Statistical issues in QTL mapping in mice

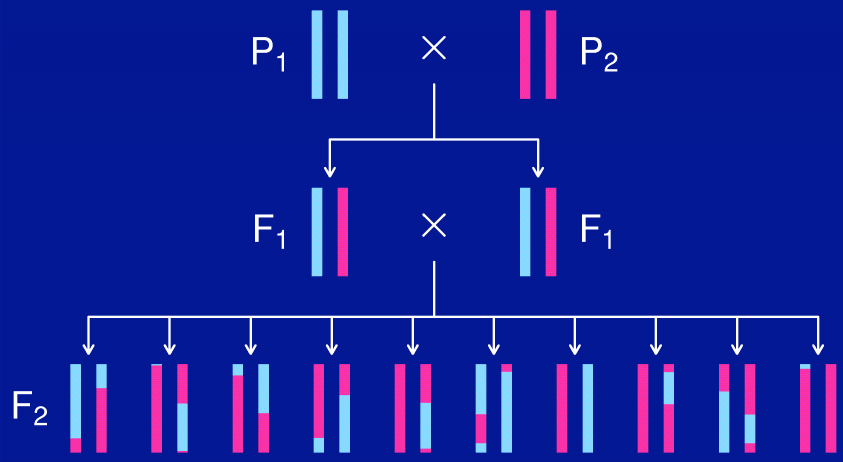
Karl W Broman
Department of Biostatistics
Johns Hopkins University

http://www.biostat.jhsph.edu/~kbroman

Outline

- Overview of QTL mapping
- The X chromosome
- Mapping multiple QTLs
- Recombinant inbred lines
- Heterogeneous stock and 8-way RILs

The intercross



3

The data

- Phenotypes, y_i
- Genotypes, x_{ij} = AA/AB/BB, at genetic markers
- A genetic map, giving the locations of the markers.

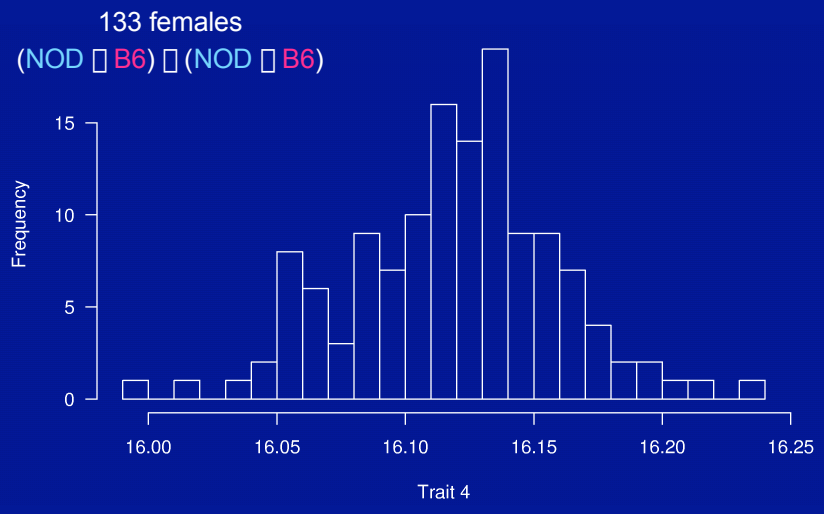
4

Goals

- Identify genomic regions (QTLs) that contribute to variation in the trait.
- Obtain interval estimates of the QTL locations.
- Estimate the effects of the QTLs.

5

Phenotypes



6

NOD

7

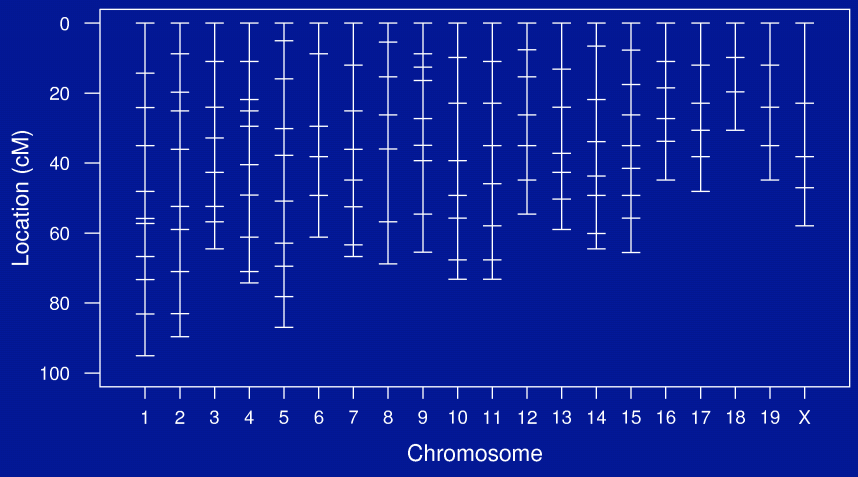
C57BL/6

8

Agouti coat

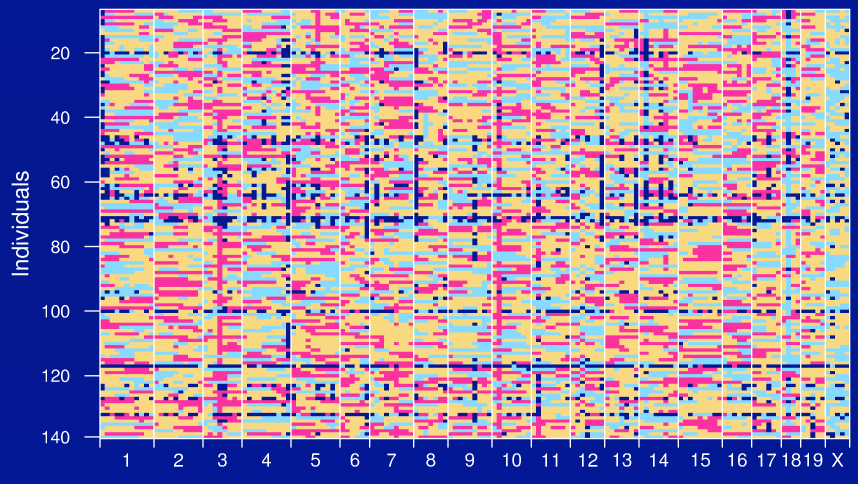
9

Genetic map



10

Genotype data



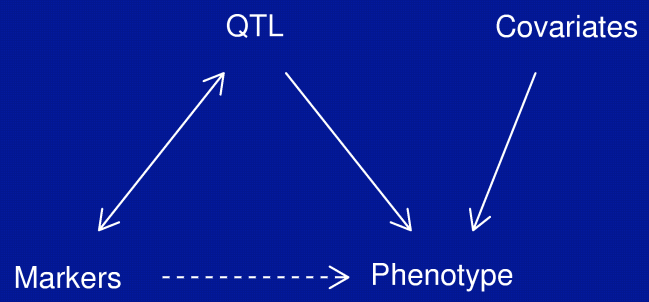
11

Goals

- Identify genomic regions (QTLs) that contribute to variation in the trait.
- Obtain interval estimates of the QTL locations.
- Estimate the effects of the QTLs.

12

Statistical structure



- Missing data: markers → QTL
- Model selection: genotypes → phenotype

13

Models: recombination

- No crossover interference
  - Locations of breakpoints according to a Poisson process.
  - Genotypes along chromosome follow a Markov chain.
- Clearly wrong, but super convenient.

14

Models: gen □ phe

Phenotype = y, whole-genome genotype = g

Imagine that p sites are all that matter.

$E(y \mid g) = \mu(g_1,\ldots,g_p)$ $\quad$ $SD(y \mid g) = \sigma(g_1,\ldots,g_p)$

Simplifying assumptions:

- $SD(y \mid g) = \sigma$, independent of g
- $y \mid g \sim$ normal($\mu(g_1,\ldots,g_p)$, σ)
- $\mu(g_1,\ldots,g_p) = \mu + \sum \alpha_j 1\{g_j = AB\} + \beta_j 1\{g_j = BB\}$

15

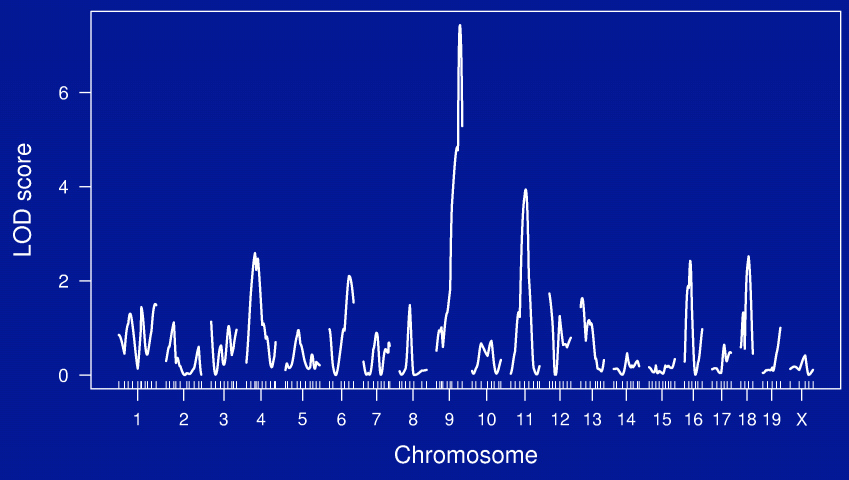
Interval mapping

Lander and Botstein 1989

- Imagine that there is a single QTL, at position z.
- Let q_i = genotype of mouse i at the QTL, and assume
  $y_i \mid q_i \sim$ normal($\mu(q_i)$, σ)
- We won't know q_i, but we can calculate
  $p_{ig} = Pr(q_i = g \mid$ marker data$)$
- y_i, given the marker data, follows a mixture of normal distributions with known mixing proportions (the p_{ig}).
- Use an EM algorithm to get MLEs of $\theta = (\mu_{AA}, \mu_{AB}, \mu_{BB}, \sigma)$.
- Measure the evidence for a QTL via the LOD score, which is the $\log_{10}$ likelihood ratio comparing the hypothesis of a single QTL at position z to the hypothesis of no QTL anywhere.

16

LOD curves



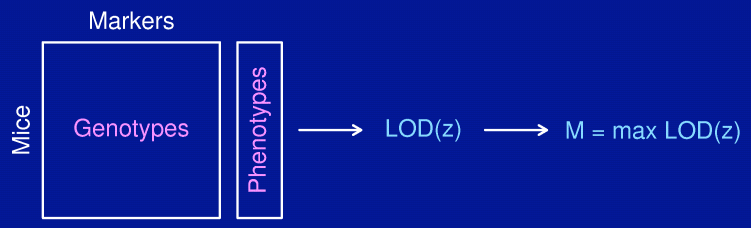
17

LOD thresholds

- To account for the genome-wide search, compare the observed LOD scores to the distribution of the maximum LOD score, genome-wide, that would be obtained if there were no QTL anywhere.
- The 95th percentile of this distribution is used as a significance threshold.
- Such a threshold may be estimated via permutations (Churchill and Doerge 1994).

18

Permutation test

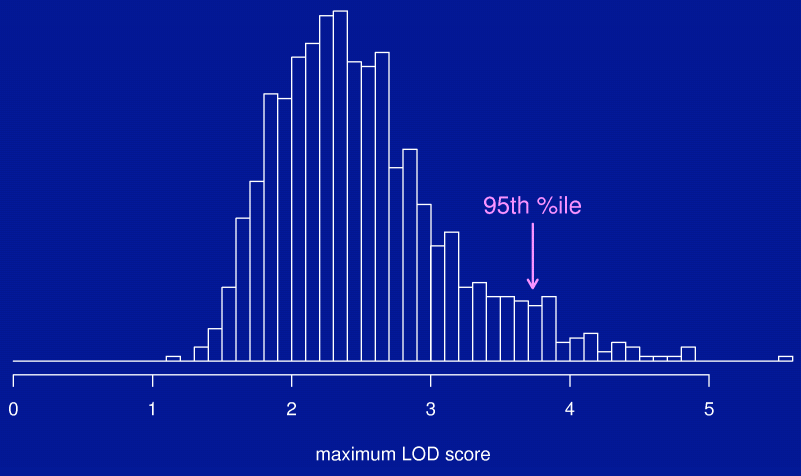


- Shuffle the phenotypes relative to the genotypes.
- Calculate M* = max LOD*, with the shuffled data.
- Repeat many times.

- LOD threshold = 95th percentile of M*.
- P-value = Pr(M* ≥ M)

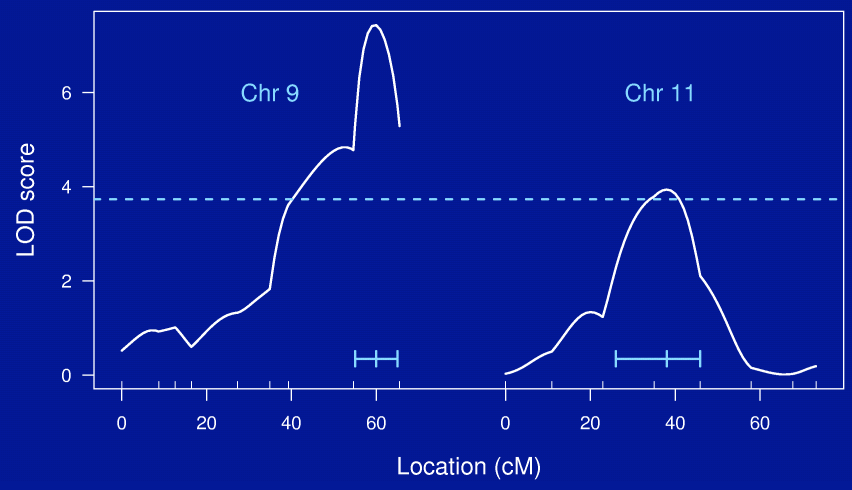
19

Permutation distribution



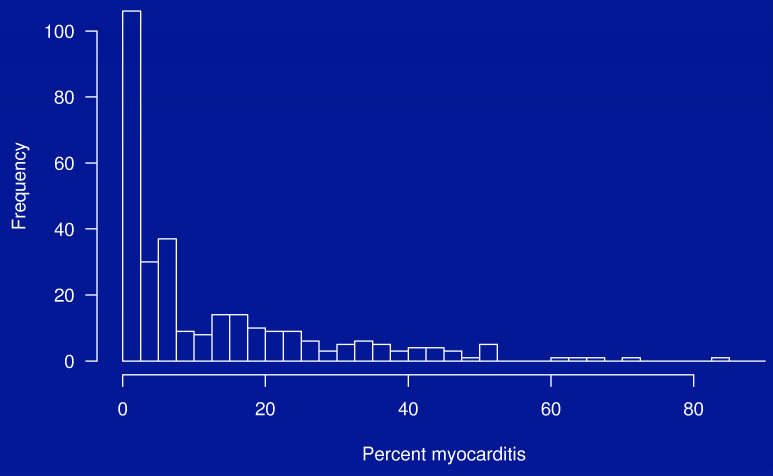
20

Chr 9 and 11



21

Non-normal traits



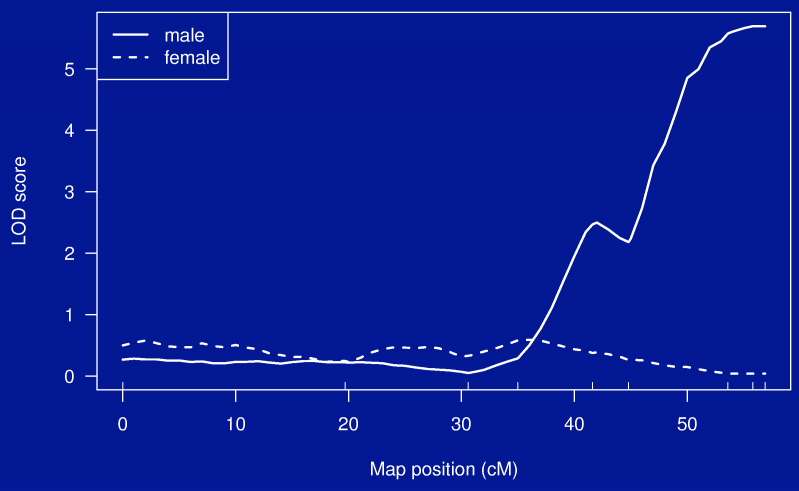
22

Non-normal traits

- Standard interval mapping assumes that the residual variation is normally distributed (and so the phenotype distribution follows a mixture of normal distributions).
- In reality: we see binary traits, counts, skewed distributions, outliers, and all sorts of odd things.
- Interval mapping, with LOD thresholds derived via permutation tests, often performs fine anyway.
- Alternatives to consider:
  - Nonparametric linkage analysis (Kruglyak and Lander 1995).
  - Transformations (e.g., log or square root).
  - Specially-tailored models (e.g., a generalized linear model, the Cox proportional hazards model, the model of Broman 2003).

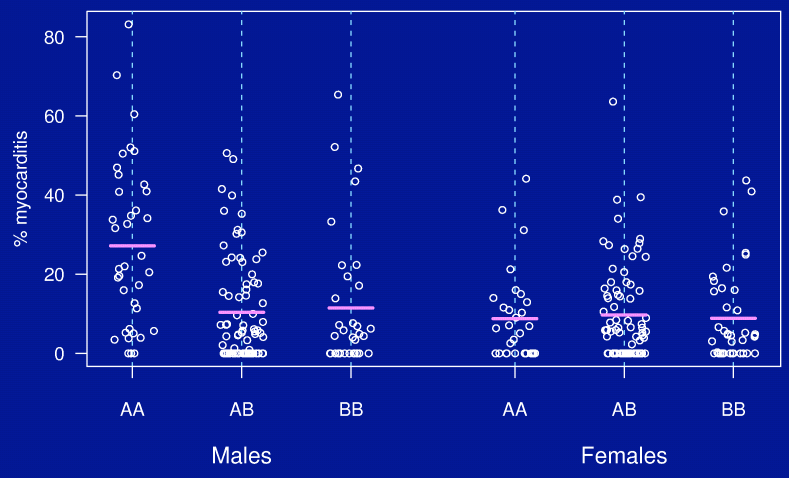
23

Split by sex



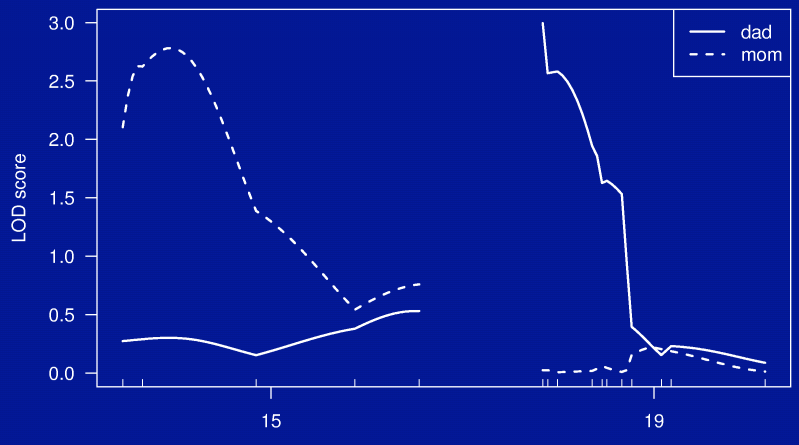
24

Split by sex



25

Split by parent-of-origin



26

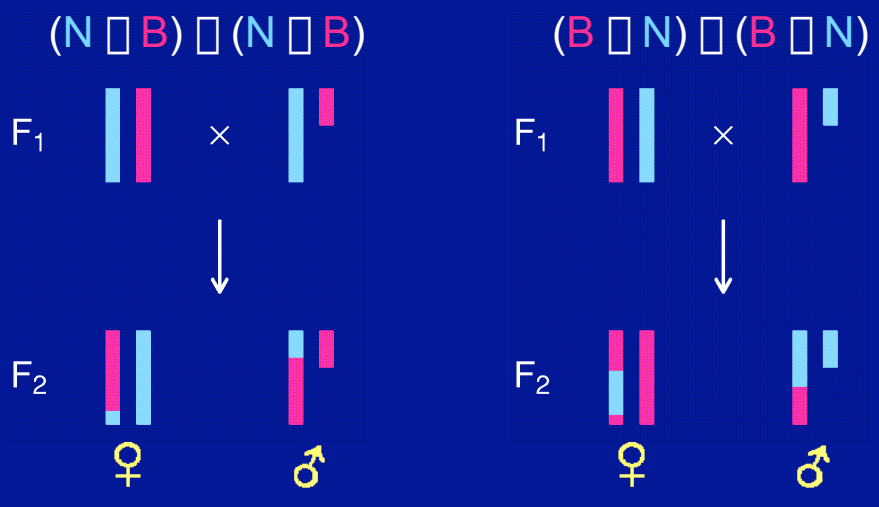
Split by parent-of-origin

Percent of individuals with phenotype

	Genotype at D15Mit252		Genotype at D19Mit59	
P-O-O	AA	AB	AA	AB
Dad	63%	54%	75%	43%
Mom	57%	23%	38%	40%

27

The X chromosome



28

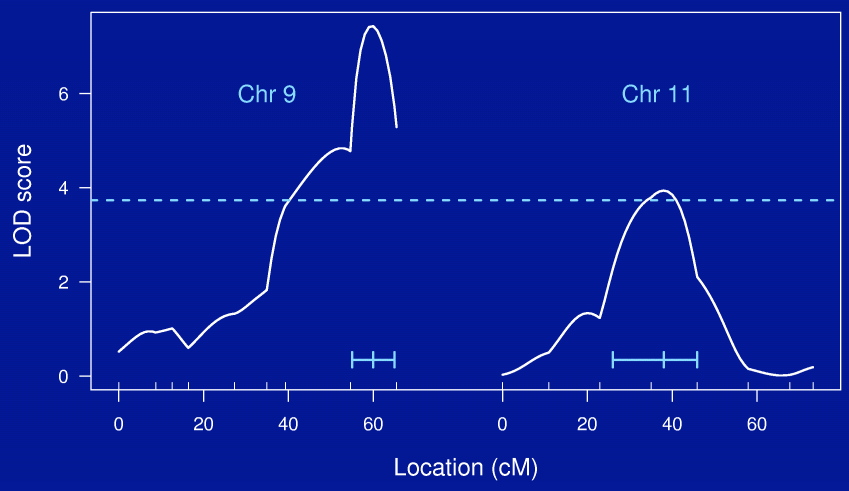
The X chromosome

- BB ≡ BY? NN ≡ NY?
- Different "degrees of freedom"
  - Autosome NN : NB : BB
  - Females, one direction NN : NB
  - Both sexes, both dir. NY : NN : NB : BB : BY

  ⇒ Need an X-chr-specific LOD threshold.
- "Null model" should include a sex effect.

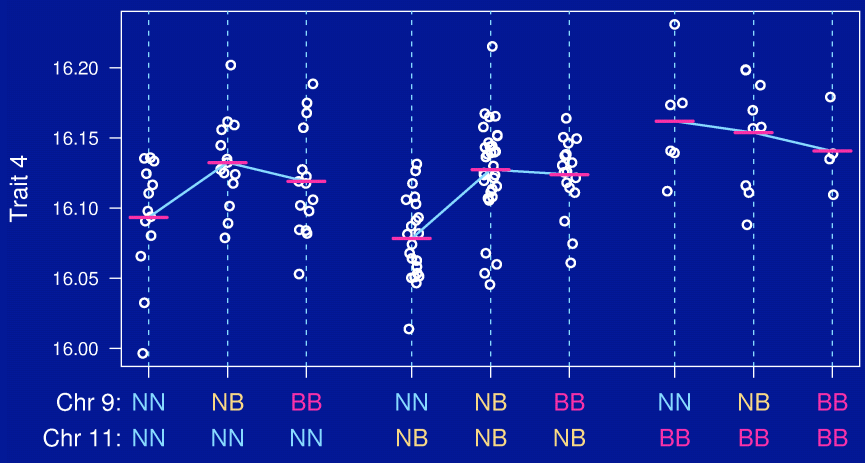
29

Chr 9 and 11



30

Epistasis



31

Going after multiple QTLs

- Greater ability to detect QTLs.
- Separate linked QTLs.
- Learn about interactions between QTLs (epistasis).

32

Model selection

- Choose a class of models.
  - Additive; pairwise interactions; regression trees
- Fit a model (allow for missing genotype data).
  - Linear regression; ML via EM; Bayes via MCMC
- Search model space.
  - Forward/backward/stepwise selection; MCMC
- Compare models.
  - $BIC_{\gamma}(\gamma) = \log L(\gamma) + (\gamma 2)\,|\gamma| \log n$

Miss important loci ⇔ include extraneous loci.

33

Special features

- Relationship among the covariates.
- Missing covariate information.
- Identify the key players vs. minimize prediction error.

34

Opportunities for improvements

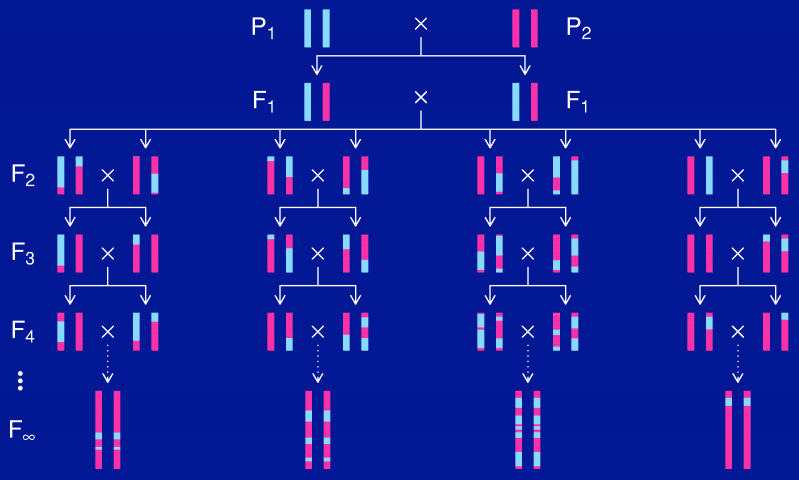
- Each individual is unique.
  - Must genotype each mouse.
  - Unable to obtain multiple invasive phenotypes (e.g., in multiple environmental conditions) on the same genotype.
- Relatively low mapping precision.

 Design a set of inbred mouse strains.

  - Genotype once.
  - Study multiple phenotypes on the same genotype.

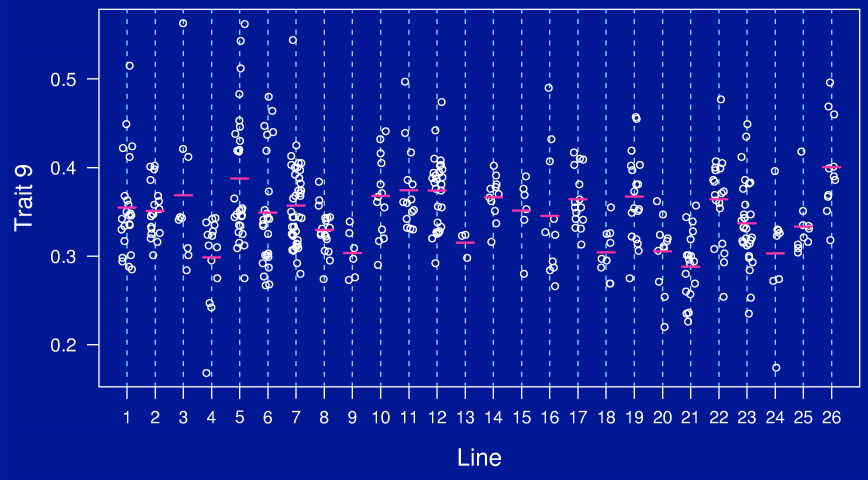
35

Recombinant inbred lines



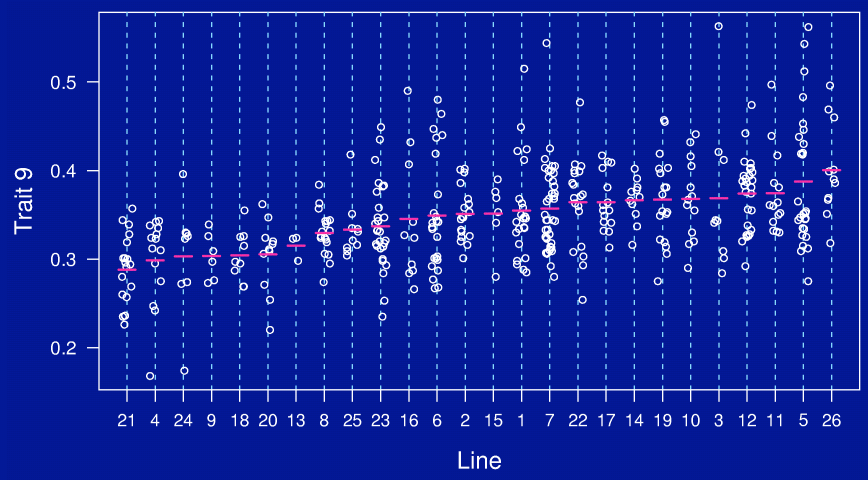
36

AXB/BXA panel



37

AXB/BXA panel



38

The usual analysis

- Calculate the phenotype average within each strain.
- Use these strain averages for QTL mapping as with a backcross (taking account of the map expansion in RILs).
- Can we do better?

  With the above data:

  Ave. no. mice per strain = 15.8 (SD = 8.4)

  Range of no. mice per strain = 3 – 39

39

A simple model for RILs

$y_{si} = \mu + \beta x_s + \alpha_s + \varepsilon_{si}$

- x_s = 0 or 1, according to genotype at putative QTL
- α_s = strain (polygenic) effect ~ normal(0, σ^2_s)
- ε_{si} = residual environment effect ~ normal(0, σ^2_e)

$\bar{y}_s = \sum_i y_{si} / n_s$

$\text{var}(\bar{y}_s) = \sigma^2_s + \sigma^2_e / n_s$

40

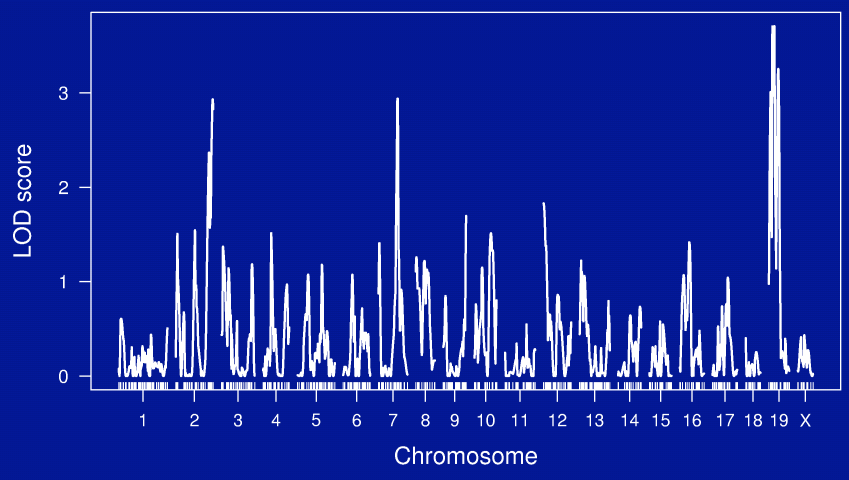
RIL analysis

If σ^2_s and σ^2_e were known:

- Work with the strain averages, $\bar{y}_s$
- Weight by $1/\left\{\sigma^2_s + \sigma^2_e/n_s\right\}$
- Equivalently, weight by $n_s/\left\{n_s h^2 + (1-h^2)\right\}$
  where $h^2 = \sigma^2_s/(\sigma^2_s + \sigma^2_e)$
- Equal n_s: The usual analysis is fine.
- h^2 large: Weight the strains equally.
- h^2 small: Weight the strains by n_s.

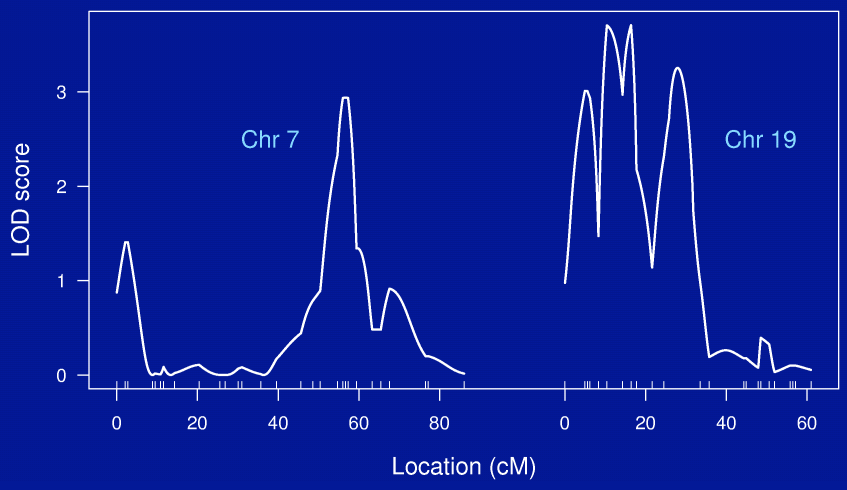
41

LOD curves



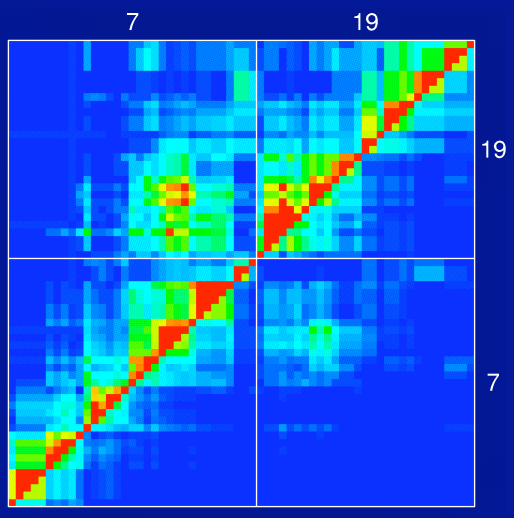
42

Chr 7 and 19



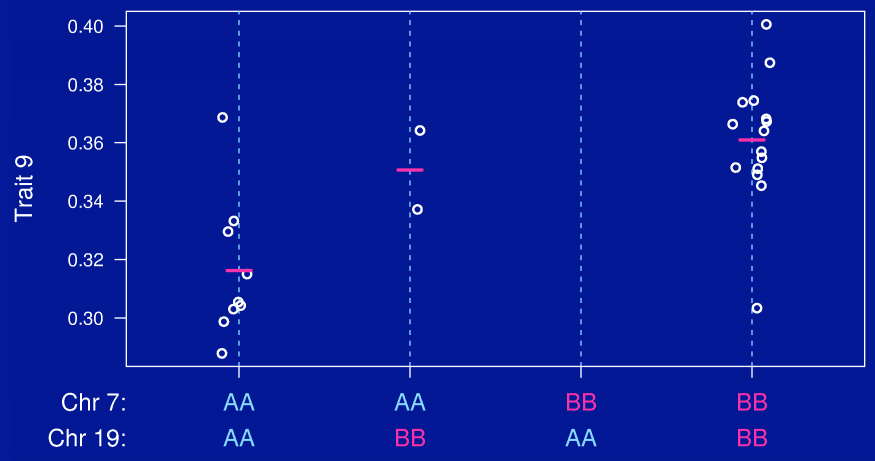
43

Recombination fractions



44

Chr 7 and 19



45

RI lines

Advantages

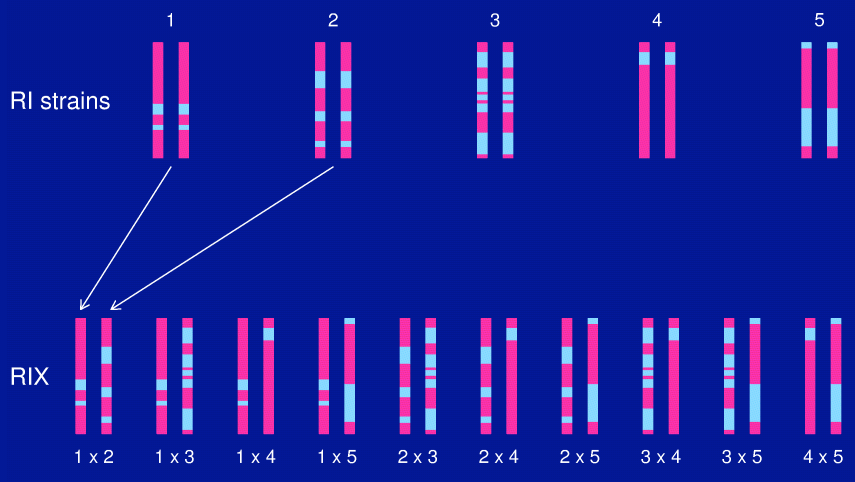
- Each strain is a eternal resource.
  - Only need to genotype once.
  - Reduce individual variation by phenotyping multiple individuals from each strain.
  - Study multiple phenotypes on the same genotype.
- Greater mapping precision.

Disadvantages

- Time and expense.
- Available panels are generally too small (10-30 lines).
- Can learn only about 2 particular alleles.
- All individuals homozygous.

46

The RIX design



47

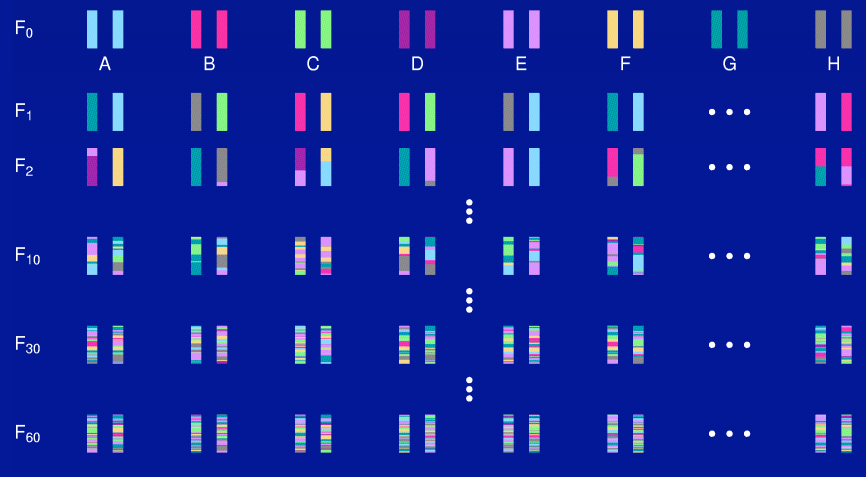
Heterogeneous stock

McClearn et al. (1970)
Mott et al. (2000); Mott and Flint (2002)

- Start with 8 inbred strains.
- Randomly breed 40 pairs.
- Repeat the random breeding of 40 pairs for each of ~60 generations (30 years).
- The genealogy (and protocol) is not completely known.

48

Heterogeneous stock



49

Heterogeneous stock

Advantages

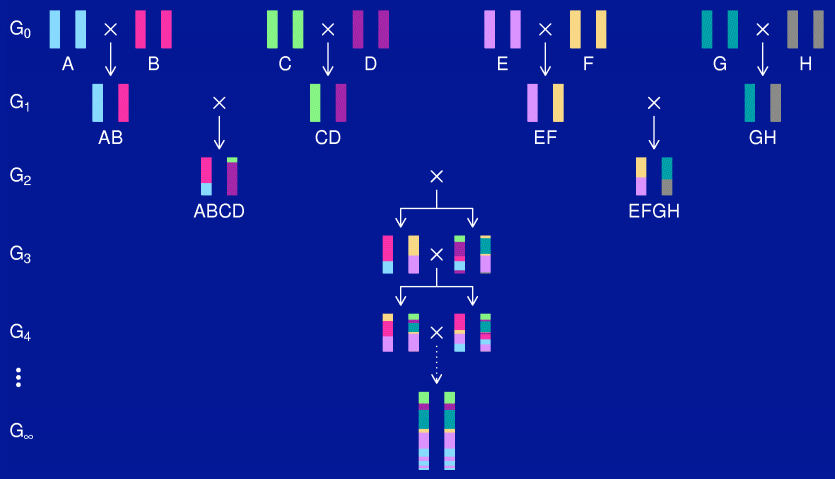
- Great mapping precision.
- Learn about 8 alleles.

Disadvantages

- Time.
- Each individual is unique.
- Need extremely dense markers.

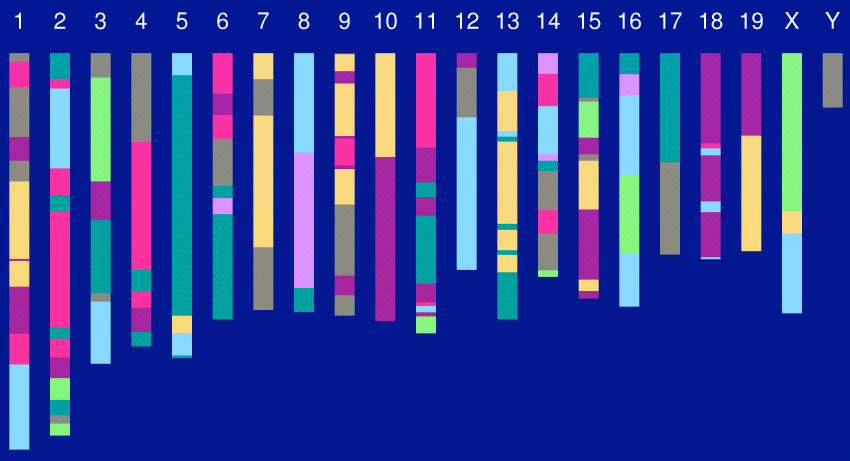
50

The “Collaborative Cross”



51

Genome of an 8-way RI



52

The "Collaborative Cross"

Advantages

- Great mapping precision.
- Eternal resource.
  - Genotype only once.
  - Study multiple invasive phenotypes on the same genotype.

Barriers

- Advantages not widely appreciated.
  - Ask one question at a time, or Ask many questions at once?
- Time.
- Expense.
- Requires large-scale collaboration.

53

To be worked out

- Breakpoint process along an 8-way RI chromosome.
- Reconstruction of genotypes given multipoint marker data.
- Single-QTL analyses.
  - Mixed models, with random effects for strains and genotypes/alleles.
- Power and precision (relative to an intercross).

54

Haldane & Waddington 1930

r = recombination fraction per meiosis between two loci

Autosomes

$Pr(G_1=AA) = Pr(G_1=BB) = 1/2$

$Pr(G_2=BB \mid G_1=AA) = Pr(G_2=AA \mid G_1=BB) = 4r / (1+6r)$

X chromosome

$Pr(G_1=AA) = 2/3 \quad Pr(G_1=BB) = 1/3$

$Pr(G_2=BB \mid G_1=AA) = 2r / (1+4r)$

$Pr(G_2=AA \mid G_1=BB) = 4r / (1+4r)$

$Pr(G_2 \neq G_1) = (8/3)\, r / (1+4r)$

55

8-way RILs

Autosomes

$Pr(G_1 = i) = 1/8$

$Pr(G_2 = j \mid G_1 = i) = r / (1+6r) \quad \text{for } i \neq j$

$Pr(G_2 \neq G_1) = 7r / (1+6r)$

X chromosome

$Pr(G_1=AA) = Pr(G_1=BB) = Pr(G_1=EE) = Pr(G_1=FF) = 1/6$

$Pr(G_1=CC) = 1/3$

$Pr(G_2=AA \mid G_1=CC) = r / (1+4r)$

$Pr(G_2=CC \mid G_1=AA) = 2r / (1+4r)$

$Pr(G_2=BB \mid G_1=AA) = r / (1+4r)$

$Pr(G_2 \neq G_1) = (14/3)\, r / (1+4r)$

56

Acknowledgments

- Terry Speed, Univ. of California, Berkeley and WEHI
- Tom Brodnicki, WEHI
- Gary Churchill, The Jackson Laboratory
- Joe Nadeau, Case Western Reserve Univ.

57