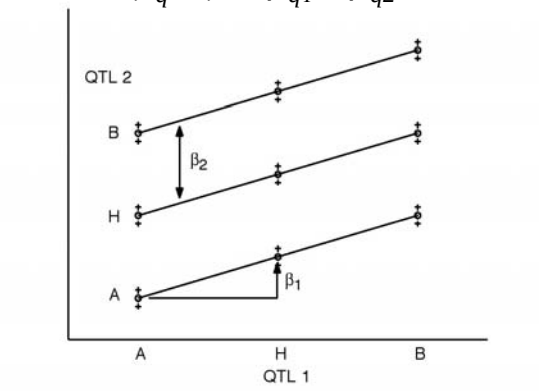
2. Gene Action and Epistasis

additive, dominant, recessive, general effects of a single QTL (Gary Churchill)



additive effects of two QTL (Gary Churchill)

$$\mu_q = \mu + \beta_{q1} + \beta_{q2}$$



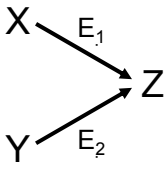
Epistasis (Gary Churchill)

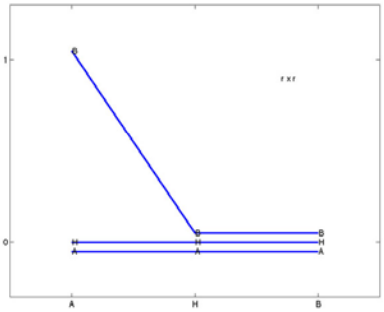
The allelic state at one locus can mask or uncover the effects of allelic variation at another.

- W. Bateson, 1907.

epistasis in parallel pathways (GAC)

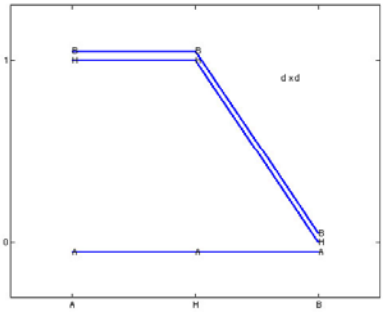
- Z keeps trait value low
- neither E_1 nor E_2 is rate limiting
- loss of function alleles are segregating from parent A at E_1 and from parent B at E_2





epistasis in a serial pathway (GAC)

- Z keeps trait value high
- either E_1 or E_2 is rate limiting
- loss of function alleles are segregating from parent B at E_1 or from parent A at E_2



epistatic interactions

- model space issues
  - 2-QTL interactions only?
    - or general interactions among multiple QTL?
  - partition of effects
    - Fisher-Cockerham or tree-structured or ?
- model search issues
  - epistasis between significant QTL
    - check all possible pairs when QTL included?
    - allow higher order epistasis?
  - epistasis with non-significant QTL
    - whole genome paired with each significant QTL?
    - pairs of non-significant QTL?
- see papers of Nengjun Yi (2000-7) in *Genetics*

limits of epistatic inference

- power to detect effects
  - epistatic model sizes grow quickly
    - $|A| = 3^{n.qtl}$ for general interactions
  - power tradeoff
    - depends sample size *vs.* model size
    - want $n / |A|$ to be fairly large (say > 5)
    - 3 QTL, $n = 100$ F2: $n / |A| \approx 4$
- rare genotypes may not be observed
  - *aa/BB & AA/bb* rare for linked loci
  - empty cells mess up balance
    - adjusted tests (type III) are wrong
  - confounds main effects & interactions

2 linked QTL
empty cell
with $n = 100$

	bb	*bB*	*BB*
aa	6	15	0
aA	15	25	15
AA	3	15	6

limits of multiple QTL?

- limits of statistical inference
  - power depends on sample size, heritability, environmental variation
  - "best" model balances fit to data and complexity (model size)
  - genetic linkage = correlated estimates of gene effects
- limits of biological utility
  - sampling: only see some patterns with many QTL
  - marker assisted selection (Bernardo 2001 *Crop Sci*)
    - 10 QTL ok, 50 QTL are too many
    - phenotype better predictor than genotype when too many QTL
    - increasing sample size may not give multiple QTL any advantage
  - hard to select many QTL simultaneously
    - 3^m possible genotypes to choose from

QTL below detection level?

- problem of selection bias
  - QTL of modest effect only detected sometimes
  - effects overestimated when detected
  - repeat studies may fail to detect these QTL
- think of probability of detecting QTL
  - avoids sharp in/out dichotomy
  - avoid pitfalls of one "best" model
  - examine "better" models with more probable QTL
- rethink formal approach for QTL
  - directly allow uncertainty in genetic architecture
  - QTL model selection over genetic architecture

3. Bayesian vs. classical QTL study

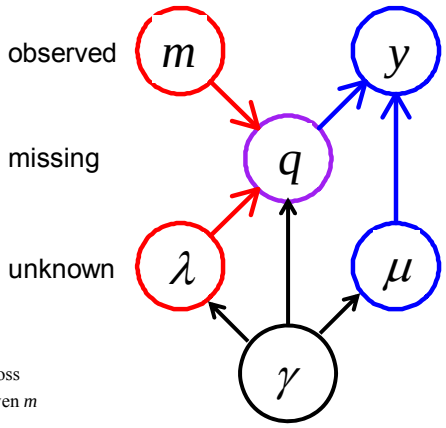
- classical study
  - ***maximize*** over unknown effects
  - ***test*** for detection of QTL at loci
  - model selection in stepwise fashion
- Bayesian study
  - ***average*** over unknown effects
  - ***estimate*** chance of detecting QTL
  - sample all possible models
- both approaches
  - average over missing QTL genotypes
  - scan over possible loci

Bayesian idea

- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetary, Moongate, London
  - famous paper in 1763 *Phil Trans Roy Soc London*
  - was Bayes the first with this idea? (Laplace?)
- basic idea (from Bayes' original example)
  - two billiard balls tossed at random (uniform) on table
  - where is first ball if the second is to its left?
    - prior: anywhere on the table
    - posterior: more likely toward right end of table

QTL model selection: key players

- observed measurements
  - y = phenotypic trait
  - m = markers & linkage map
  - i = individual index (1,…,n)
- missing data
  - missing marker data
  - q = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - λ = QT locus (or loci)
  - μ = phenotype model parameters
  - γ = QTL model/genetic architecture
- pr(q|m,λ,γ) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for q given m
- pr(y|q,μ,γ) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters μ (could be non-parametric)



after
Sen Churchill (2001)

Bayes posterior vs. maximum likelihood

- LOD: classical Log ODds
  - maximize likelihood over effects μ
  - `R/qtl scanone/scantwo: method = "em"`
- *LPD*: Bayesian *L*og *P*osterior *D*ensity
  - average posterior over effects μ
  - `R/qtl scanone/scantwo: method = "imp"`

$$\mathrm{LOD}(\lambda) = \log_{10}\{\max_{\mu} \mathrm{pr}(y \mid m, \mu, \lambda)\} + c$$

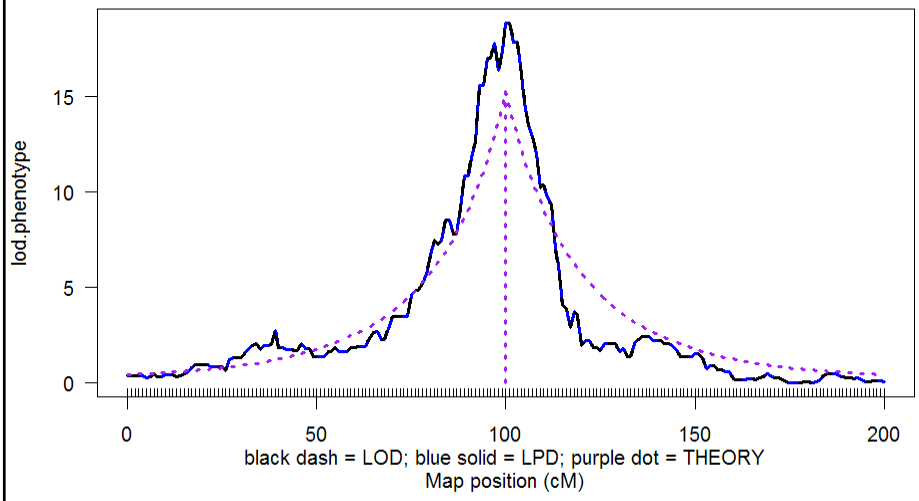
$$\mathrm{LPD}(\lambda) = \log_{10}\{\mathrm{pr}(\lambda \mid m)\int \mathrm{pr}(y \mid m, \mu, \lambda)\mathrm{pr}(\mu)d\mu\} + C$$

likelihood mixes over missing QTL genotypes:

$$\mathrm{pr}(y \mid m, \mu, \lambda) = \sum_q \mathrm{pr}(y \mid q, \mu)\mathrm{pr}(q \mid m, \lambda)$$

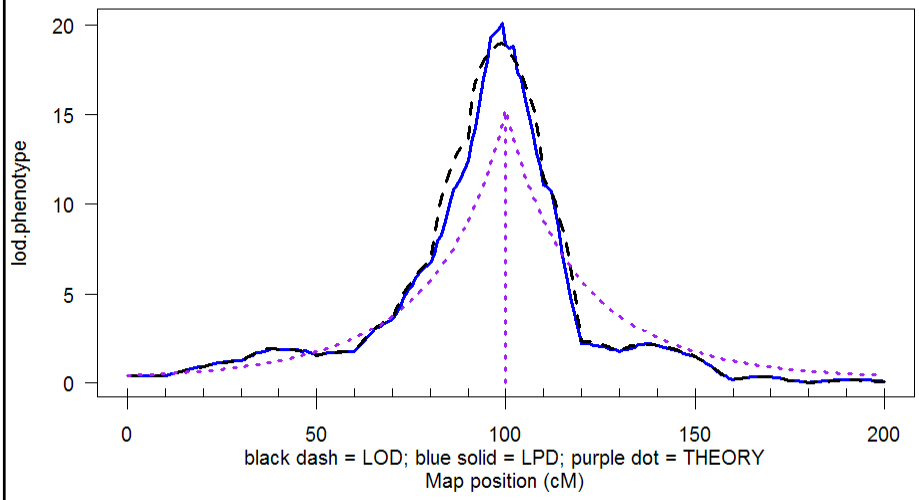
LOD & LPD: 1 QTL

n.ind = 100, 1 cM marker spacing



LOD & LPD: 1 QTL

n.ind = 100, 10 cM marker spacing



marginal LOD or LPD

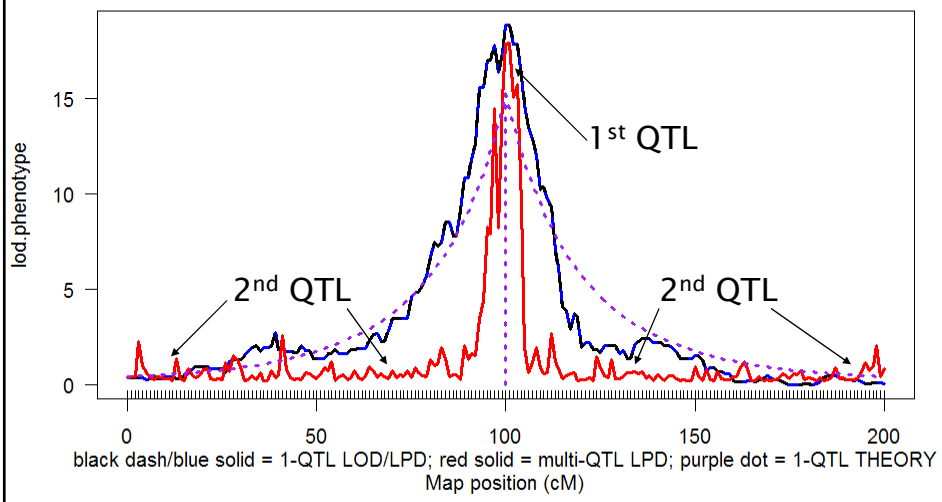
- compare two genetic architectures (γ_2,γ_1) at each locus
  - with (γ_2) or without (γ_1) another QTL at locus λ
    - preserve model hierarchy (e.g. drop any epistasis with QTL at λ)
  - with (γ_2) or without (γ_1) epistasis with QTL at locus λ
  - γ_2 contains γ_1 as a sub-architecture
- allow for multiple QTL besides locus being scanned
  - architectures γ_1 and γ_2 may have QTL at several other loci
  - use marginal LOD, LPD or other diagnostic
  - posterior, Bayes factor, heritability

$$\mathrm{LOD}(\lambda \mid \gamma_2) - \mathrm{LOD}(\lambda \mid \gamma_1)$$

$$\mathrm{LPD}(\lambda \mid \gamma_2) - \mathrm{LPD}(\lambda \mid \gamma_1)$$

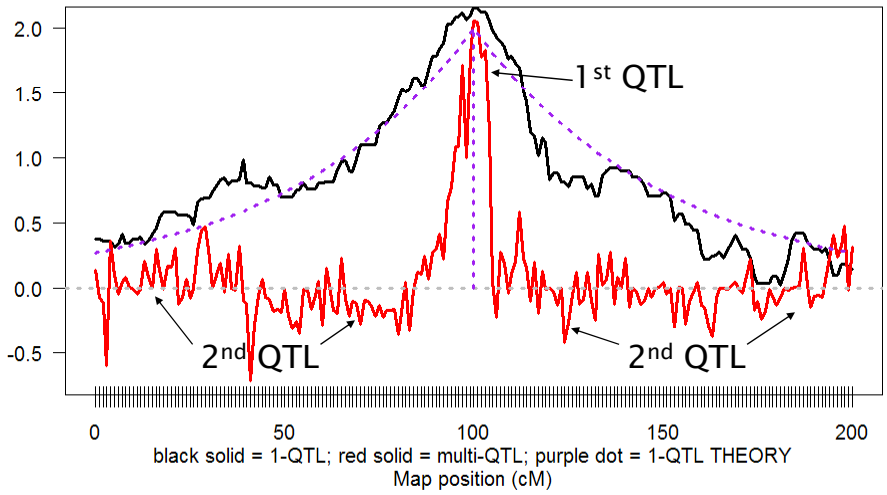
LPD: 1 QTL vs. multi-QTL

marginal contribution to LPD from QTL at λ



substitution effect: 1 QTL vs. multi-QTL

single QTL effect vs. marginal effect from QTL at λ



why use a Bayesian approach?

- first, do *both* classical and Bayesian
  - always nice to have a separate validation
  - each approach has its strengths and weaknesses
- classical approach works quite well
  - selects large effect QTL easily
  - directly builds on regression ideas for model selection
- Bayesian approach is comprehensive
  - samples most probable genetic architectures
  - formalizes model selection within one framework
  - readily (!) extends to more complicated problems

4. QTL model selection

- select class of models
  - see earlier slides above
- decide how to compare models
  - (Bayesian interval mapping talk later)
- search model space
  - (Bayesian interval mapping talk later)
- assess performance of procedure
  - see Kao (2000), Broman and Speed (2002)
  - Manichaukul, Moon, Yandell, Broman (in prep)
  - be wary of HK regression assessments

pragmatics of multiple QTL

- evaluate some objective for model given data
  - classical likelihood
  - Bayesian posterior
- search over possible genetic architectures (models)
  - number and positions of loci
  - gene action: additive, dominance, epistasis
- estimate "features" of model
  - means, variances & covariances, confidence regions
  - marginal or conditional distributions
- art of model selection
  - how select "best" or "better" model(s)?
  - how to search over useful subset of possible models?

comparing models

- balance model fit against model complexity
  - want to fit data well (maximum likelihood)
  - without getting too complicated a model

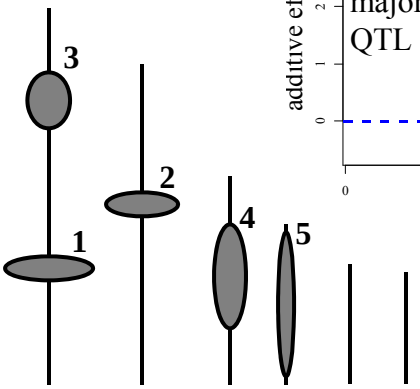
	smaller model	**bigger model**
fit model	miss key features	fits better
estimate phenotype	may be biased	no bias
predict new data	may be biased	no bias
interpret model	easier	more complicated
estimate effects	low variance	high variance

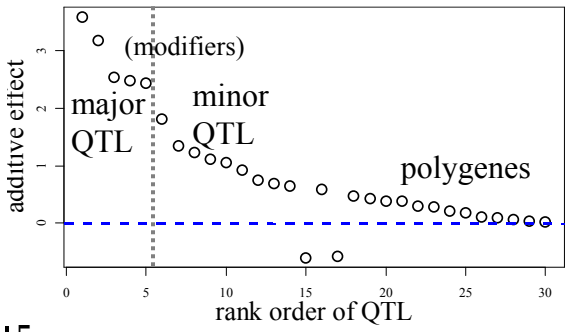
Bayesian model averaging

- average summaries over multiple architectures
- avoid selection of "best" model
- focus on "better" models
- examples in data talk later

Pareto diagram of QTL effects

major QTL on linkage map





5. QTL software options

- methods
  - approximate QTL by markers
  - exact multiple QTL interval mapping
- software platforms
  - MapMaker/QTL (obsolete)
  - QTLCart (statgen.ncsu.edu/qtlcart**)**
  - R/qtl (www.rqtl.org)
  - R/qtlbim (www.qtlbim.org)
  - Yandell, Bradbury (2007) book chapter

approximate QTL methods

- marker regression
  - locus & effect confounded
  - lose power with missing data
- Haley-Knott (least squares) regression
  - correct mean, wrong variance
  - biased by pattern of missing data (Kao 2000)
- extended HK regression
  - correct mean and variance
  - minimizes bias issue (R/qtl "ehk" method)
- composite interval mapping (QTLCart)
  - use markers to approximate other QTL
  - properties depend on marker spacing, missing data

exact QTL methods

- interval mapping (Lander, Botstein 1989)
  - scan whole genome for single QTL
  - bias for linked QTL, low power
- multiple interval mapping (Kao, Zeng, Teasdale 1999)
  - sequential scan of all QTL
  - stepwise model selection
- multiple imputation (Sen, Churchill 2001)
  - fill in (impute) missing genotypes along genome
  - average over multiple imputations
- Bayesian interval mapping (Yi et al. 2005)
  - sample most likely models
  - marginal scans conditional on other QTL

QTL software platforms

- QTLCart (statgen.ncsu.edu/qtlcart**)**
  - includes features of original MapMaker/QTL
    - not designed for building a linkage map
  - easy to use Windows version WinQTLCart
  - based on Lander-Botstein maximum likelihood LOD
    - extended to marker cofactors (CIM) and multiple QTL (MIM)
    - epistasis, some covariates (GxE)
    - stepwise model selection using information criteria
  - some multiple trait options
  - OK graphics
- R/qtl (www.rqtl.org)
  - includes functionality of classical interval mapping
  - many useful tools to check genotype data, build linkage maps
  - excellent graphics
  - several methods for 1-QTL and 2-QTL mapping
    - epistasis, covariates (GxE)
  - tools available for multiple QTL model selection

Bayesian QTL software options

- Bayesian Haley-Knott approximation: no epistasis
  - Berry C (1998)
    - R/bqtl (www.r-project.org contributed package)
- multiple imputation: epistasis, mostly 1-2 QTL but some multi-QTL
  - Sen and Churchill (2000)
    - matlab/pseudomarker (www.jax.org/staff/churchill/labsite/software)
  - Broman et al. (2003)
    - R/qtl (www.rqtl.org)
- Bayesian interval mapping via MCMC: no epistasis
  - Satagopan et al. (1996); Satagopan, Yandell (1996) Gaffney (2001)
    - R/bim (www.r-project.org contributed package)
    - WinQTLCart/bmapqtl (statgen.ncsu.edu/qtlcart)
  - Stephens & Fisch (1998): no code release
  - Sillanpää Arjas (1998)
    - multimapper (www.rni.helsinki.fi/~mjs)
- Bayesian interval mapping via MCMC: epistasis
  - Yandell et al. (2007)
    - R/qtlbim (www.qtlbim.org)
- Bayesian shrinkage: no epistasis
  - Wang et al. Xu (2005): no code release

R/qtlbim: www.qtlbim.org

- Properties
  - cross-compatible with R/qtl
  - new MCMC algorithms
    - Gibbs with loci indicators; no reversible jump
  - epistasis, fixed & random covariates, GxE
  - extensive graphics
- Software history
  - initially designed (Satagopan Yandell 1996)
  - major revision and extension (Gaffney 2001)
  - R/bim to CRAN (Wu, Gaffney, Jin, Yandell 2003)
  - R/qtlbim to CRAN (Yi, Yandell et al. 2006)
- Publications
  - Yi et al. (2005); Yandell et al. (2007); …

many thanks

U AL Birmingham
Nengjun Yi
Tapan Mehta
Samprit Banerjee
Daniel Shriner
Ram Venkataraman
David Allison

Jackson Labs
Gary Churchill
Hao Wu
Hyuna Yang
Randy von Smith

Alan Attie
Jonathan Stoehr
Hong Lan
Susie Clee
Jessica Byers
Mark Gray-Keller

Tom Osborn
David Butruille
Marcio Ferrera
Josh Udahl
Pablo Quijada

UW-Madison Stats
Yandell lab
Jaya Satagopan
Fei Zou
Patrick Gaffney
Chunfang Jin
Elias Chaibub
W Whipple Neely
Jee Young Moon
Elias Chaibub

Michael Newton
Karl Broman
Christina Kendziorski
Daniel Gianola
Liang Li
Daniel Sorensen

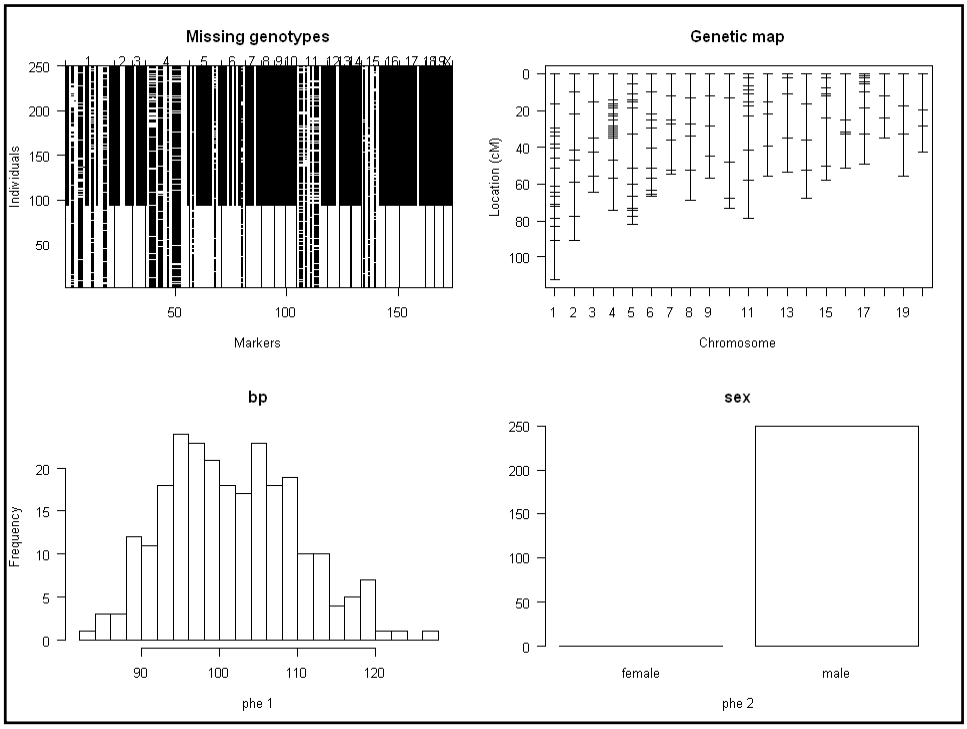
USDA Hatch, NIH/NIDDK (Attie), NIH/R01s (Yi, Broman)

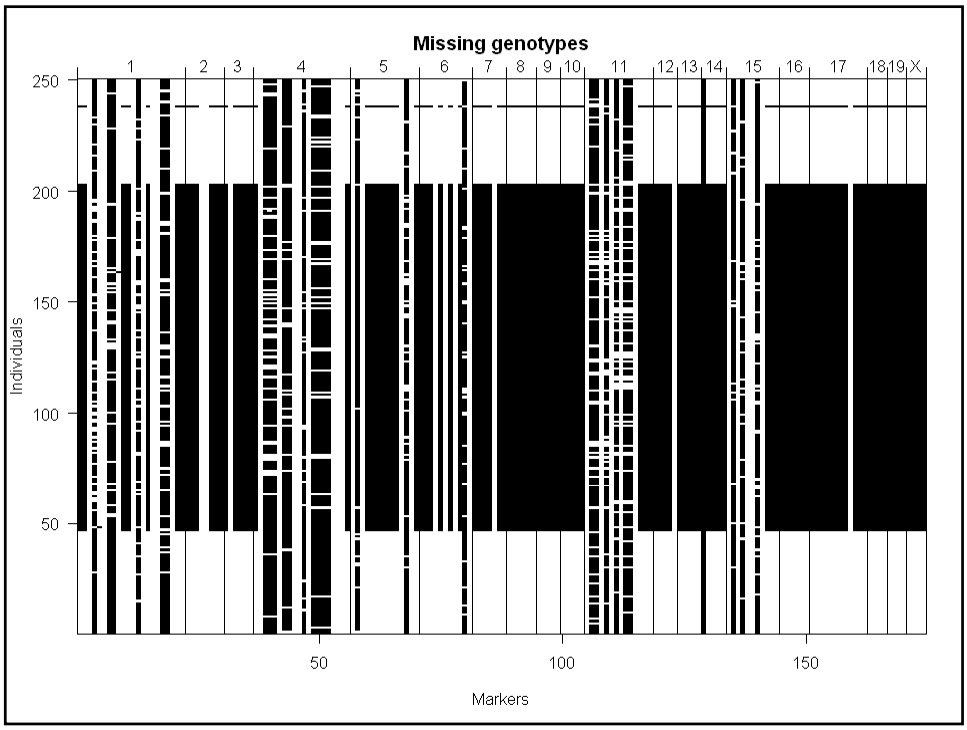
R/qtl & R/qtlbim Tutorials

- R statistical graphics & language system
- R/qtl tutorial
  - R/qtl web site: www.rqtl.org
  - Tutorial: www.rqtl.org/tutorials/rqtltour.pdf
  - R code: www.stat.wisc.edu/~yandell/qtlbim/rqtltour.R
  - url.show("http://www.stat.wisc.edu/~yandell/qtlbim/rqtltour.R")
- R/qtlbim tutorial
  - R/qtlbim web site: www.qtlbim.org
  - Tutorial and R code:
    - www.stat.wisc.edu/~yandell/qtlbim/rqtlbimtour.pdf
    - www.stat.wisc.edu/~yandell/qtlbim/rqtlbimtour.R

R/qtl tutorial (www.rqtl.org)

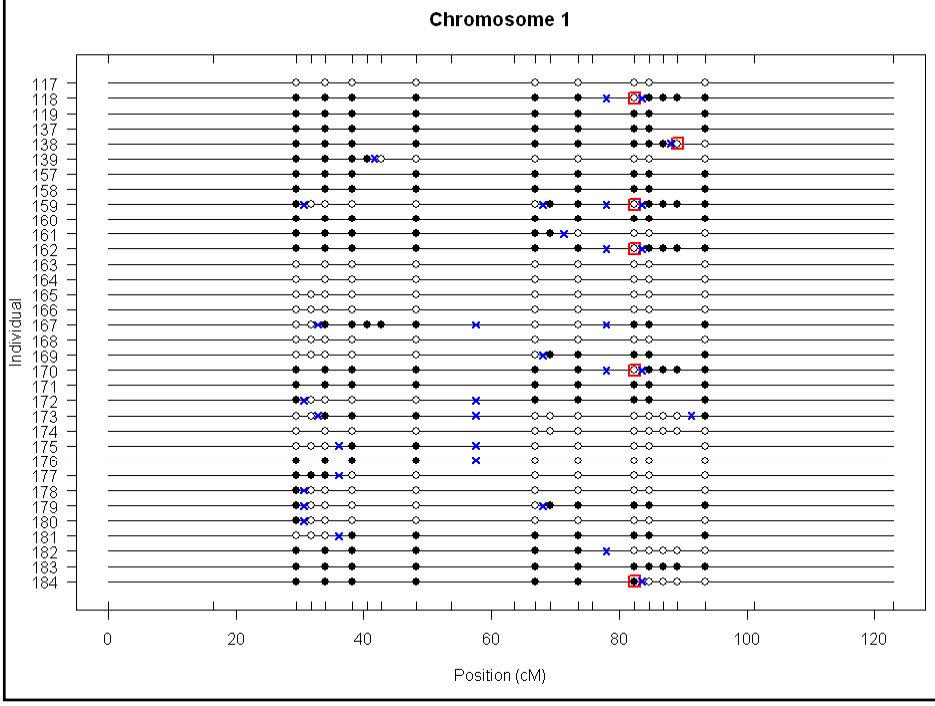
```
> library(qtl)
> data(hyper)
> summary(hyper)
    Backcross

    No. individuals:    250

    No. phenotypes:     2
    Percent phenotyped: 100 100

    No. chromosomes:    20
        Autosomes:      1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
        X chr:          X

    Total markers:      174
    No. markers:        22 8 6 20 14 11 7 6 5 5 14 5 5 5 11 6 12 4 4 4
    Percent genotyped:  47.7
    Genotypes (%):      AA:50.2  AB:49.8
> plot(hyper)
> plot.missing(hyper, reorder = TRUE)
```


Missing genotypes
Individuals
Markers
Genetic map
Location (cM)
Chromosome
bp
Frequency
phe 1
sex
female
male
phe 2


Missing genotypes
Individuals
Markers

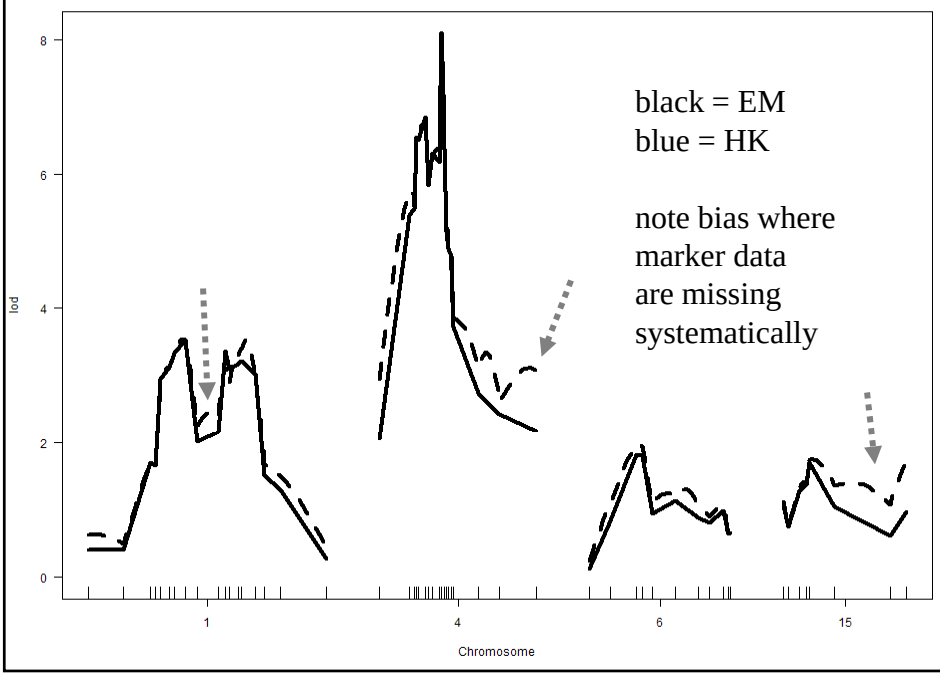
R/qtl: find genotyping errors

```
> hyper <- calc.errorlod(hyper, error.prob=0.01)
> top.errorlod(hyper)

   chr  id    marker errorlod
1    1 118   D1Mit14 8.372794
2    1 162   D1Mit14 8.372794
3    1 170   D1Mit14 8.372794
4    1 159   D1Mit14 8.350341
5    1  73   D1Mit14 6.165395
6    1  65   D1Mit14 6.165395
7    1  88   D1Mit14 6.165395
8    1 184   D1Mit14 6.151606
9    1 241   D1Mit14 6.151606
...
16   1 215  D1Mit267 5.822192
17   1 108  D1Mit267 5.822192
18   1 138  D1Mit267 5.822192
19   1 226  D1Mit267 5.822192
20   1 199  D1Mit267 5.819250
21   1  84  D1Mit267 5.808400

> plot.geno(hyper, chr=1, ind=c(117:119,137:139,157:184))
```

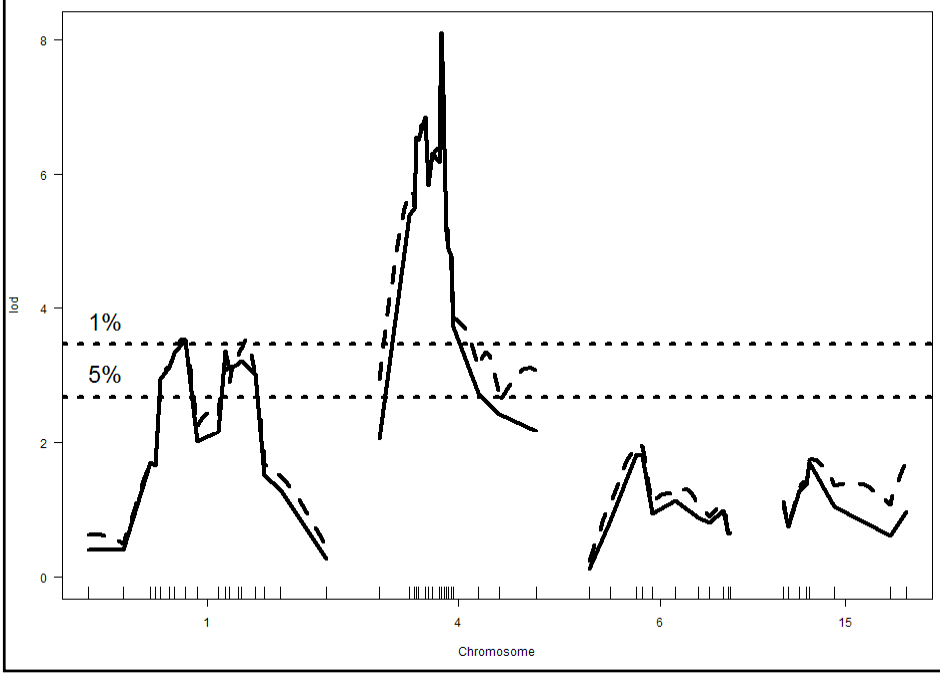
R/qtl: 1 QTL interval mapping

```
> hyper <- calc.genoprob(hyper, step=1,
  error.prob=0.01)
> out.em <- scanone(hyper)
> out.hk <- scanone(hyper, method="hk")
> summary(out.em, threshold=3)
         chr  pos  lod
c1.loc45   1 48.3 3.52
D4Mit164   4 29.5 8.02

> summary(out.hk, threshold=3)
         chr  pos  lod
c1.loc45   1 48.3 3.55
D4Mit164   4 29.5 8.09

> plot(out.em, chr = c(1,4,6,15))
> plot(out.hk, chr = c(1,4,6,15), add = TRUE, lty = 2)
```

R/qtl: permutation threshold

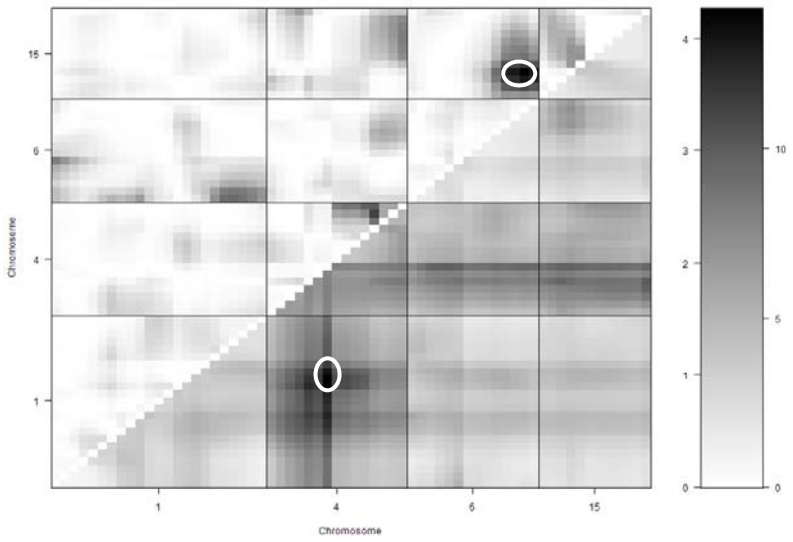
```
> operm.hk <- scanone(hyper, method="hk",
  n.perm=1000)
Doing permutation in batch mode ...
> summary(operm.hk, alpha=c(0.01,0.05))
LOD thresholds (1000 permutations)
    lod
1% 3.79
5% 2.78

> summary(out.hk, perms=operm.hk, alpha=0.05,
  pvalues=TRUE)
  chr  pos  lod  pval
1   1 48.3 3.55 0.015
2   4 29.5 8.09 0.000
```

R/qtl: 2 QTL scan

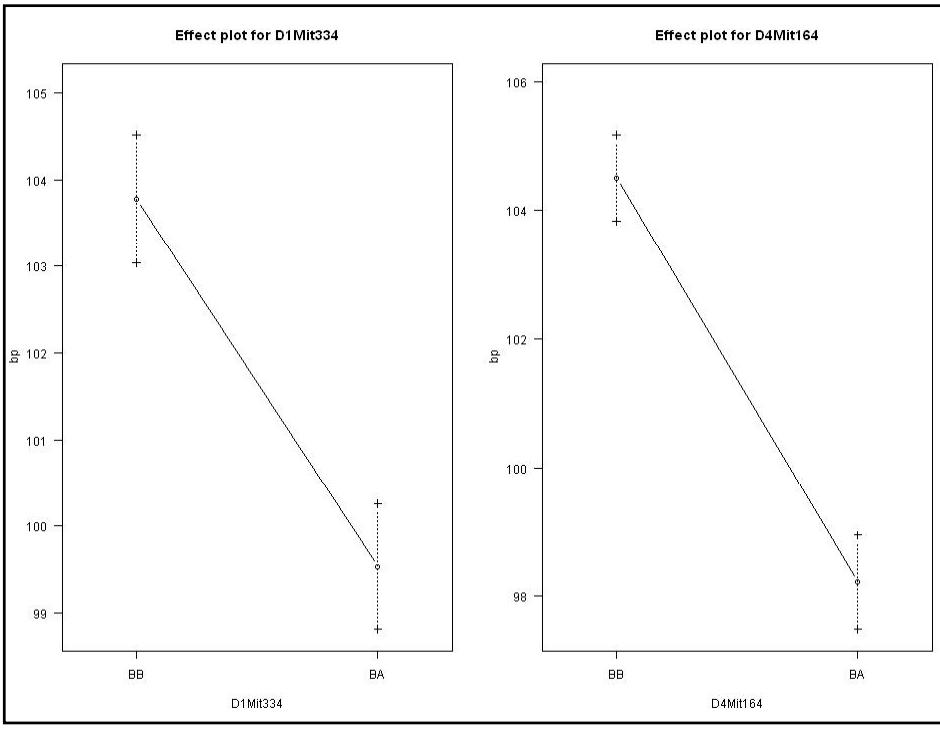
```
> hyper <- calc.genoprob(hyper, step=5, error.prob=0.01)
>
> out2.hk <- scantwo(hyper, method="hk")
 --Running scanone
 --Running scantwo
 (1,1)
 (1,2)
...
 (19,19)
 (19,X)
 (X,X)
> summary(out2.hk, thresholds=c(6.0, 4.7, 4.4, 4.7, 2.6))

        pos1f pos2f lod.full lod.fv1 lod.int     pos1a pos2a lod.add lod.av1
c1 :c4   68.3  30.0    14.13    6.51   0.225      68.3  30.0   13.90   6.288
c2 :c19  47.7   0.0     6.71    5.01   3.458      52.7   0.0    3.25   1.552
c3 :c3   37.2  42.2     6.10    5.08   0.226      37.2  42.2    5.87   4.853
c6 :c15  60.0  20.5     7.17    5.22   3.237      25.0  20.5    3.93   1.984
c9 :c18  67.0  37.2     6.31    4.79   4.083      67.0  12.2    2.23   0.708
c12:c19   1.1  40.0     6.48    4.79   4.090       1.1   0.0    2.39   0.697

> plot(out2.hk, chr=c(1,4,6,15))
```

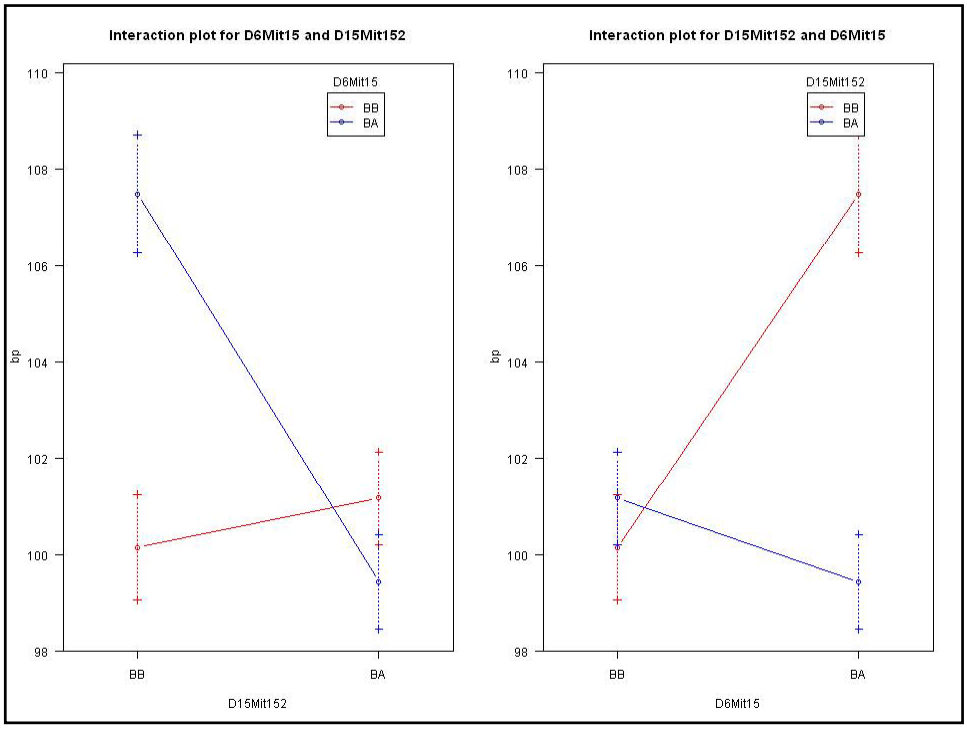

upper triangle/left scale: epistasis LOD
lower triangle/right scale: 2-QTL LOD

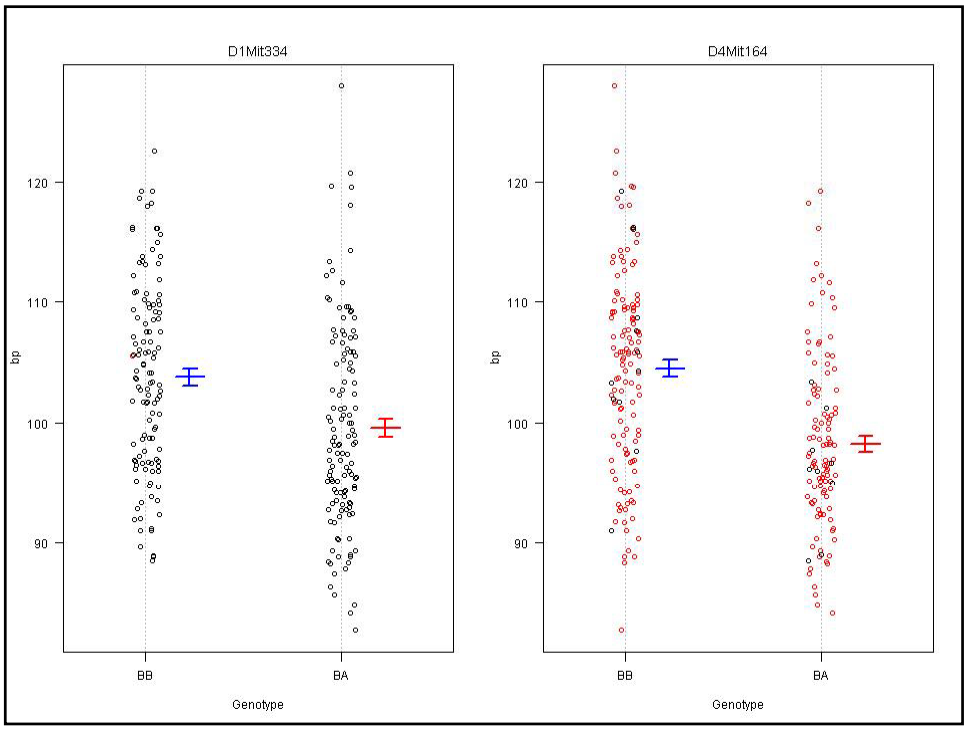


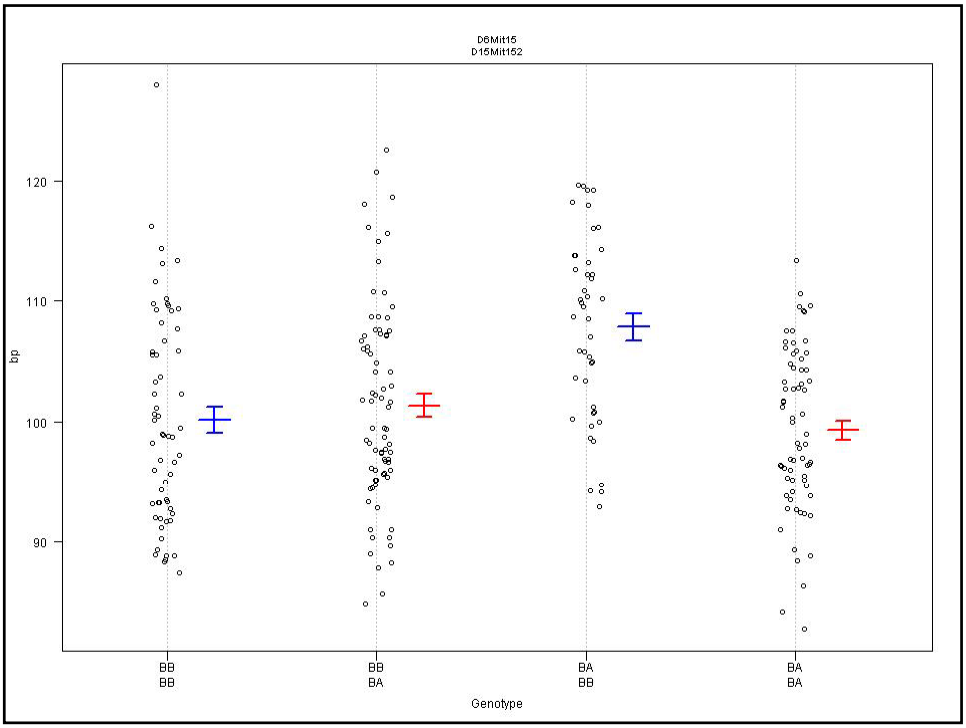
Effect & Interaction Plots

```
## Effect plots and interaction plot.
hyper <- sim.geno(hyper, step=2, n.draws=16, error.prob=0.01)
effectplot(hyper, pheno.col = 1, mname1 = "D1Mit334")
effectplot(hyper, pheno.col = 1, mname1 = "D4Mit164")
markers <- find.marker(hyper, chr = c(6,15), pos = c(70,20))
effectplot(hyper, pheno.col = 1,
   mname1 = markers[1], mname2 = markers[2])
effectplot(hyper, pheno.col = 1,
   mname1 = markers[2], mname2 = markers[1])

## Strip plot of data (phenotype by genotype).
plot.pxg(hyper, "D1Mit334")
plot.pxg(hyper, "D4Mit164")
plot.pxg(hyper, markers)
```


Interaction plot for D6Mit15 and D15Mit152
D6Mit15
BB
BA
bp
D15Mit152
Interaction plot for D15Mit152 and D6Mit15
D15Mit152
BB
BA
bp
D6Mit15


D1Mit334
bp
Genotype
D4Mit164
bp
Genotype



R/qtl: ANOVA imputation at QTL

```
> hyper <- sim.geno(hyper, step=2, n.draws=16, error.prob=0.01)
> qtl <- makeqtl(hyper, chr = c(1, 1, 4, 6, 15), pos = c(50, 76, 30, 70, 20))

> my.formula <- y ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q4:Q5
> out.fitqtl <- fitqtl(hyper, pheno.col = 1, qtl, formula = my.formula)
> summary(out.fitqtl)

Full model result
----------------------------------
Model formula is:  y ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q4:Q5

       df        SS        MS      LOD     %var Pvalue(Chi2) Pvalue(F)
Model   6  5789.089 964.84822 21.54994 32.76422           0         0
Error 243 11879.847  48.88826
Total 249 17668.936

Drop one QTL at a time ANOVA table:
----------------------------------
                 df Type III SS      LOD     %var F value Pvalue(F)
Chr1@50           1     297.149    1.341    1.682   6.078   0.01438 *
Chr1@76           1     520.664    2.329    2.947  10.650   0.00126 **
Chr4@30           1    2842.089   11.644   16.085  58.134  5.50e-13 ***
Chr6@70           2    1435.721    6.194    8.126  14.684  9.55e-07 ***
Chr15@20          2    1083.842    4.740    6.134  11.085  2.47e-05 ***
Chr6@70:Chr15@20  1     955.268    4.199    5.406  19.540  1.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

selected R/qtl publications

www.stat.wisc.edu/~yandell/statgen

- www.rqtl.org
- tutorials and code at web site
  - www.rqtl.org/tutorials
- Broman et al. (2003 *Bioinformatics*)
  - R/qtl introduction
- Broman (2001 *Lab Animal*)
  - nice overview of QTL issues
- Broman & Sen 2009 book (*Springer*)

R/qtlbim (www.qtlbim.org)

- cross-compatible with R/qtl
- model selection for genetic architecture
  - epistasis, fixed & random covariates, GxE
  - samples multiple genetic architectures
  - examines summaries over nested models
- extensive graphics

```
> url.show("http://www.stat.wisc.edu/~yandell/qtlbim/rqtlbimtour.R")
```

R/qtlbim: tutorial

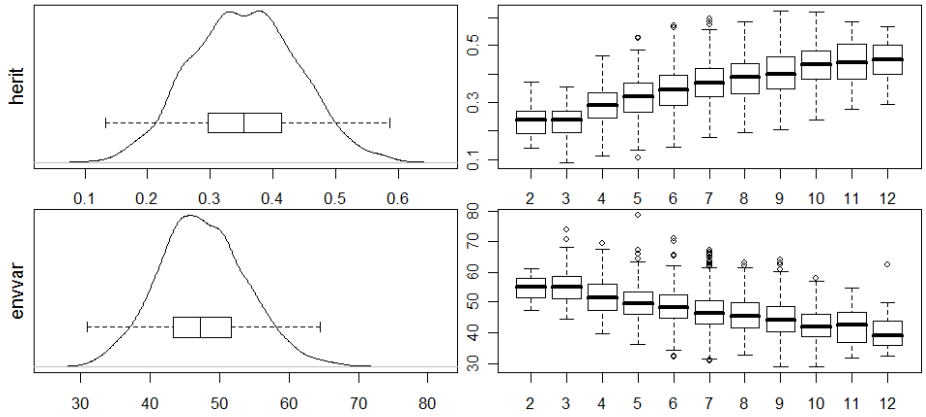
(www.stat.wisc.edu/~yandell/qtlbim)

```
> data(hyper)
## Drop X chromosome (for now).
> hyper <- subset(hyper, chr=1:19)
> hyper <- qb.genoprob(hyper, step=2)
## This is the time-consuming step:
> qbHyper <- qb.mcmc(hyper, pheno.col = 1)
## Here we get stored samples.
> data(qbHyper)
> summary(qbHyper)
```


R/qtlbim: initial summaries

```
> summary(qbHyper)

Bayesian model selection QTL mapping object qbHyper on cross object hyper
had 3000 iterations recorded at each 40 steps with 1200 burn-in steps.

Diagnostic summaries:
         nqtl   mean envvar varadd  varaa    var
Min.    2.000  97.42  28.07  5.112  0.000  5.112
1st Qu. 5.000 101.00  44.33 17.010  1.639 20.180
Median  7.000 101.30  48.57 20.060  4.580 25.160
Mean    6.543 101.30  48.80 20.310  5.321 25.630
3rd Qu. 8.000 101.70  53.11 23.480  7.862 30.370
Max.   13.000 103.90  74.03 51.730 34.940 65.220

Percentages for number of QTL detected:
 2  3  4  5  6  7  8  9 10 11 12 13
 2  3  9 14 21 19 17 10  4  1  0  0

Percentages for number of epistatic pairs detected:
pairs
 1  2  3  4  5  6
29 31 23 11  5  1

Percentages for common epistatic pairs:
 6.15  4.15   4.6   1.7 15.15   1.4   1.6   4.9  1.15  1.17   1.5  5.11   1.2  7.15   1.1
   63    18    10     6     6     5     4     4     3     3     3     2     2     2     2

> plot(qb.diag(qbHyper, items = c("herit", "envvar")))
```

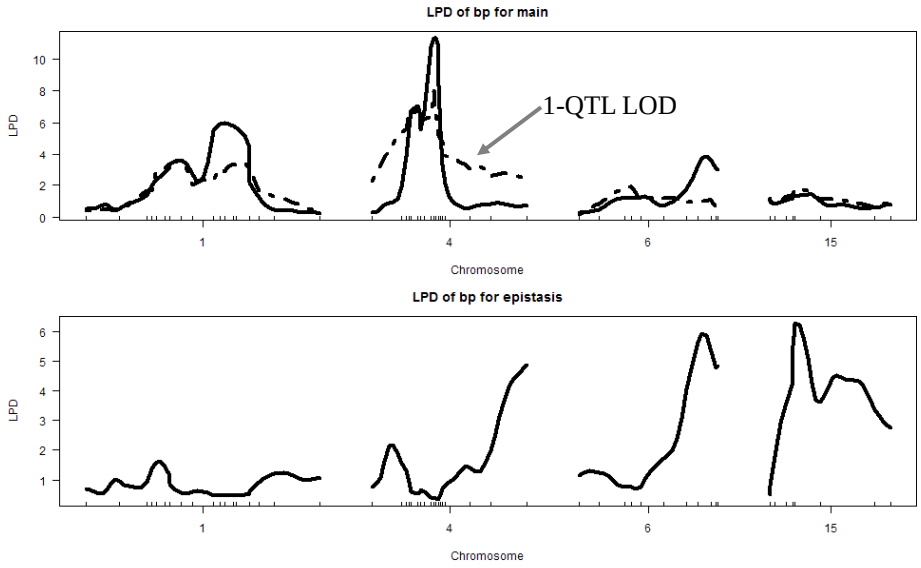
diagnostic summaries



R/qtlbim: 1-D (*not* 1-QTL!) scan

```
> one <- qb.scanone(qbHyper, chr = c(1,4,6,15), type =
   "LPD")
> summary(one)

LPD of bp for main,epistasis,sum

    n.qtl  pos m.pos e.pos  main epistasis   sum
c1  1.331 64.5  64.5  67.8  6.10     0.442  6.27
c4  1.377 29.5  29.5  29.5 11.49     0.375 11.61
c6  0.838 59.0  59.0  59.0  3.99     6.265  9.60
c15 0.961 17.5  17.5  17.5  1.30     6.325  7.28

> plot(one, scan = "main")
> plot(out.em, chr=c(1,4,6,15), add = TRUE, lty = 2)
> plot(one, scan = "epistasis")
```

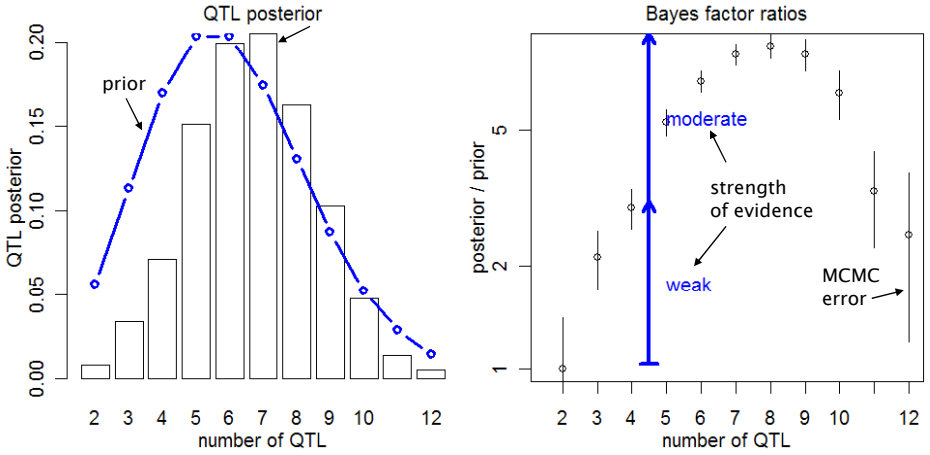
1-QTL LOD vs. marginal LPD



most probable patterns

```
> summary(qb.BayesFactor(qbHyper, item = "pattern"))

                  nqtl posterior    prior    bf  bfse
1,4,6,15,6:15        5   0.03400 2.71e-05 24.30 2.360
1,4,6,6,15,6:15      6   0.00467 5.22e-06 17.40 4.630
1,1,4,6,15,6:15      6   0.00600 9.05e-06 12.80 3.020
1,1,4,5,6,15,6:15    7   0.00267 4.11e-06 12.60 4.450
1,4,6,15,15,6:15     6   0.00300 4.96e-06 11.70 3.910
1,4,4,6,15,6:15      6   0.00300 5.81e-06 10.00 3.330
1,2,4,6,15,6:15      6   0.00767 1.54e-05  9.66 2.010
1,4,5,6,15,6:15      6   0.00500 1.28e-05  7.56 1.950
1,2,4,5,6,15,6:15    7   0.00267 6.98e-06  7.41 2.620
1,4                  2   0.01430 1.51e-04  1.84 0.279
1,1,2,4              4   0.00300 3.66e-05  1.59 0.529
1,2,4                3   0.00733 1.03e-04  1.38 0.294
1,1,4                3   0.00400 6.05e-05  1.28 0.370
1,4,19               3   0.00300 5.82e-05  1.00 0.333

> plot(qb.BayesFactor(qbHyper, item = "nqtl"))
```

hyper: number of QTL
posterior, prior, Bayes factors

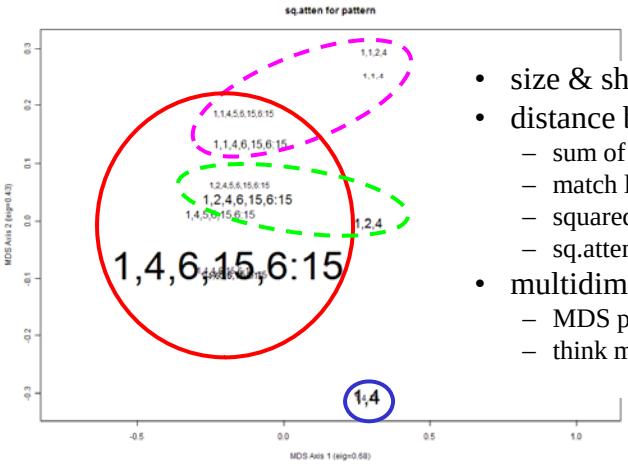


what is best estimate of QTL?

- **find most probable pattern**
  - **1,4,6,15,6:15 has posterior of 3.4%**
- **estimate locus across all nested patterns**
  - **Exact pattern seen ~100/3000 samples**
  - **Nested pattern seen ~2000/3000 samples**
- **estimate 95% confidence interval using quantiles**

```
> best <- qb.best(qbHyper)
> summary(best)$best

    chrom locus locus.LCL locus.UCL     n.qtl
247     1  69.9  24.44875   95.7985 0.8026667
245     4  29.5  14.20000   74.3000 0.8800000
248     6  59.0  13.83333   66.7000 0.7096667
246    15  19.5  13.10000   55.7000 0.8450000

> plot(best)
```

what patterns are “near” the best?

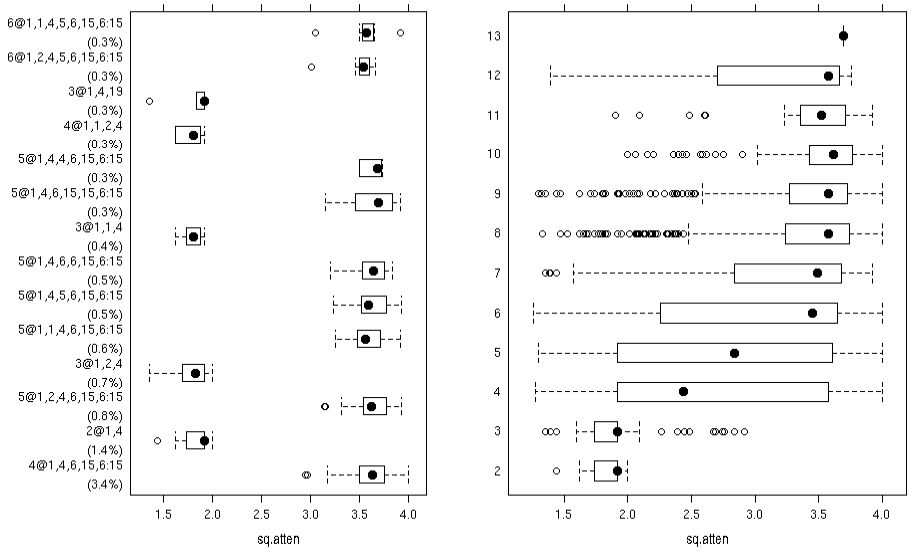


- size & shade ~ posterior
- distance between patterns
  - sum of squared attenuation
  - match loci between patterns
  - squared attenuation = $(1\text{-}2r)^2$
  - sq.atten in scale of LOD & LPD
- multidimensional scaling
  - MDS projects distance onto 2-D
  - think mileage between cities

how close are other patterns?

```
> target <- qb.best(qbHyper)$model[[1]]
> summary(qb.close(qbHyper, target))

 score by sample number of qtl
    Min. 1st Qu. Median  Mean 3rd Qu.  Max.
2  1.437   1.735  1.919 1.834   1.919 2.000
3  1.351   1.735  1.916 1.900   1.919 2.916
4  1.270   1.916  2.437 2.648   3.574 4.000
5  1.295   1.919  2.835 2.798   3.611 4.000
6  1.257   2.254  3.451 3.029   3.648 4.000
...
13 3.694   3.694  3.694 3.694   3.694 3.694

 score by sample chromosome pattern
                     Percent  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
4@1,4,6,15,6:15          3.4 2.946   3.500  3.630 3.613   3.758 4.000
2@1,4                    1.4 1.437   1.735  1.919 1.832   1.919 2.000
5@1,2,4,6,15,6:15        0.8 3.137   3.536  3.622 3.611   3.777 3.923
3@1,2,4                  0.7 1.351   1.700  1.821 1.808   1.919 2.000
5@1,1,4,6,15,6:15        0.6 3.257   3.484  3.563 3.575   3.698 3.916
5@1,4,5,6,15,6:15        0.5 3.237   3.515  3.595 3.622   3.777 3.923
5@1,4,6,6,15,6:15        0.5 3.203   3.541  3.646 3.631   3.757 3.835
...

> plot(close)
> plot(close, category = "nqtl")
```

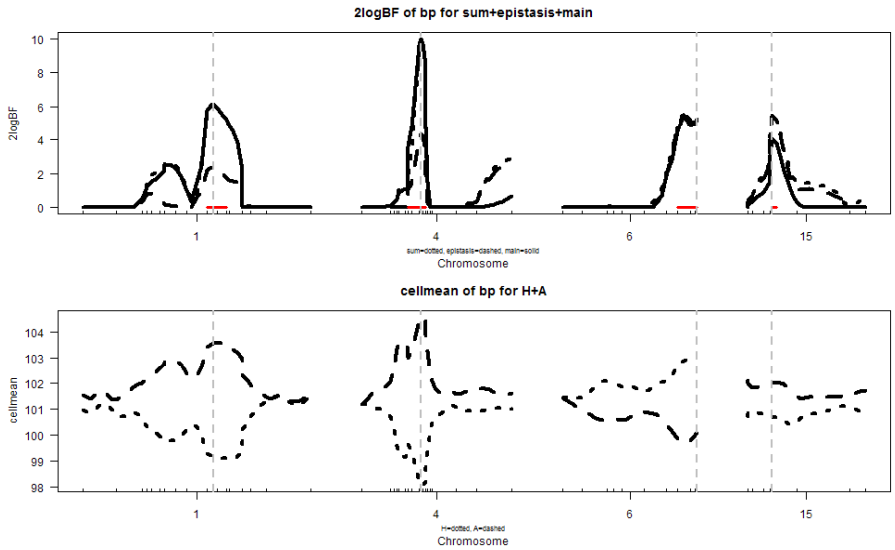
how close are other patterns?



R/qtlbim: automated QTL selection

```
> hpd <- qb.hpdone(qbHyper, profile = "2logBF")
> summary(hpd)

   chr n.qtl  pos lo.50% hi.50% 2logBF       A       H
1    1 0.829 64.5   64.5   72.1  6.692 103.611  99.090
4    4 3.228 29.5   25.1   31.7 11.169 104.584  98.020
6    6 1.033 59.0   56.8   66.7  6.054  99.637 102.965
15  15 0.159 17.5   17.5   17.5  5.837 101.972 100.702

> plot(hpd)
```

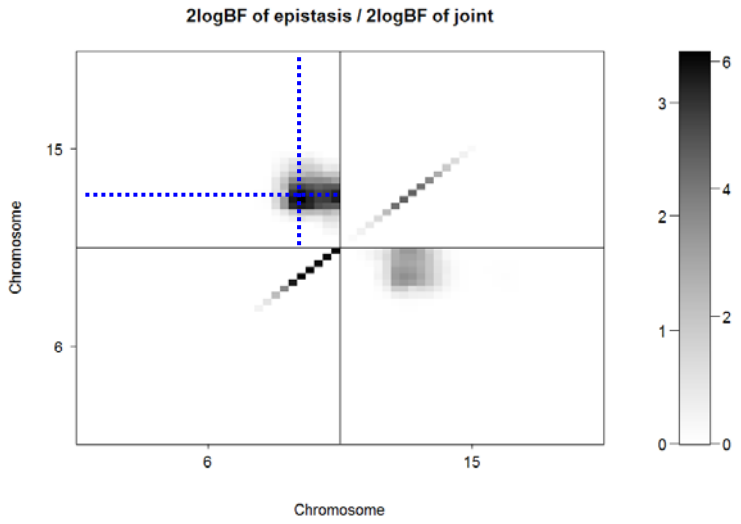
2log(BF) scan with 50% HPD region



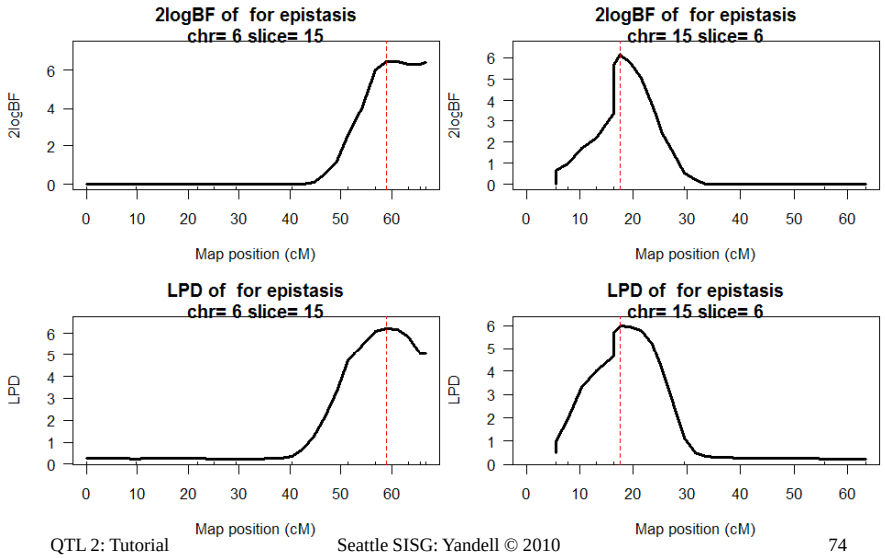
R/qtlbim: 2-D (*not* 2-QTL) scans

```
> two <- qb.scantwo(qbHyper, chr = c(6,15),
  type = "2logBF")
> plot(two)

> plot(two, chr = 6, slice = 15)
> plot(two, chr = 15, slice = 6)

> two.lpd <- qb.scantwo(qbHyper, chr = c(6,15),
  type = "LPD")
> plot(two.lpd, chr = 6, slice = 15)
> plot(two.lpd, chr = 15, slice = 6)
```

2-D plot of 2logBF: chr 6 & 15

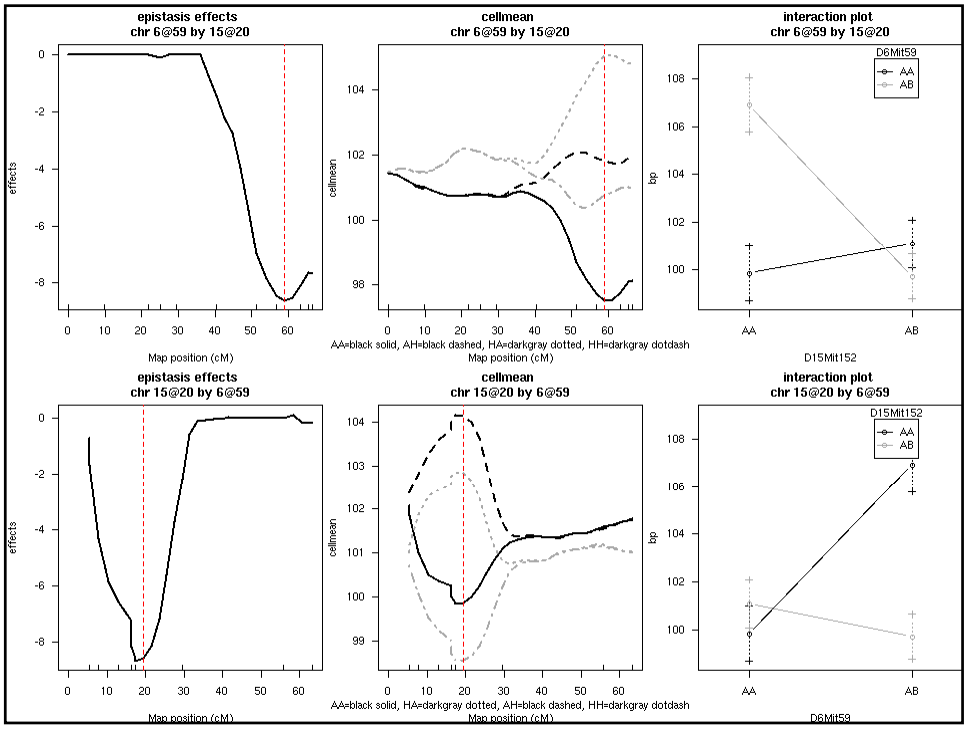


1-D Slices of 2-D scans: chr 6 & 15



R/qtlbim: slice of epistasis

```
> slice <- qb.slicetwo(qbHyper, c(6,15), c(59,19.5))
> summary(slice)

2logBF of bp for epistasis

    n.qtl  pos m.pos e.pos epistasis slice
c6  0.838 59.0  59.0  66.7      15.8  18.1
c15 0.961 17.5  17.5  17.5      15.5  60.6

cellmean of bp for AA,HA,AH,HH

    n.qtl  pos m.pos   AA  HA  AH    HH slice
c6  0.838 59.0  59.0 97.4 105 102 100.8  18.1
c15 0.961 17.5  17.5 99.8 103 104  98.5  60.6

estimate of bp for epistasis

    n.qtl  pos m.pos e.pos epistasis slice
c6  0.838 59.0  59.0  66.7     -7.86  18.1
c15 0.961 17.5  17.5  17.5     -8.72  60.6

> plot(slice, figs = c("effects", "cellmean", "effectplot"))
```

selected publications

www.stat.wisc.edu/~yandell/statgen

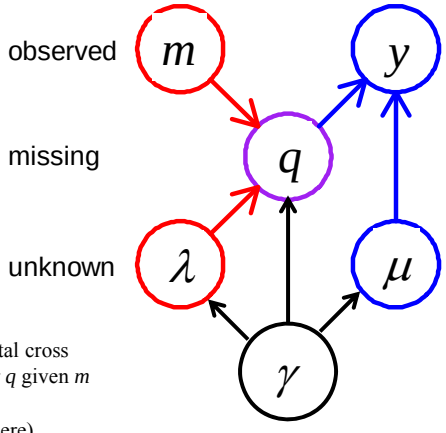
- www.qtlbim.org
- vignettes in R/qtlbim package
- Yandell, Bradbury (2007) *Plant Map* book chapter
  - overview/comparison of QTL methods
- Yandell et al. (2007 *Bioinformatics*)
  - R/qtlbim introduction
- Yi et al. (2005 *Genetics*, 2007 *Genetics*)
  - methodology of R/qtlbim

Bayesian Interval Mapping

1. Bayesian strategy
2. Markov chain sampling
3. sampling genetic architectures
4. criteria for model selection

QTL model selection: key players

- observed measurements
  - y = phenotypic trait
  - m = markers & linkage map
  - i = individual index (1,…,n)
- missing data
  - missing marker data
  - q = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - λ = QT locus (or loci)
  - μ = phenotype model parameters
  - γ = QTL model/genetic architecture
- pr($q|m,\lambda,\gamma$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for q given m
- pr($y|q,\mu,\gamma$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters μ (could be non-parametric)



after
Sen Churchill (2001)

1. Bayesian strategy for QTL study

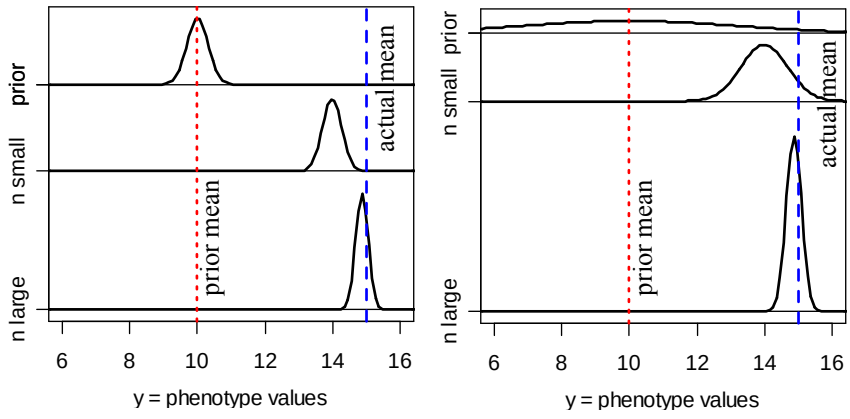
- augment data (y,m) with missing genotypes q
- study unknowns (μ,λ,γ) given augmented data (y,m,q)
  - find better genetic architectures γ
  - find most likely genomic regions = QTL = λ
  - estimate phenotype parameters = genotype means = μ
- sample from posterior in some clever way
  - multiple imputation (Sen Churchill 2002)
  - Markov chain Monte Carlo (MCMC)
    - (Satagopan et al. 1996; Yi et al. 2005, 2007)

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{constant}}$$

$$\text{posterior for } q,\mu,\lambda,\gamma = \frac{\text{phenotype likelihood} * [\text{prior for } q,\mu,\lambda,\gamma]}{\text{constant}}$$

$$\text{pr}(q,\mu,\lambda,\gamma \mid y,m) = \frac{\text{pr}(y \mid q,\mu,\gamma) * [\text{pr}(q \mid m,\lambda,\gamma)\text{pr}(\mu \mid \gamma)\text{pr}(\lambda \mid m,\gamma)\text{pr}(\gamma)]}{\text{pr}(y \mid m)}$$

Bayes posterior for normal data



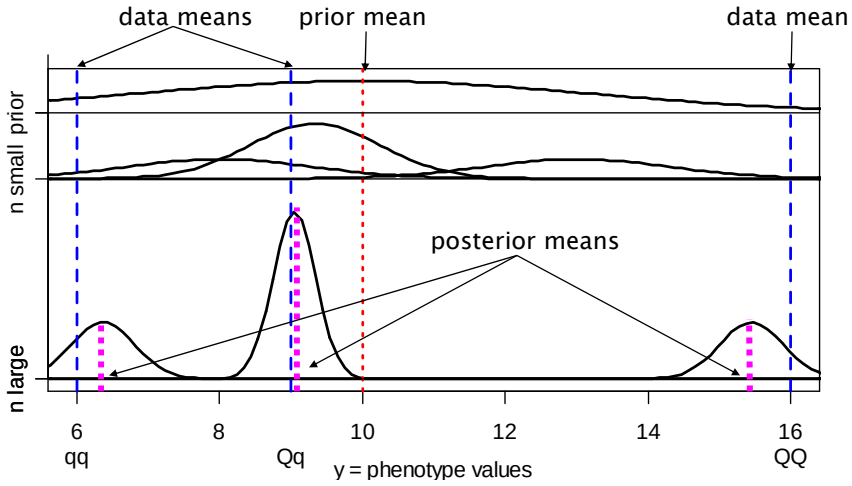
small prior variance large prior variance

Bayes posterior for normal data

model	$y_i = \mu + e_i$
environment	e ~ $N(0, \sigma^2)$, σ^2 known
likelihood	$y \sim N(\mu, \sigma^2)$
prior	$\mu \sim N(\mu_0, \kappa\sigma^2)$, κ known
posterior:	mean tends to sample mean
single individual	$\mu \sim N(\mu_0 + b_1(y_1 - \mu_0), b_1\sigma^2)$
sample of n individuals	$\mu \sim N\left(b_n\bar{y}_\bullet + (1-b_n)\mu_0, b_n\sigma^2/n\right)$ with $\bar{y}_\bullet = \underset{\{i=1,\ldots,n\}}{\text{sum}} y_i/n$
shrinkage factor (shrinks to 1)	$b_n = \frac{\kappa n}{\kappa n + 1} \rightarrow 1$

what values are the genotypic means?

phenotype model pr($y|q,\mu$)



Bayes posterior QTL means

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

phenotype mean: $E(y \mid q) = \mu_q \qquad V(y \mid q) = \sigma^2$

genotypic prior: $E(\mu_q) = \bar{y}_\bullet \qquad V(\mu_q) = \kappa\sigma^2$

posterior: $E(\mu_q \mid y) = b_q \bar{y}_q + (1 - b_q)\bar{y}_\bullet \qquad V(\mu_q \mid y) = b_q \sigma^2 / n_q$

$n_q = \text{count}\{q_i = q\} \qquad \bar{y}_q = \underset{\{q_i = q\}}{\text{sum}} y_i / n_q$

shrinkage: $b_q = \dfrac{\kappa n_q}{\kappa n_q + 1} \to 1$

partition genotypic effects on phenotype

- phenotype depends on genotype
- genotypic value partitioned into
  - main effects of single QTL
  - epistasis (interaction) between pairs of QTL

$$\mu_q = \beta_0 + \beta_q = E(Y;q)$$
$$\beta_q = \beta(q_2) + \beta(q_2) + \beta(q_1, q_2)$$

partitition genotypic variance

- consider same 2 QTL + epistasis

- centering variance $V(\beta_0) = \kappa_0 \sigma^2 = s^2$

- genotypic variance $V(\beta_q) = \kappa_1 \sigma^2 = \sigma_q^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}^2$

- heritability $h_q^2 = \frac{\sigma_q^2}{\sigma_q^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2$

posterior mean ≈ LS estimate

$$\beta_q \mid y \sim N(b_q \hat{\beta}_q, b_q C_q \sigma^2)$$

$$\approx N(\hat{\beta}_q, C_q \sigma^2)$$

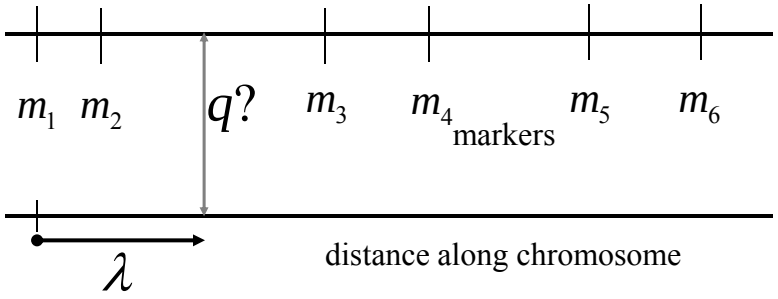
LS estimate $\hat{\beta}_q = \text{sum}_i[\text{sum}_j \hat{\beta}(q_{ij})] = \text{sum}_i w_{qi} y_i$

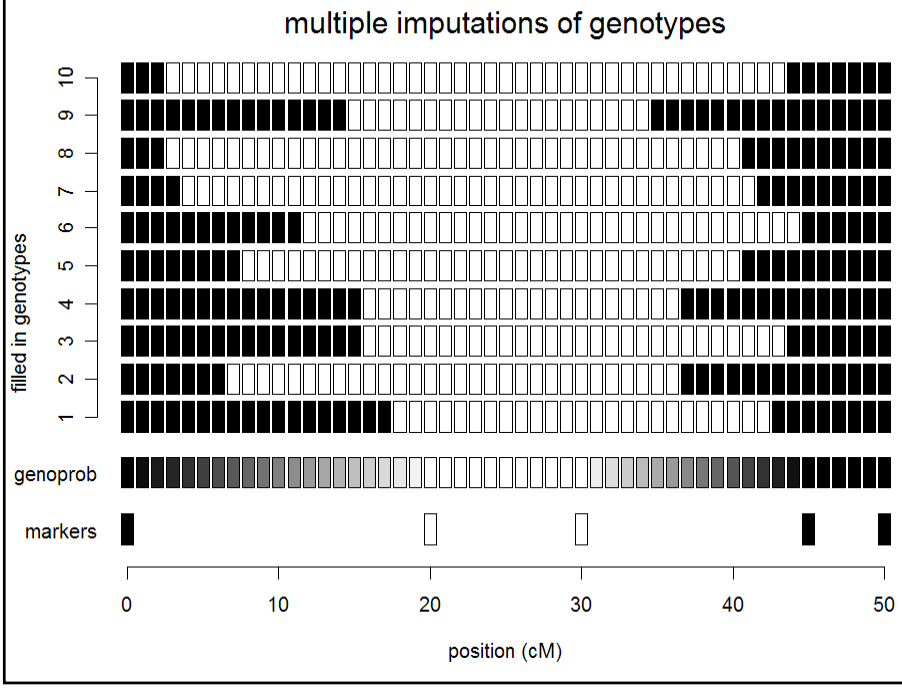
variance $V(\hat{\beta}_q) = \text{sum}_i w_{qi}^2 \sigma^2 = C_q \sigma^2$

shrinkage $b_q = \kappa_1 / (\kappa_1 + C_q) \to 1$

pr($q|m,\lambda$) recombination model

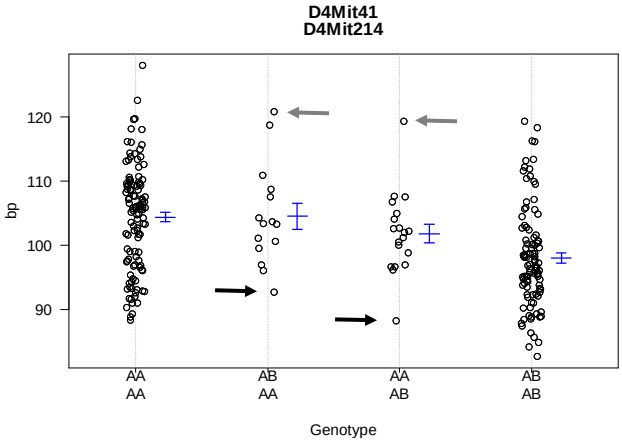
pr($q|m,\lambda$) = pr(geno | map, locus) ≈
pr(geno | flanking markers, locus)





what are likely QTL genotypes *q*?

how does phenotype *y* improve guess?



what are probabilities for genotype *q* between markers?

recombinants AA:AB

all 1:1 if ignore *y* and if we use *y*?

posterior on QTL genotypes *q*

- full conditional of *q* given data, parameters
  - proportional to prior pr($q \mid m, \lambda$)
    - weight toward *q* that agrees with flanking markers
  - proportional to likelihood pr($y \mid q, \mu$)
    - weight toward *q* with similar phenotype values
  - posterior recombination model balances these two
- this *is* the E-step of EM computations

$$\text{pr}(q \mid y, m, \mu, \lambda) = \frac{\text{pr}(y \mid q, \mu) * \text{pr}(q \mid m, \lambda)}{\text{pr}(y \mid m, \mu, \lambda)}$$

Where are the loci λ on the genome?

- prior over genome for QTL positions
  - flat prior = no prior idea of loci
  - or use prior studies to give more weight to some regions
- posterior depends on QTL genotypes q

  $\text{pr}(\lambda \mid m,q) = \text{pr}(\lambda)\ \text{pr}(q \mid m,\lambda) / \text{constant}$
  - constant determined by averaging
    - over all possible genotypes q
    - over all possible loci λ on entire map
- no easy way to write down posterior

what is the genetic architecture γ?

- which positions correspond to QTLs?
  - priors on loci (previous slide)
- which QTL have main effects?
  - priors for presence/absence of main effects
    - same prior for all QTL
    - can put prior on each d.f. (1 for BC, 2 for F2)
- which pairs of QTL have epistatic interactions?
  - prior for presence/absence of epistatic pairs
    - depends on whether 0,1,2 QTL have main effects
    - epistatic effects less probable than main effects

γ = genetic architecture:

loci:

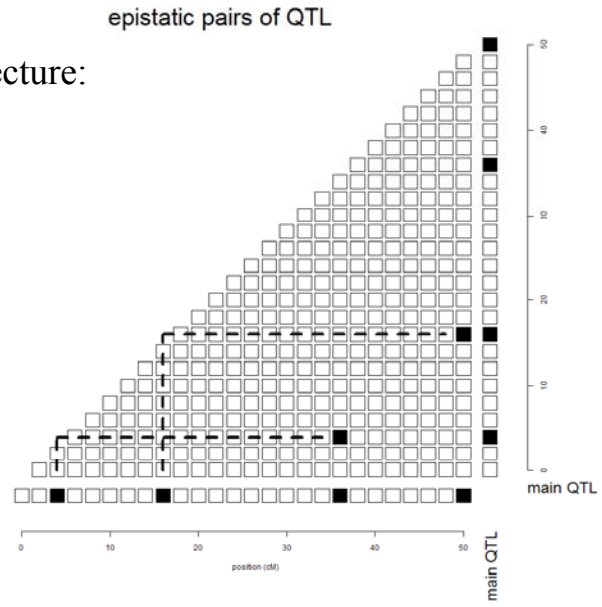
main QTL

epistatic pairs

effects:

add, dom

aa, ad, dd



Bayesian priors & posteriors

- augmenting with missing genotypes q
  - prior is recombination model
  - posterior is (formally) E step of EM algorithm
- sampling phenotype model parameters μ
  - prior is "flat" normal at grand mean (no information)
  - posterior shrinks genotypic means toward grand mean
  - (details for unexplained variance omitted here)
- sampling QTL loci λ
  - prior is flat across genome (all loci equally likely)
- sampling QTL genetic architecture model γ
  - number of QTL
    - prior is Poisson with mean from previous IM study
  - genetic architecture of main effects and epistatic interactions
    - priors on epistasis depend on presence/absence of main effects

2. Markov chain sampling

- construct Markov chain around posterior
  - want posterior as stable distribution of Markov chain
  - in practice, the chain tends toward stable distribution
    - initial values may have low posterior probability
    - burn-in period to get chain mixing well
- sample QTL model components from full conditionals
  - sample locus λ given q,γ (using Metropolis-Hastings step)
  - sample genotypes q given λ,μ,y,γ (using Gibbs sampler)
  - sample effects μ given q,y,γ (using Gibbs sampler)
  - sample QTL model γ given λ,μ,y,q (using Gibbs or M-H)

$$(\lambda,q,\mu,\gamma) \sim \text{pr}(\lambda,q,\mu,\gamma \mid y,m)$$

$$(\lambda,q,\mu,\gamma)_1 \rightarrow (\lambda,q,\mu,\gamma)_2 \rightarrow \cdots \rightarrow (\lambda,q,\mu,\gamma)_N$$

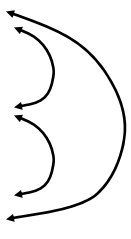
MCMC sampling of unknowns (q,μ,λ)
for given genetic architecture γ

- Gibbs sampler
  - genotypes q
  - effects μ
  - *not* loci λ

$$q \sim \text{pr}(q \mid y_i, m_i, \mu, \lambda)$$

$$\mu \sim \frac{\text{pr}(y \mid q,\mu)\text{pr}(\mu)}{\text{pr}(y \mid q)}$$

$$\lambda \sim \frac{\text{pr}(q \mid m,\lambda)\text{pr}(\lambda \mid m)}{\text{pr}(q \mid m)}$$



- Metropolis-Hastings sampler
  - extension of Gibbs sampler
  - does not require normalization
    - pr($q \mid m$) = sum$_\lambda$ pr($q \mid m, \lambda$) pr(λ)

Gibbs sampler
for two genotypic means

- want to study two correlated effects
  - could sample directly from their bivariate distribution
  - assume correlation ρ is known
- instead use Gibbs sampler:
  - sample each effect from its full conditional given the other
  - pick order of sampling at random
  - repeat many times

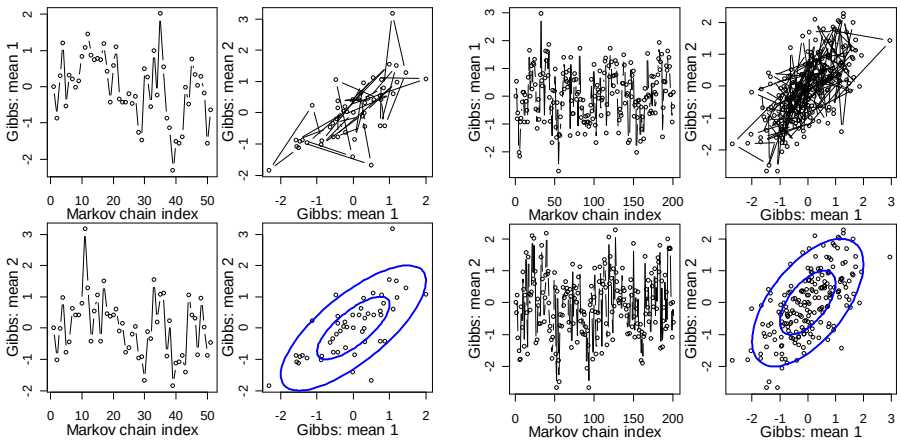
$$\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

$$\mu_1 \sim N\left(\rho\mu_2, 1-\rho^2\right)$$

$$\mu_2 \sim N\left(\rho\mu_1, 1-\rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples $N = 200$ samples



full conditional for locus

- cannot easily sample from locus full conditional

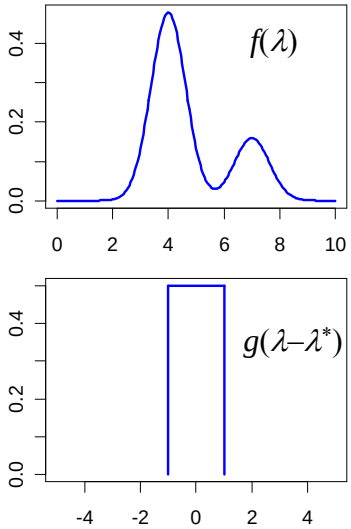
  $\text{pr}(\lambda \mid y,m,\mu,q) = \text{pr}(\lambda \mid m,q)$

  $= \text{pr}(q \mid m, \lambda)\ \text{pr}(\lambda) / \text{constant}$
- constant is very difficult to compute explicitly
  - must average over all possible loci λ over genome
  - must do this for every possible genotype q
- Gibbs sampler will not work in general
  - but can use method based on ratios of probabilities
  - Metropolis-Hastings is extension of Gibbs sampler

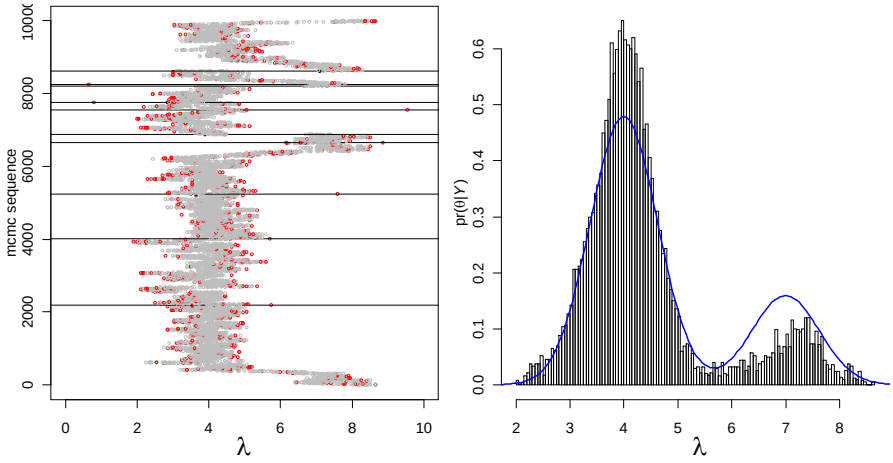
Metropolis-Hastings idea

- want to study distribution $f(\lambda)$
  - take Monte Carlo samples
    - unless too complicated
  - take samples using ratios of f
- Metropolis-Hastings samples:
  - propose new value λ^*
    - near (?) current value λ
    - from some distribution g
  - accept new value with prob a
    - Gibbs sampler: $a = 1$ always

$$a = \min\left(1, \frac{f(\lambda^*)g(\lambda^* - \lambda)}{f(\lambda)g(\lambda - \lambda^*)}\right)$$

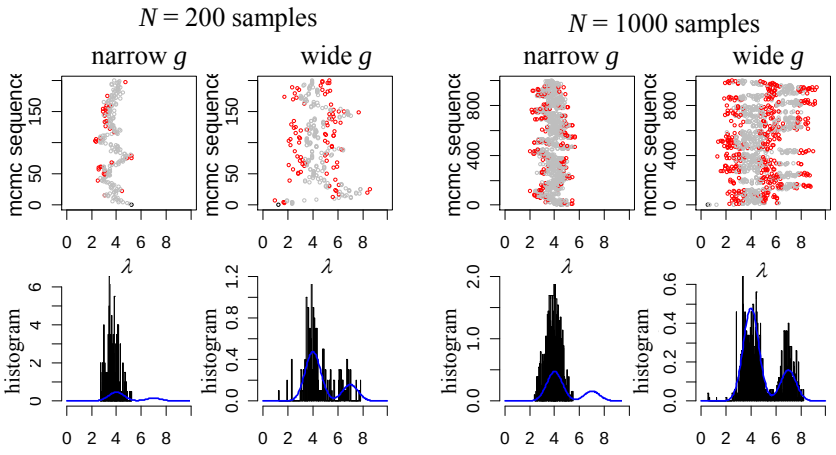


Metropolis-Hastings for locus λ



added twist: occasionally propose from entire genome

Metropolis-Hastings samples

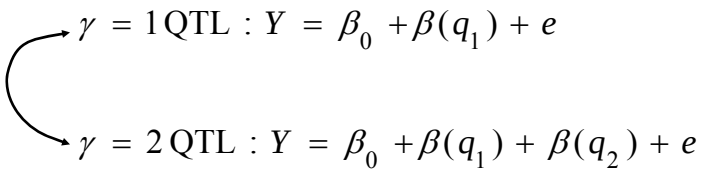


3. sampling genetic architectures

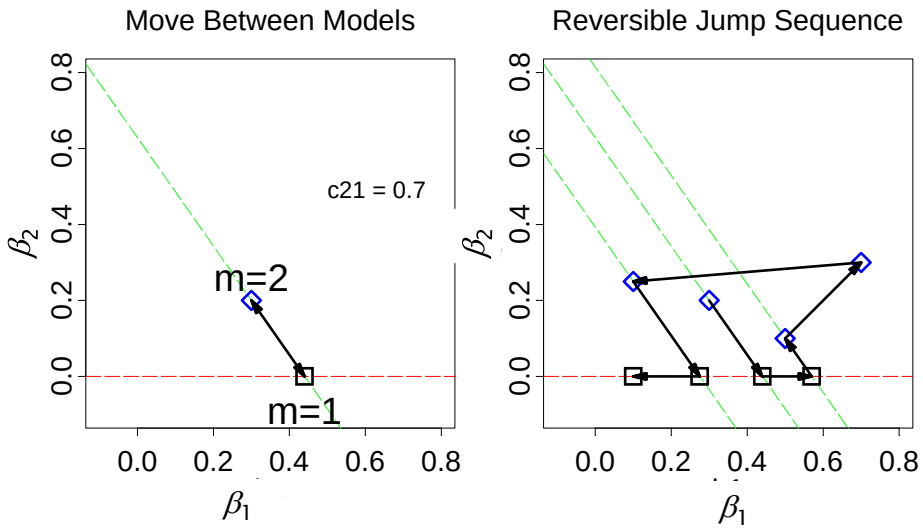
- search across genetic architectures γ of various sizes
  - allow change in number of QTL
  - allow change in types of epistatic interactions
- methods for search
  - reversible jump MCMC
  - Gibbs sampler with loci indicators
- complexity of epistasis
  - Fisher-Cockerham effects model
  - general multi-QTL interaction & limits of inference

reversible jump MCMC

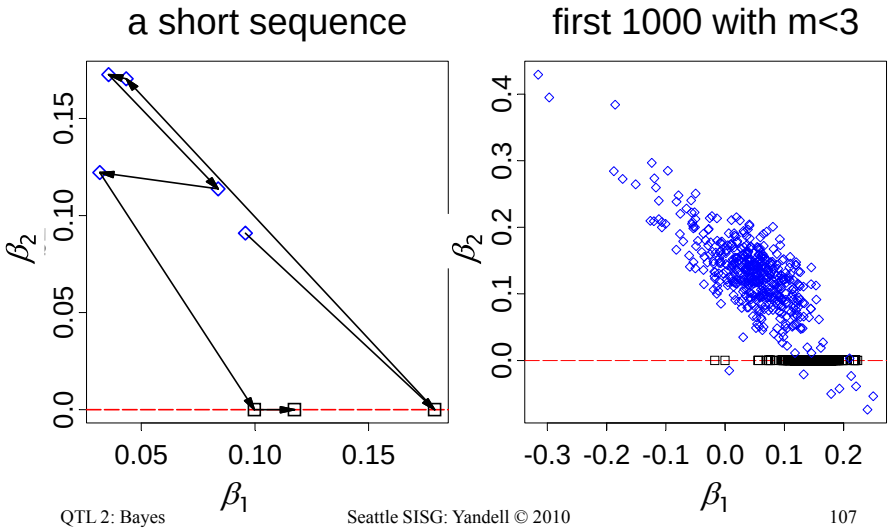
- consider known genotypes q at 2 known loci λ
  - models with 1 or 2 QTL
- M-H step between 1-QTL and 2-QTL models
  - model changes dimension (via careful bookkeeping)
  - consider mixture over QTL models H



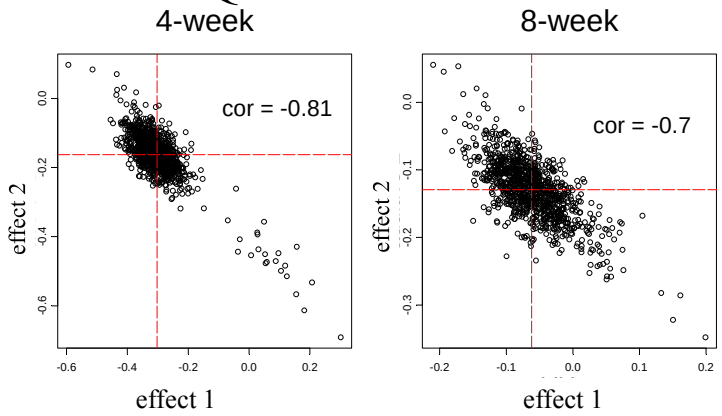
geometry of reversible jump



geometry allowing q and λ to change

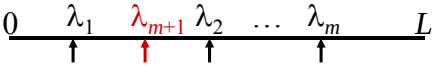


collinear QTL = correlated effects



- linked QTL = collinear genotypes
  - ➢ correlated estimates of effects (negative if in coupling phase)
  - ➢ sum of linked effects usually fairly constant

sampling across QTL models γ



action steps: draw one of three choices

- update QTL model γ with probability 1-$b(\gamma)$-$d(\gamma)$
  - update current model using full conditionals
  - sample QTL loci, effects, and genotypes
- add a locus with probability $b(\gamma)$
  - propose a new locus along genome
  - innovate new genotypes at locus and phenotype effect
  - decide whether to accept the "birth" of new locus
- drop a locus with probability $d(\gamma)$
  - propose dropping one of existing loci
  - decide whether to accept the "death" of locus

Gibbs sampler with loci indicators

- consider only QTL at pseudomarkers
  - every 1-2 cM
  - modest approximation with little bias
- use loci indicators in each pseudomarker
  - $\gamma = 1$ if QTL present
  - $\gamma = 0$ if no QTL present
- Gibbs sampler on loci indicators γ
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$\mu_q = \mu + \gamma_1 \beta(q_1) + \gamma_2 \beta(q_2), \quad \gamma_k = 0,1$$

Bayesian shrinkage estimation

- soft loci indicators
  - strength of evidence for λ_j depends on γ
  - $0 \le \gamma \le 1$ (grey scale)
  - shrink most γs to zero
- Wang et al. (2005 *Genetics*)
  - Shizhong Xu group at U CA Riverside

$$\mu_q = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_1), \quad 0 \le \gamma_k \le 1$$

other model selection approaches

- include all potential loci in model
- assume "true" model is "sparse" in some sense
- Sparse partial least squares
  - Chun, Keles (2009 *Genetics*; 2010 *JRSSB*)
- LASSO model selection
  - Foster (2006); Foster Verbyla Pitchford (2007 *JABES*)
  - Xu (2007 *Biometrics*); Yi Xu (2007 *Genetics*)
  - Shi Wahba Wright Klein Klein (2008 *Stat & Infer*)

4. criteria for model selection
balance fit against complexity

- classical information criteria
  - penalize likelihood L by model size $|\gamma|$
  - IC = $-2 \log L(\gamma \mid y)$ + penalty(γ)
  - maximize over unknowns
- Bayes factors
  - marginal posteriors pr($y \mid \gamma$)
  - average over unknowns

classical information criteria

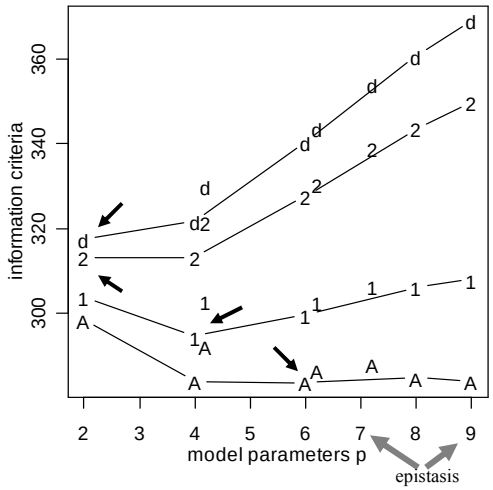
- start with likelihood $L(\gamma \mid y, m)$
  - measures fit of architecture (γ) to phenotype (y)
    - given marker data (m)
  - genetic architecture (γ) depends on parameters
    - have to estimate loci (μ) and effects (λ)
- complexity related to number of parameters
  - $|\gamma|$ = size of genetic architecture
    - BC: $|\gamma| = 1 + n.qtl + n.qtl(n.qtl - 1) = 1 + 4 + 12 = 17$
    - F2: $|\gamma| = 1 + 2n.qtl + 4n.qtl(n.qtl - 1) = 1 + 8 + 48 = 57$

classical information criteria

- construct information criteria
  - balance fit to complexity
  - Akaike $\text{AIC} = -2 \log(L) + 2 |\gamma|$
  - Bayes/Schwartz $\text{BIC} = -2 \log(L) + |\gamma| \log(n)$
  - Broman $\text{BIC}_\delta = -2 \log(L) + \delta |\gamma| \log(n)$
  - general form: $\text{IC} = -2 \log(L) + |\gamma| D(n)$
- compare models
  - hypothesis testing: designed for one comparison
    - $2 \log[LR(\gamma_1, \gamma_2)] = L(y|m, \gamma_2) - L(y|m, \gamma_1)$
  - model selection: penalize complexity
    - $\text{IC}(\gamma_1, \gamma_2) = 2 \log[LR(\gamma_1, \gamma_2)] + (|\gamma_2| - |\gamma_1|) D(n)$

information criteria vs. model size

- WinQTL 2.0
- SCD data on F2
- A=AIC
- 1=BIC(1)
- 2=BIC(2)
- d=BIC(δ)
- models
  - 1,2,3,4 QTL
    - 2+5+9+2
  - epistasis
    - 2:2 AD



Bayes factors

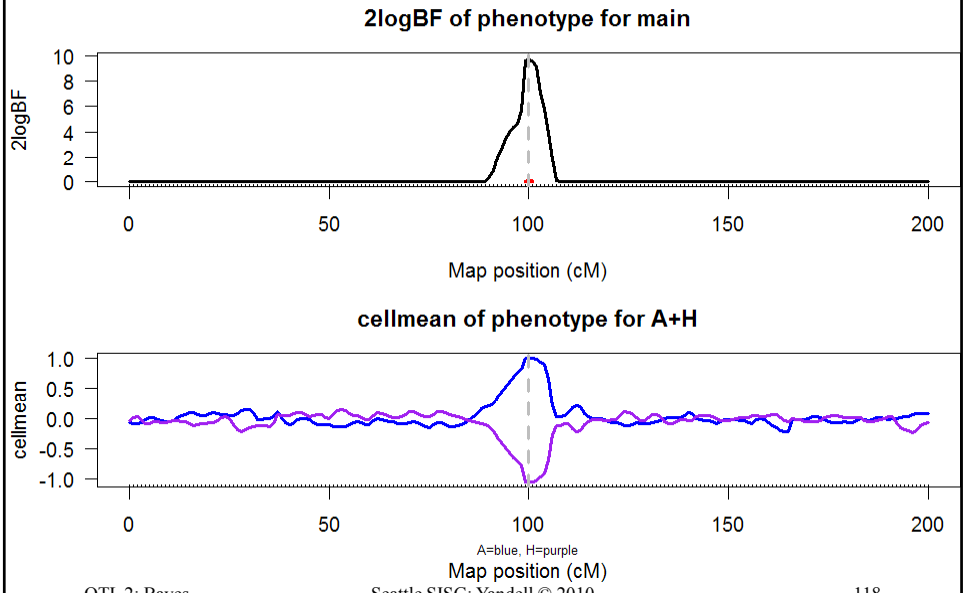
- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - averaged over unknowns

$$B_{12} = \frac{\text{pr}(\gamma_1 \mid y,m)/\text{pr}(\gamma_2 \mid y,m)}{\text{pr}(\gamma_1)/\text{pr}(\gamma_2)} = \frac{\text{pr}(y \mid m,\gamma_1)}{\text{pr}(y \mid m,\gamma_2)}$$

- roughly equivalent to BIC
  - BIC maximizes over unknowns
  - BF averages over unknowns

$$-2\log(B_{12}) = -2\log(LR) - (|\gamma_2| - |\gamma_1|)\log(n)$$

scan of marginal Bayes factor & effect

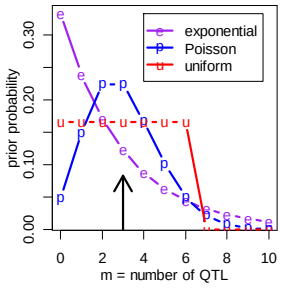


issues in computing Bayes factors

- *BF* insensitive to shape of prior on γ
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects θ
  - prior variance should reflect data variability
  - resolved by using hyper-priors
    - automatic algorithm; no need for user tuning
- easy to compute Bayes factors from samples
  - sample posterior using MCMC
  - posterior pr($\gamma \mid y, m$) is marginal histogram

Bayes factors & genetic architecture γ

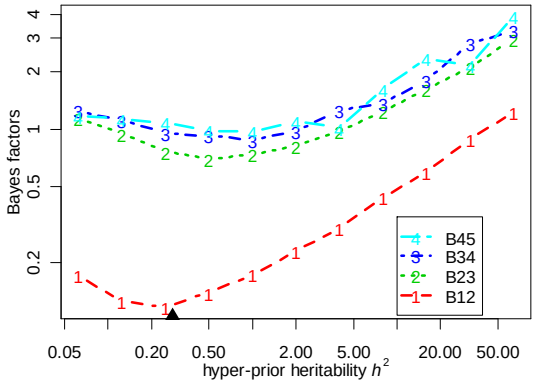
- $|\gamma|$ = number of QTL
  - prior pr(γ) chosen by user
  - posterior pr($\gamma|y,m$)
    - sampled marginal histogram
    - shape affected by prior pr(A)



$$BF_{\gamma_1,\gamma_2} = \frac{\text{pr}(\gamma_1|y,m)/\text{pr}(\gamma_1)}{\text{pr}(\gamma_2|y,m)/\text{pr}(\gamma_2)}$$

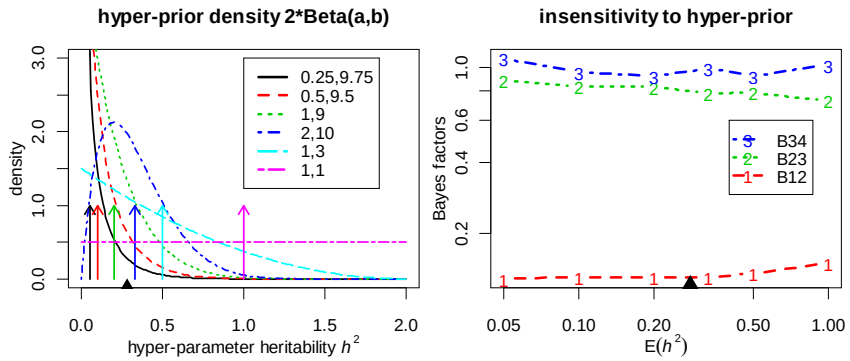
- pattern of QTL across genome
- gene action and epistasis

BF sensitivity to fixed prior for effects



$$\beta_{qj} \sim \text{N}\left(0, \sigma_G^2/m\right), \sigma_G^2 = h^2\sigma^2_{\text{total}}, h^2 \text{ fixed}$$

BF insensitivity to random effects prior



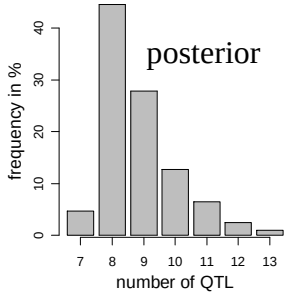
$$\beta_{qj} \sim \mathrm{N}\left(0, \sigma_G^2 / m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, \tfrac{1}{2} h^2 \sim \mathrm{Beta}(a,b)$$

examples in detail

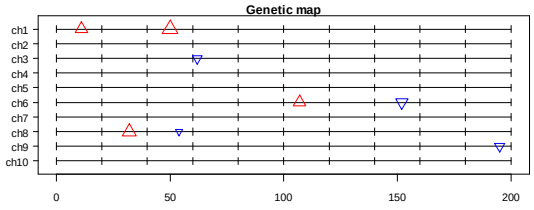
- simulation study (after Stephens & Fisch (1998)
- obesity in mice (*n* = 421)
  - epistatic QTLs with no main effects
- expression phenotype (SCD1) in mice (*n* = 108)
  - multiple QTL and epistasis
- mapping two correlated phenotypes
  - Jiang & Zeng 1995 paper
  - *Brassica napus* vernalization
- gonad shape in *Drosophila* spp. (insect) (*n* = 1000)
  - multiple traits reduced by PC
  - many QTL and epistasis

simulation with 8 QTL

- simulated F2 intercross, 8 QTL
  - (Stephens, Fisch 1998)
  - *n*=200, heritability = 50%
  - detected 3 QTL
- increase to detect all 8
  - *n*=500, heritability to 97%



QTL	chr	loci	effect
1	1	11	–3
2	1	50	–5
3	3	62	+2
4	6	107	–3
5	6	152	+3
6	8	32	–4
7	8	54	+1
8	9	195	+2



loci pattern across genome

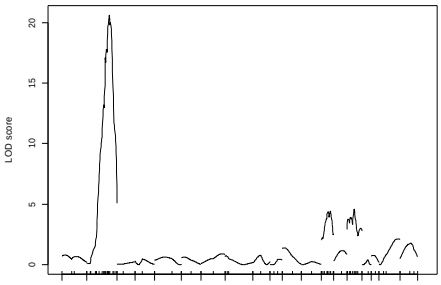
- notice which chromosomes have persistent loci
- best pattern found 42% of the time

	Chromosome										
m	**1**	2	**3**	4	5	**6**	7	**8**	**9**	10	**Count of 8000**
8	**2**	**0**	**1**	**0**	**0**	**2**	**0**	**2**	**1**	**0**	3371
9	*3*	0	1	0	0	2	0	2	1	0	751
7	2	0	1	0	0	2	0	*1*	1	0	377
9	2	0	1	0	0	2	0	2	1	0	218
9	2	0	1	0	0	*3*	0	2	1	0	218
9	2	0	1	0	0	2	0	2	*2*	0	198

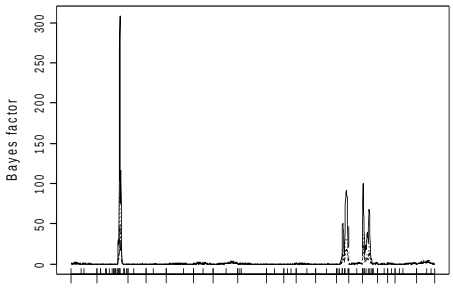
obesity in CAST/Ei BC onto M16i

- 421 mice (Daniel Pomp)
  - (213 male, 208 female)
- 92 microsatellites on 19 chromosomes
  - 1214 cM map
- subcutaneous fat pads
  - pre-adjusted for sex and dam effects
- Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005) *Genetics*

non-epistatic analysis

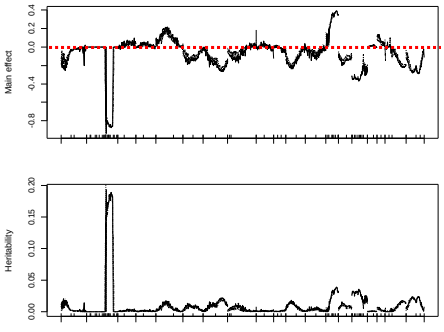


single QTL LOD profile

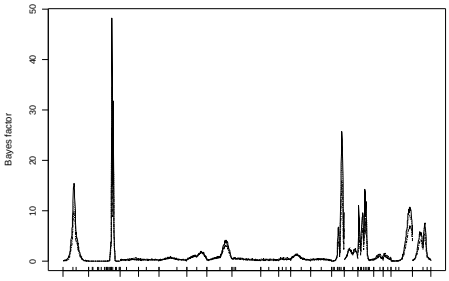


multiple QTL
Bayes factor profile

posterior profile of main effects in epistatic analysis

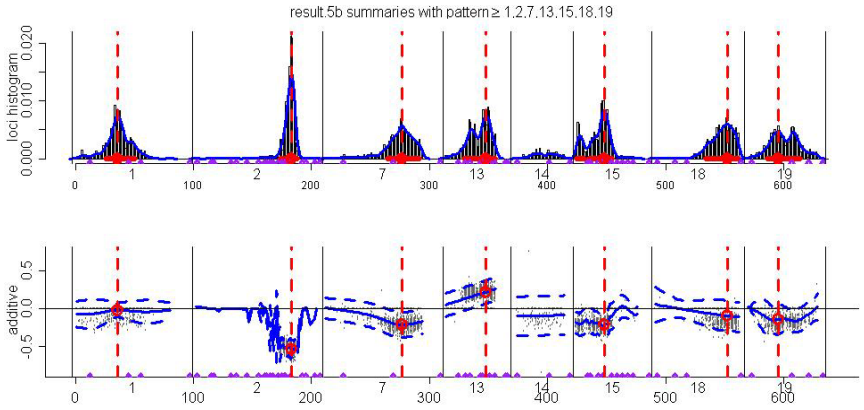


main effects & heritability profile

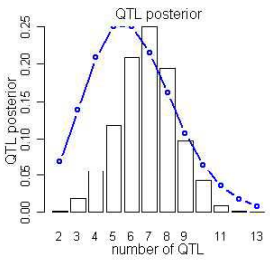


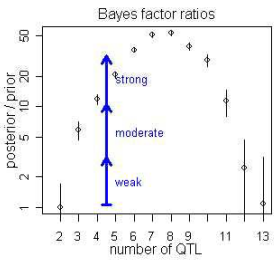
Bayes factor profile

posterior profile of main effects in epistatic analysis



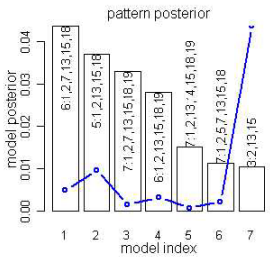
model selection via Bayes factors for epistatic model

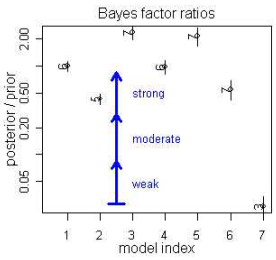




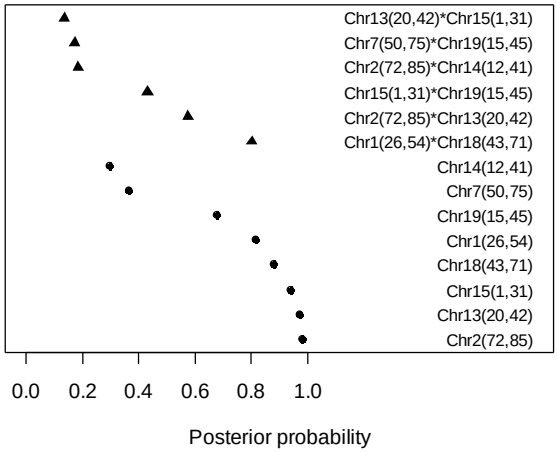
number of QTL

QTL pattern

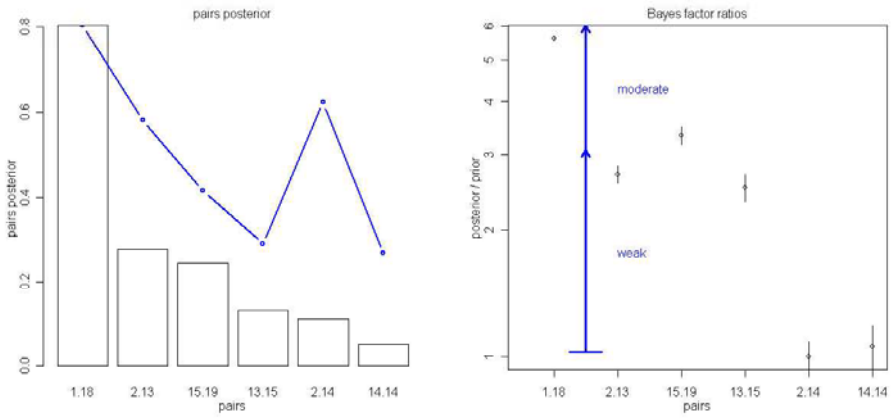




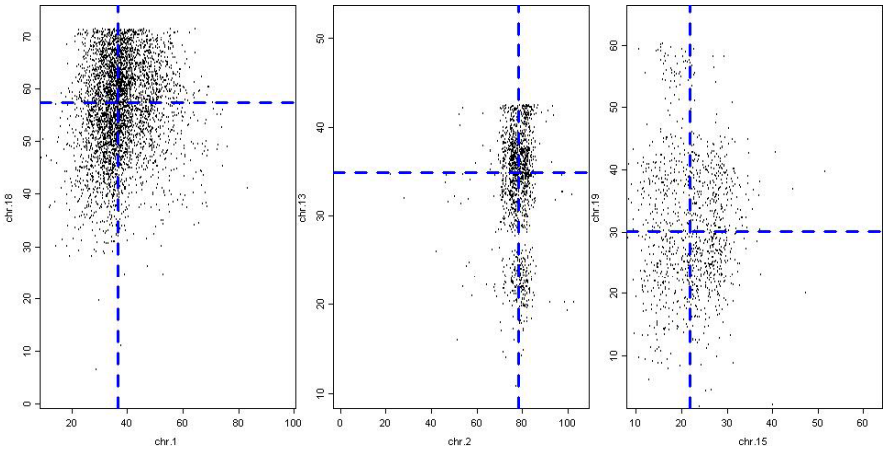
posterior probability of effects



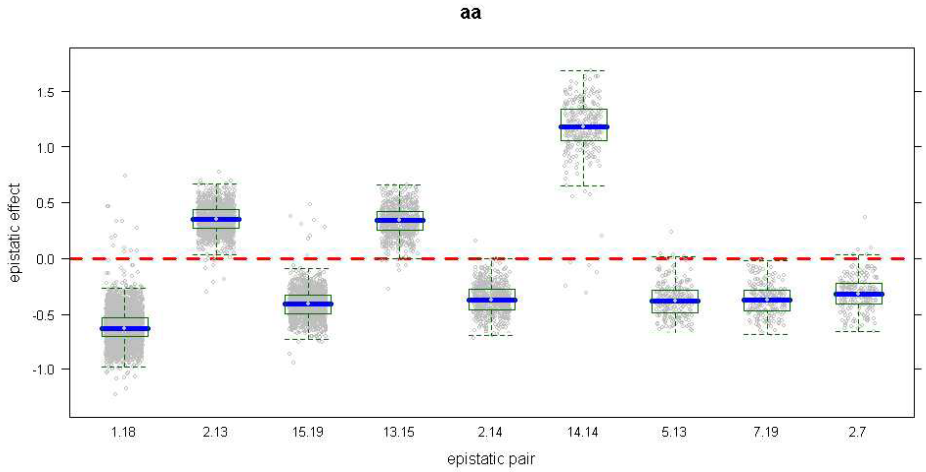
model selection for pairs



scatterplot estimates of epistatic loci



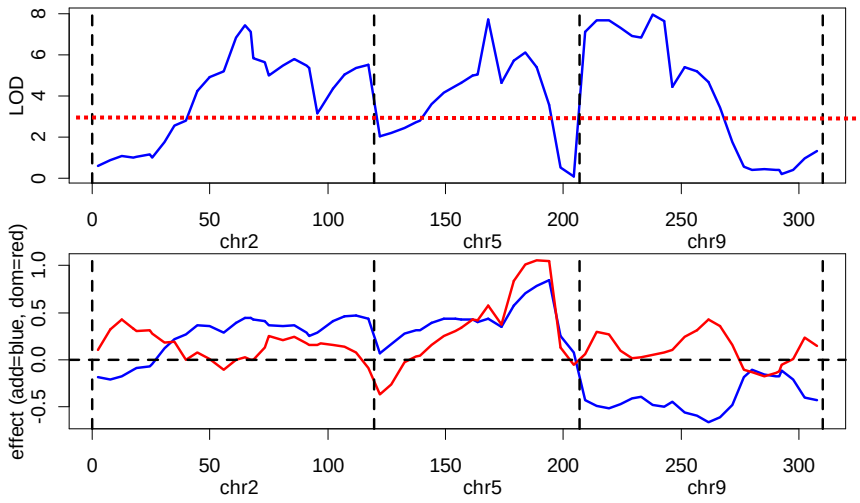
stronger epistatic effects



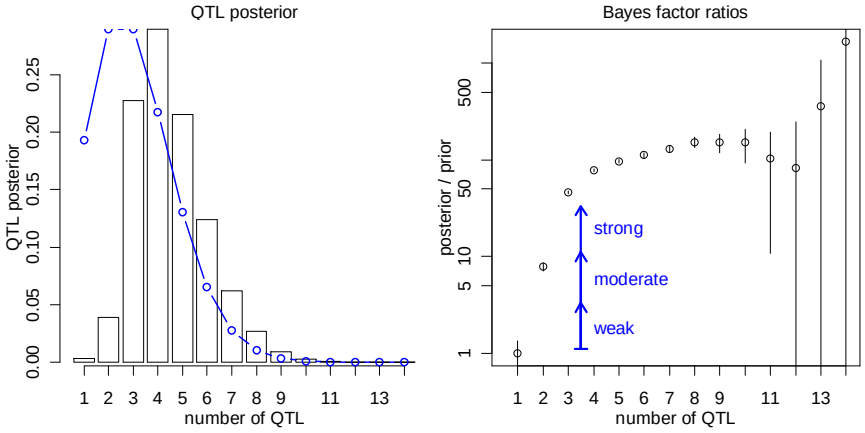
studying diabetes in an F2

- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  - Affymetrix microarrays on parental strains, F1
    - key tissues: adipose, liver, muscle, β-cells
    - novel discoveries of differential expression (Nadler et al. 2000 PNAS; Lan et al. 2002 in review; Ntambi et al. 2002 PNAS)
  - RT-PCR on 108 F2 mice liver tissues
    - 15 genes, selected as important in diabetes pathways
    - SCD1, PEPCK, ACO, FAS, GPAT, PPARgamma, PPARalpha, G6Pase, PDI,…

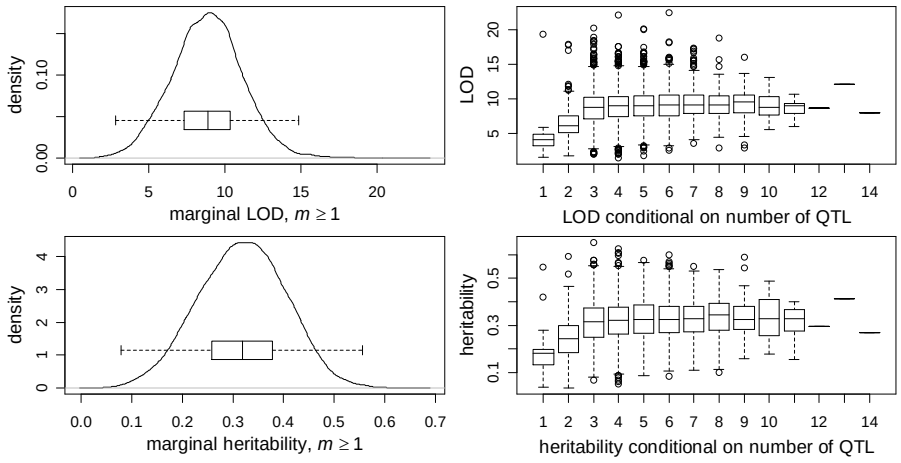
Multiple Interval Mapping (QTLCart)
SCD1: multiple QTL plus epistasis!



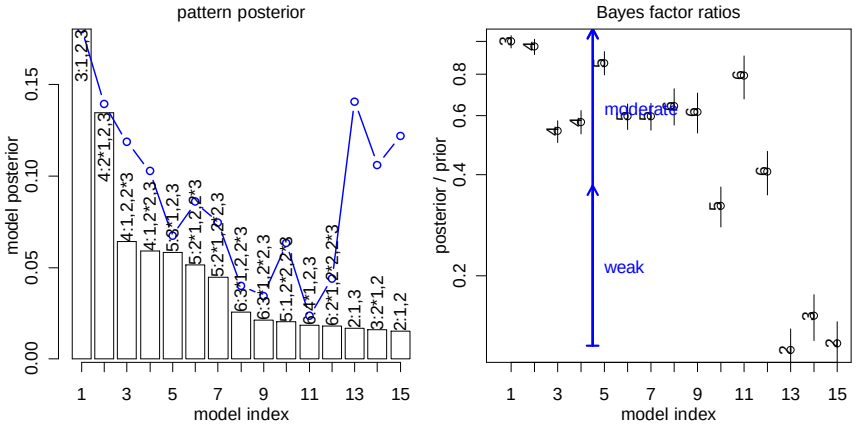
Bayesian model assessment: number of QTL for SCD1



Bayesian LOD and h^2 for SCD1

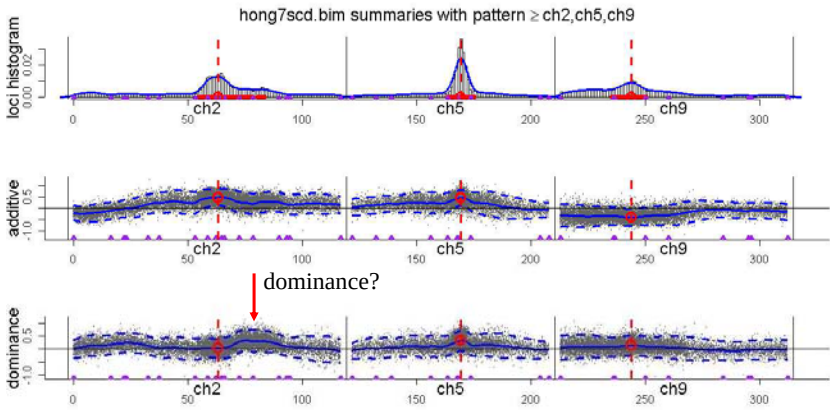


Bayesian model assessment: chromosome QTL pattern for SCD1

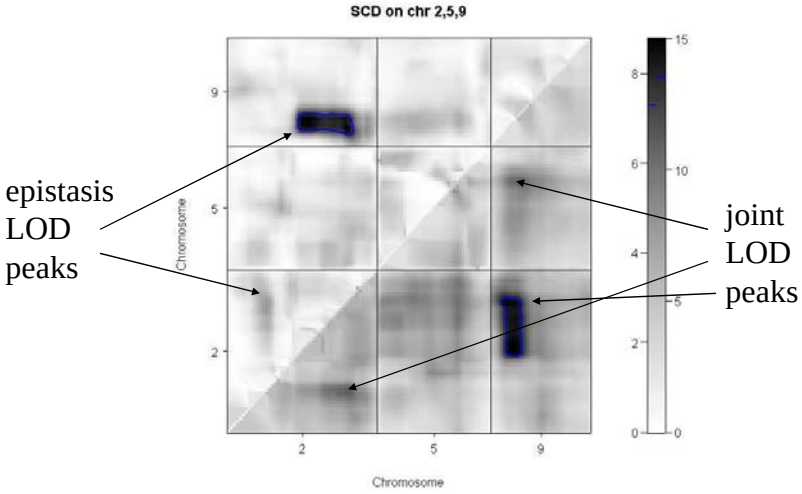


trans-acting QTL for SCD1

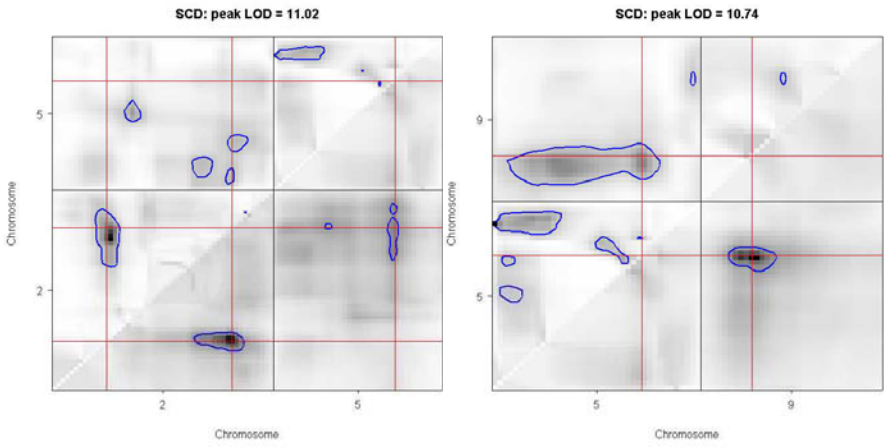
(no epistasis yet: see Yi, Xu, Allison 2003)



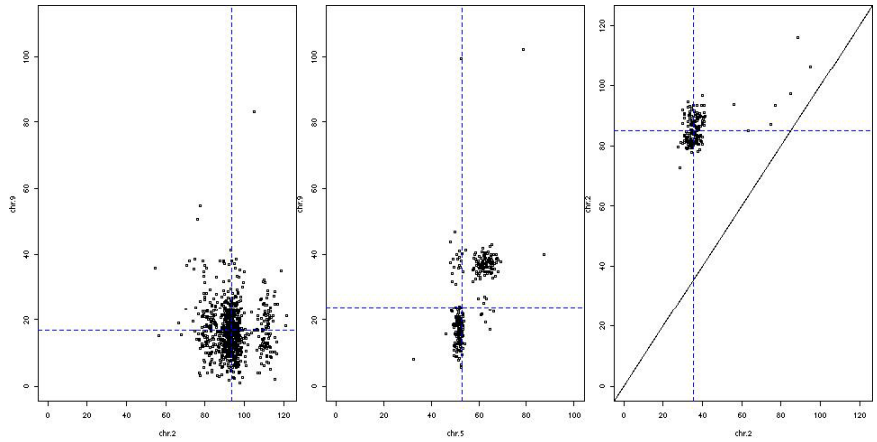
2-D scan: assumes only 2 QTL!



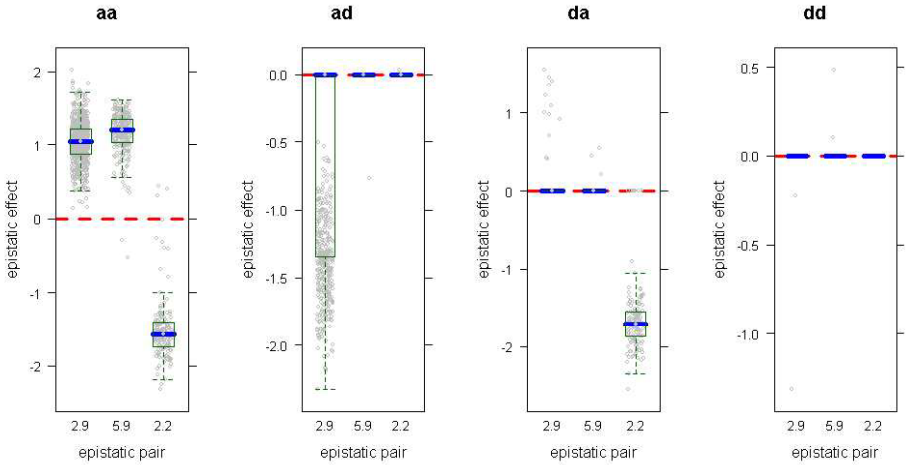
sub-peaks can be easily overlooked!



epistatic model fit



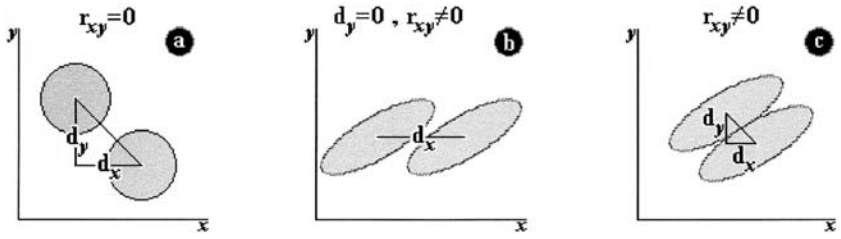
Cockerham epistatic effects



co-mapping multiple traits

- avoid reductionist approach to biology
  - address physiological/biochemical mechanisms
  - Schmalhausen (1942); Falconer (1952)
- separate close linkage from pleiotropy
  - 1 locus or 2 linked loci?
- identify epistatic interaction or canalization
  - influence of genetic background
- establish QTL x environment interactions
- decompose genetic correlation among traits
- increase power to detect QTL

interplay of pleiotropy & correlation



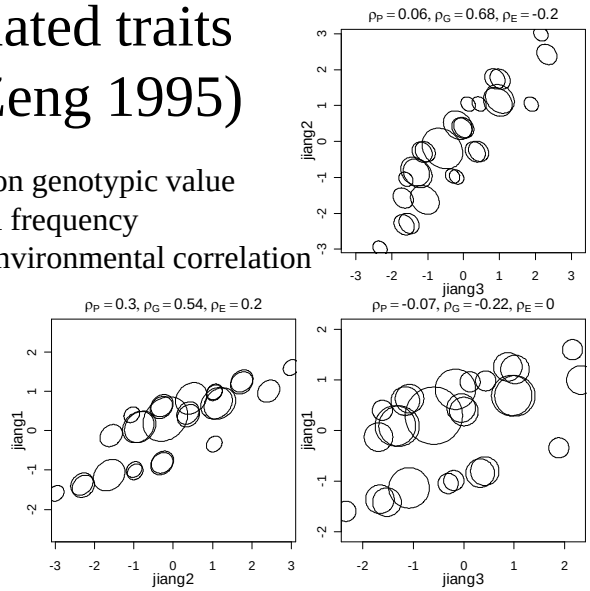
pleiotropy only correlation only both

Korol et al. (2001)

3 correlated traits (Jiang Zeng 1995)

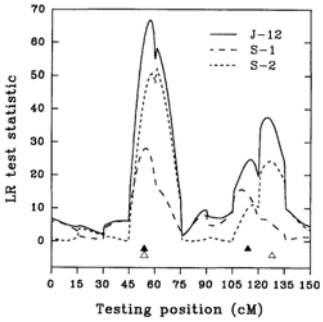
ellipses centered on genotypic value
width for nominal frequency
main axis angle environmental correlation
3 QTL, F2
27 genotypes

note signs of genetic and environmental correlation



pleiotropy or close linkage?

2 traits, 2 qtl/trait
pleiotropy @ 54cM
linkage @ 114,128cM
Jiang Zeng (1995)



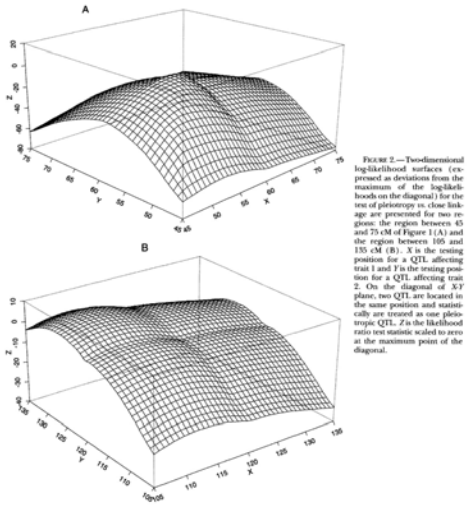


FIGURE 2.—Two-dimensional log-likelihood surfaces (expressed as deviations from the maximum of the log-likelihoods on the diagonal) for the test of pleiotropy vs. close linkage are presented for two regions: the region between 45 and 75 cM of Figure 1(A) and the region between 105 and 135 cM (B). X is the testing position for a QTL affecting trait 1 and Y is the testing position for a QTL affecting trait 2. On the diagonal of XY plane, two QTL are located in the same position and statistically are treated as one pleiotropic QTL. Z is the likelihood ratio test statistic scaled to zero at the maximum point of the diagonal.

Brassica napus: 2 correlated traits

- 4-week & 8-week vernalization effect
  - log(days to flower)
- genetic cross of
  - Stellar (annual canola)
  - Major (biennial rapeseed)
- 105 F1-derived double haploid (DH) lines
  - homozygous at every locus (*QQ* or *qq*)
- 10 molecular markers (RFLPs) on LG9
  - two QTLs inferred on LG9 (now chromosome N2)
  - corroborated by Butruille (1998)
  - exploiting synteny with *Arabidopsis thaliana*

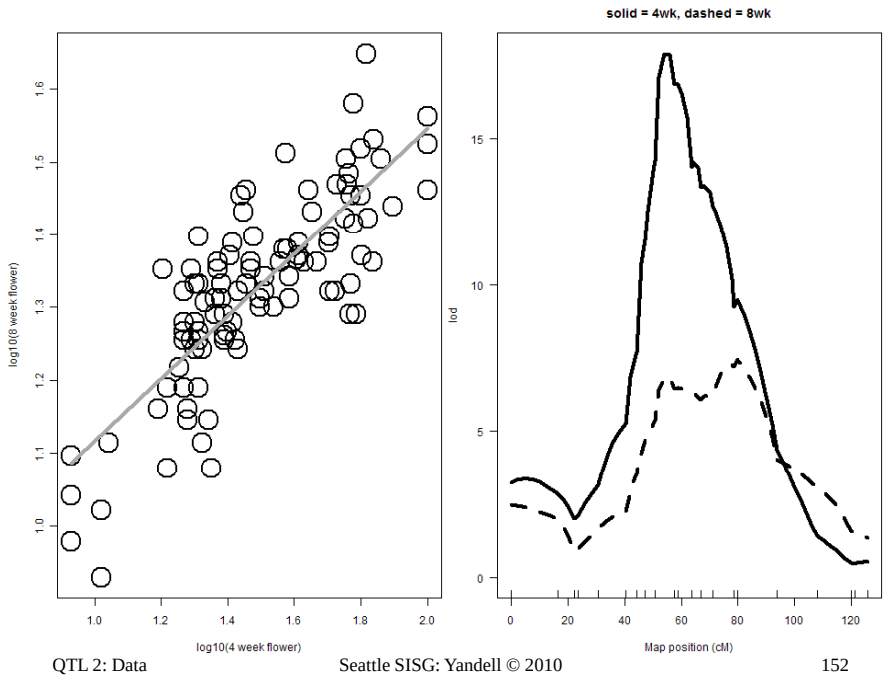
QTL with GxE or Covariates

- adjust phenotype by covariate
  - covariate(s) = environment(s) or other trait(s)
- additive covariate
  - covariate adjustment same across genotypes
  - "usual" analysis of covariance (ANCOVA)
- interacting covariate
  - address GxE
  - capture genotype-specific relationship among traits
- another way to think of multiple trait analysis
  - examine single phenotype adjusted for others

R/qtl & covariates

- additive and/or interacting covariates
- test for QTL after adjusting for covariates

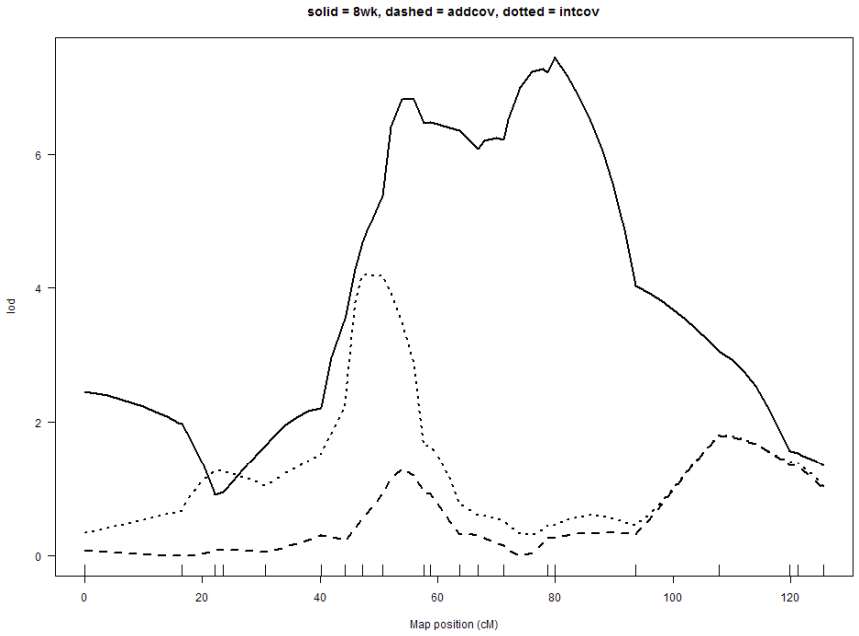
```
## Get Brassica data.
library(qtlbim)
data(Bnapus)
Bnapus <- calc.genoprob(Bnapus, step = 2, error = 0.01)

## Scatterplot of two phenotypes: 4wk & 8wk flower time.
plot(Bnapus$pheno$log10flower4,Bnapus$pheno$log10flower8)

## Unadjusted IM scans of each phenotype.
fl8 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower8"))
fl4 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower4"))
plot(fl4, fl8, chr = "N2", col = rep(1,2), lty = 1:2,
  main = "solid = 4wk, dashed = 8wk", lwd = 4)
```

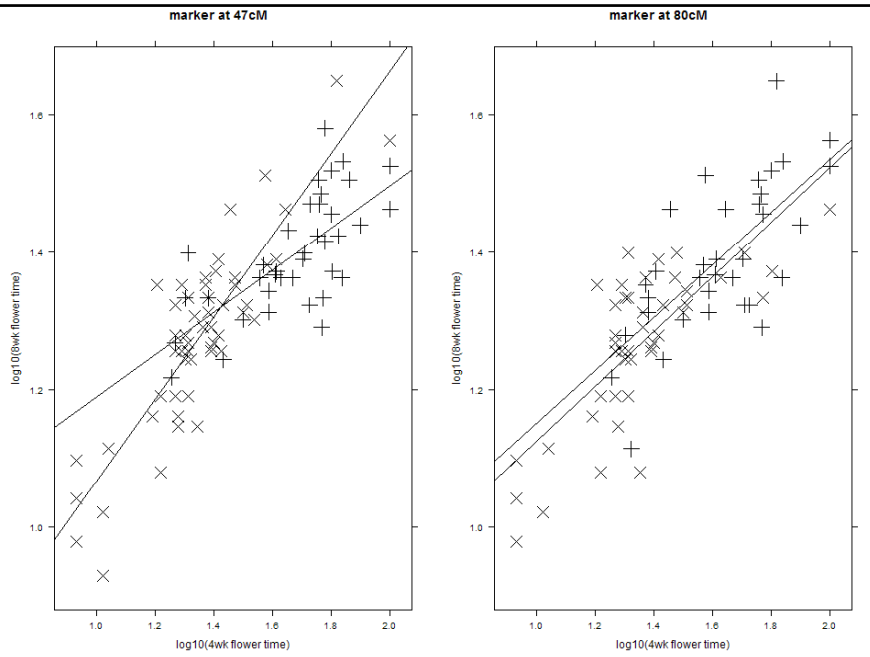



R/qtl & covariates

- additive and/or interacting covariates
- test for QTL after adjusting for covariates

```
## IM scan of 8wk adjusted for 4wk.
## Adjustment independent of genotype
fl8.4 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower8"),
  addcov = Bnapus$pheno$log10flower4)

## IM scan of 8wk adjusted for 4wk.
## Adjustment changes with genotype.
fl8.4 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower8"),
  intcov = Bnapus$pheno$log10flower4)

plot(fl8, fl8.4a, fl8.4, chr = "N2",
  main = "solid = 8wk, dashed = addcov, dotted = intcov")
```



scatterplot adjusted for covariate

```
## Set up data frame with peak markers, traits.
markers <- c("E38M50.133","ec2e5a","wg7f3a")
tmpdata <- data.frame(pull.geno(Bnapus)[,markers])
tmpdata$fl4 <- Bnapus$pheno$log10flower4
tmpdata$fl8 <- Bnapus$pheno$log10flower8

## Scatterplots grouped by marker.
library(lattice)
xyplot(fl8 ~ fl4, tmpdata, group = wg7f3a,
  col = "black", pch = 3:4, cex = 2, type = c("p","r"),
  xlab = "log10(4wk flower time)",
  ylab = "log10(8wk flower time)",
  main = "marker at 47cM")
xyplot(fl8 ~ fl4, tmpdata, group = E38M50.133,
  col = "black", pch = 3:4, cex = 2, type = c("p","r"),
  xlab = "log10(4wk flower time)",
  ylab = "log10(8wk flower time)",
  main = "marker at 80cM")
```

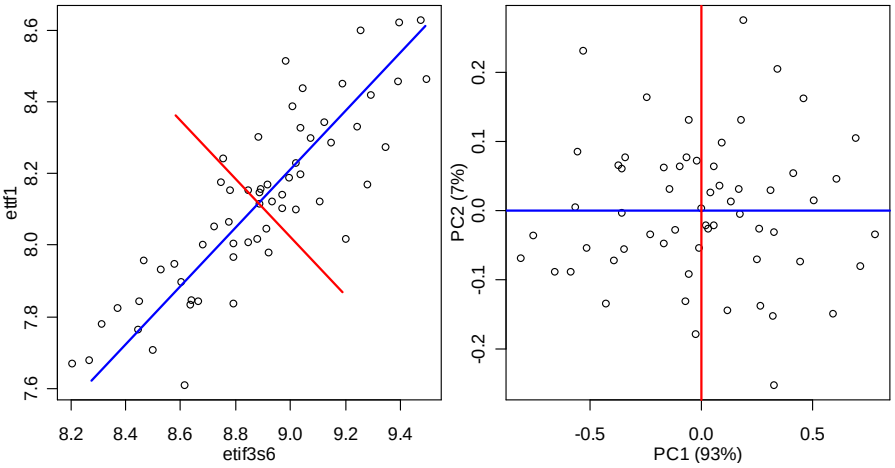
R/qtlbim and GxE

- similar idea to R/qtl
  - fixed and random additive covariates
  - GxE with fixed covariate
- multiple trait analysis tools coming soon
  - theory & code mostly in place
  - properties under study
  - expect in R/qtlbim later this year
  - Samprit Banerjee (N Yi, advisor)

reducing many phenotypes to 1

- *Drosophila mauritiana* x *D. simulans*
  - reciprocal backcrosses, ~500 per bc
- response is “shape” of reproductive piece
  - trace edge, convert to Fourier series
  - reduce dimension: first principal component
- many linked loci
  - brief comparison of CIM, MIM, BIM

PC for two correlated phenotypes



shape phenotype via PC

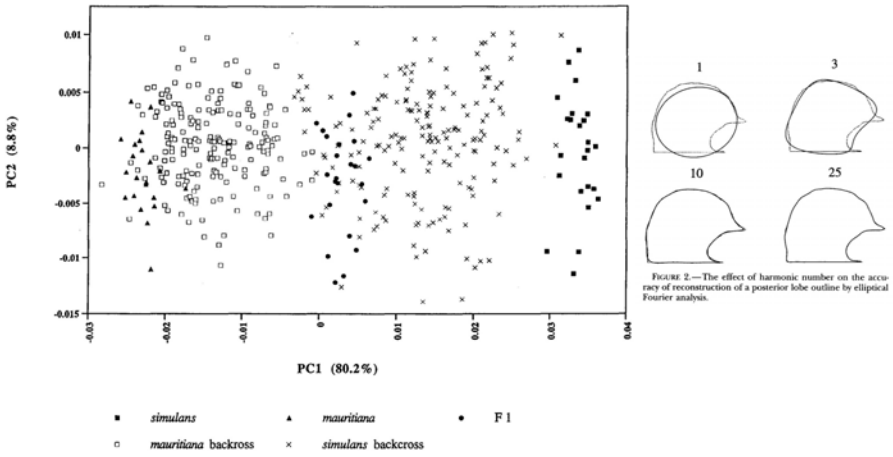


FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

FIGURE 5.—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of five genotypic classes are represented. Each point represents an average of scores from the left and right sides of an individual (with a few exceptions for which the score is from one side only). The percentage of variation in the Fourier coefficients accounted for by each principal component is given in parentheses. Liu et al. (1996) *Genetics*

shape phenotype in BC study indexed by PC1

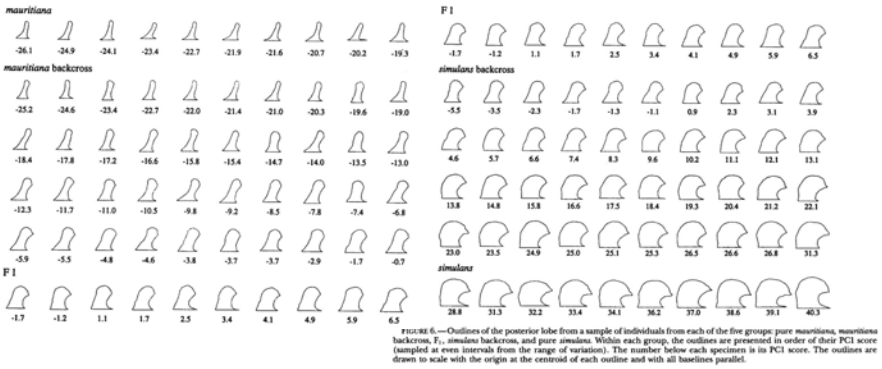


FIGURE 6.—Outlines of the posterior lobe from a sample of individuals from each of the five groups: pure *mauritiana*, *mauritiana* backcross, F_1, *simulans* backcross, and pure *simulans*. Within each group, the outlines are presented in order of their PC1 score (sampled at even intervals from the range of variation). The number below each specimen is its PC1 score. The outlines are drawn to scale with the origin at the centroid of each outline and with all baselines parallel.

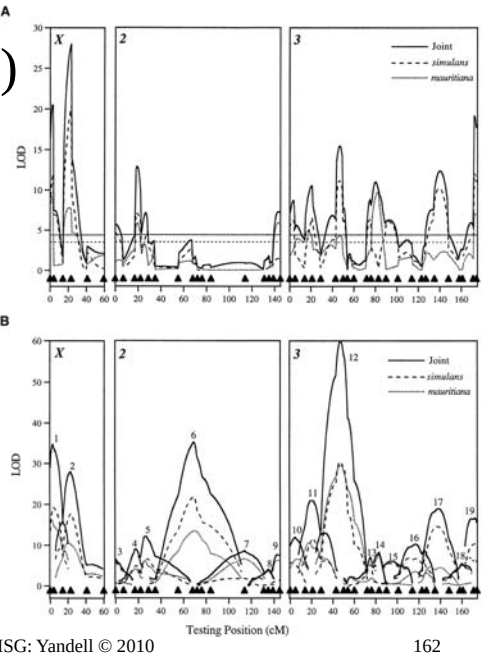
Liu et al. (1996) *Genetics*

Zeng et al. (2000) CIM vs. MIM

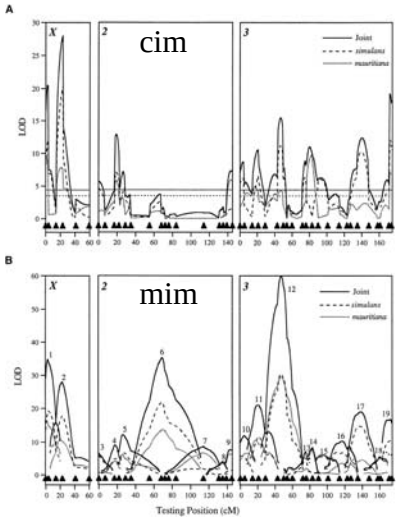
composite interval mapping
- (Liu et al. 1996)
- narrow peaks
- miss some QTL

multiple interval mapping
- (Zeng et al. 2000)
- triangular peaks

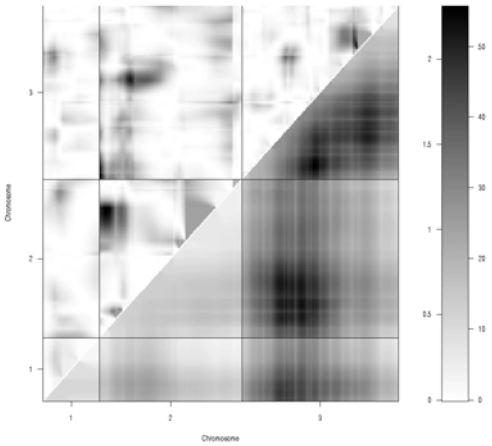
both conditional 1-D scans
- fixing all other "QTL"



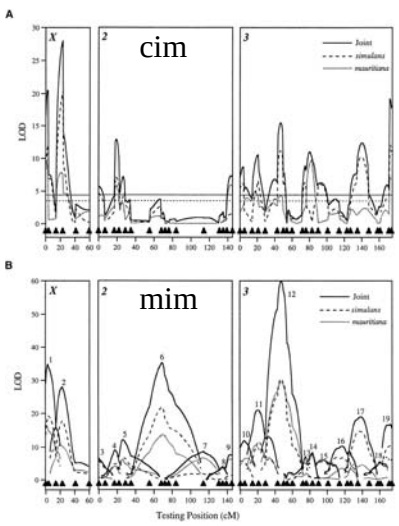
CIM, MIM and IM pairscan

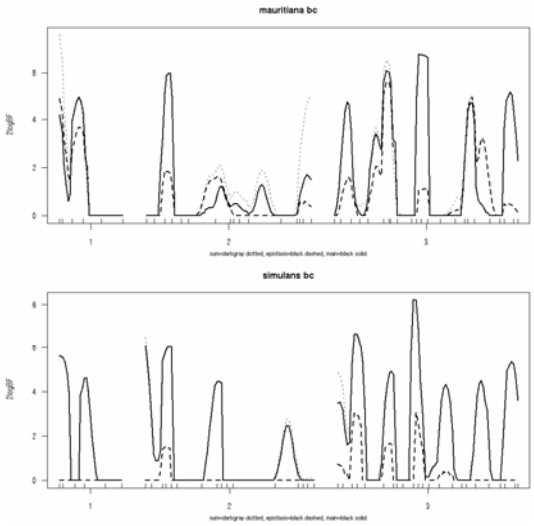


2-D im



multiple QTL: CIM, MIM and BIM





Computational Infrastructure for Systems Genetics Analysis

Brian Yandell, UW-Madison

high-throughput analysis of systems data
enable biologists & analysts to share tools

UW-Madison: Yandell,Attie,Broman,Kendziorski
Jackson Labs: Churchill
U Groningen: Jansen,Swertz
UC-Denver: Tabakoff
LabKey: Igra

www.stat.wisc.edu/~yandell/statgen
byandell@wisc.edu

- UW-Madison
  - Alan Attie
  - Christina Kendziorski
  - Karl Broman
  - Mark Keller
  - Andrew Broman
  - Aimee Broman
  - YounJeong Choi
  - Elias Chaibub Neto
  - Jee Young Moon
  - John Dawson
  - Ping Wang
  - NIH Grants DK58037, DK66369, GM74244, GM69430 , EY18869
- Jackson Labs (HTDAS)
  - Gary Churchill
  - Ricardo Verdugo
  - Keith Sheppard
- UC-Denver (PhenoGen)
  - Boris Tabakoff
  - Cheryl Hornbaker
  - Laura Saba
  - Paula Hoffman
- Labkey Software
  - Mark Igra
- U Groningen (XGA)
  - Ritsert Jansen
  - Morris Swertz
  - Pjotr Pins
  - Danny Arends
- Broad Institute
  - Jill Mesirov
  - Michael Reich

experimental context

- B6 x BTBR obese mouse cross
  - model for diabetes and obesity
  - 500+ mice from intercross (F2)
  - collaboration with Rosetta/Merck
- genotypes
  - 5K SNP Affymetrix mouse chip
  - care in curating genotypes! (map version, errors, …)
- phenotypes
  - clinical phenotypes (>100 / mouse)
  - gene expression traits (>40,000 / mouse / tissue)
  - other molecular phenotypes

how does one filter traits?

- want to reduce to “manageable” set
  - 10/100/1000: depends on needs/tools
  - How many can the biologist handle?
- how can we create such sets?
  - data-driven procedures
    - correlation-based modules
      - Zhang & Horvath 2005 *SAGMB*, Keller et al. 2008 *Genome Res*
      - Li et al. 2006 *Hum Mol Gen*
    - mapping-based focus on genome region
  - function-driven selection with database tools
    - GO, KEGG, etc
    - Incomplete knowledge leads to bias
  - random sample

why build Web eQTL tools?

- common storage/maintainence of data
  - one well-curated copy
  - central repository
  - reduce errors, ensure analysis on same data
- automate commonly used methods
  - biologist gets immediate feedback
  - statistician can focus on new methods
  - codify standard choices

how does one build tools?

- no one solution for all situations
- use existing tools wherever possible
  - new tools take time and care to build!
  - downloaded databases must be updated regularly
- human component is key
  - need informatics expertise
  - need continual dialog with biologists
- build bridges (interfaces) between tools
  - Web interface uses PHP
  - commands are created dynamically for R
- continually rethink & redesign organization

perspectives for building a community where disease data and models are shared

Benefits of wider access to datasets and models:

1- catalyze new insights on disease & methods

2- enable deeper comparison of methods & results

Lessons Learned:

1- need quick feedback between biologists & analysts

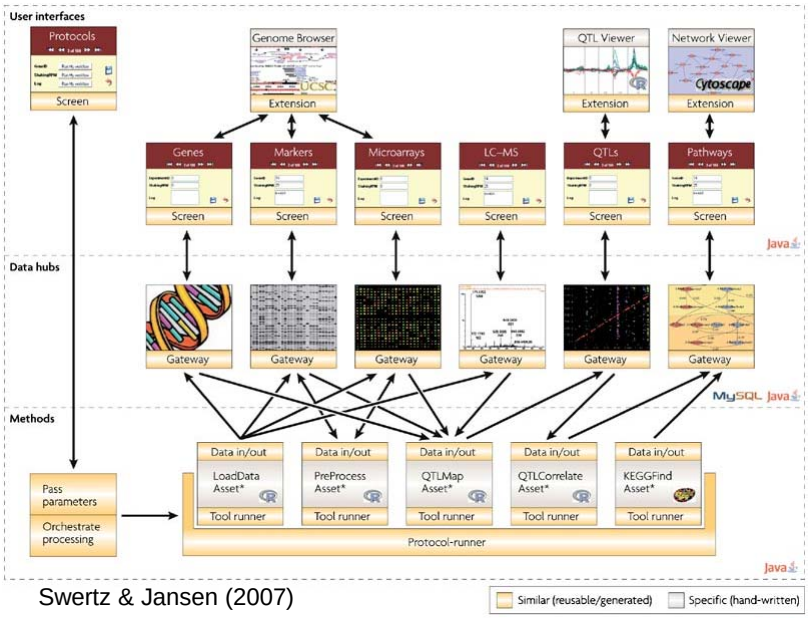
2- involve biologists early in development

3- repeated use of pipelines leads to
documented learning from experience
increased rigor in methods

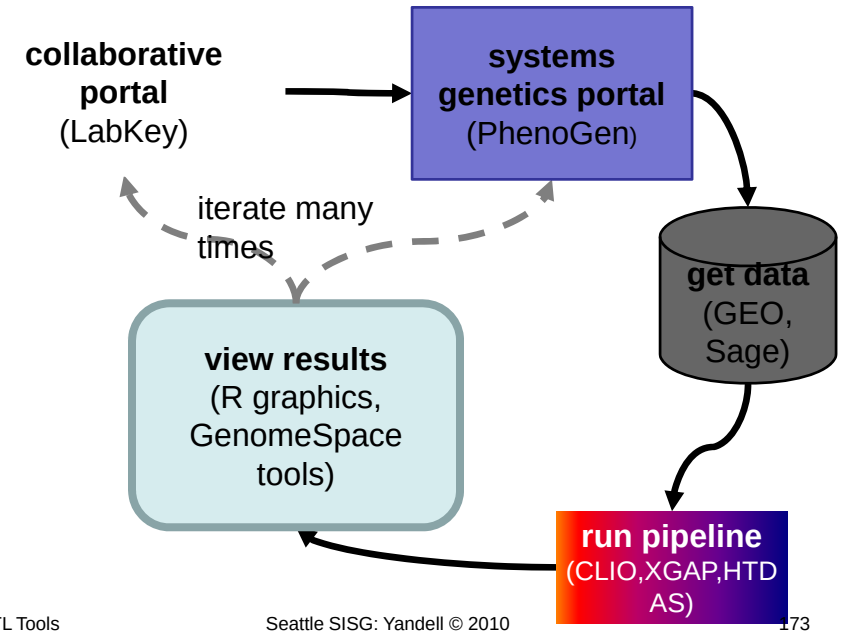
Challenges Ahead:

1- stitching together components as coherent system

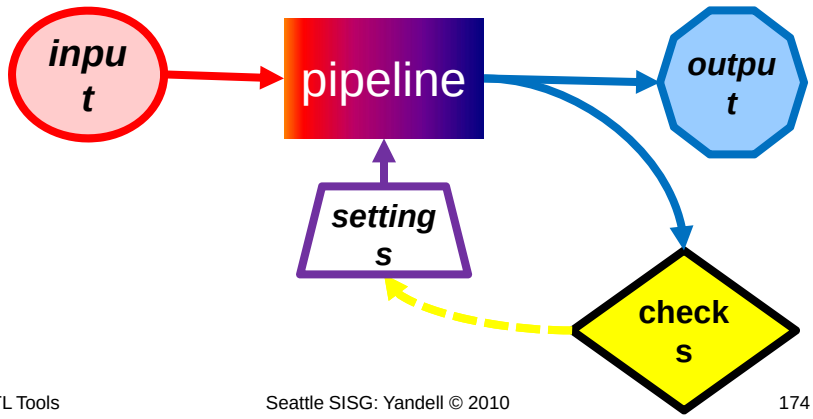
2- ramping up to ever larger molecular datasets



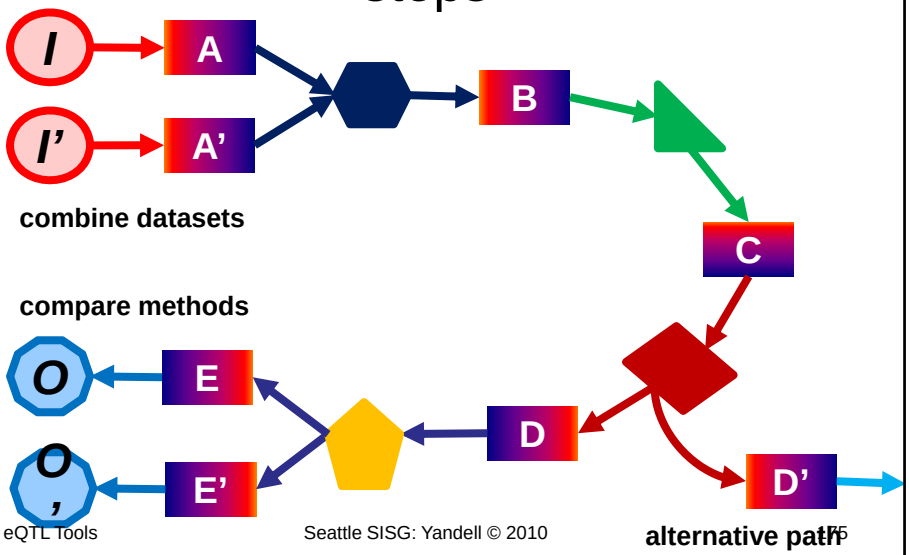
Swertz & Jansen (2007)



analysis pipeline acts on objects (extends concept of GenePattern)

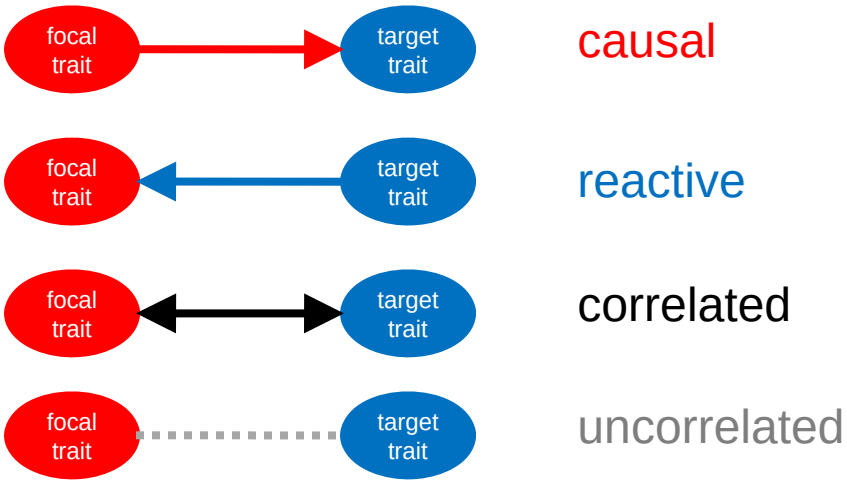


pipeline is composed of many steps

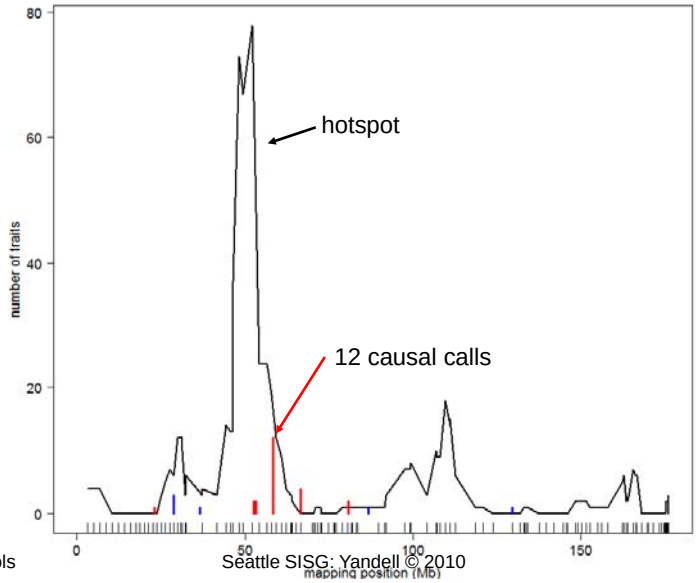


causal model selection choices

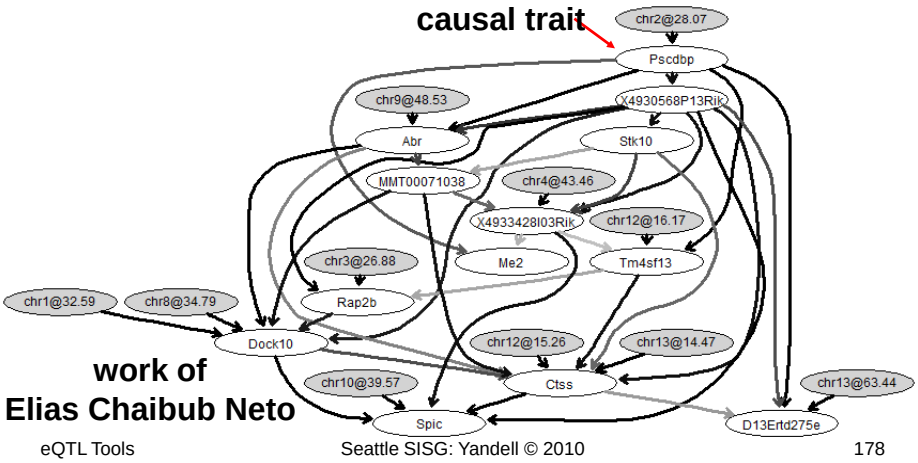
in context of larger, unknown network



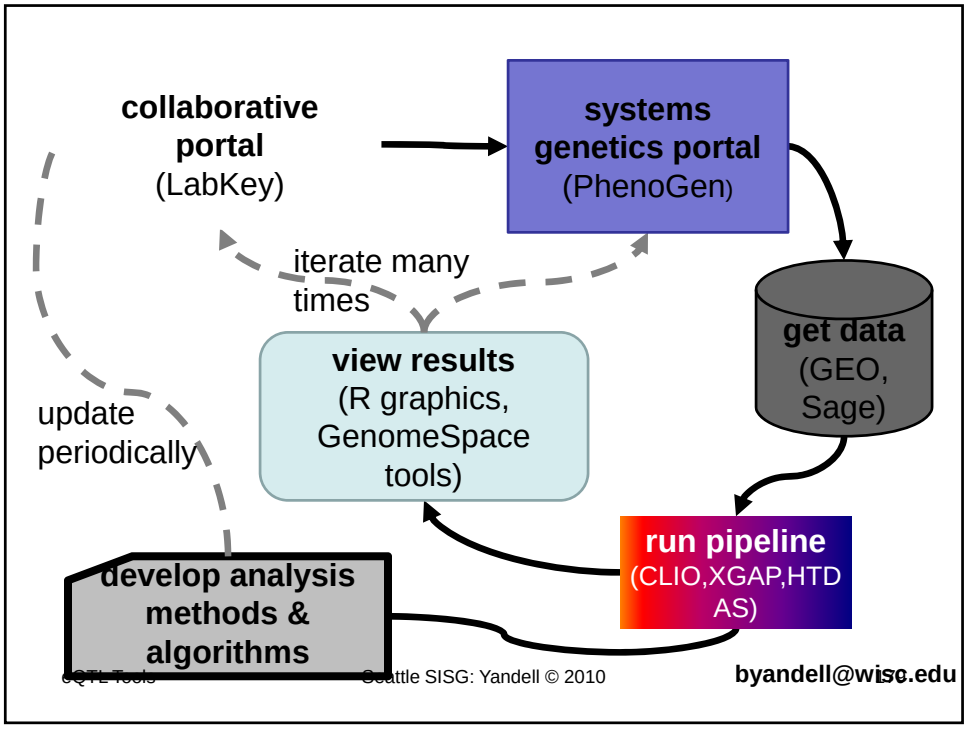
BxH ApoE-/- chr 2: causal architecture

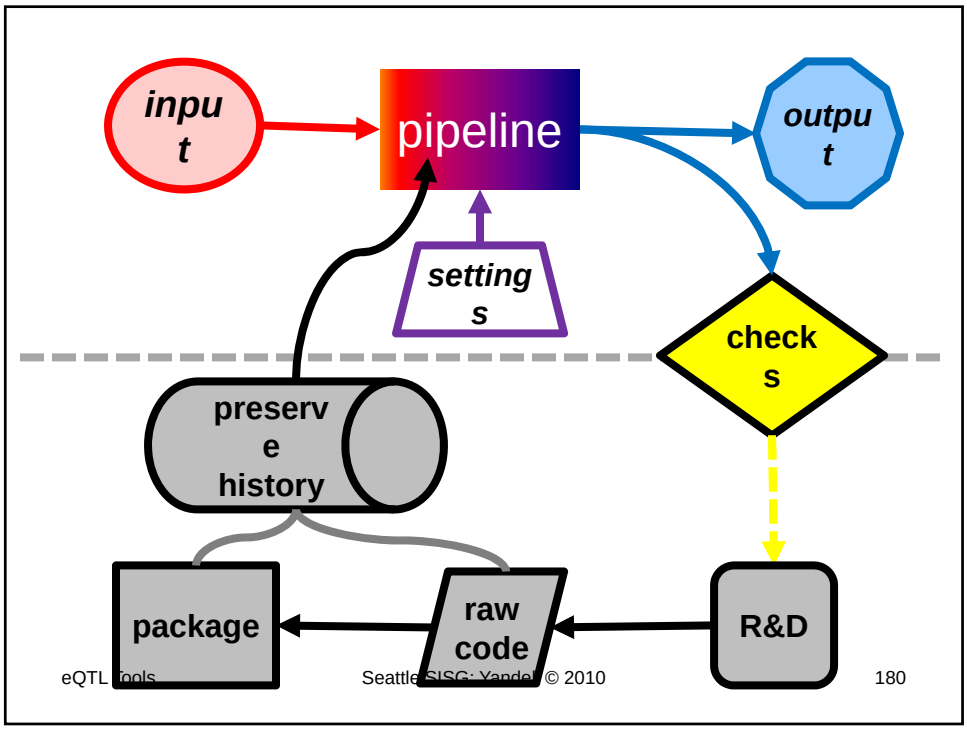


BxH ApoE-/- causal network for transcription factor Pscdbp



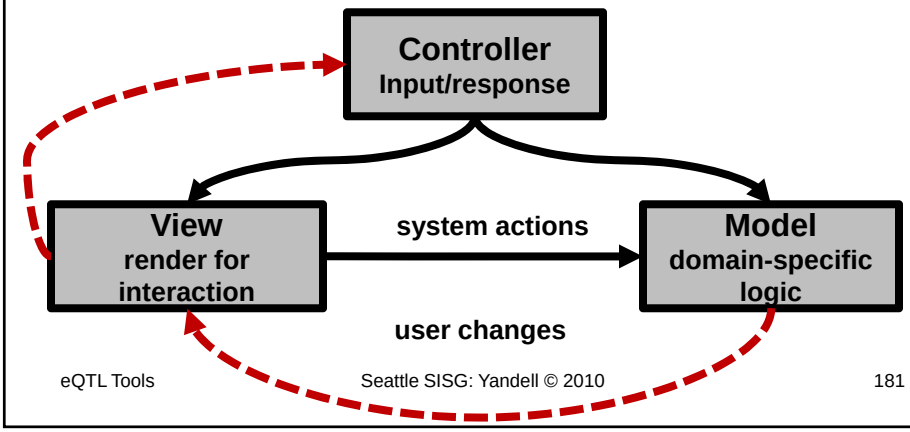
work of Elias Chaibub Neto

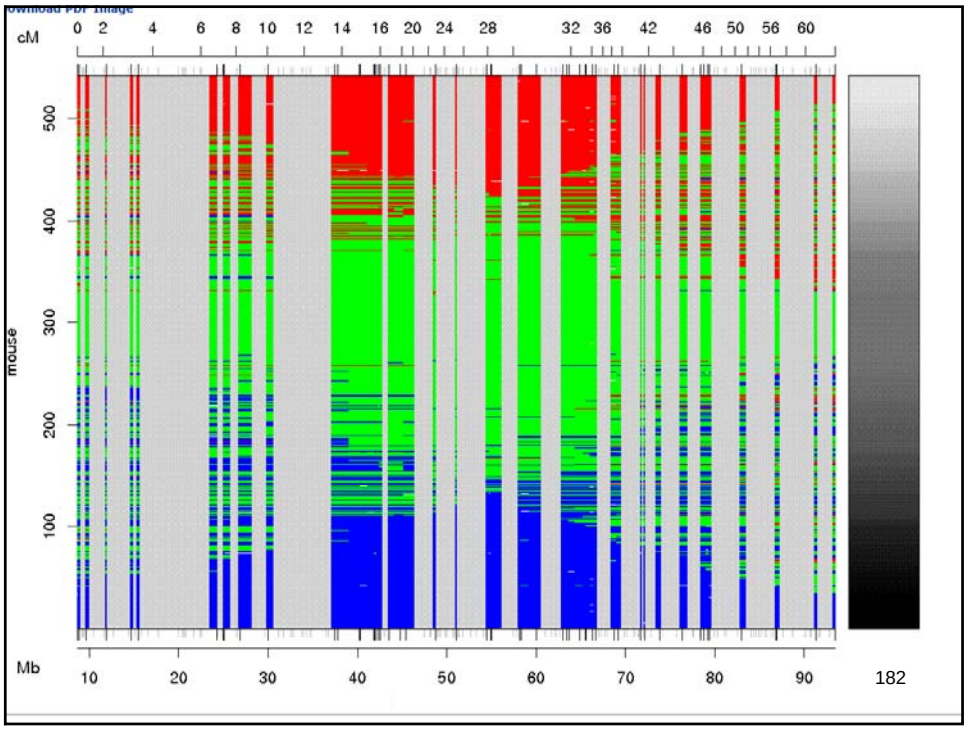

collaborative portal (LabKey)
systems genetics portal (PhenoGen)
iterate many times
get data (GEO, Sage)
view results (R graphics, GenomeSpace tools)
update periodically
run pipeline (CLIO,XGAP,HTDAS)
develop analysis methods & algorithms
eQTL Tools
Seattle SISG: Yandell © 2010
byandell@wisc.edu

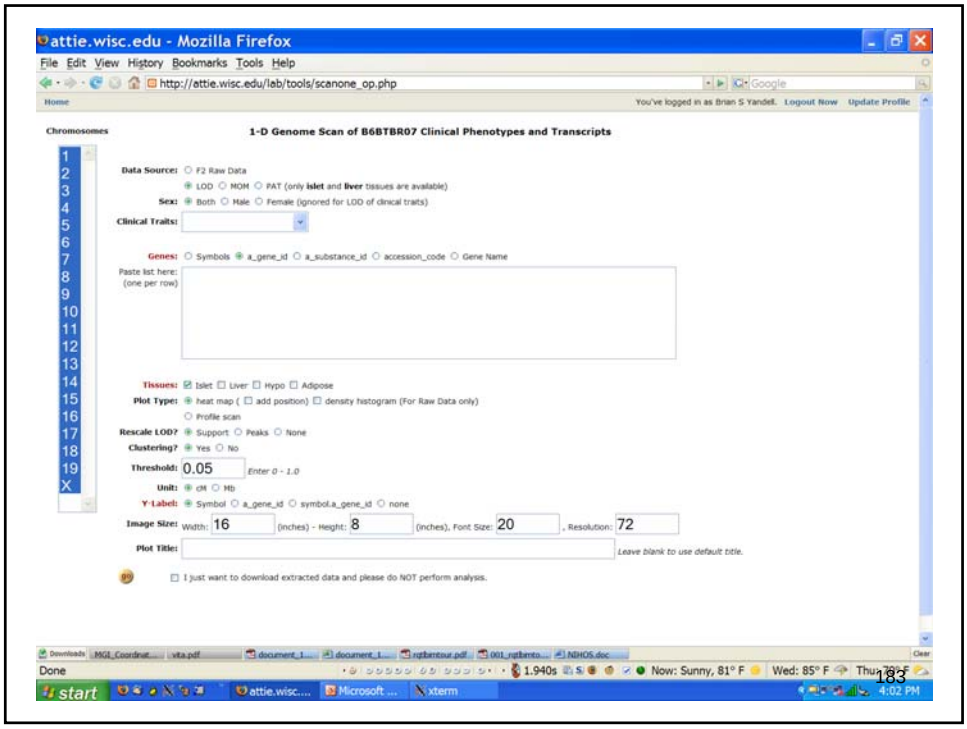

input
pipeline
output
setting s
checks
preserve history
package
raw code
R&D
eQTL Tools
Seattle SISG: Yandell © 2010
180

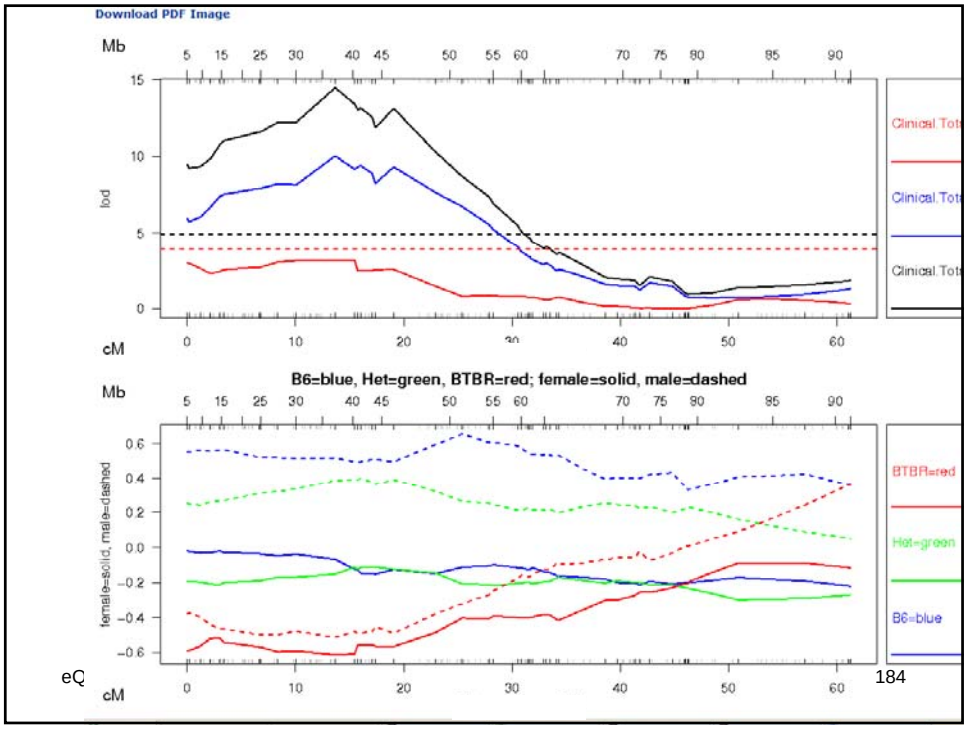
Model/View/Controller (MVC)
software architecture

- isolate domain logic from input and presentation
- permit independent development, testing, maintenance



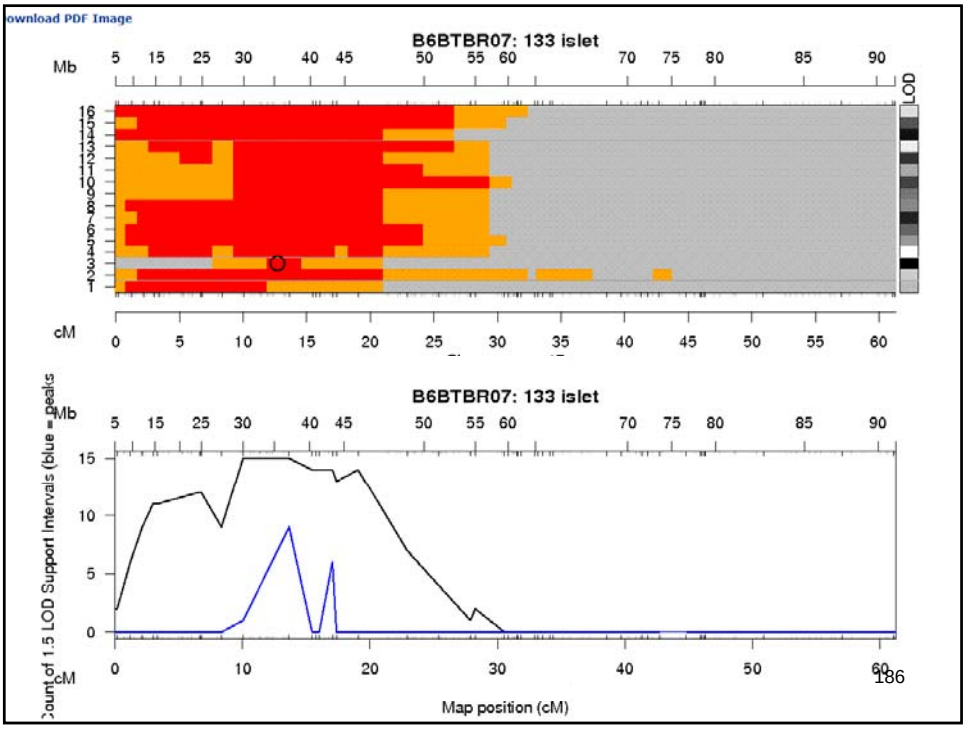


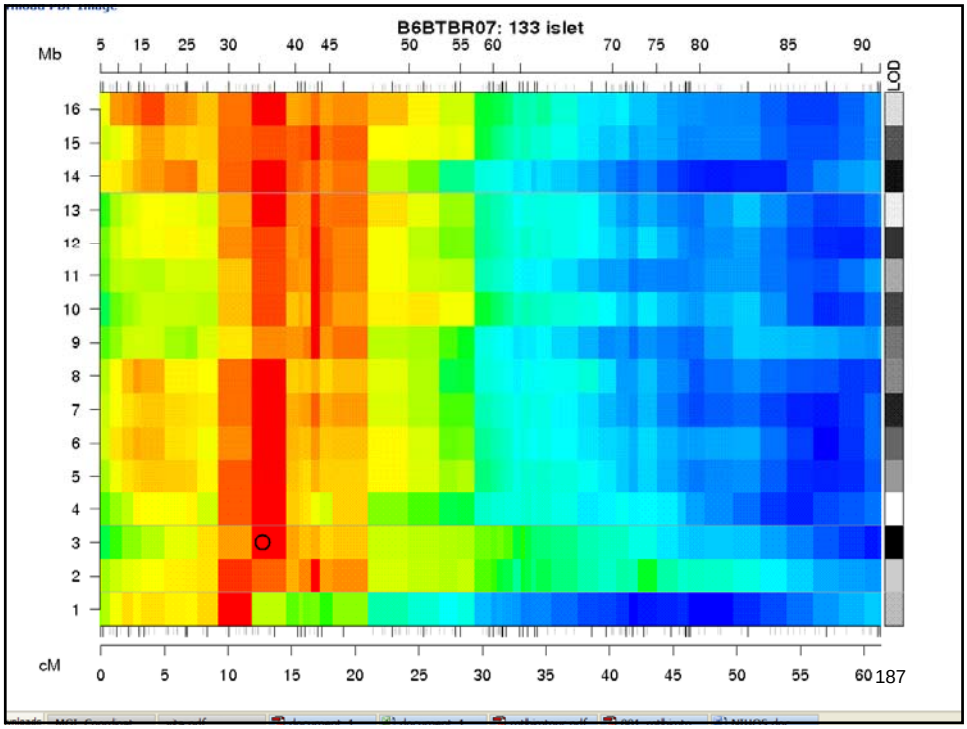

attie.wisc.edu - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://attie.wisc.edu/lab/tools/scanone_op.php
Home
You've logged in as Brian S Yandell. Logout Now Update Profile
Chromosomes
1-D Genome Scan of B6BTBR07 Clinical Phenotypes and Transcripts
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Data Source: F2 Raw Data
LOD MOM PAT (only islet and liver tissues are available)
Sex: Both Male Female (ignored for LOD of clinical traits)
Clinical Traits:
Genes: Symbols a_gene_id a_substance_id accession_code Gene Name
Paste list here: (one per row)
Tissues: Islet Liver Hypo Adipose
Plot Type: heat map (add position) density histogram (For Raw Data only)
Profile scan
Rescale LOD? Support Peaks None
Clustering? Yes No
Threshold: 0.05 Enter 0 - 1.0
Unit: cM Mb
Y Label: Symbol a_gene_id symbol.a_gene_id none
Image Size: Width: 16 (inches) - Height: 8 (inches), Font Size: 20 , Resolution: 72
Plot Title: Leave blank to use default title.
I just want to download extracted data and please do NOT perform analysis.
Done
1.940s
Now: Sunny, 81° F
Wed: 85° F
start
attie.wisc....
Microsoft ...
xterm
4:02 PM

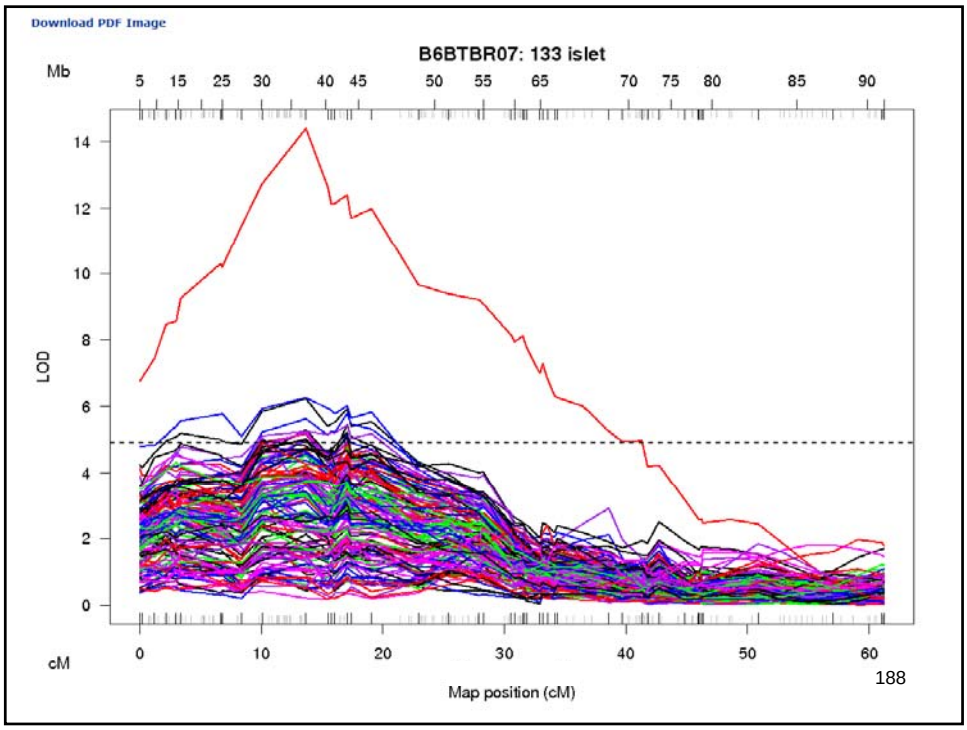

Download PDF Image
Mb
lod
cM
Clinical.Tot
Clinical.Tot
Clinical.Tot
B6=blue, Het=green, BTBR=red; female=solid, male=dashed
Mb
female=solid, male=dashed
cM
BTBR=red
Het=green
B6=blue

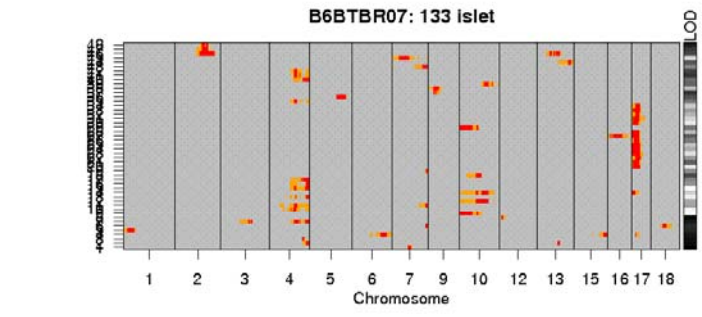
automated R script

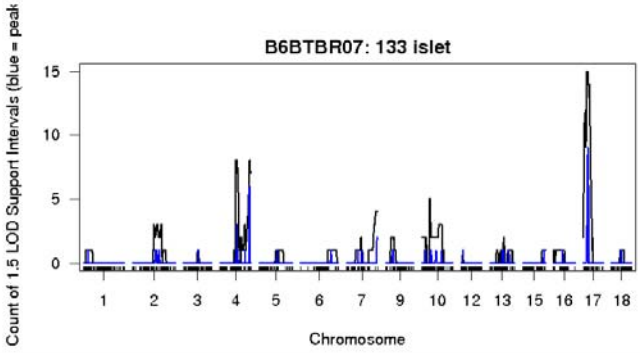
```
library('B6BTBR07')

out <- multtrait(cross.name='B6BTBR07',
  filename = 'scanone_1214952578.csv',
  category = 'islet', chr = c(17),
  threshold.level = 0.05, sex = 'both',)

sink('scanone_1214952578.txt')
print(summary(out))
sink()

bitmap('scanone_1214952578%03d.bmp',
  height = 12, width = 16, res = 72, pointsize = 20)
plot(out, use.cM = TRUE)
dev.off()
```

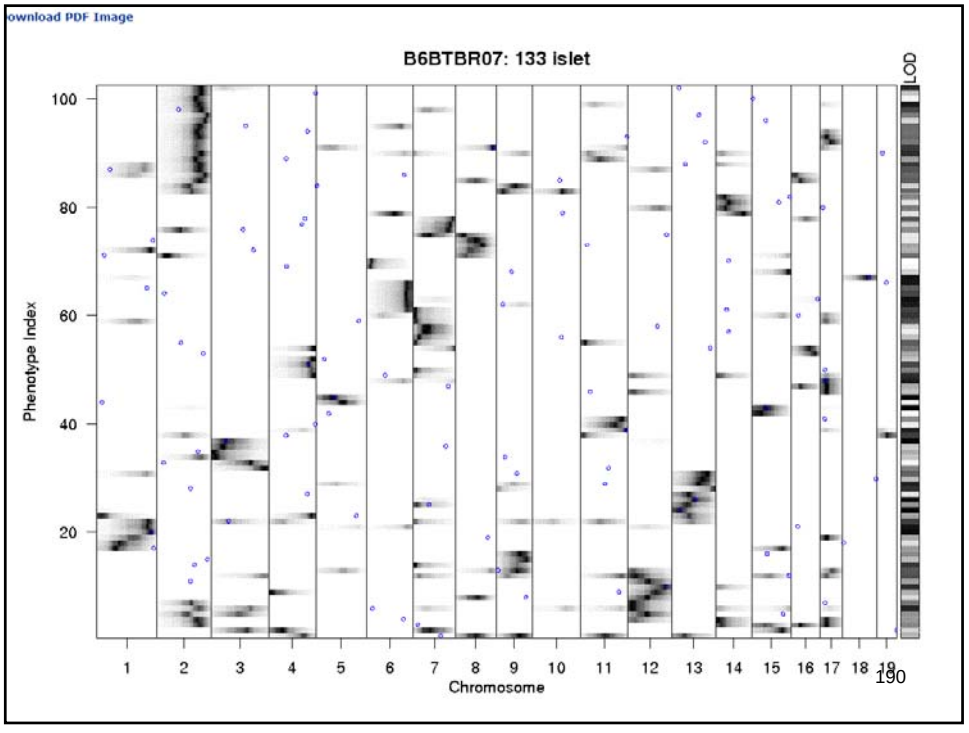

B6BTBR07: 133 islet
Mb
5 15 25 30 40 45 50 55 60 70 75 80 85 90
LOD
16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
cM
0 5 10 15 20 25 30 35 40 45 50 55 60
187


Download PDF Image
B6BTBR07: 133 islet
Mb
5 15 25 30 40 45 50 55 65 70 75 80 85 90
LOD
14 12 10 8 6 4 2 0
cM
0 10 20 30 40 50 60
Map position (cM)
188


B6BTBR07: 133 islet
LOD
1 2 3 4 5 6 7 9 10 12 13 15 16 17 18
Chromosome


B6BTBR07: 133 islet
Count of 1.5 LOD Support Intervals (blue = peak
15
10
5
0
1 2 3 4 5 6 7 9 10 12 13 15 16 17 18
Chromosome

189


ownload PDF Image
B6BTBR07: 133 islet
LOD
Phenotype Index
100
80
60
40
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Chromosome

190

Inferring Causal Phenotype Networks

Elias Chaibub Neto & Brian S. Yandell

UW-Madison

June 2010

outline

- QTL-driven directed graphs
  - Assume QTLs known, network unknown
  - Infer links (edges) between pairs of phenotypes (nodes)
    - Based on partial correlation
  - Infer causal direction for edges
  - Chaibub et al. (2008 *Genetics*)
  - Software R/qdg available on CRAN
- Causal graphical models in systems genetics
  - QTLs unknown, network unknown
  - Infer both genetic architecture (QTLs) and pathways (networks)
  - Chaibub et al. (2010 *Ann Appl Statist*)
  - Software R/qtlnet (www.stat.wisc.edu/~yandell/sysgen/qtlnet)

QTL-driven directed graphs

- See edited slides by Elias Chaibub Neto
  - BIOCOMP 2008 talk
  - Chaibub Neto, Ferrara, Attie, Yandell (2008) Inferring causal phenotype networks from segregating populations. *Genetics 179*: 1089-1100.
  - Ferrara et al. Attie (2008) Genetic networks of liver metabolism revealed by integration of metabolic and transcriptomic profiling. *PLoS Genet 4*: e1000034.

Introduction

- Our objective is to learn metabolic pathways from data.
- We represent these pathways by directed networks composed by transcripts, metabolites and clinical traits.
- These phenotypes are quantitative in nature, and can be analyzed using quantitative genetics techniques.

Introduction

- In particular, we use Quantitative Trait Loci (QTL) mapping methods to identify genomic regions affecting the phenotypes.

- Since variations in the genotypes (QTLs) cause variations in the phenotypes, but not the other way around, we can unambiguously determine the causal direction

$$\text{QTL} \Rightarrow \text{phenotype}$$

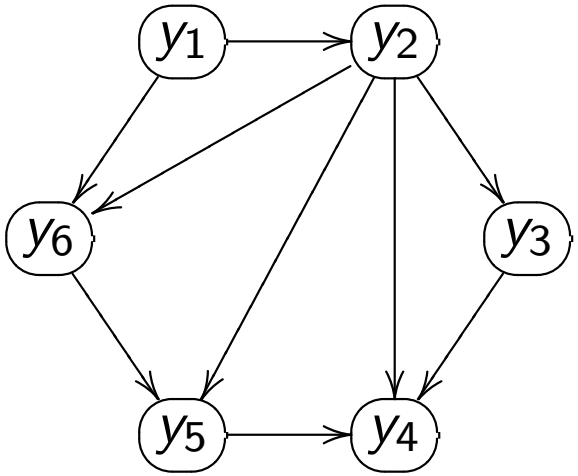
- Knowing that a QTL causally affects a phenotype will allow us to infer causal direction between phenotypes.

PC algorithm

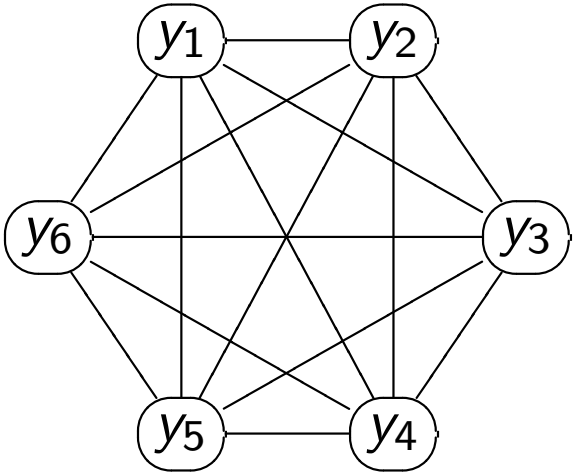
- Causal discovery algorithm developed by Spirtes et al 1993.
- It is composed of two parts:
  1. Infers the skeleton of the causal model.
  2. Partially orient the graph (orient some but not all edges).
- We are only interested in the first part (the "PC skeleton algorithm"). We do **not** use the PC algorithm to edge orientation (we use the QDG algorithm instead).

Step 1 (PC skeleton algorithm)

Suppose the true network describing the causal relationships between six transcripts is



The PC-algorithm starts with the complete undirected graph



and progressively eliminates edges based on conditional independence tests.

Step 1 (PC skeleton algorithm)

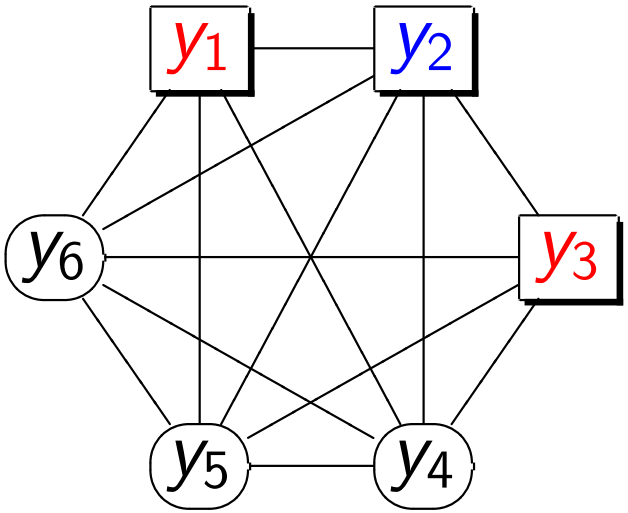
The algorithm performs several rounds of conditional independence tests of increasing order.

It starts with all zero order tests, then performs all first order, second order ...

- Notation: $\perp\!\!\!\perp \equiv$ independence. We read $i \perp\!\!\!\perp j \mid k$ as *i is conditionally independent from j given k*.
- Remark: in the Gaussian case zero partial correlation implies conditional independence, thus

$$i \perp\!\!\!\perp j \mid k \quad \Leftrightarrow \quad cor(i, j \mid k) = 0 \quad \Rightarrow \quad \text{drop } (i, j) \text{ edge}$$

Example (order 1)

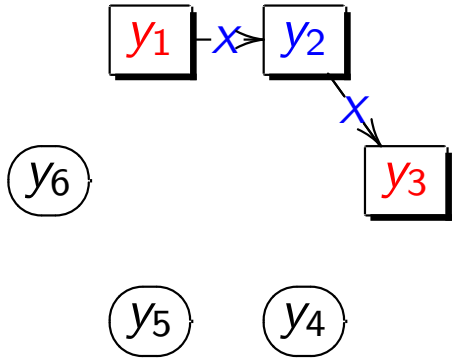


$A(1) \setminus 2 = \{2, 4, 5, 6\}$

$1 \perp\!\!\!\perp 3 \mid 2$

vs

$1 \not\perp\!\!\!\perp 3 \mid 2$



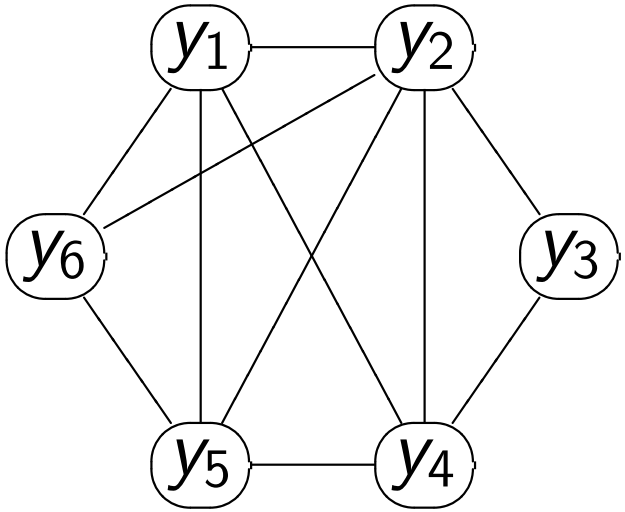
y_2 d-separates y_1 from y_3

$1 \perp\!\!\!\perp 3 \mid 2$

drop edge

move to next edge

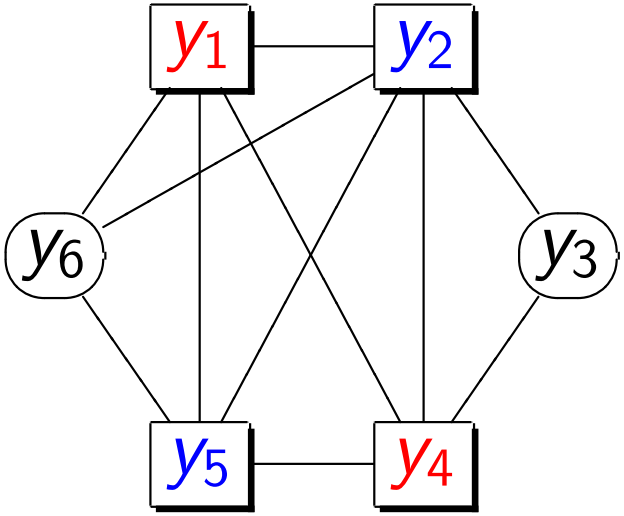
Example (order 1)



After all first order conditional independence tests.

The algorithm then moves to second order conditional independence tests.

Example (order 2)

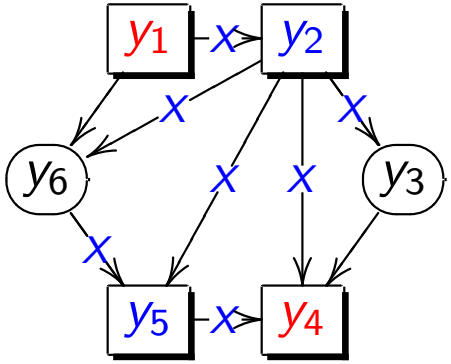


$A(1) \setminus 4 = \{2, 5, 6\}$

$1 \perp\!\!\!\perp 4 \mid 2, 5$

vs

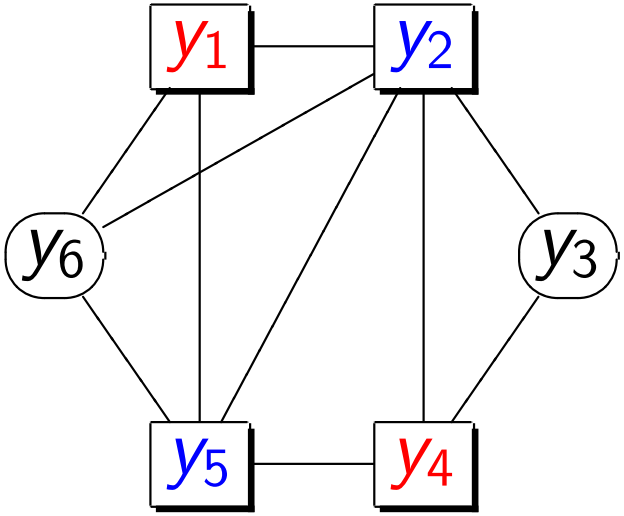
$1 \not\perp\!\!\!\perp 4 \mid 2, 5$



(y_2, y_5) d-separate y_1 from y_4

$1 \perp\!\!\!\perp 4 \mid 2, 5$

Example (order 2)

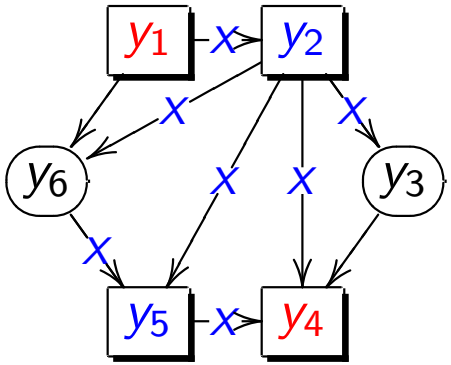


$A(1) \setminus 4 = \{2, 5, 6\}$

$1 \perp\!\!\!\perp 4 \mid 2, 5$

vs

$1 \not\perp\!\!\!\perp 4 \mid 2, 5$



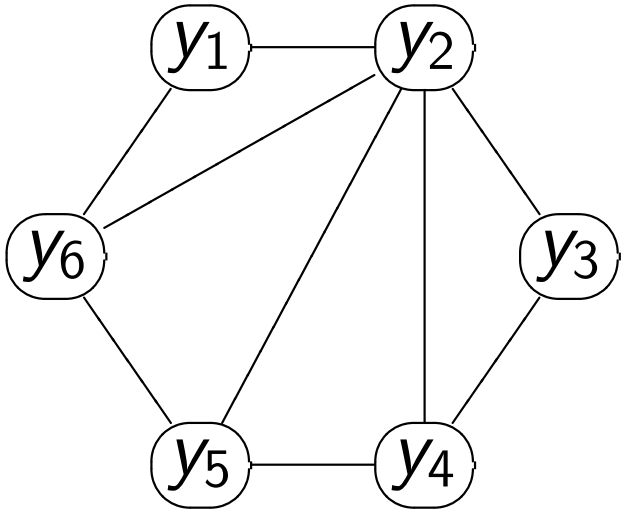
(y_2, y_5) d-separate y_1 from y_4

$1 \perp\!\!\!\perp 4 \mid 2, 5$

drop edge

move to next edge

Example (order 2)



After all second order conditional independence tests.

The algorithm than moves to third order, fourth order ...

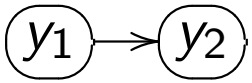
It stops when for each pair (i, j) the cardinality of

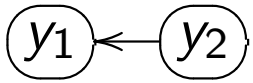
$$A(i) \setminus j$$

is smaller then the order of the algorithm.

Edge orientation

Consider two traits y_1 and y_2. Our problem is to decide between models:

M_1 : 

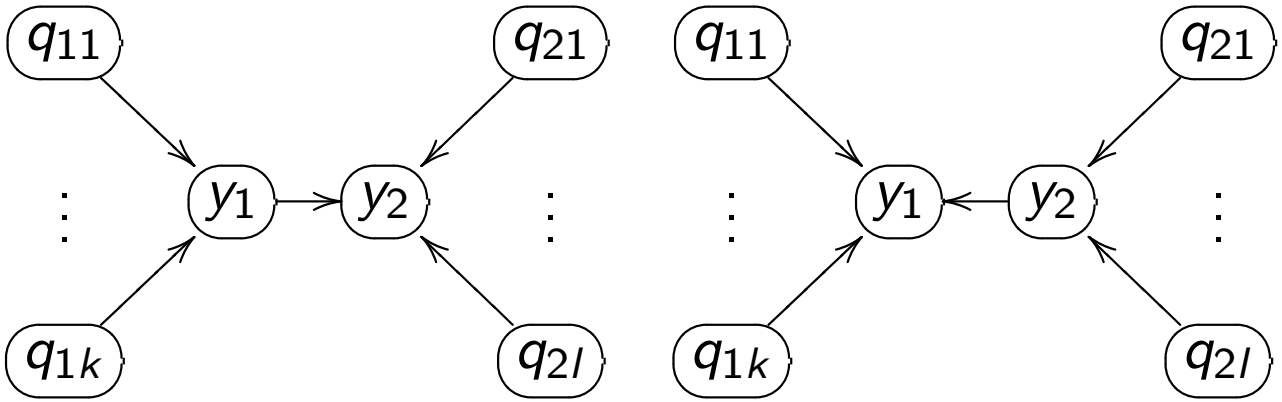
$M2$: 

Problem: the above models are likelihood equivalent,

$$f(y_1)f(y_2 \mid y_1) = f(y_1, y_2) = f(y_2)f(y_1 \mid y_2) \ .$$

Edge orientation

However, models



are *not* likelihood equivalent because

$$f(\mathbf{q}_1)f(y_1 \mid \mathbf{q}_1)f(y_2 \mid y_1, \mathbf{q}_2)f(\mathbf{q}_2)$$
$$\neq$$
$$f(\mathbf{q}_2)f(y_2 \mid \mathbf{q}_2)f(y_1 \mid y_2, \mathbf{q}_1)f(\mathbf{q}_1)$$

Edge orientation

We perform model selection using a direction LOD score

$$LOD = \log_{10} \left\{ \frac{\prod_{i=1}^{n} f(y_{1i} \mid \mathbf{q}_{1i}) f(y_{2i} \mid y_{1i}, \mathbf{q}_{2i})}{\prod_{i=1}^{n} f(y_{2i} \mid \mathbf{q}_{2i}) f(y_{1i} \mid y_{2i}, \mathbf{q}_{1i})} \right\}$$

where $f()$ represents the predictive density, that is, the sampling model with parameters replaced by the corresponding maximum likelihood estimates.

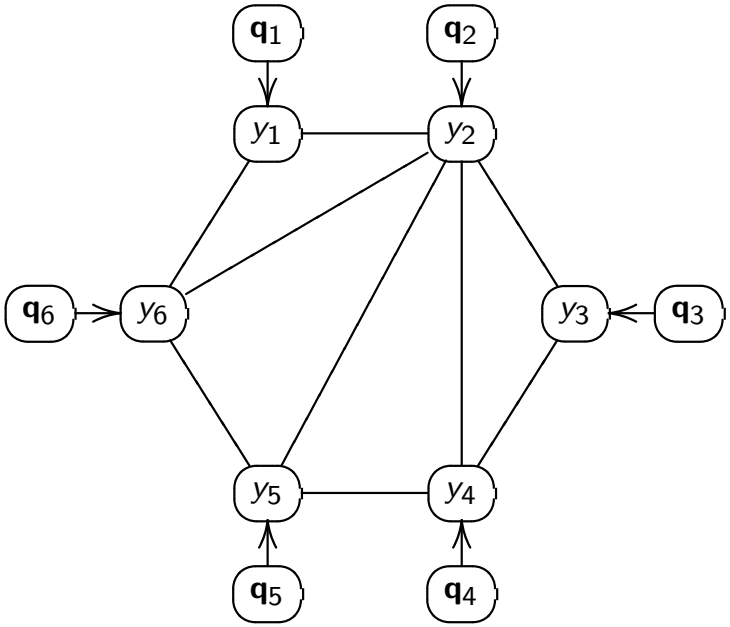
QDG algorithm

QDG stands for QTL-directed dependency graph.
The QDG algorithm is composed of 7 steps:

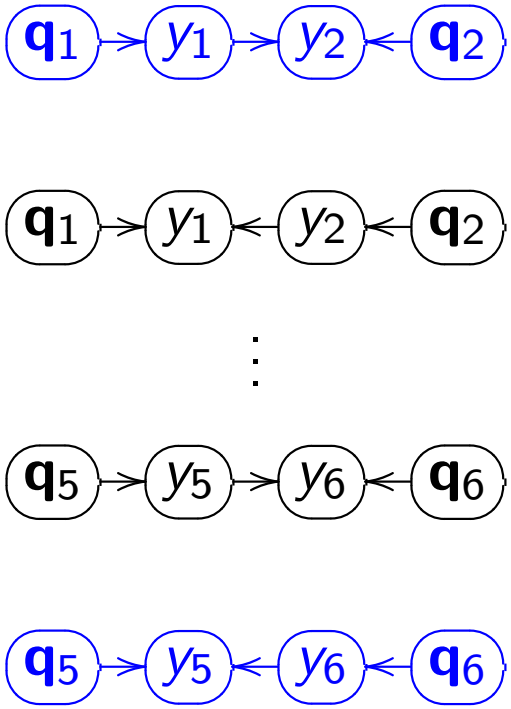
1. Get the causal skeleton (with the PC skeleton algorithm).
2. Use QTLs to orient the edges in the skeleton.
3. Choose a random ordering of edges, and
4. Recompute orientations incorporating causal phenotypes in the models (update the causal model according to changes in directions).
5. Repeat 4 iteratively until no more edges change direction (the resulting graph is one solution).
6. Repeat steps 3, 4, and 5 many times and store all different solutions.
7. Score all solutions and select the graph with best score.

Step 2

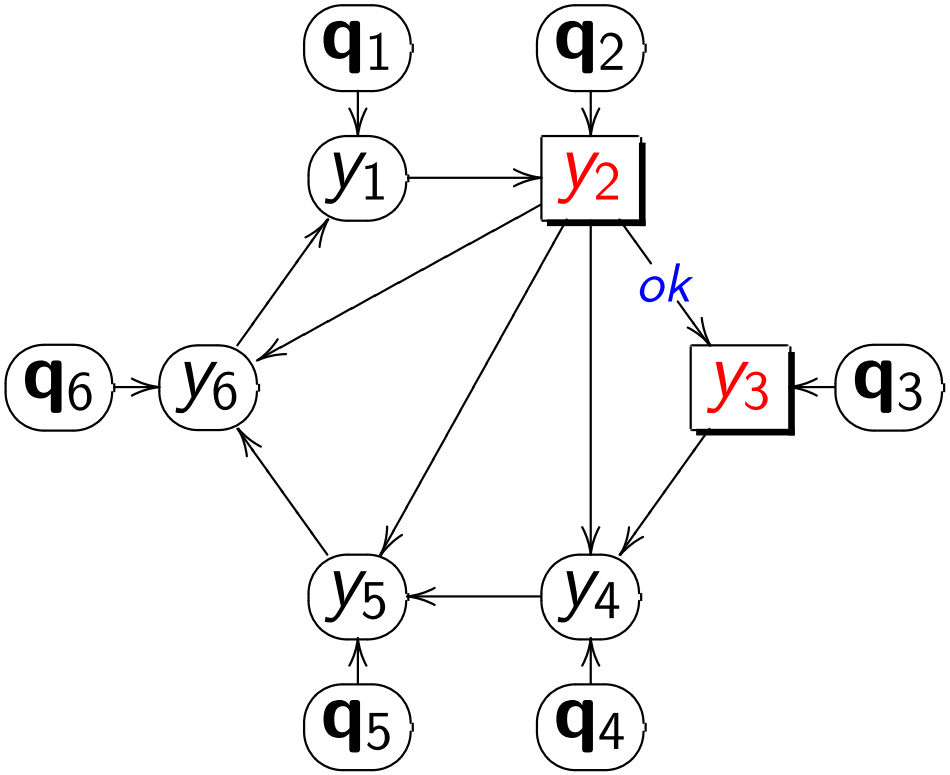
Now suppose that for each transcript we have a set of e-QTLs

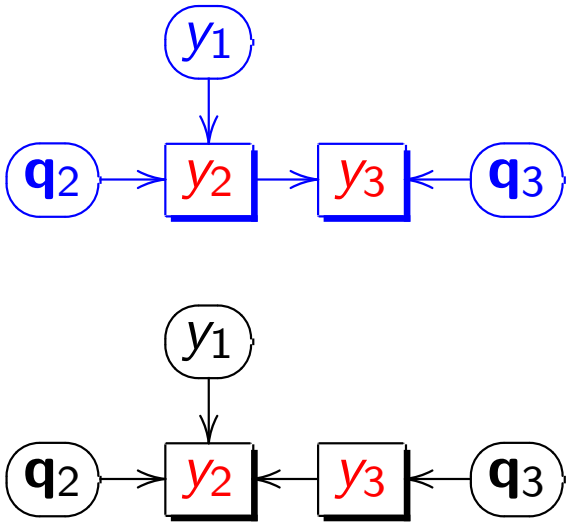


Given the QTLs we can distinguish causal direction:

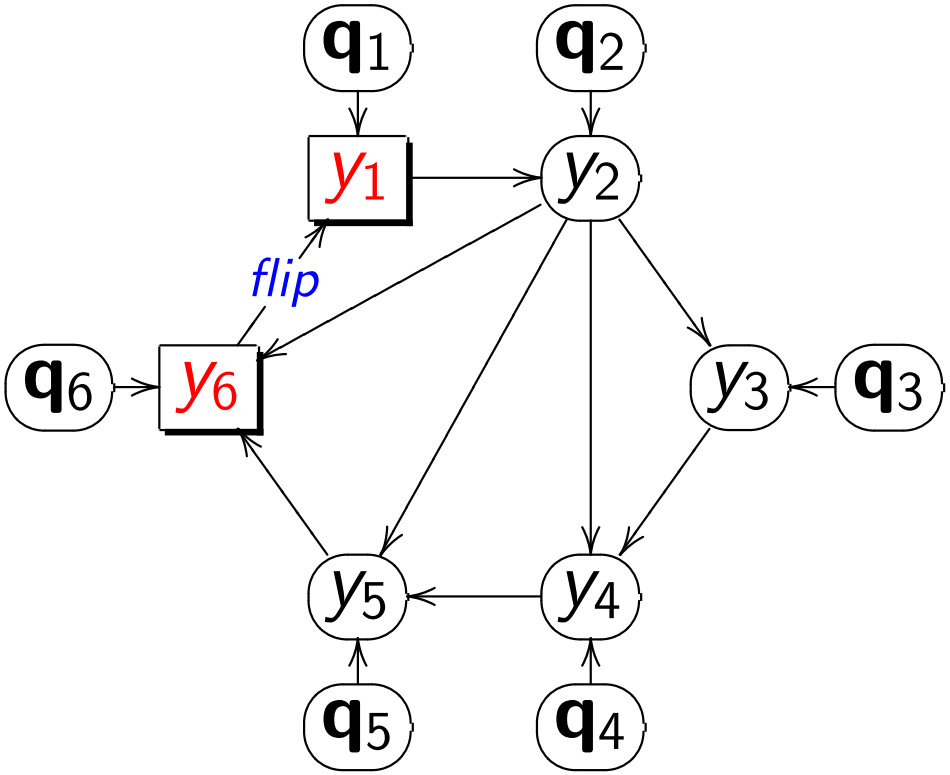


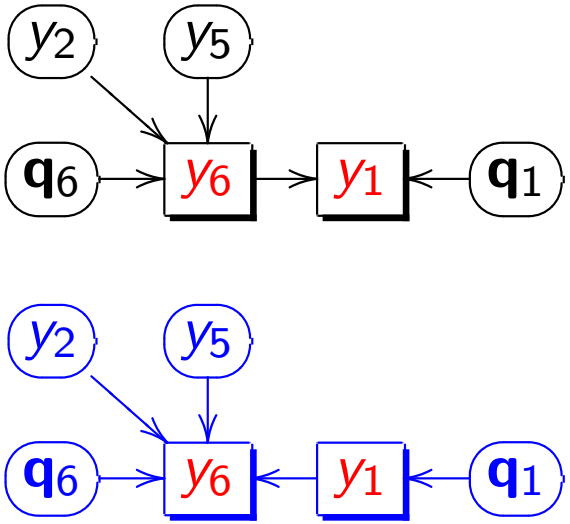
Steps 4 and 5 (first iteration)





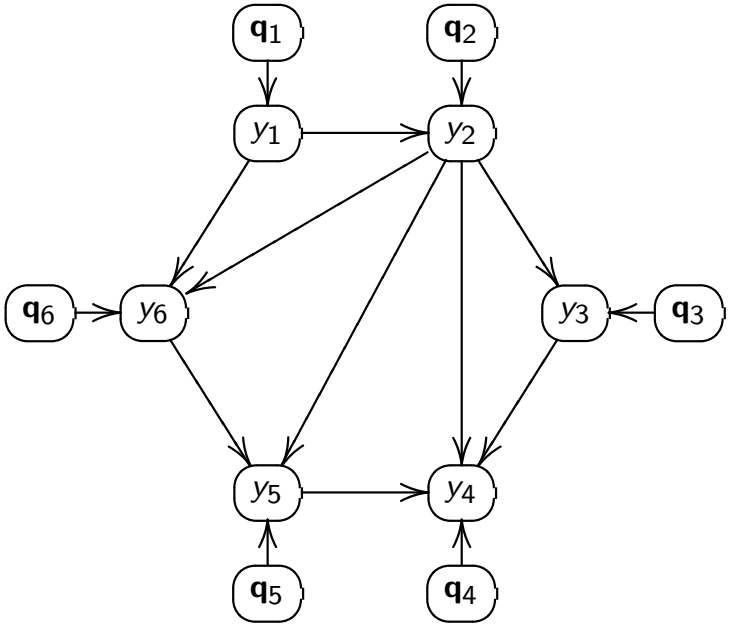
Steps 4 and 5 (first iteration)





Steps 4 and 5 (first iteration)

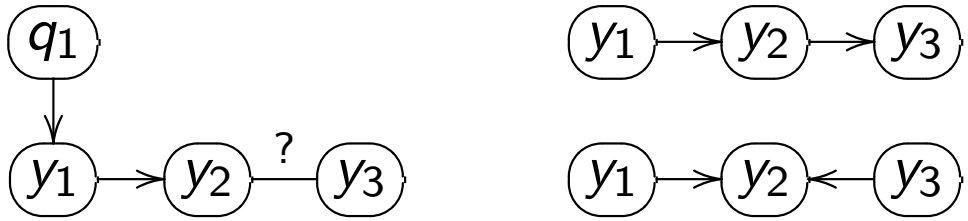
Suppose the updated causal model after the first iteration (DG_1) is



Since some arrows changed direction ($DG_1 \neq DG_0$), the algorithm goes for another round of re-computations.

Directing edges without QTLs

- In general we need to have at least one QTL per pair of phenotypes to infer causal direction.
- In some situations, however, we may be able to infer causal direction for a pair of phenotypes without QTLs. Eg.

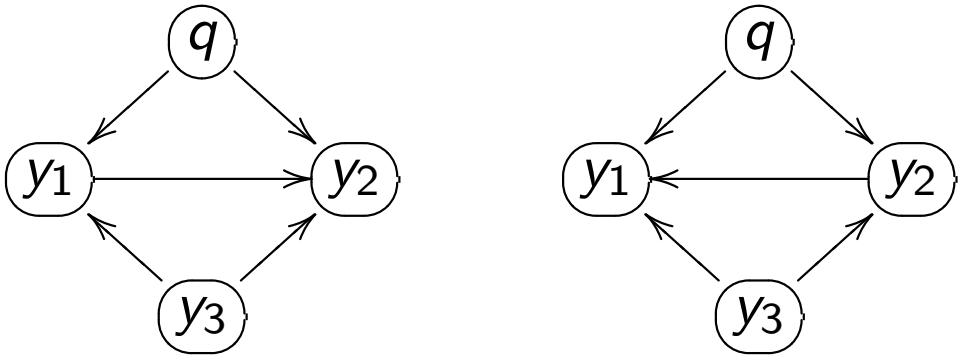


since $f(y_1)\, f(y_2 \mid y_1)\, f(y_3 \mid y_2) \neq f(y_1)\, f(y_2 \mid y_1, y_3)\, f(y_3)$.

- So both QTLs and phenotypes play important roles in the orientation process.

Unresolvable situation

- We cannot infer direction when the phenotypes have exactly same set of QTLs and causal phenotypes

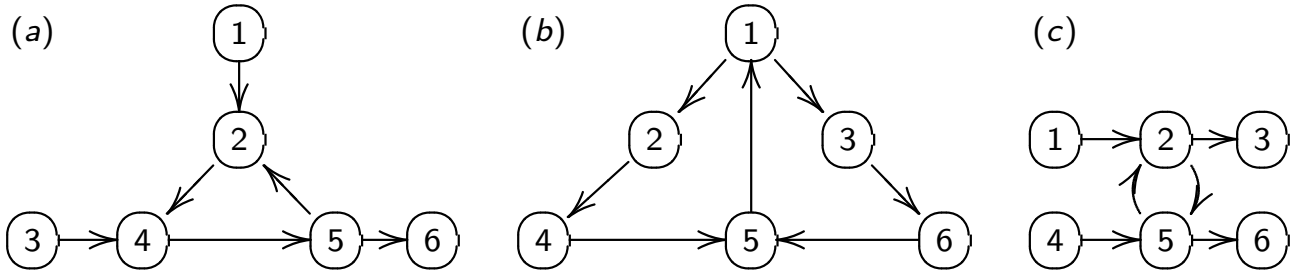


since

$$f(y_1 \mid y_3, q)\, f(y_2 \mid y_1, y_3, q) \;=\; f(y_1 \mid y_2, y_3, q)\, f(y_2 \mid y_3, q)$$

Cyclic networks

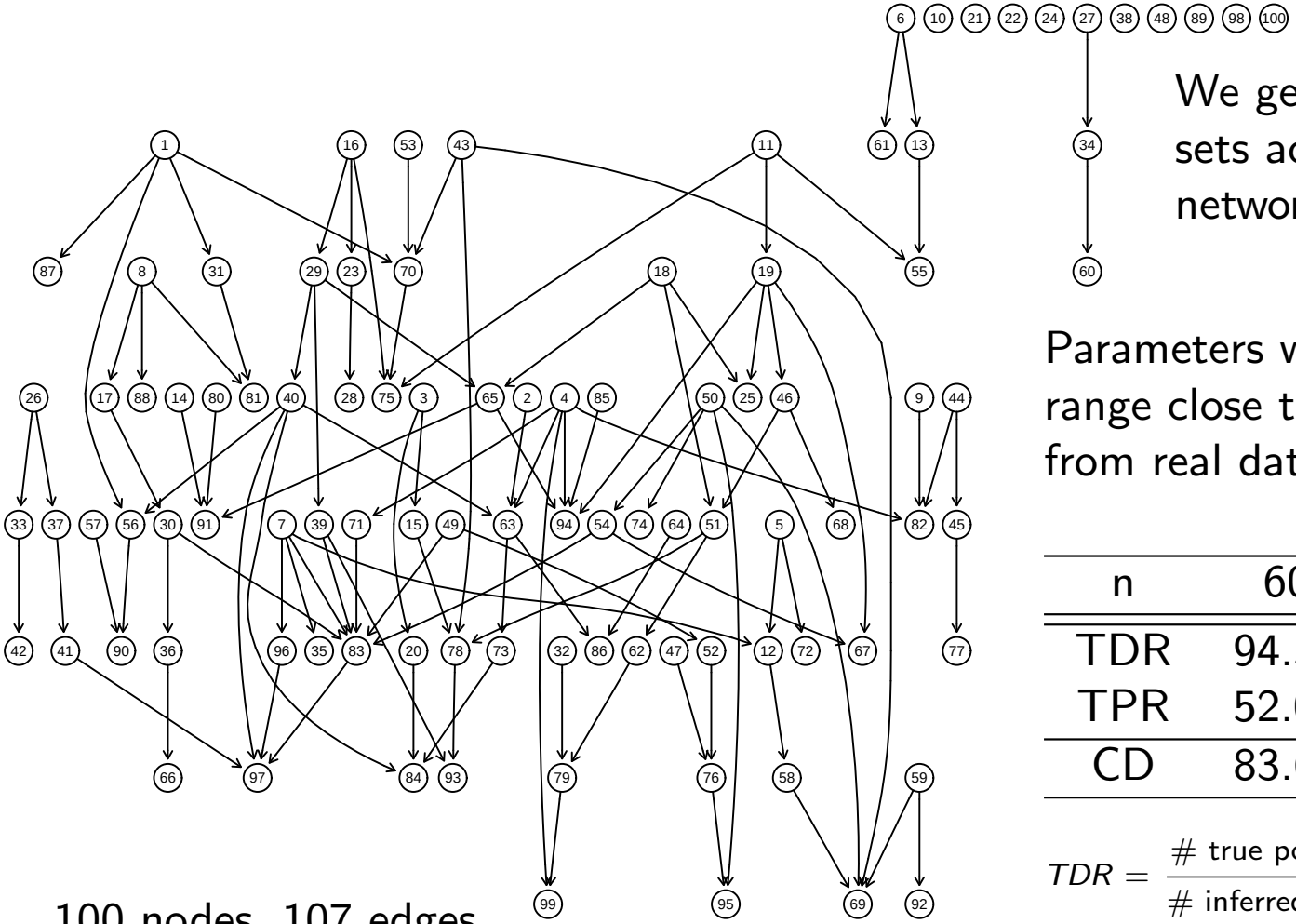
- Our simulations showed good performance with toy cyclic graphs, though.



- The spurious edges in graph (c) were detected at low rates.

- QDG approach cannot detect reciprocal interactions. In graph (c) it orients the edge (2)—(5) in the direction with higher strength.

Simulations



100 nodes, 107 edges

2 or 3 QTLs per phenotype (not shown)

We generated 100 data sets according to this network.

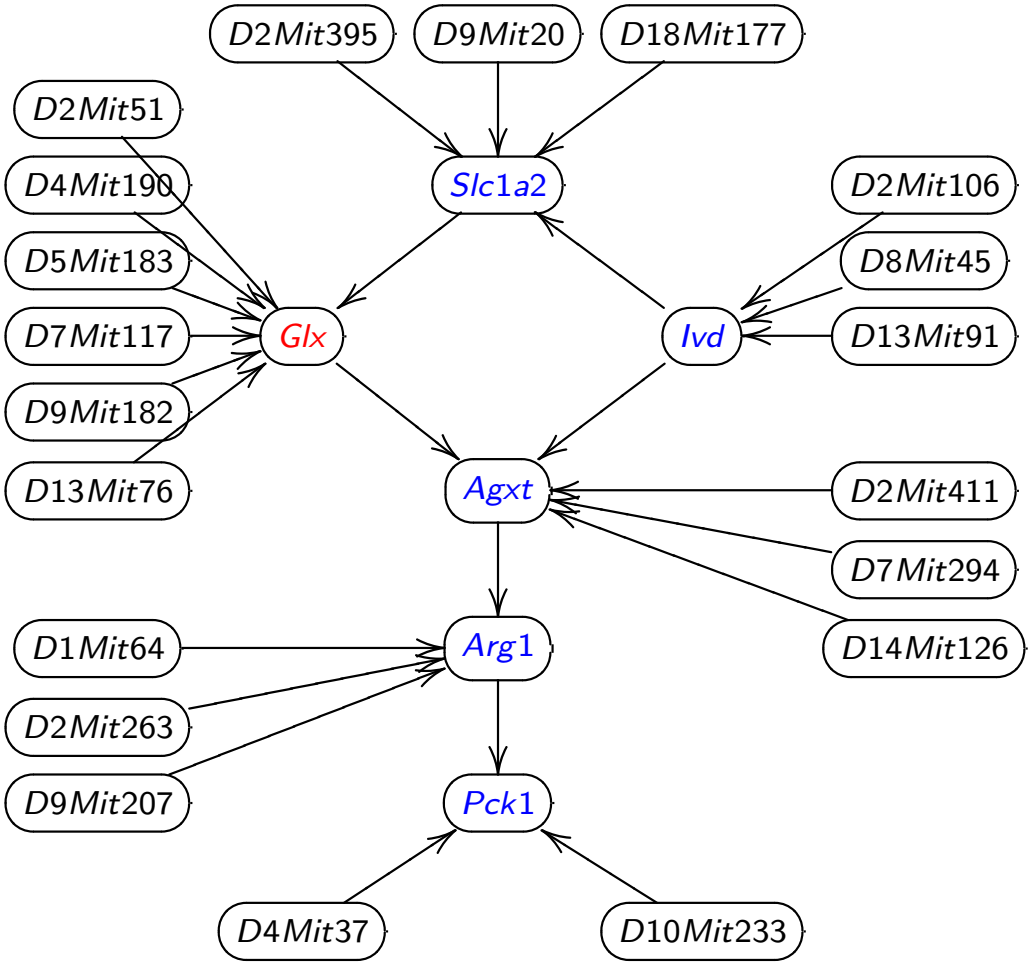
Parameters were chosen in a range close to values estimated from real data.

n	60	300	500
TDR	94.53	95.18	91.22
TPR	52.07	87.33	93.64
CD	83.65	98.58	99.63

$$TDR = \frac{\#\text{ true positives}}{\#\text{ inferred edges}}, \quad TPR = \frac{\#\text{ true positives}}{\#\text{ true edges}}$$

CD: correct direction

Real data example



- We constructed a network from metabolites and transcripts involved in liver metabolism.
- We validated this network with *in vitro experiments* (Ferrara et al 2008). Four out of six predictions were confirmed.

Software and references

The *qdg* R package is available at CRAN.

References:

- Chaibub Neto et al 2008. Inferring causal phenotype networks from segregating populations. Genetics 179: 1089-1100.
- Ferrara et al 2008. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptomic profiling. PLoS Genetics 4: e1000034.
- Spirtes et al 1993. Causation, prediction and search. MIT press.

Acknowledgements

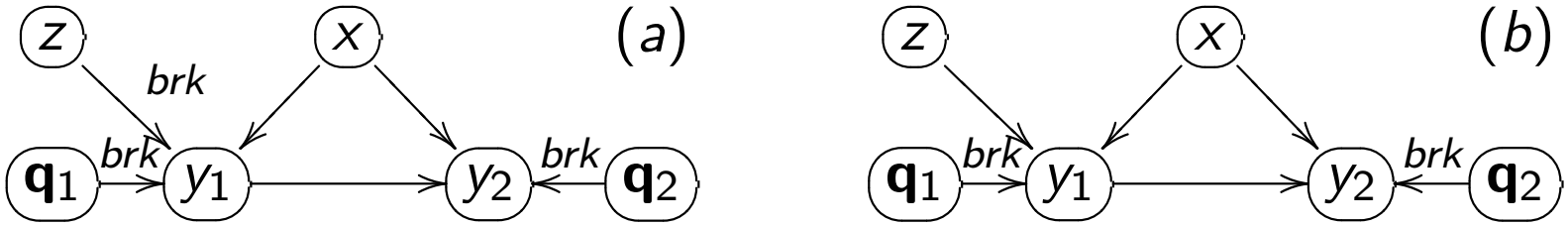
Co-authors:

- Alan D. Attie
- Brian S. Yandell
- Christine T. Ferrara

Funding:

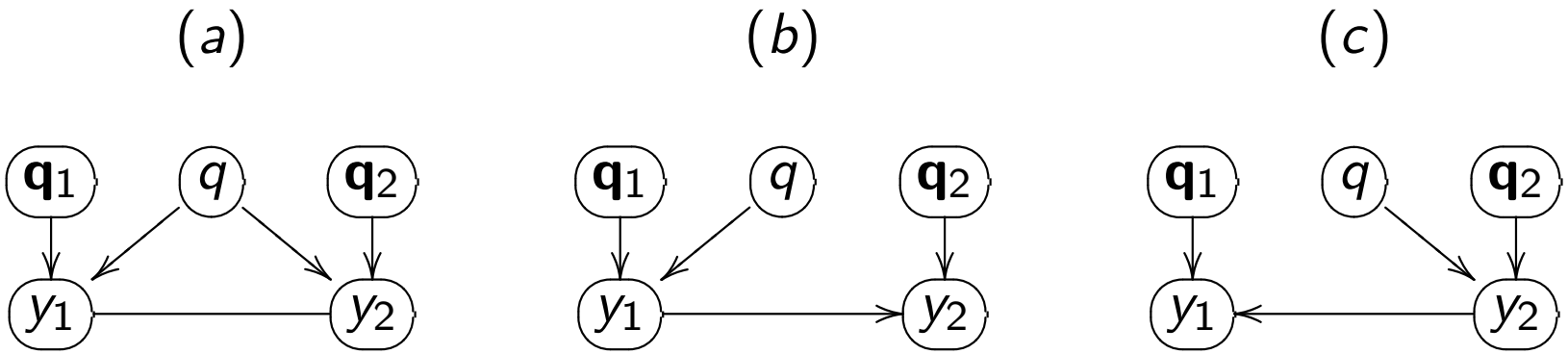
- CNPq Brazil
- NIH grants DK66369, DK58037 and DK06639

Permutation p-values



- To break the connections (brk) that affect direction of an edge, we permute the corresponding pair of nodes (and their common covariates) as a block.
- In panel (a) we permute (y_1, y_2, x) as a block breaking the connections with z, $\mathbf{q}_1$ and $\mathbf{q}_2$;
- In panel (b) we incorrectly keep z in the permutation block.

Direct versus indirect effects of a common QTL



- A strong QTL directly affecting an upstream trait may also be (incorrectly) detected as a QTL for a downstream phenotype.
- To resolve this situation we apply a generalization of Schadt et al. 2005 allowing for multiple QTLs.
- Model (a) supports both traits being directly affected by the common QTL q. Model (b) implies that q directly affects y_1 but should not be included as a QTL of phenotype y_2. Model (c) supports the reverse situation.

causal graphical models in systems genetics

- Chaibub Neto, Keller, Attie , Yandell (2010) Causal Graphical Models in Systems Genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann Appl Statist 4*: 320-339)
- Related references
  - Schadt et al. Lusis (2005 *Nat Genet*); Li et al. Churchill (2006 *Genetics*); Chen Emmert-Streib Storey(2007 *Genome Bio*); Liu de la Fuente Hoeschele (2008 *Genetics*); Winrow et al. Turek (2009 *PLoS ONE*)
- Jointly infer unknowns of interest
  - genetic architecture
  - causal network

Basic idea of QTLnet

- Genetic architecture given causal network
  - Trait y depends on parents pa(y) in network
  - QTL for y found conditional on pa(y)
    - Parents pa(y) are interacting covariates for QTL scan
- Causal network given genetic architecture
  - Build (adjust) causal network given QTL

MCMC for QTLnet

- Propose new causal network with simple changes to current network
  - Change edge direction
  - Add or drop edge
- Find any new genetic architectures (QTLs)
  - Update phenotypes whose parents pa(y) change in new network
- Compute likelihood for new network and QTL
- Accept or reject new network and QTL
  - Usual Metropolis-Hastings idea

Future work

- Incorporate latent variables
  - Aten et al. Horvath (2008 *BMC Sys Biol*)
- Allow for prior information about network
  - Werhli and Husmeier (2007 *SAGMB*); Dittrich et al. Müller (2008 *Bioinfo*); Zhu et al. Schadt (2008 *Nat Genet*); Lee et al. Koller (2009 *PLoS Genet*); Thomas et al. Portier (2009 *Genome Bio*); Wu et al. Lin (2009 *Bioinfo*)
- Improve algorithm efficiency
  - Ramp up to 1000s of phenotypes
- Extend to outbred crosses, humans