Efficient and Robust Statistical Methods for Quantitative Trait Loci Analysis

Brian S. Yandell

University of Wisconsin-Madison

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

with Fei Zou, University of North Carolina,

and Patrick J. Gaffney, Lubrizol

Jackson Laboratory

October 2001

Many Thanks

Michael Newton
Daniel Sorensen
Daniel Gianola
Jaya Satagopan
Patrick Gaffney
Fei Zou
Liang Li

Tom Osborn
David Butruille
Marcio Ferrera
Josh Udahl
Pablo Quijada
Alan Attie
Jonathan Stoehr

USDA Hatch Grants

Goals

away from normality

- fewer assumptions
- extended phenotypes
- check robustness
- multiple crosses

how many QTL?

- inferring the number
- sampling all QT loci
- estimating heritability

Interval Mapping Basics

- known measurements
  - phenotypic trait Y
  - markers X (and linkage map)
- unknown quantities
  - QT locus (or loci) λ
  - QT genotypes Q
- known segregation model $P(Q|X,\lambda)$
  - based on recombination, map function
- unknown aspects of phenotype model $P(Y|Q)$
  - distribution shape (could be assumed normal)
  - parameters θ, σ^2, if used (could be non-parametric)

Interval Mapping Mixture (BC)

- what shape histogram does trait *Y* have?
  - shape $P(Y|\mathrm{Qq})$ with genotype Qq
  - shape $P(Y|\mathrm{QQ})$ with genotype QQ
- is the QTL at a given locus λ?
  - no QTL: $P(Y|\mathrm{Qq}) = P(Y|\mathrm{QQ})$
  - yes QTL: mixture if genotype unknown
- mixture of shapes across possible genotypes

$$P(Y|X,\lambda) = P(\mathrm{Qq}|X,\lambda)P(Y|\mathrm{QQ}) + P(\mathrm{QQ}|X,\lambda)P(Y|\mathrm{QQ})$$

$P(Y|X,\lambda)$ = sum over possible Q of $P(Q|X,\lambda)P(Y|Q)$

Interval Mapping Likelihood

- likelihood: basis for scanning the genome

  $L(\lambda|Y)$ = product of $P(Y_i|X_i,\lambda)$ over $i = 1,\ldots,n$

  $L(\lambda|Y)$ = product_i of sum_Q of $P(Q|X_i,\lambda)P(Y_i|Q)$

- problem: unknown phenotype model $P(Y|Q)$
  - parametric $P(Y|Q) = Normal(Q\theta, \sigma^2)$
  - semi-parametric $P(Y|Q) = f(Y)\exp(Q\beta)$
  - non-parametric $P(Y|Q) = F_Q(Y)$ unspecified

Limitations of Parametric Models

- measurements not normal
  - counts (*e.g.* number of tumors)
  - survival time (*e.g.* days to flowering)
- false positives due to miss-specified model
  - check model assumptions?
- want more robust estimates of effects
  - parametric: only center (mean), spread (SD)
  - shape of distribution may be important

Semi-parametric QTL

- phenotype model $P(Y|Q) = f(Y)\exp(Q\beta)$
- test for QTL at locus λ
  - $\beta = 0$ implies $P(Y|\mathrm{QQ}) = P(Y|\mathrm{Qq})$
- includes many standard phenotype models

  normal $P(Y|Q) = N(\mu_Q,\sigma^2)$

  Poisson $P(Y|Q) = \mathrm{Poisson}(\mu_Q)$

  exponential, binomial, …

  (exercise: verify these are special cases for BC)

Semi-parametric Empirical Likelihood

- phenotype model $P(Y|Q) = f(Y)\exp(Q\beta)$
- non-parametric empirical likelihood (Owen 1988)

$$L(\lambda,\beta,f|Y,X) = \text{product}_i\ [\text{sum}_Q\ P(Q|X_i,\lambda)f(Y_i)\exp(Q\beta)]$$
$$= \text{product}_i\ f(Y_i)\ [\text{sum}_Q\ P(Q|X_i,\lambda)\exp(Q\beta)]$$
$$= \text{product}_i\ f(Y_i)\ w(X_i|\lambda,\beta)$$

  - weights $w(X_i|\lambda,\beta)$ rely only on flanking markers
- "point mass" at each measured phenotype
  - subject to constraints to be a distribution

  $\text{sum}_i\ P(Y_i|Q) = 1$ for all possible genotypes Q
- profile likelihood: $L(\lambda|Y,X) = \max_{\beta,f} L(\lambda,\beta,f|Y,X)$

Semi-parametric Formal Tests

- clever trick: use partial empirical LOD
  - Zou, Fine, Yandell (2001 *Biometrika*)
  - $\mathrm{LOD}(\lambda) \approx \log_{10} L(\lambda|Y,X)$
- has same formal behavior as parametric LOD
  - single locus test: approximately χ^2 with 1 d.f.
  - genome-wide scan: can use same critical values
  - permutation test: possible with some work
- can estimate cumulative distributions
  - nice properties (converge to Gaussian processes)

Non-parametric Methods

- phenotype model $P(Y|Q) = F_Q(Y)$
- Kruglyak, Lander (1995)
  - formal rank-sum test, replacing Y by rank(Y)
  - claimed no estimator of QTL effects
- estimators are indeed possible
  - semi-parametric shift (Hodges-Lehmann)
    - Zou (2001) thesis
  - non-parametric cumulative distribution
    - Fine, Zou, Yandell (2001 in review)

Rank-Sum QTL Methods

- phenotype model $P(Y|Q) = F_Q(Y)$
- replace Y by rank(Y) and perform IM
  - extension of Wilcoxon rank-sum test
  - fully non-parametric
- Hodges-Lehmann estimator of shift β
  - most efficient if $P(Y|Q) = F(Y+Q\beta)$
  - find β that matches medians
    - problem: genotypes Q unknown
    - resolution: Haley-Knott (1992) regression scan

Non-Parametric QTL CDFs

- estimate non-parametric phenotype model
  - cumulative distributions $F_Q(y) = P(Y \leq y \,|Q)$
  - can use to check parametric model validity
- basic idea:

  $P(Y \leq y \,|X,\lambda) = \text{sum}_Q \, P(Q|X,\lambda)F_Q(y)$
  - depends on X only through flanking markers
  - few possible flanking marker genotypes
    - 4 for BC, 9 for F2, etc.

Finding NP QTL CDFs

- cumulative distributions $F_Q(y) = P(Y \leq y \,|Q)$
- $F = \{F_Q$, all possible QT genotypes $Q\}$
  - BC: $F = \{F_{QQ}=P(Y \leq y|QQ), F_{Qq}=P(Y \leq y|Qq)\}$
- find F to minimize over all phenotypes y

  $\text{sum}_i\, [I(Y_i \leq y) - \text{sum}_Q\, P(Q|X,\lambda)F_Q(y)]^2$
- looks complicated, but simple to implement

Non-parametric CDF Properties

- readily extended to censored data
  - time to flowering for non-vernalized plants
- nice large sample properties
  - estimates of $F(y) = \{F_Q(y)\}$ jointly normal
  - point-wise, experiment-wise confidence bands
- more robust to heavy tails and outliers
- can use to assess parametric assumptions

Combining Multiple Crosses

- combining inbred lines in search of QTL
  - most IM methods limited to single cross
  - animal model largely focuses on polygenes
  - individuals no longer independent given *Q*
- recent work in plant sciences
  - Bernardo (1994) Wright's relationship matrix *A*
  - Rebai *et al.* (1994) regression method
  - Xu Atchley (1995) IBD & *A* for QTLs & polygenes
  - Liu Zeng (2000) multiple inbred lines, fixed effect IM
  - Zou, Yandell, Fine (2001 *Genetics*) power, threshold

Thresholds for Multiple Crosses

- permutation test
  - Churchill Doerge (1994); Doerge Churchill (1996)
  - computationally intensive
  - difficult to compare different designs
- theoretical approximation
  - Lander Botstein (1989) Dupuis Siegmund (1999)
    - single cross, dense linkage map
  - Rebai *et al.* (1994, 1995) approximate extension
    - Piepho (2001) improved calculation of efficiency
  - Zou, Yandell, Fine (2001) extend original theory

Extension of Threshold Theory

- likelihood for multiple crosses of inbred lines with *m* founders
  - approximately χ^2 with *m* degrees of freedom
  - genome-wide threshold theory
    - extends naturally based on Ornstein-Uhlenbeck
  - threshold based on dense or sparse linkage map
- some calculations based on BC1, F2, BC2
  - Liu Zeng (2000) ECM method to estimate

$Y_{ij} \sim Normal(Q_{ij}\theta_j+\mu_j, \sigma_j^2)$, j = cross

Bayesian Interval Mapping for Inbred Lines

- return to single inbred cross
  - parametric phenotype model (normal)
- connection between likelihood and posterior
  - maximize *L*, sample from posterior
- how many QTL?
  - model selection: number of QTL as unknown
- learning from data

Bayesian Interval Mapping

- recall likelihood for inbred lines

  $L(\lambda|Y) = \text{product}_i\ [\text{sum}_Q\ P(Q|X_i,\lambda)P(Y_i|Q,\theta)]$

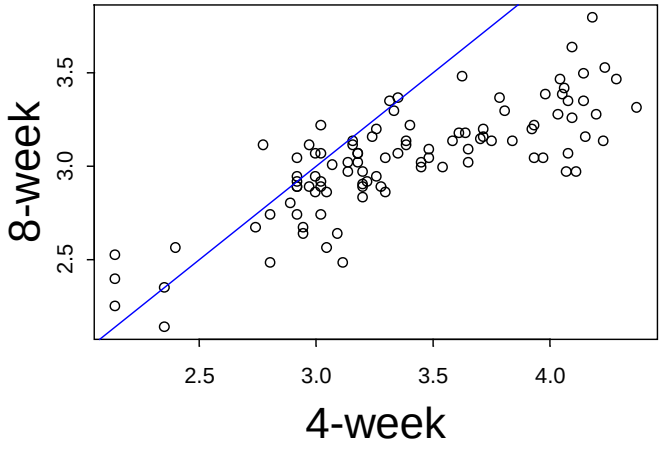
- Bayesian posterior idea
  - sample unknown data instead

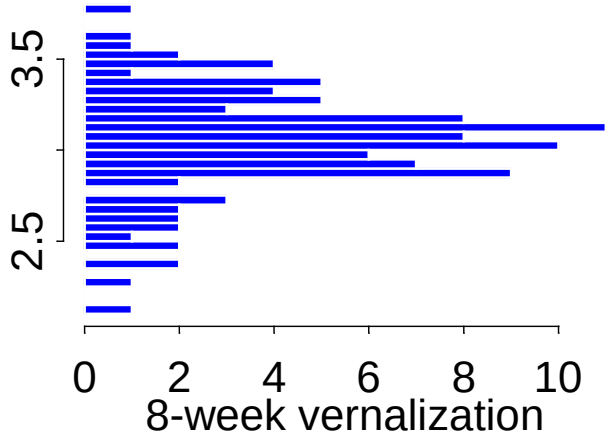
    $P(\lambda,Q,\theta|Y,X) = [\text{product}_i\ P(Q_i|X_i,\lambda)\ P(Y_i|Q_i,\theta)]P(\lambda,\theta|X)$
  - marginal summaries provide key information
    - loci: $P(\lambda|Y,X) = \text{sum}_{Q,\theta}\ P(\lambda,Q,\theta|Y,X)$
    - effects: $P(\theta|Y,X) = \text{sum}_{Q,\lambda}\ P(\lambda,Q,\theta|Y,X)$
  - Satagopan *et al.* (1996 *Genetics*)

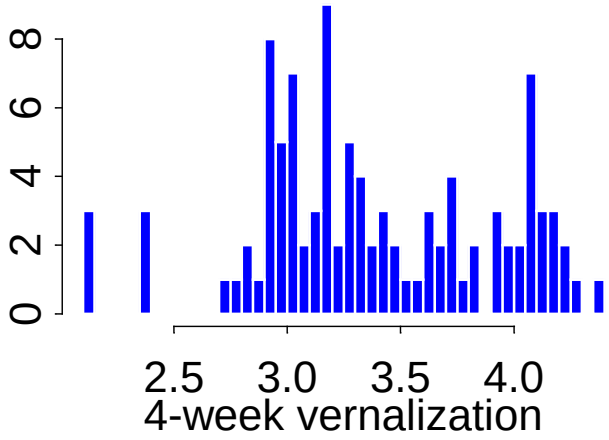
Brassica napus Data

- 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*

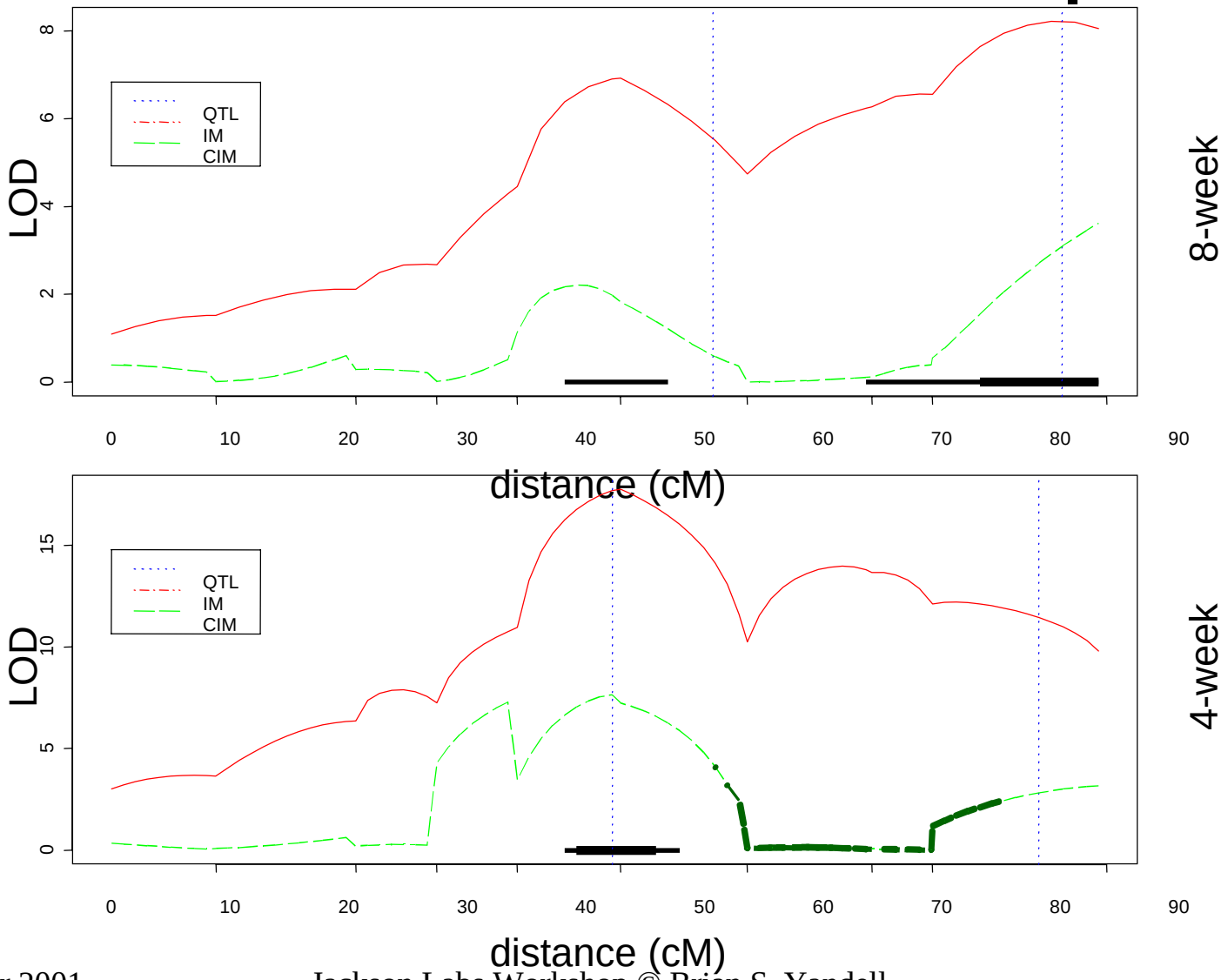
Brassica 4- & 8-week Data





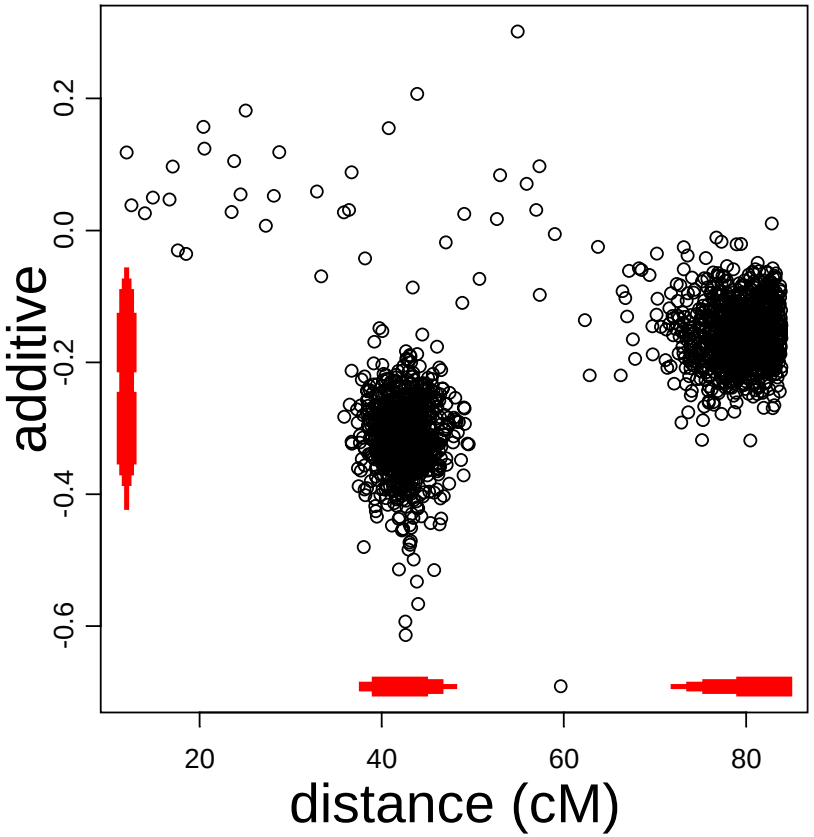


Brassica Data LOD Maps

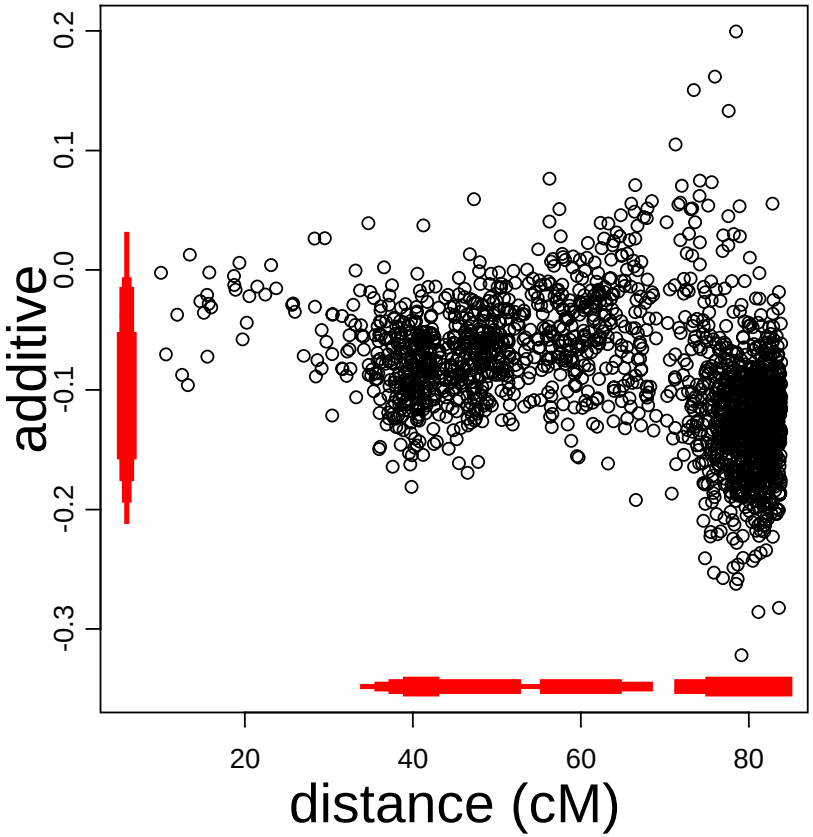


Brassica Sampled Summaries

4-week



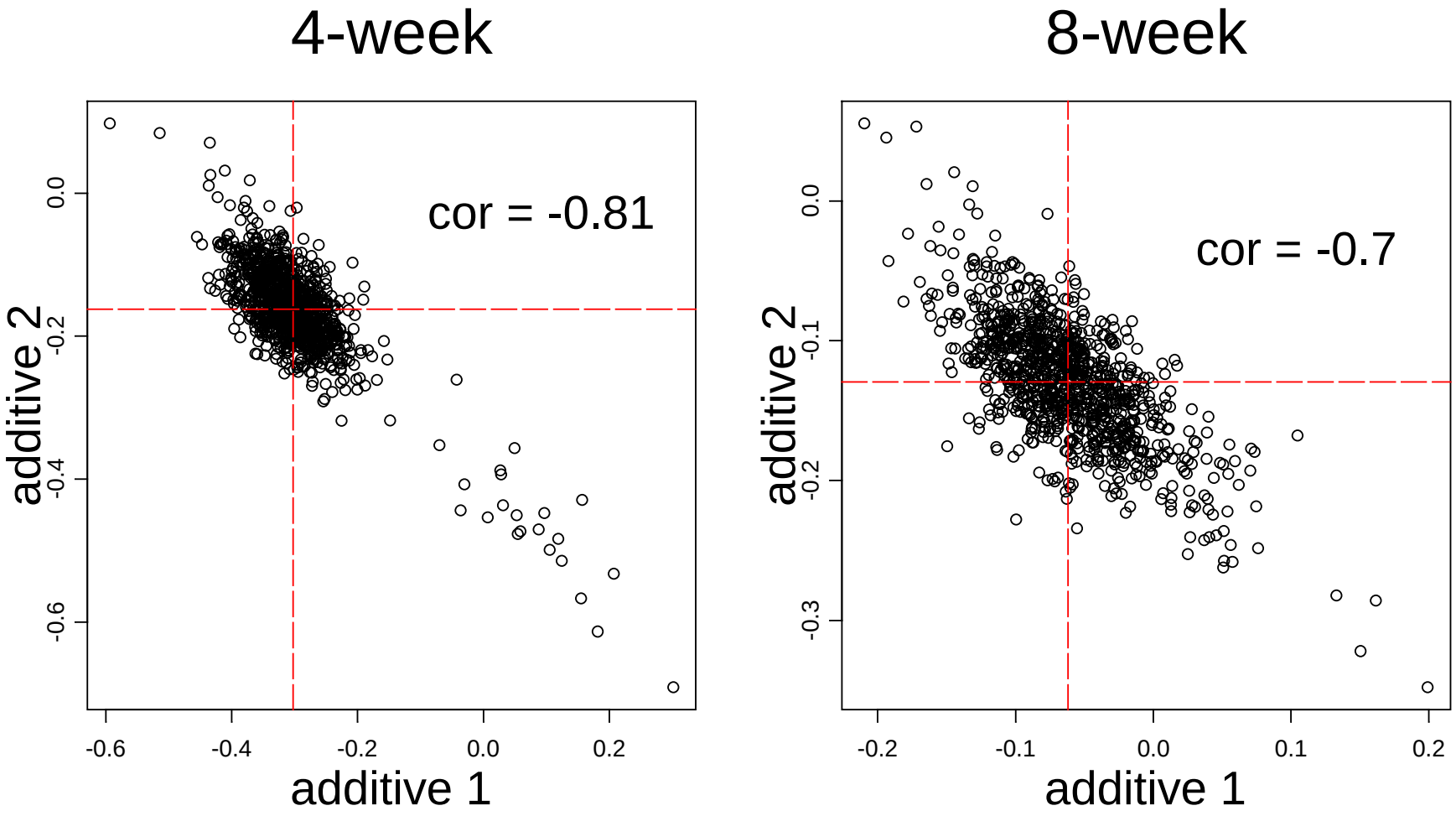
8-week



Collinearity of QTLs

- multiple QT genotypes are correlated
  - QTL linked on same chromosome
  - difficult to distinguish if close
- estimates of QT effects are correlated
  - poor identifiability of effects parameters
  - correlations give clue of how much to trust
- which QTL to go after in breeding?
  - largest effect?
  - may be biased by nearby QTL

Brassica effect Correlations



How many (detectable) QTL?

- build m = number of QTL into model

  $P(\lambda,Q,\theta|Y,X,m) = P(Q|X,\lambda,m)\ P(Y|Q,\theta,m)P(\lambda,\theta|X,m)$

  – prior on number of QTL
    - Poisson or exponential seem to work best
    - uniform can push posterior to more complicated model
- model selection
  – Bayes factors (Jaya Satagopan talk)
  – sample m as part of a bigger model
- many, many QTL affect most any trait
  – how many detectable with these data?
  – limits to useful detection (Bernardo 2000)

Sampling the Number of QTL

- almost analogous to stepwise regression

  $P(Y_i|Q_i,\theta)$: $\quad Y_i = \mu + Q_{i1}\alpha_1 + \ldots + Q_{im}\alpha_m + e_i$

  – but regressors (QT genotypes) are unknown
  – linked loci = collinear regressors = correlated effects
- use reversible jump MCMC to change *m*
  – bookkeeping helps in comparing models
  – adjust to change of variables between models
  – Green (1995); Richardson Green (1997)
  – other approaches out there these days…

Model Selection in Regression

- consider known regressors (X = markers)
  - models with 1 or 2 regressors
- jump between models
  - centering regressors simplifies calculations

$$m = 1 : Y_i = \mu + \alpha(X_{i1} - \overline{X}_1) + e_i$$

$$m = 2 : Y_i = \mu + \alpha_1(X_{i1} - \overline{X}_1) + \alpha_2(X_{i1} - \overline{X}_1) + e_i$$

Regressor Slope Estimators

recall least squares estimators of slopes

note relation of α_1 in model 2 to α in model 1

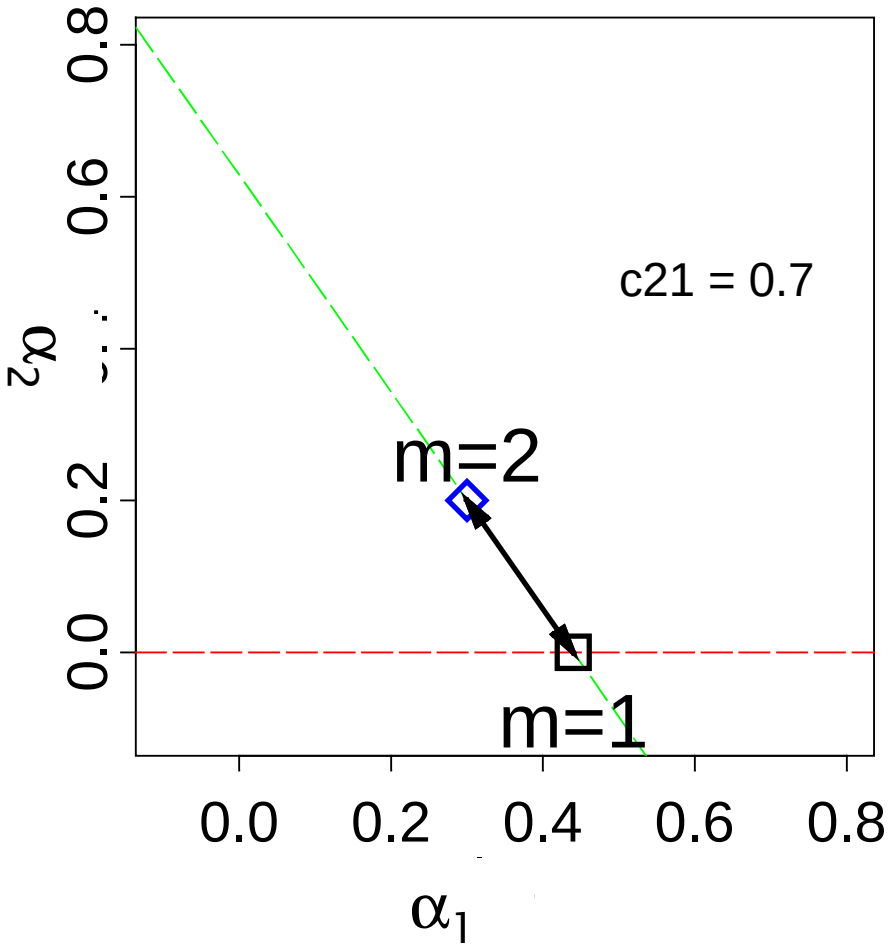
$$m=1: \quad \hat{\alpha}=\frac{c_{1Y}\, s_Y}{s_1}, \quad c_{1Y}=\mathrm{corr}(X_1,Y), s_Y=\mathrm{SD}(Y),...$$

$$m=2: \quad \hat{\alpha}_1=\frac{(c_{1Y}-c_{12}c_{2Y})s_Y}{s_1}=\hat{\alpha}-\frac{c_{12}c_{2Y}s_Y}{s_1}$$

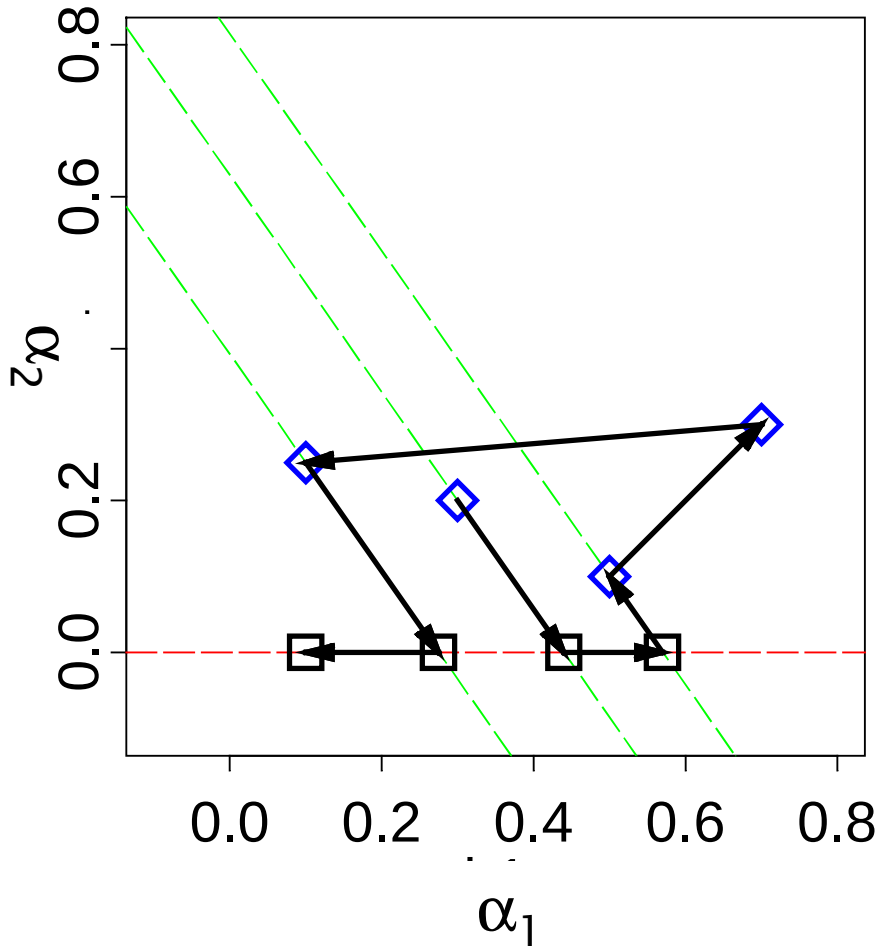
$$m=2: \quad \hat{\alpha}_2=\frac{(c_{2Y}-c_{12}c_{1Y})s_Y}{s_2}$$

Geometry of Reversible Jump

Move Between Models

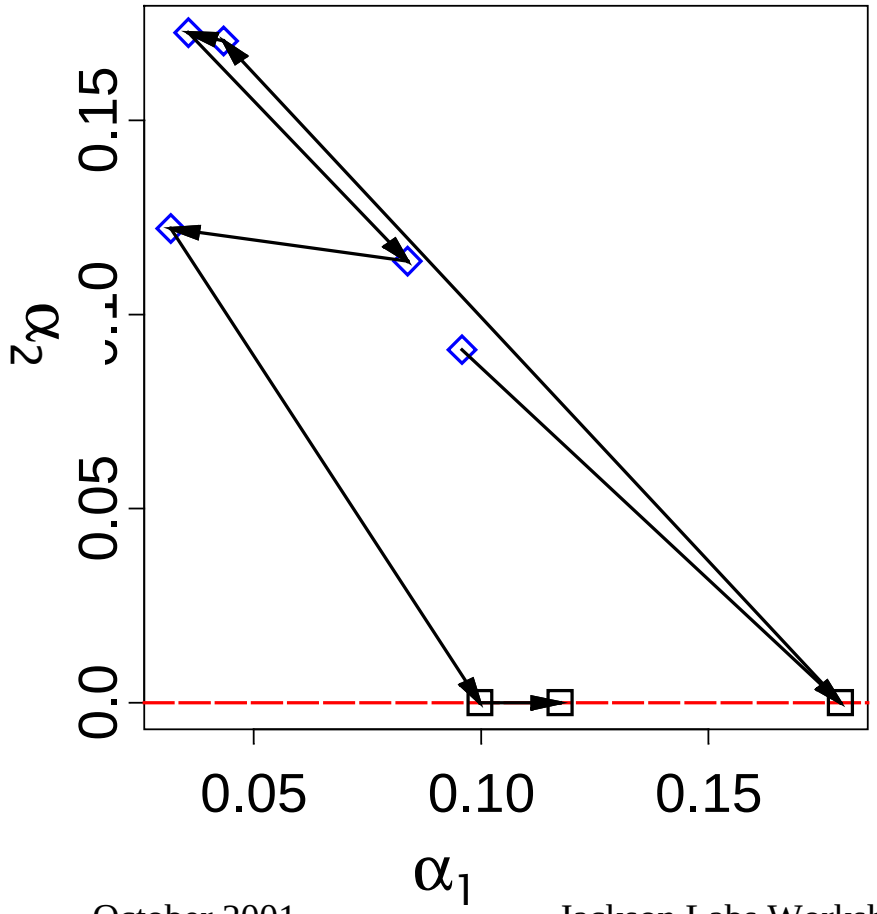


Reversible Jump Sequence

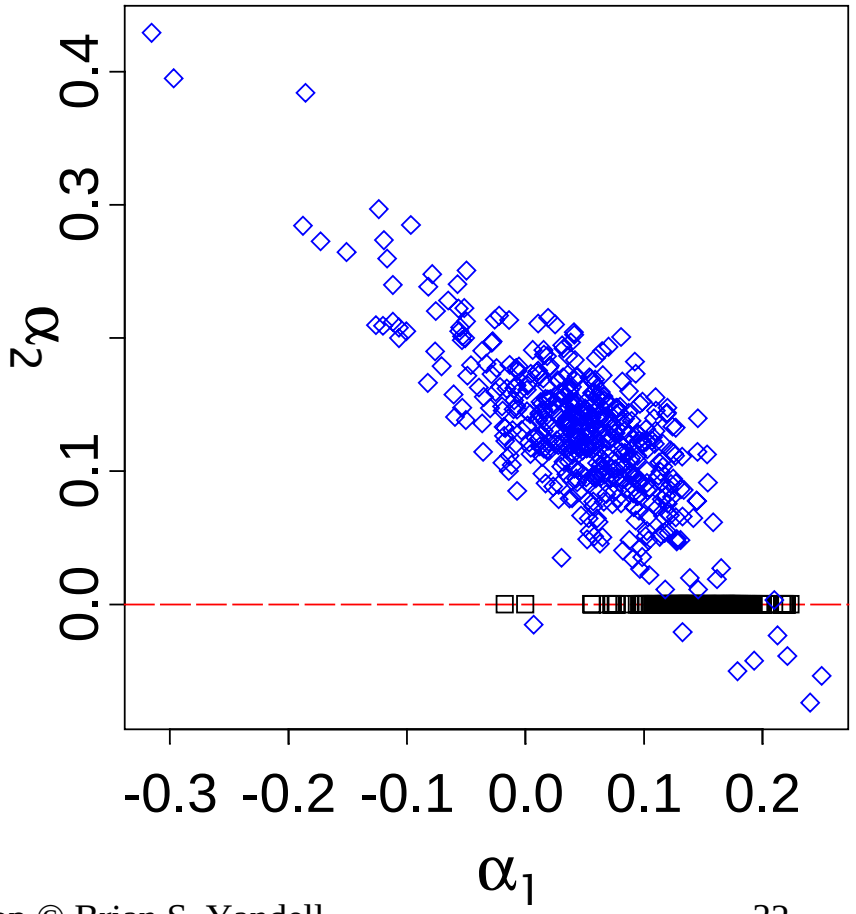


QT additive Reversible Jump

a short sequence



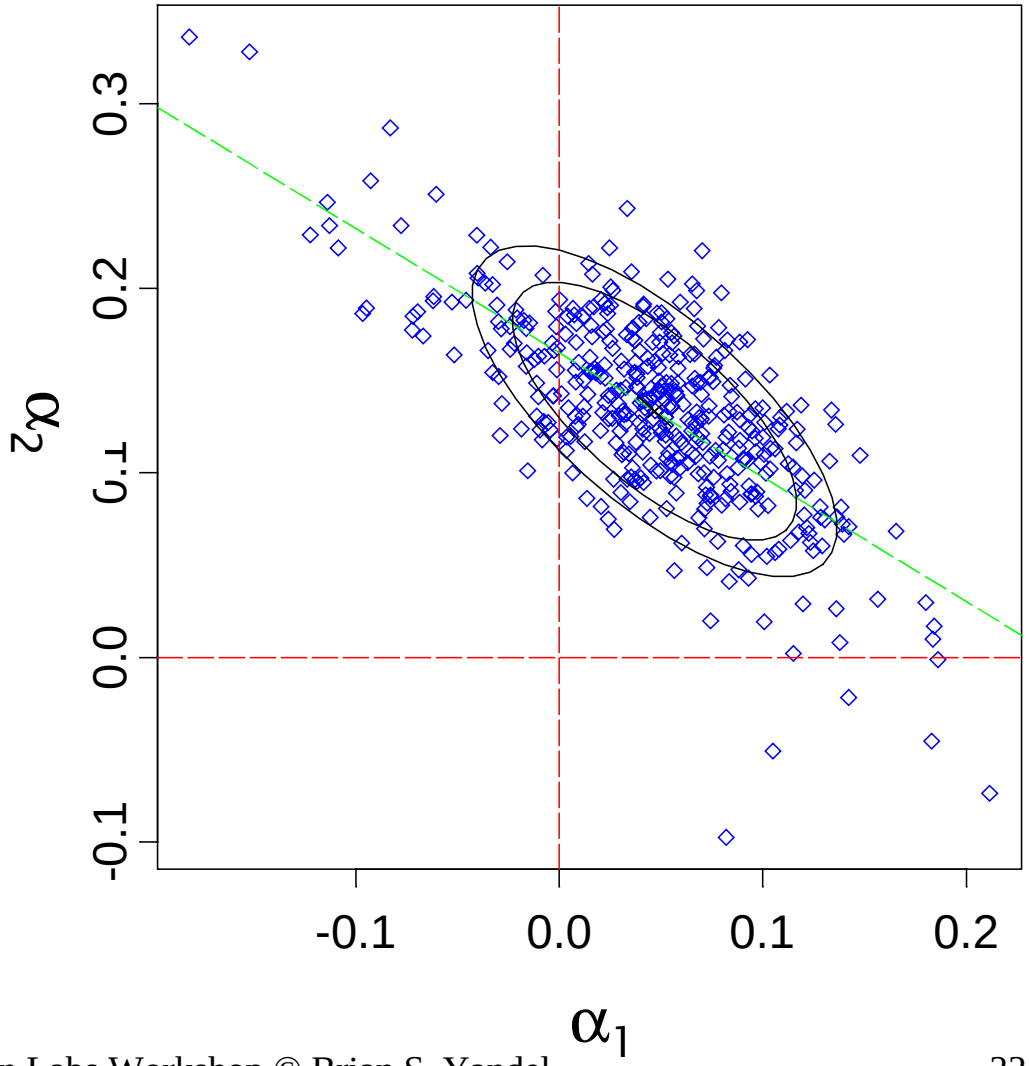
first 1000 with m<3



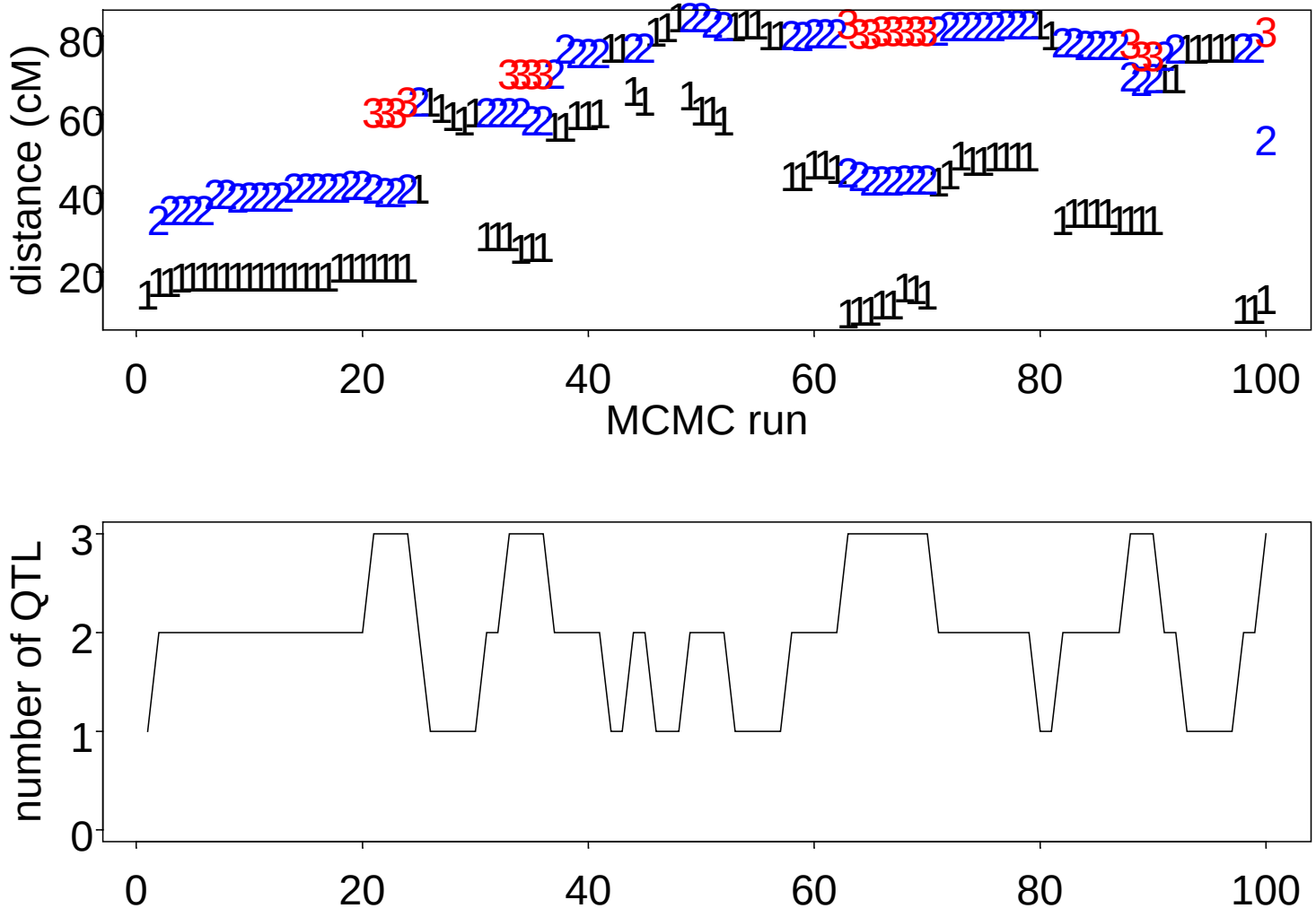
Credible Set for additive

90% & 95% sets based on normal

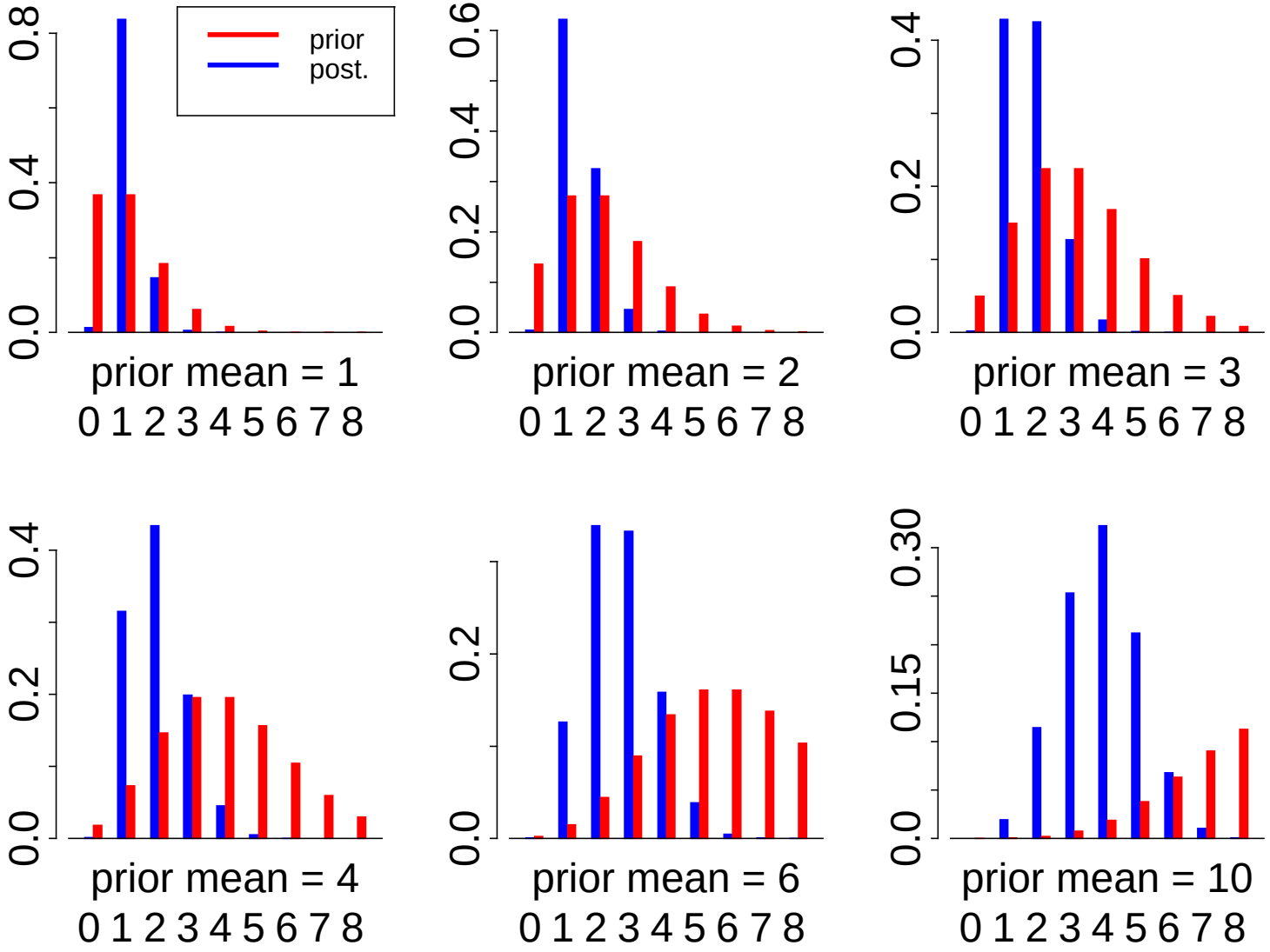
regression line corresponds to slope of updates



Jumping QTL number & loci



#QTL for *Brassica* 8-week



Bayes Factor Sensitivity

- Bayes factors computed from RJ-MCMC
  - posterior $P(m|Y,X)$ affected by prior $P(m)$
  - *BF* insensitive to prior

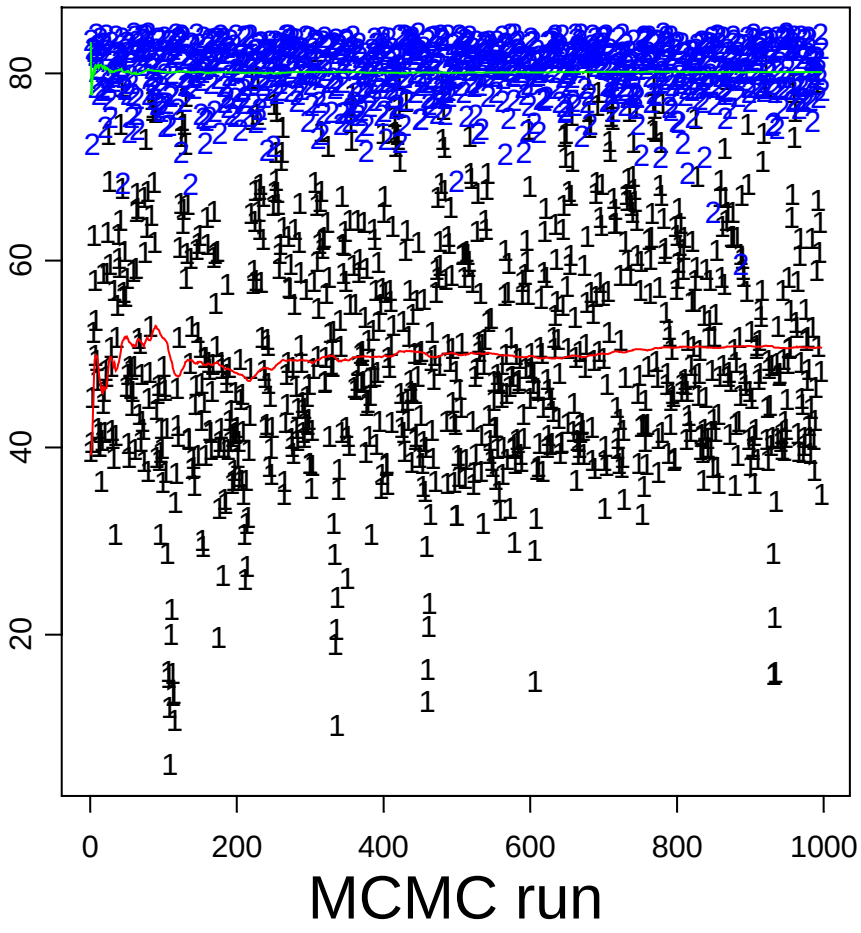
$$BF_{m,m+1} = \frac{P(m|Y,X)/P(m)}{P(m+1|Y,X)/P(m+1)}$$

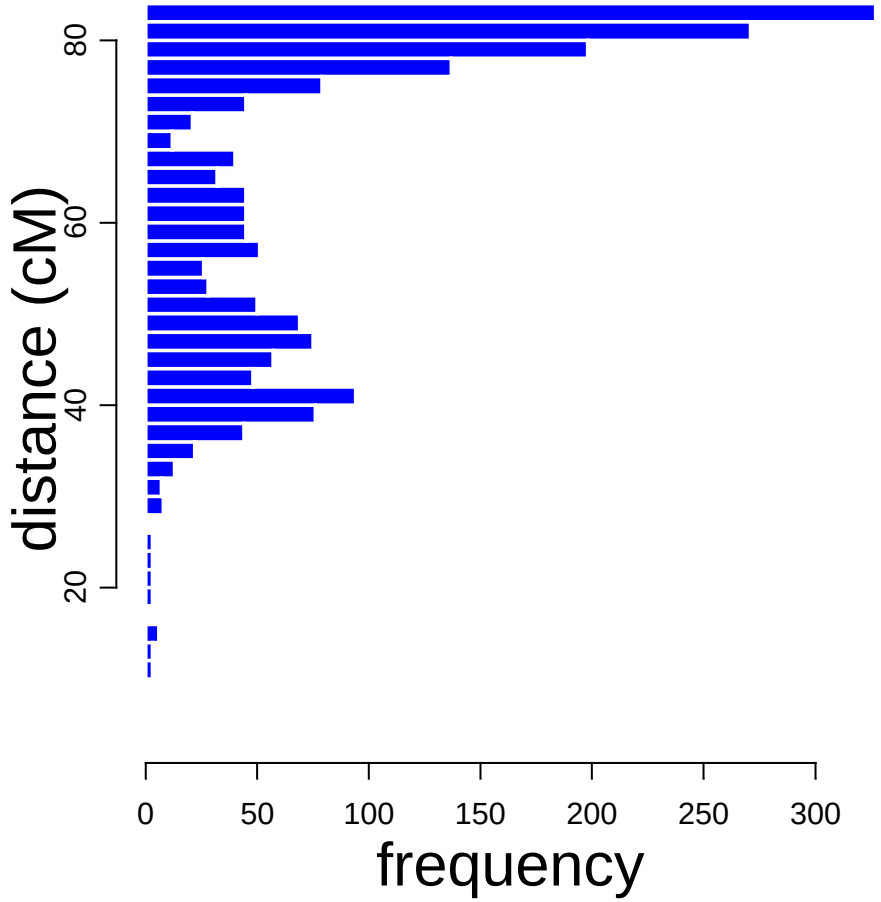
    - exponential, Poisson, uniform
  - *BF* sensitivity to prior variance on effects θ
    - prior variance should reflect data variability
    - resolved by using hyperpriors
      - automatic algorithm; no need for tuning by user

How To Infer Loci?

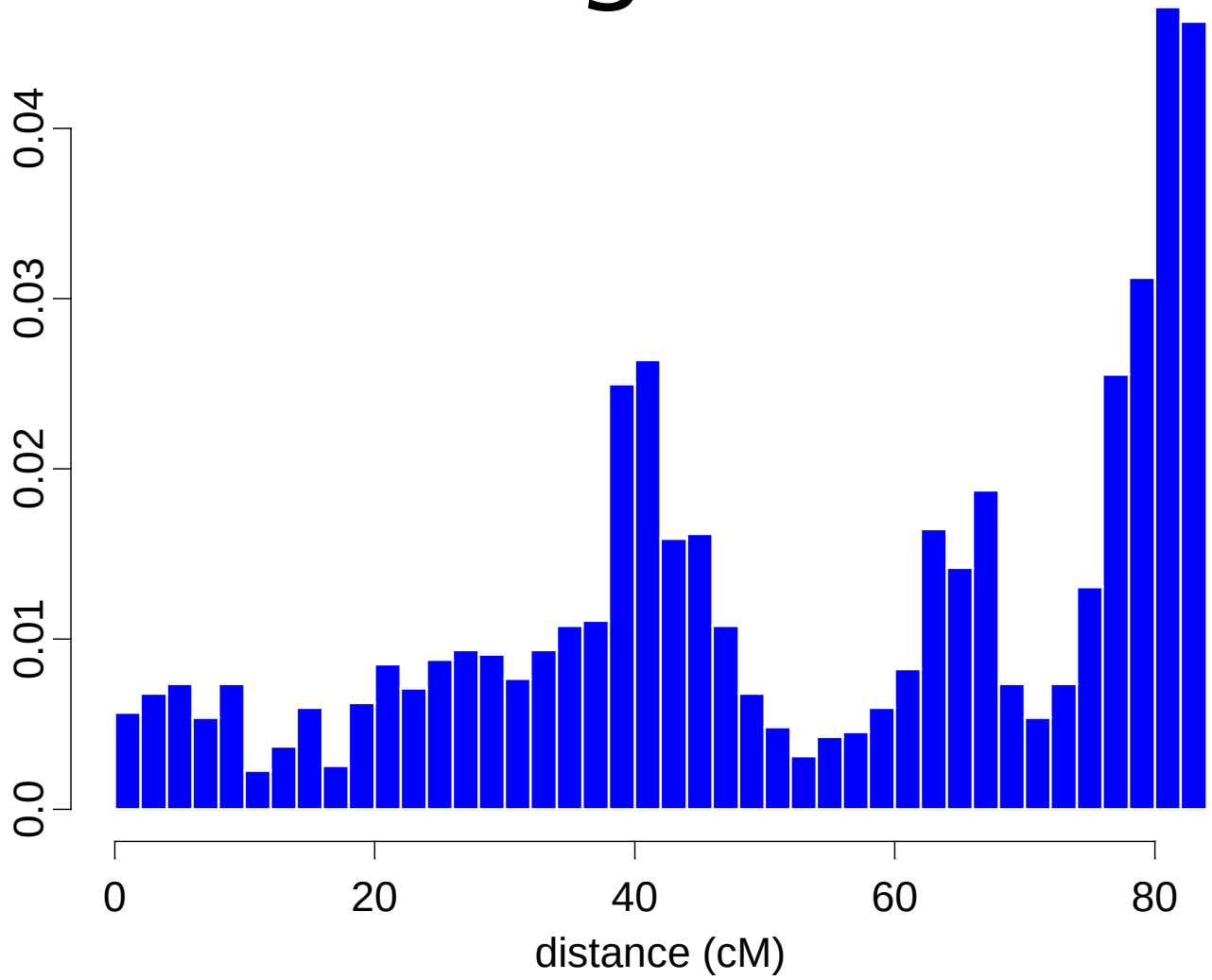
- if *m* is known, use fixed MCMC
  - histogram of loci
  - issue of bump hunting
- combining loci estimates in RJ-MCMC
  - some steps are from wrong model
    - too few loci (bias)
    - too many loci (variance/identifiability)
  - condition on number of loci
    - subsets of Markov chain

Brassica 8-week Data `locus` MCMC with $m=2$

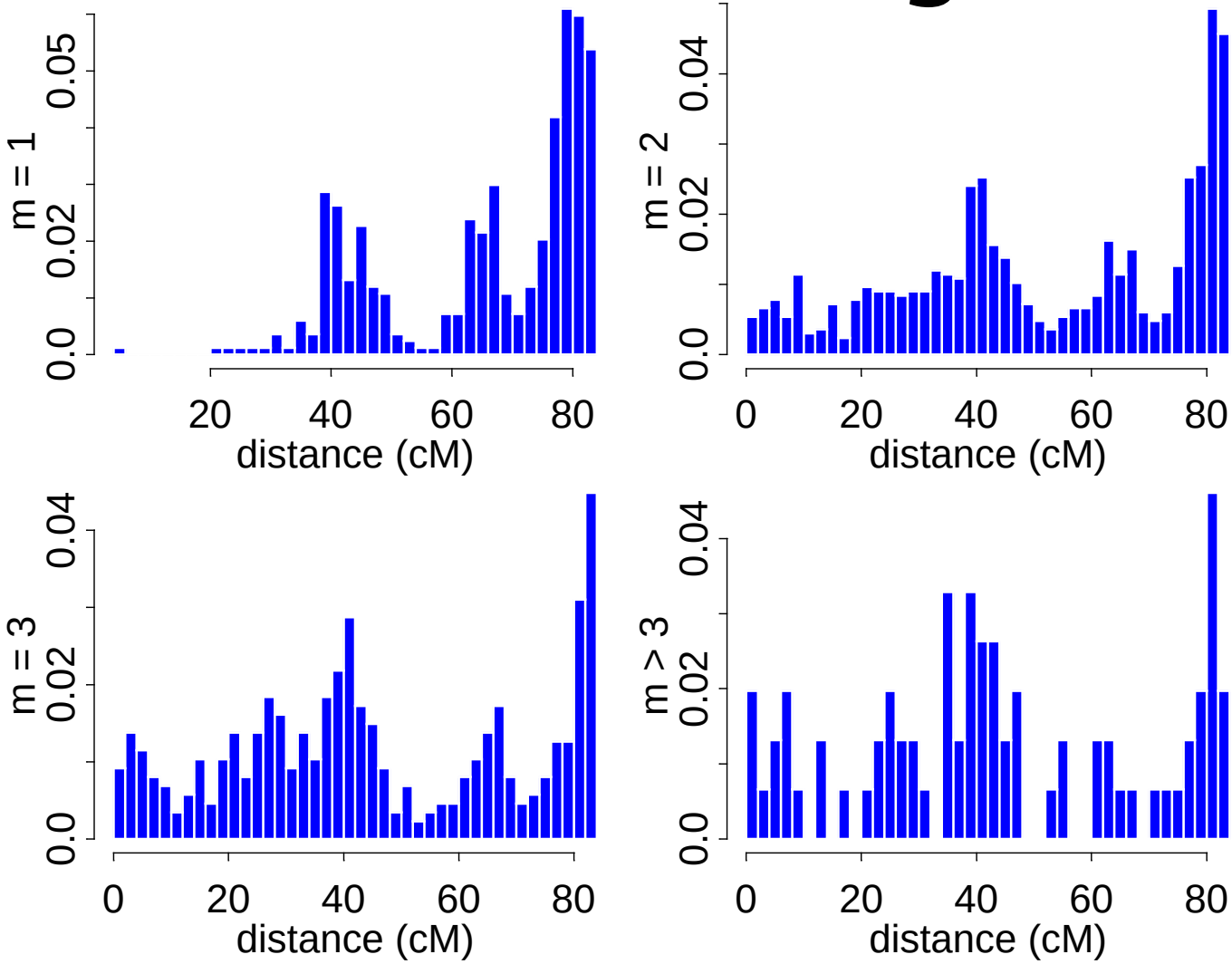




Raw Histogram of loci



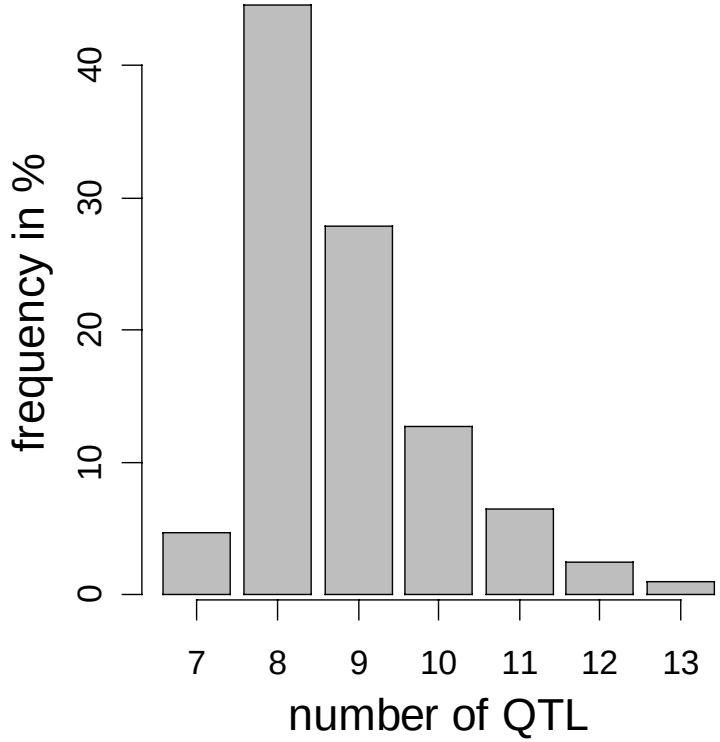
Conditional Histograms



A Complicated Example

- simulated 200 individuals (Stephens, Fisch 1998)
- 8 QTL, heritability = 50%; detected 3 QTL
- increase heritability to 97% to detect all 8

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**



Best Loci Model by Chromosome

- 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

Reanalysis of other *Brassica* data

- *Brassica napus*
  - 19 chromosomes, 480 markers
  - infer 4 QTL (2 linked, 2 unlinked)
- *Brassica rapa*
  - 9 chromosomes
  - infer 4 QTL with added *FLC* marker
    - 2 tightly linked in repulsion
- Ferreira *et al.* (1994), Kole *et al.* (1997, 2001)

Computational Issues

- more complicated when $m > 2$
  - avoid matrix inverses: Cholesky decomposition
  - multivariate updates: all effects, all loci at once
- improvements in sampling efficiency
  - pre-burnin to overshoot *m*, burnin to wash out
  - occasional long distance loci update
- bump hunting to sort out loci
- Gaffney (2001 PhD thesis)

Bayesian IM Software

- General MCMC software
  - U Bristol links
    - http://www.stats.bris.ac.uk/MCMC/pages/links.html
  - BUGS (Bayesian inference Using Gibbs Sampling)
    - http://www.mrc-bsu.cam.ac.uk/bugs/
- Our MCMC software for QTLs
  - our C code using LAPACK
    - ftp://ftp.stat.wisc.edu/pub/yandell/revjump.tar.gz
  - our QTL Cart module
    - Bmapqtl 3rd party module (Windows available)
    - R post processing

QTL References

- Bernardo R (2000) What if we knew all the genes for a quantitative trait in hybrid crops? *Crop Sci.* (submitted).
- Fine JP, Zou F, Yandell BS (2001) Nonparametric estimation of mixture distributions with known mixture proportions. (submitted).
- Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. *Genetics 139*: 1421-1428.
- Liu YF, Zeng ZB (2000) A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. *Genet Res 75*: 345-355.
- Zou F (2001) Efficient and robust statistical methodologies for quantitative trait loci analysis. *PhD thesis, UW-Madison Statistics.*
- Zou F, Fine JP, Yandell BS (2001) On empirical likelihood for a semiparametric mixture model. *Biometrika 00*: 000-000.
- Zou F, Yandell BS, Fine JP (2001) Threshold and power calculations for QTL analysis of combined lines. *Genetics 00*: 000-000.

QTL MCMC References

- Gaffney PJ (2001) An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. *PhD thesis, UW-Madison Statistics*.
- Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* 144: 805-816.
- Satagopan JM, Yandell BS (1996) Estimating the number of quantitative trait loci via Bayesian model determination. *Proc JSM Biometrics Section*.
- Sillanpaa MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data., *Genetics* 148: 1373-1388.
- Stephens DA, Fisch RD (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* 54: 1334-1347.

Reversible Jump MCMC References

- Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711-732.
- Kuo L, Mallick B (1998) Variable selection for regression models. *Sankhya, Series B, Indian J Statistics 60*: 65-81
- Mallick BK (1998) Bayesian curve estimation by polynomial of random order. *J Statistical Planning and Inference 70*: 91-109
- Richardson S, Green PJ (1997) On Bayesian analysis of mixture with an unknown of components. *J Royal Statist Soc B* 59: 731-792.