# Bayesian Interval Mapping





# 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

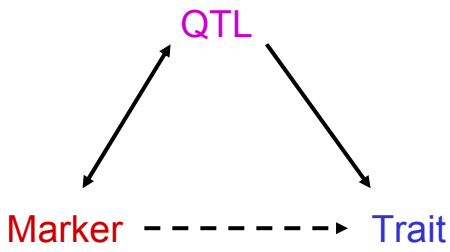
cross two inbred lines

→ linkage disequilibrium

→ associations

→ linked segregating QTL

(after Gary Churchill)






# 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?

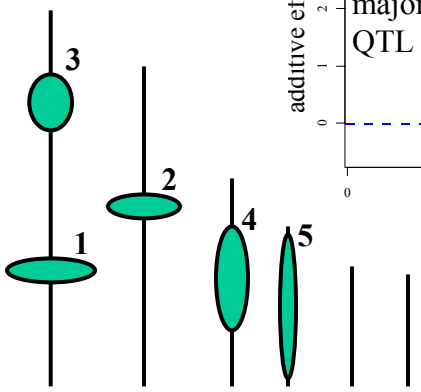
# advantages of multiple QTL approach

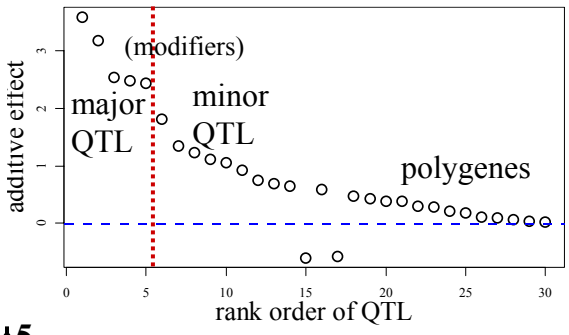
- improve statistical power, precision
  - increase number of QTL detected
  - better estimates of loci: less bias, smaller intervals
- improve inference of complex genetic architecture
  - patterns and individual elements of epistasis
  - appropriate estimates of means, variances, covariances
    - 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



# 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?

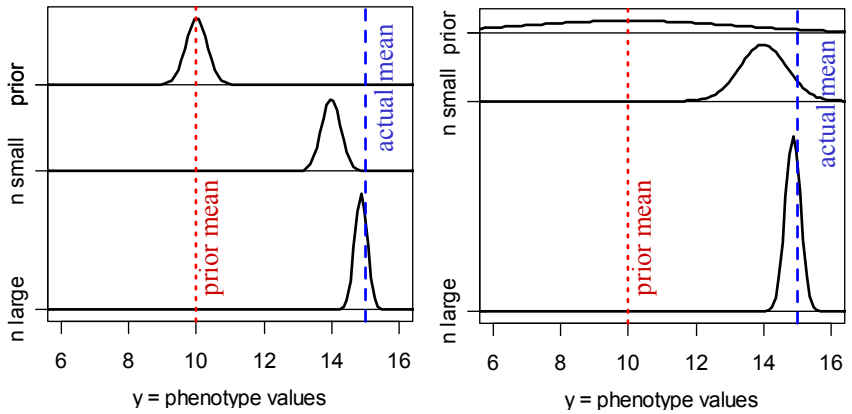
- problem of selection bias
  - QTL of modest effect only detected sometimes
  - their effects are biased upwards when detected
- probability that QTL detected
  - avoids sharp in/out dichotomy
  - avoid pitfalls of one "best" model
  - examine "better" models with more probable QTL
- build $m$ = number of QTL detected into QTL model
  - directly allow uncertainty in genetic architecture
  - model selection over genetic architecture

# 2. Bayesian QTL mapping

- 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



# Bayes posterior for normal data




small prior variance large prior variance

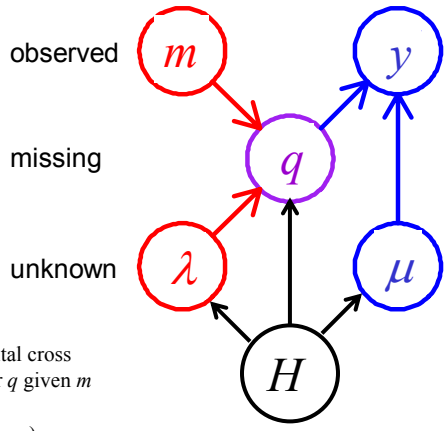
# Bayes posterior for normal data

| | |
|---|---|
| model | $y_i = \mu + e_i$ |
| environment | $e \sim N(0, \sigma^2)$, $\sigma^2$ known |
| likelihood | $y \sim N(\mu, \sigma^2)$ |
| prior | $\mu \sim N(\mu_0, \kappa\sigma^2)$, $\kappa$ 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 $\bar{y}_\bullet = \underset{\{i=1,\ldots,n\}}{\text{sum}} y_i / n$ |
| fudge factor (shrinks to 1) | $b_n = \frac{\kappa n}{\kappa n + 1} \to 1$ |



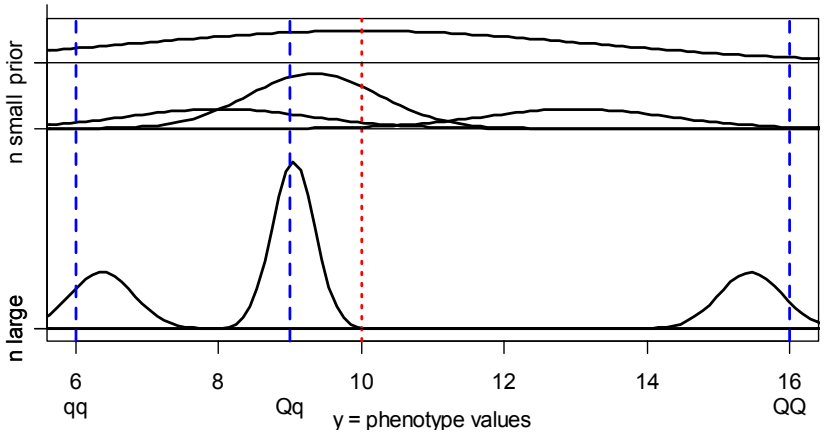
# Bayesian QTL: key players

- 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
  - $\lambda$ = QT locus (or loci)
  - $\mu$ = phenotype model parameters
  - $H$ = QTL model/genetic architecture
- pr($q|m,\lambda,H$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for $q$ given $m$
- pr($y|q,\mu,H$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters $\mu$ (could be non-parametric)




after
Sen Churchill (2001)

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






# Bayes posterior QTL means

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

prior: $\mu_q \sim N\left(\bar{y}_\bullet, \kappa\sigma^2\right)$

posterior: $\mu_q \sim N\left(b_q\bar{y}_q + (1-b_q)\bar{y}_\bullet, b_q\sigma^2/n_q\right)$

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

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

# partition of multiple QTL effects

- partition genotype-specific mean into QTL effects
  $\mu_q$ = mean + main effects + epistatic interactions
  $\mu_q = \mu + \beta_q = \mu + \text{sum}_{j \text{ in } H}\, \beta_{qj}$
- priors on mean and effects

| | | | |
|---|---|---|---|
| $\mu$ | ~ | $N(\mu_0, \kappa_0\sigma^2)$ | grand mean |
| $\beta_q$ | ~ | $N(0, \kappa_1\sigma^2)$ | model-independent genotypic effect |
| $\beta_{qj}$ | ~ | $N(0, \kappa_1\sigma^2/\lvert H\rvert)$ | effects down-weighted by size of $H$ |

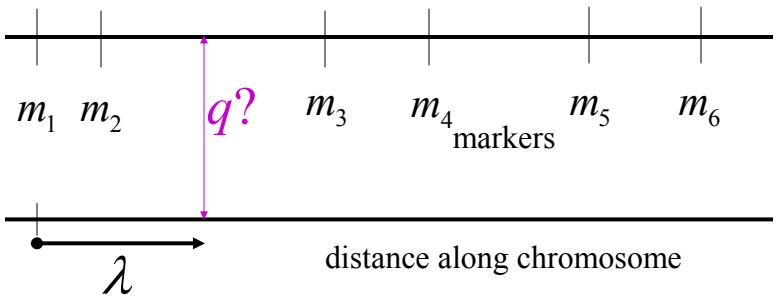
- determine hyper-parameters via empirical Bayes

$$\mu_0 \approx \bar{Y}_\bullet \text{ and } \kappa_1 \approx \frac{h^2}{1-h^2} = \frac{\sigma_G^2}{\sigma^2}$$

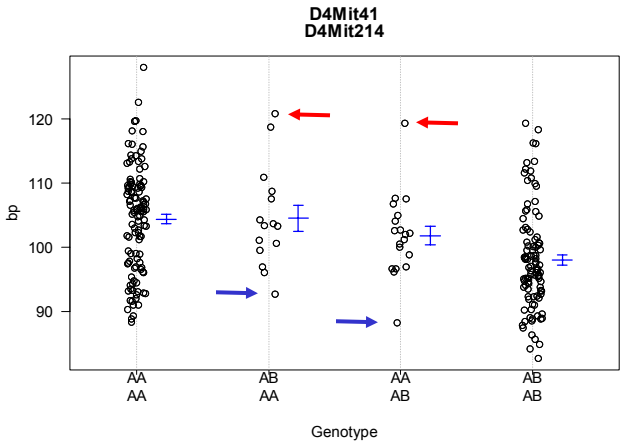


# pr($q|m,\lambda$) recombination model

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



# how does phenotype *Y* improve posterior for genotype *Q*?




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

- full conditional for $q$ depends data for individual $i$
  - proportional to prior $\text{pr}(q \mid m_i, \lambda)$
    - weight toward $q$ that agrees with flanking markers
  - proportional to likelihood $\text{pr}(y_i|q,\mu)$
    - weight toward $q$ so that group mean $\mu_q \approx y_i$
- phenotype and prior recombination may conflict
  - posterior recombination balances these two weights
  - this is "E step" in EM for classical QTL analysis

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

# Bayesian model posterior

- augment data ($y,m$) with unknowns $q$
- study unknowns ($\mu,\lambda,q$) given data ($y,m$)
  - properties of posterior pr($\mu,\lambda,q \mid y,m$ )
- sample from posterior in some clever way
  - multiple imputation or MCMC

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

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



# Bayesian priors for QTL

- missing genotypes $q$
  - pr( $q \mid m, \lambda$ )
  - recombination model is formally a prior
- effects ( $\mu$, $\sigma^2$ )
  - prior = pr( $\mu_q \mid \sigma^2$ ) pr($\sigma^2$ )
  - use conjugate priors for normal phenotype
    - pr( $\mu_q \mid \sigma^2$ ) = normal
    - pr($\sigma^2$ ) = inverse chi-square
- each locus $\lambda$ may be uniform over genome
  - pr($\lambda \mid m$ ) = 1 / length of genome
- combined prior
  - pr( $q, \mu, \lambda \mid m$ ) = pr( $q \mid m, \lambda$ ) pr( $\mu$ ) pr($\lambda \mid m$ )

# 3. Markov chain sampling of architectures

- 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
- hard to sample ($q$, $\mu$ , $\lambda$, $H$) from joint posterior
  - update ($q,\mu,\lambda$) from full conditionals for model $H$
  - update genetic architecture $H$

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

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



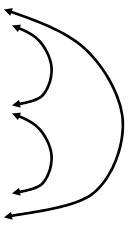
# MCMC sampling of $(\lambda,q,\mu)$

- Gibbs sampler
  - genotypes $q$
  - effects $\mu$
  - *not* loci $\lambda$

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

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

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



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

# full conditional for locus

- cannot easily sample from locus full conditional

$$\begin{aligned} \text{pr}(\lambda \mid y,m,\mu,q) &= \text{pr}(\lambda \mid m,q) \\ &= \text{pr}(q \mid m,\lambda)\,\text{pr}(\lambda) / \text{constant} \end{aligned}$$

- constant is very difficult to compute explicitly
  - must average over all possible loci $\lambda$ 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



# Gibbs sampler idea

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

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

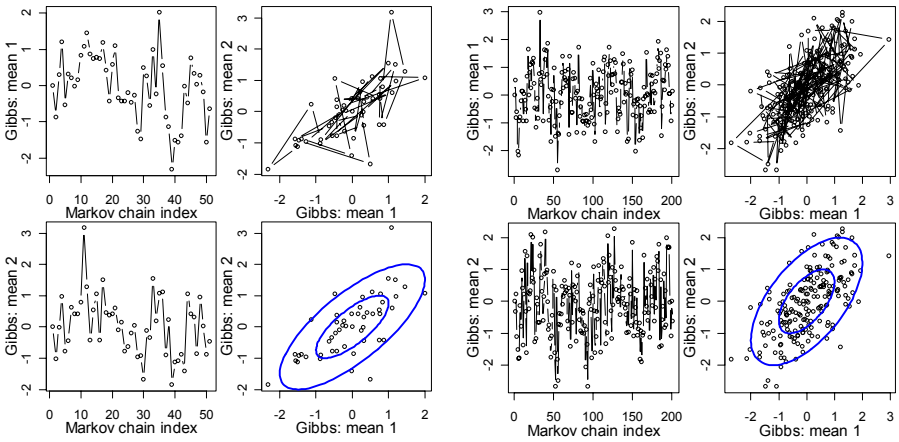
$$\mu_1 \sim N\left(\rho\mu_2, 1-\rho^2\right)$$

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

# Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

$N = 200$ samples

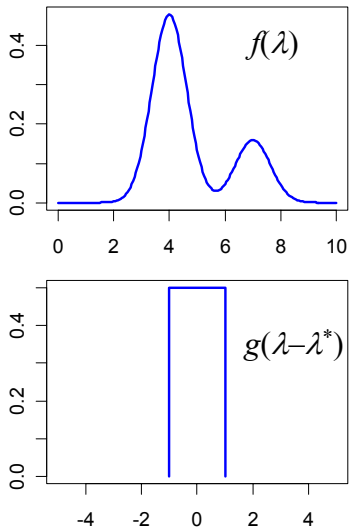




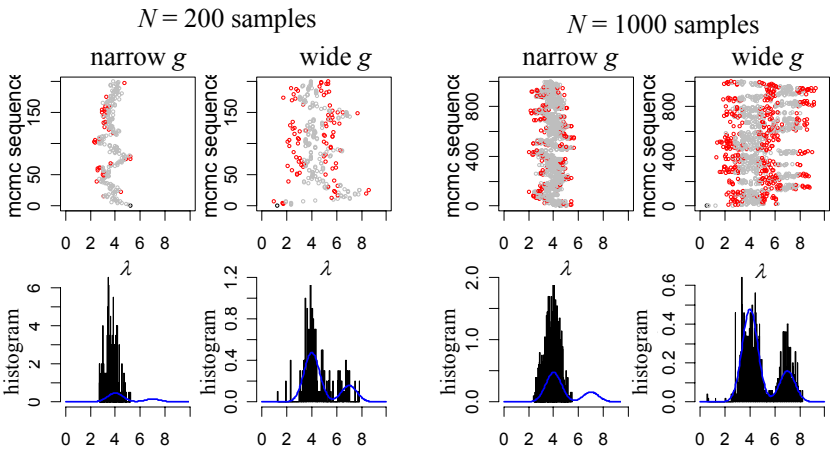

# Metropolis-Hastings idea

- 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 $\lambda^*$
    - near (?) current value $\lambda$
    - 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 samples






# 4. sampling across architectures

- search across genetic architectures *M* 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

# model selection in regression

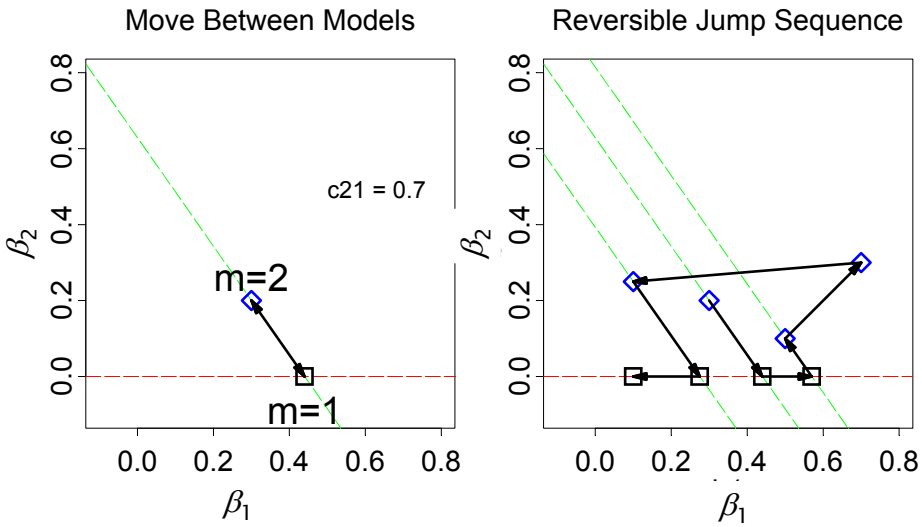
- consider known genotypes $q$ at 2 known loci $\lambda$
  - models with 1 or 2 QTL
- jump between 1-QTL and 2-QTL models
- adjust parameters when model changes
  - $\beta_{q1}$ estimate changes between models 1 and 2
  - due to collinearity of QTL genotypes

$$m = 1 : \mu_q = \mu + \beta_{q1}$$

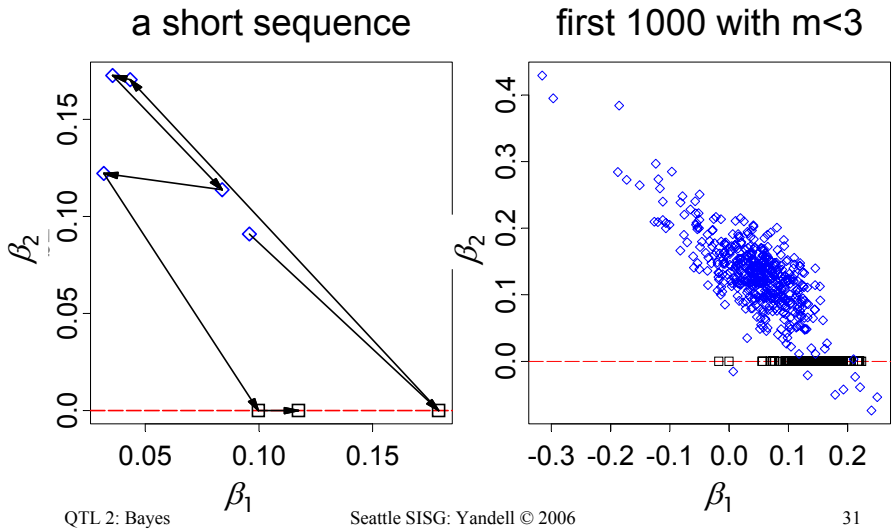
$$m = 2 : \mu_q = \mu + \beta_{q1} + \beta_{q2}$$



# geometry of reversible jump

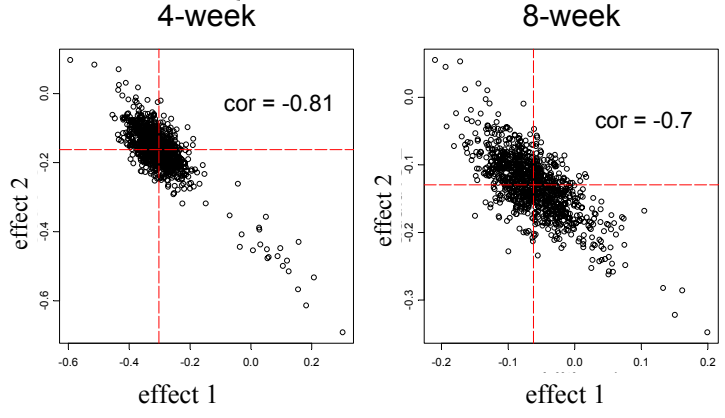


# geometry allowing $q$ and $\lambda$ to change






# collinear QTL = correlated effects




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

# reversible jump MCMC idea


- Metropolis-Hastings updates: draw one of three choices
  - update *m*-QTL model with probability 1-*b*(*m*+1)-*d*(*m*)
    - update current model using full conditionals
    - sample *m* QTL loci, effects, and genotypes
  - add a locus with probability *b*(*m*+1)
    - propose a new locus and innovate new genotypes & genotypic effect
    - decide whether to accept the "birth" of new locus
  - drop a locus with probability *d*(*m*)
    - propose dropping one of existing loci
    - decide whether to accept the "death" of locus
- Satagopan Yandell (1996, 1998); Sillanpaa Arjas (1998); Stevens Fisch (1998)
  - these build on RJ-MCMC idea of Green (1995); Richardson Green (1997)



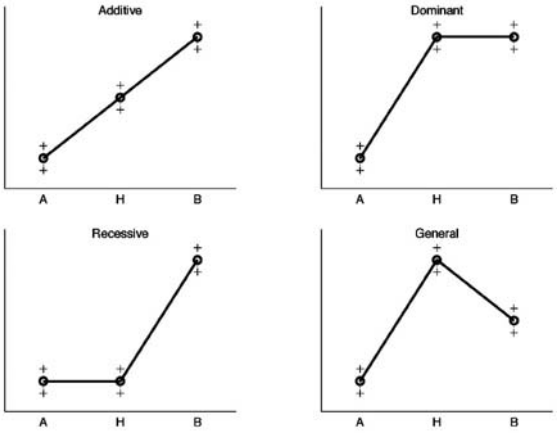
# Gibbs sampler with loci indicators

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

$$\mu_q = \mu + \delta_1 \beta_{q1} + \delta_2 \beta_{q2}$$

# 5. Gene Action and Epistasis

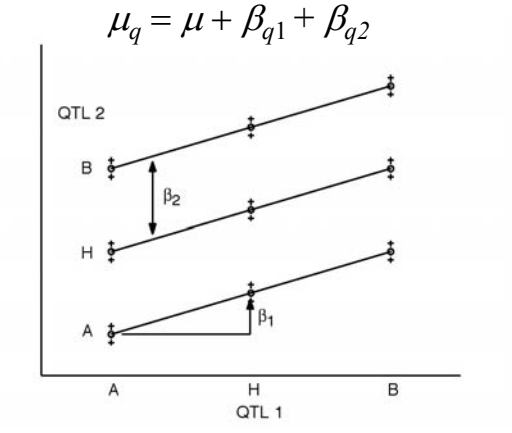
additive, dominant, recessive, general effects of a single QTL (Gary Churchill)






# additive effects of two QTL (Gary Churchill)

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



# Epistasis (Gary Churchill)

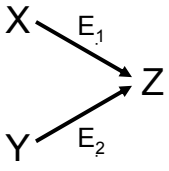
The allelic state at one locus can mask or uncover the effects of allelic variation at another.

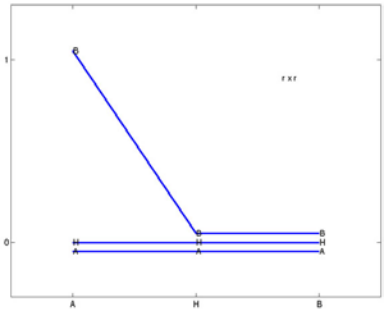
- W. Bateson, 1907.



# 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 $E_2$



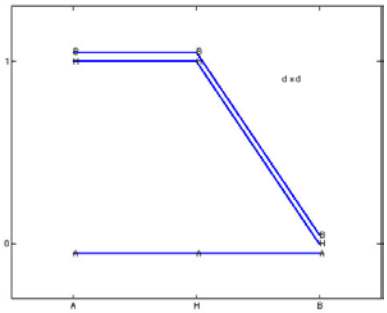



# epistasis in a serial pathway (GAC)

- Z keeps trait value high

$$X \xrightarrow{E_1} Y \xrightarrow{E_2} Z$$

- neither $E_1$ nor $E_2$ is rate limiting
- loss of function alleles are segregating from parent B at $E_1$ and from parent A at $E_2$





# QTL with epistasis

- same phenotype model overview

$$y = \mu_q + e, \text{var}(e) = \sigma^2$$

- partition of genotypic value with epistasis

$$\mu_q = \mu + \beta_{q1} + \beta_{q2} + \beta_{q12}$$

- partition of genetic variance & heritability

$$\text{var}(\mu_q) = \sigma_G^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}^2$$

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2$$

# 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?
- Yi Xu (2000) *Genetics;* Yi, Xu, Allison (2003) *Genetics;* Yi *et al.* (2005) *Genetics*



# limits of epistatic inference

- power to detect effects
  - epistatic model size grows exponentially
    - $|H| = 3^{nqtl}$ for general interactions
  - power depends on ratio of $n$ to model size
    - want $n / |H|$ to be fairly large (say > 5)
    - $n = 100$, $nqtl = 3$, $n / |H| \approx 4$
- empty cells mess up adjusted (Type 3) tests
  - missing $q_1Q_2$ / $q_1Q_2$ or $q_1Q_2q_3$ / $q_1Q_2q_3$ genotype
  - null hypotheses not what you would expect
  - can confound main effects and interactions
  - can bias AA, AD, DA, DD partition

# 6. comparing QTL models

- balance model fit with model "complexity“
  - want maximum likelihood
  - without too complicated a model
- information criteria quantifies the balance
  - Bayes information criteria (BIC) for likelihood
  - Bayes factors for Bayesian approach



# Bayes factors & BIC

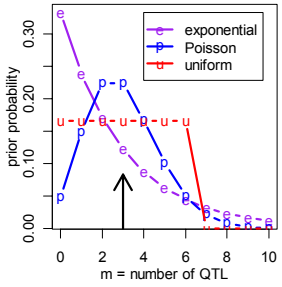
$$B_{12} = \frac{\text{pr}(\text{ model}_1 \mid Y) / \text{pr}(\text{ model}_2 \mid Y)}{\text{pr}(\text{ model}_1) / \text{pr}(\text{ model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

- what is a Bayes factor?
  - ratio of posterior odds to prior odds
  - ratio of model likelihoods
- BF is equivalent to *LR* statistic when
  - comparing two nested models
  - simple hypotheses (e.g. 1 vs 2 QTL)
- BF is equivalent to Bayes Information Criteria (BIC)
  - for general comparison of any models
  - want Bayes factor to be substantially larger than 1 (say 10 or more)

$$-2\log(B_{12}) = -2\log(LR) - (p_2 - p_1)\log(n)$$

# Bayes factors and genetic model $H$

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




$$BF_{H,H+1} = \frac{\text{pr}(H|y,m)/\text{pr}(H)}{\text{pr}(H+1|y,m)/\text{pr}(H+1)}$$

- pattern of QTL across genome
- gene action and epistasis



# issues in computing Bayes factors

- *BF* insensitive to shape of prior on *nqtl*
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects $\theta$
  - 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(*nqtl*|*y*,*m*) is marginal histogram