Gene mapping in mice

Karl W Broman
Department of Biostatistics
Johns Hopkins University

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

Goal

- Identify genes that contribute to common human diseases.

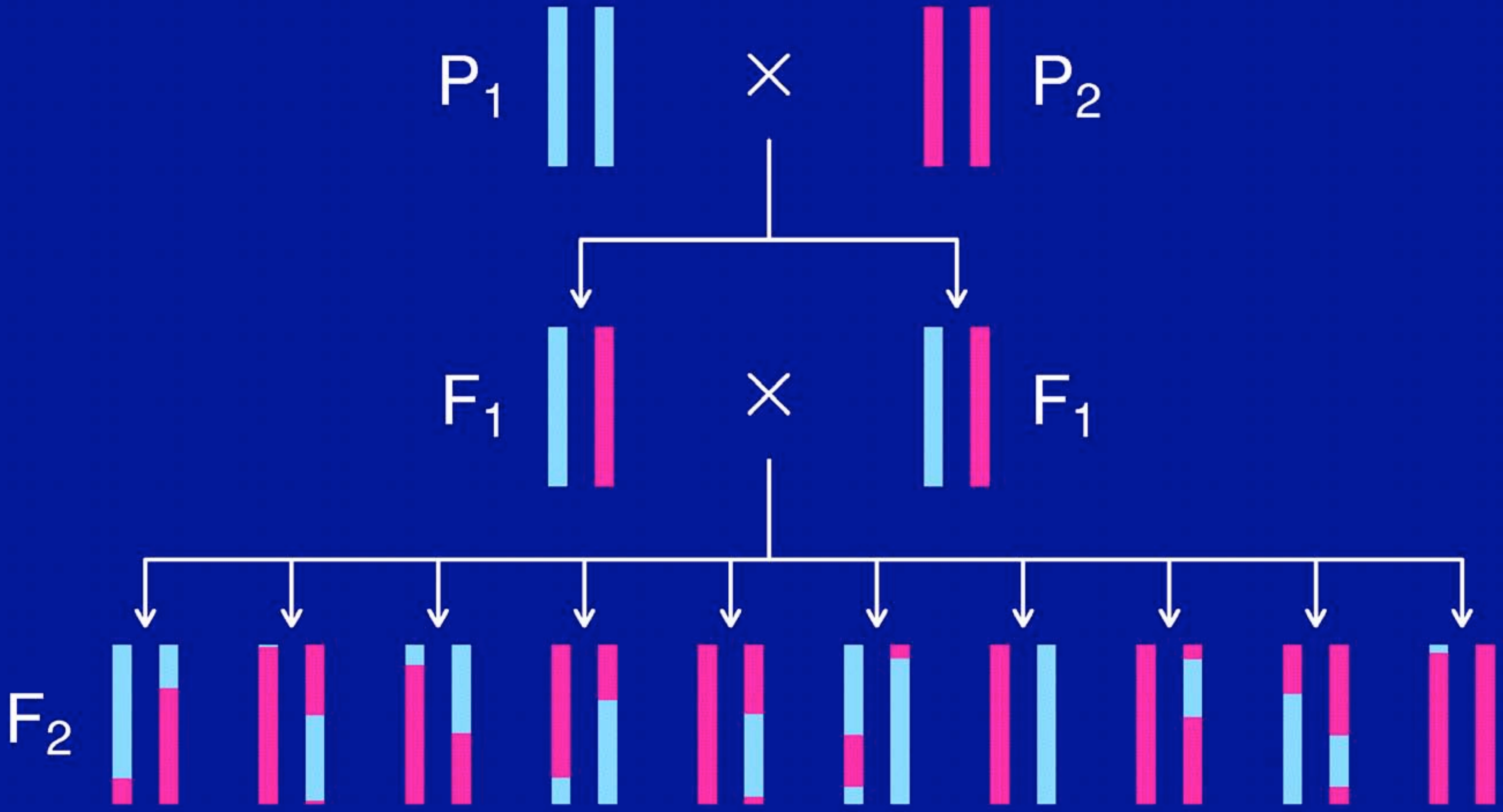
Advantages of the mouse

- Small and cheap
- Inbred lines
- Large, controlled crosses
- Experimental interventions
- Knock-outs and knock-ins

The mouse as a model

- Same genes?
  - The genes involved in a phenotype in the mouse may also be involved in similar phenotypes in the human.
- Similar complexity?
  - The complexity of the etiology underlying a mouse phenotype provides some indication of the complexity of similar human phenotypes.
- Transfer of statistical methods.
  - The statistical methods developed for gene mapping in the mouse serve as a basis for similar methods applicable in direct human studies.

The intercross



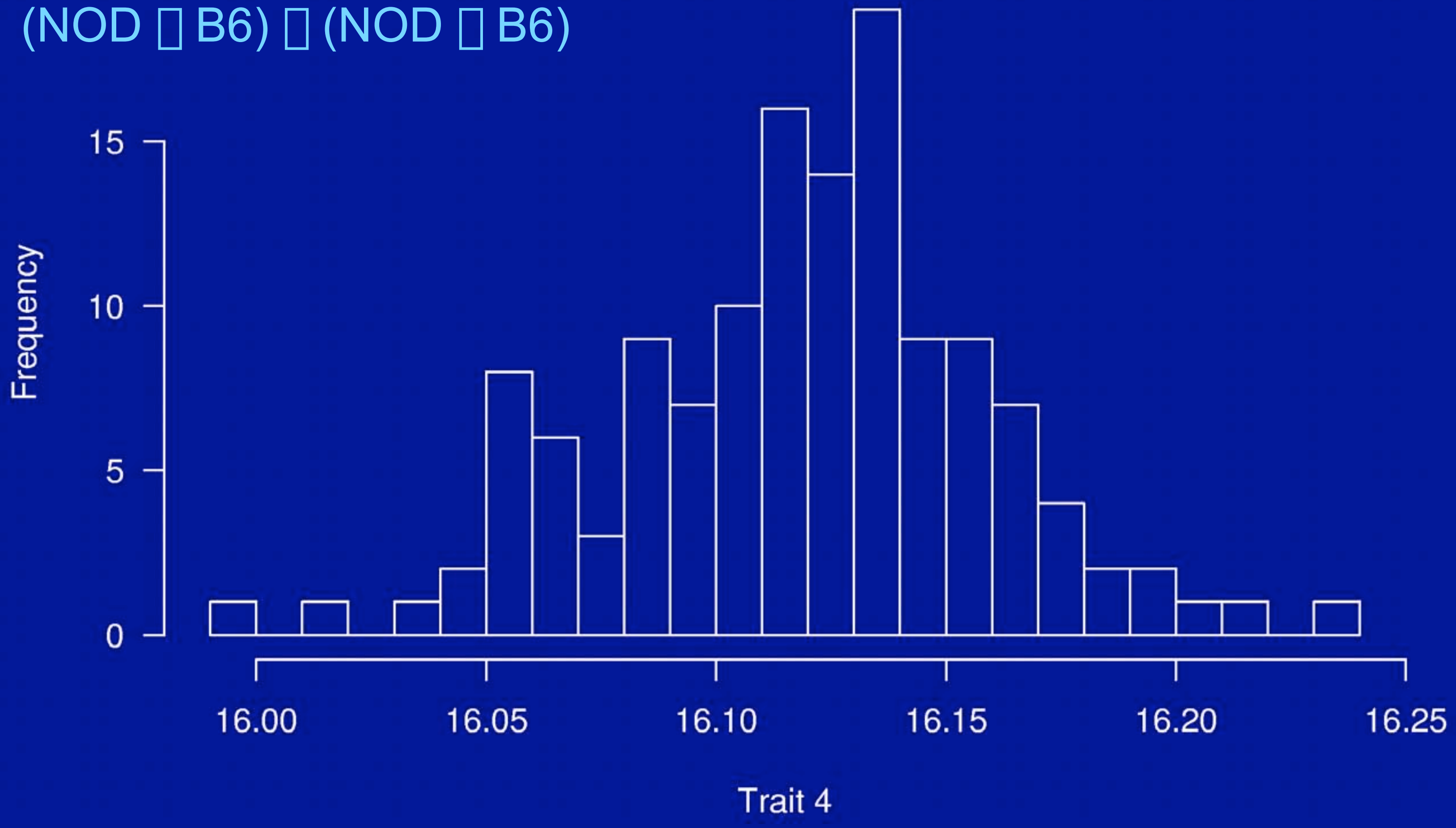
The data

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

Phenotypes

133 females

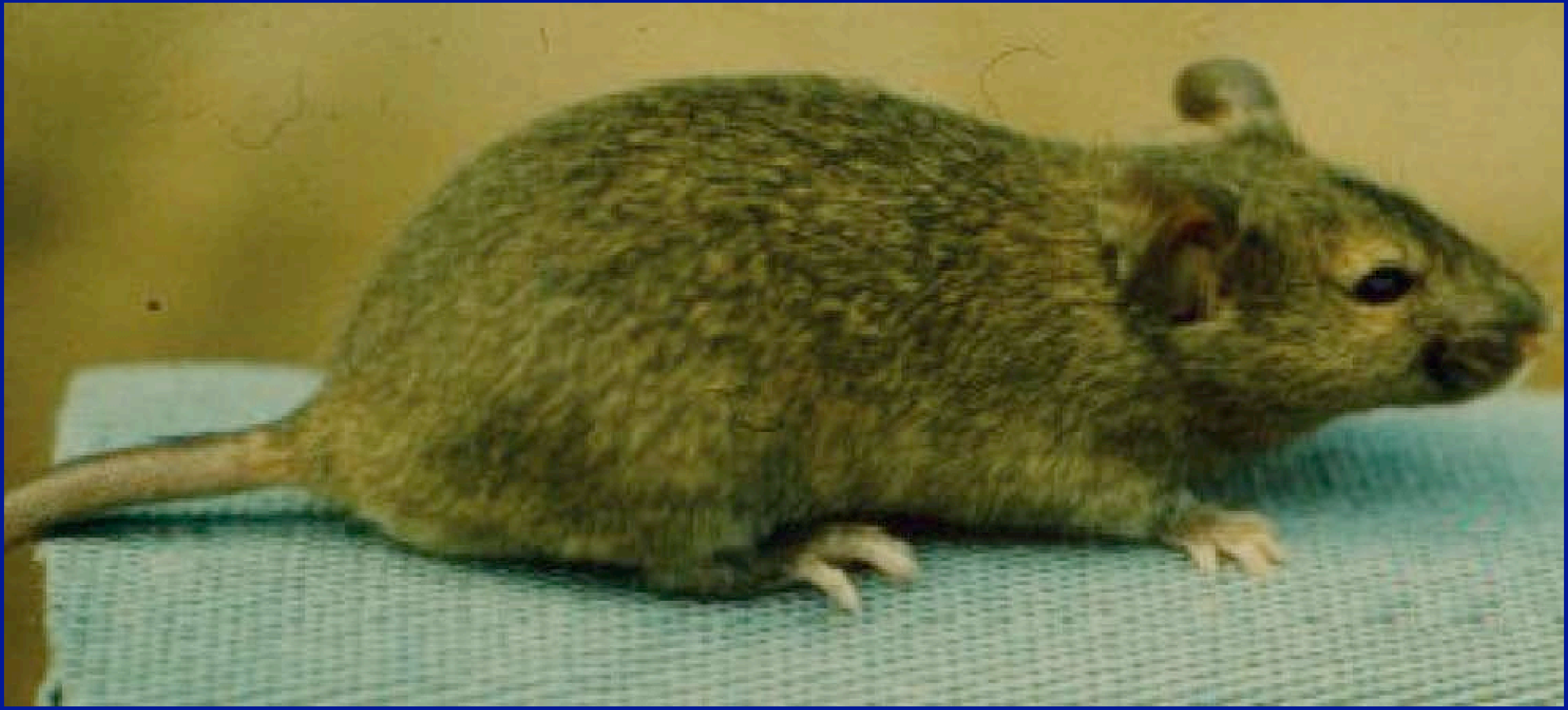
(NOD × B6) × (NOD × B6)



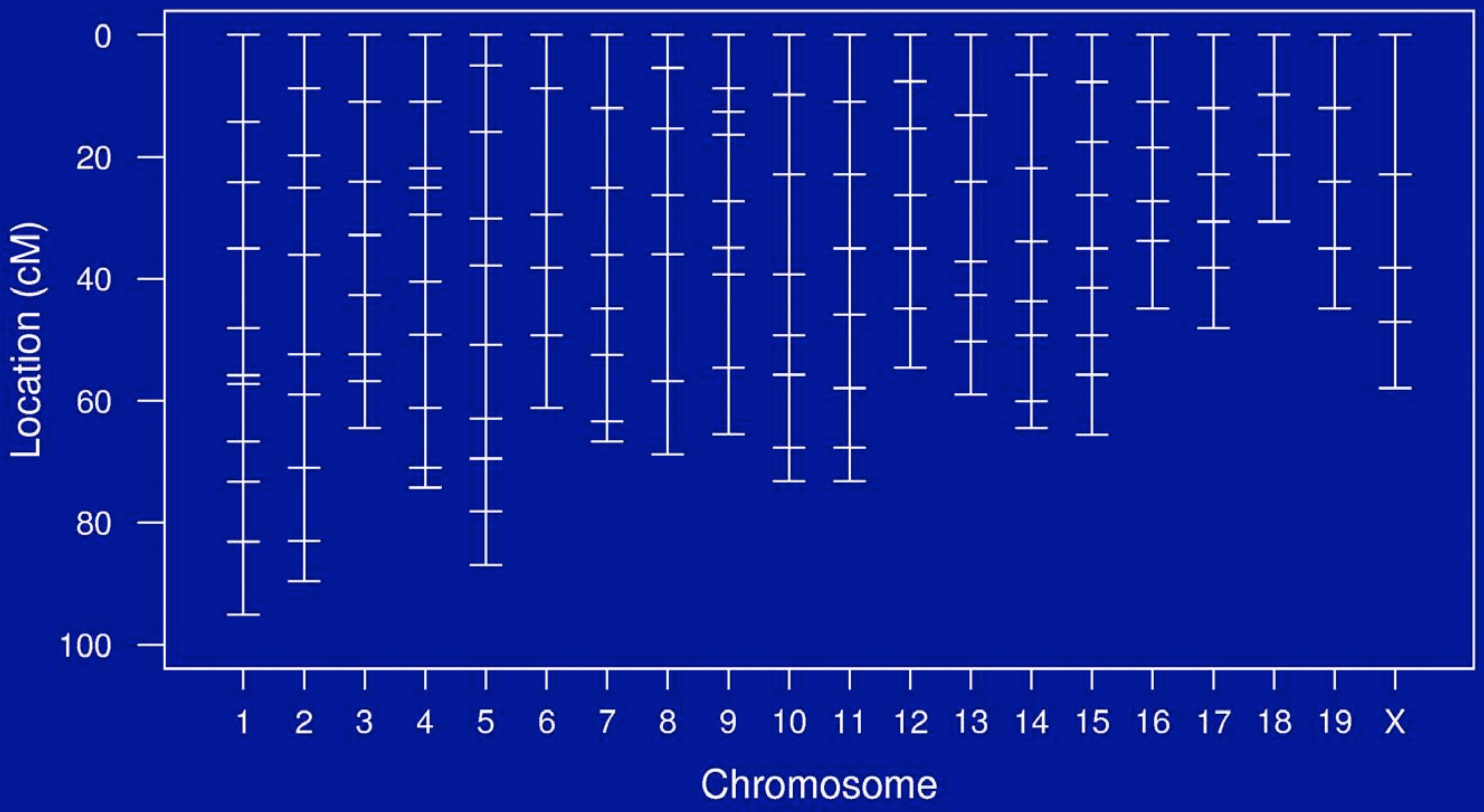
NOD

C57BL/6

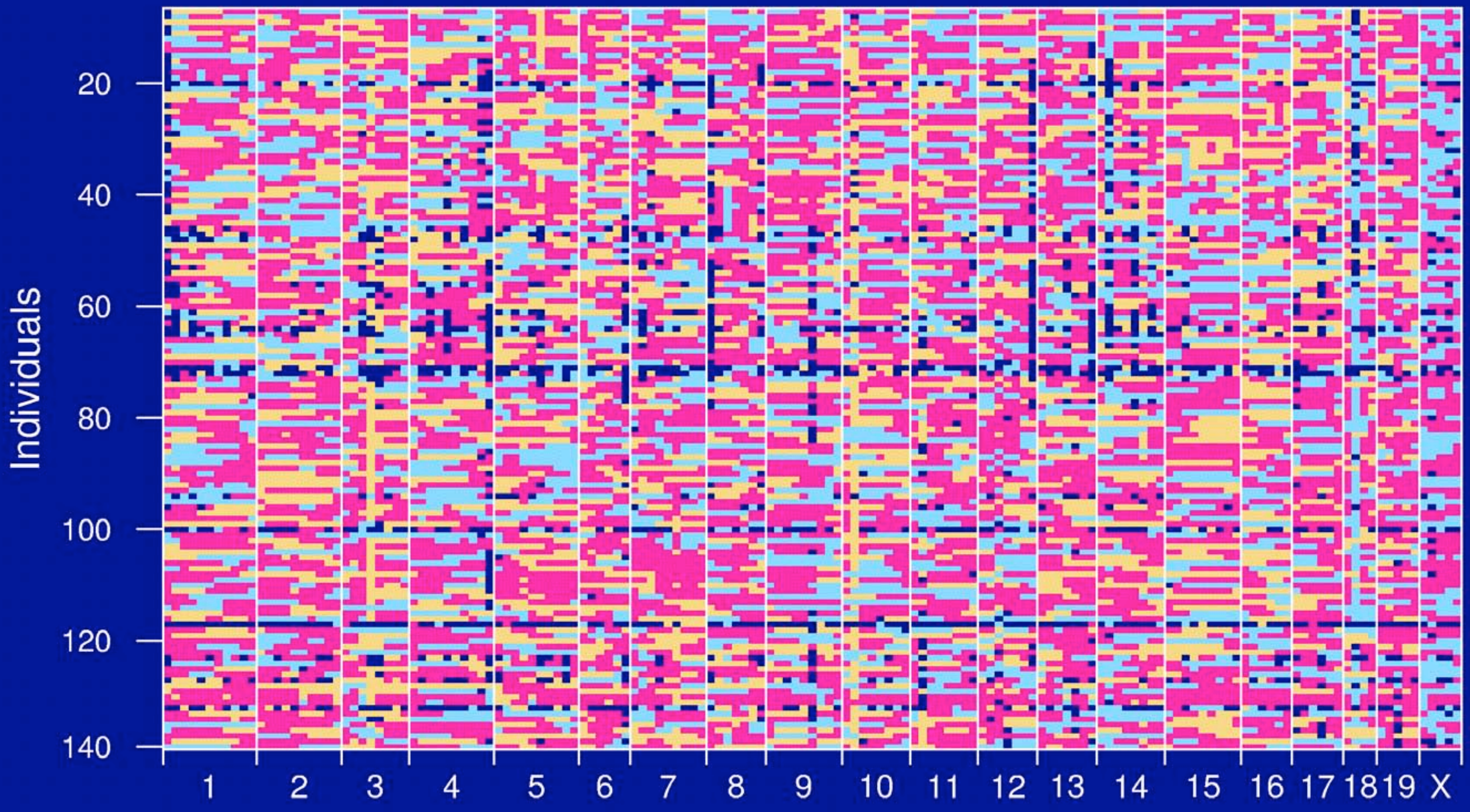
Agouti coat



Genetic map


Location (cM)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

Genotype data



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.

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.

Models: gen □ phe

Phenotype = y, whole-genome genotype = g

Imagine that p sites are all that matter.

$E(y \mid g) = \square(g_1,\ldots,g_p)$ $SD(y \mid g) = \square(g_1,\ldots,g_p)$

Simplifying assumptions:

- $SD(y \mid g) = \square$, independent of g
- $y \mid g \sim \text{normal}(\ \square(g_1,\ldots,g_p), \square\ )$
- $\square(g_1,\ldots,g_p) = \square + \sum \square_j \, 1\{g_j = \text{AB}\} + \square_j \, 1\{g_j = \text{BB}\}$

Interval mapping

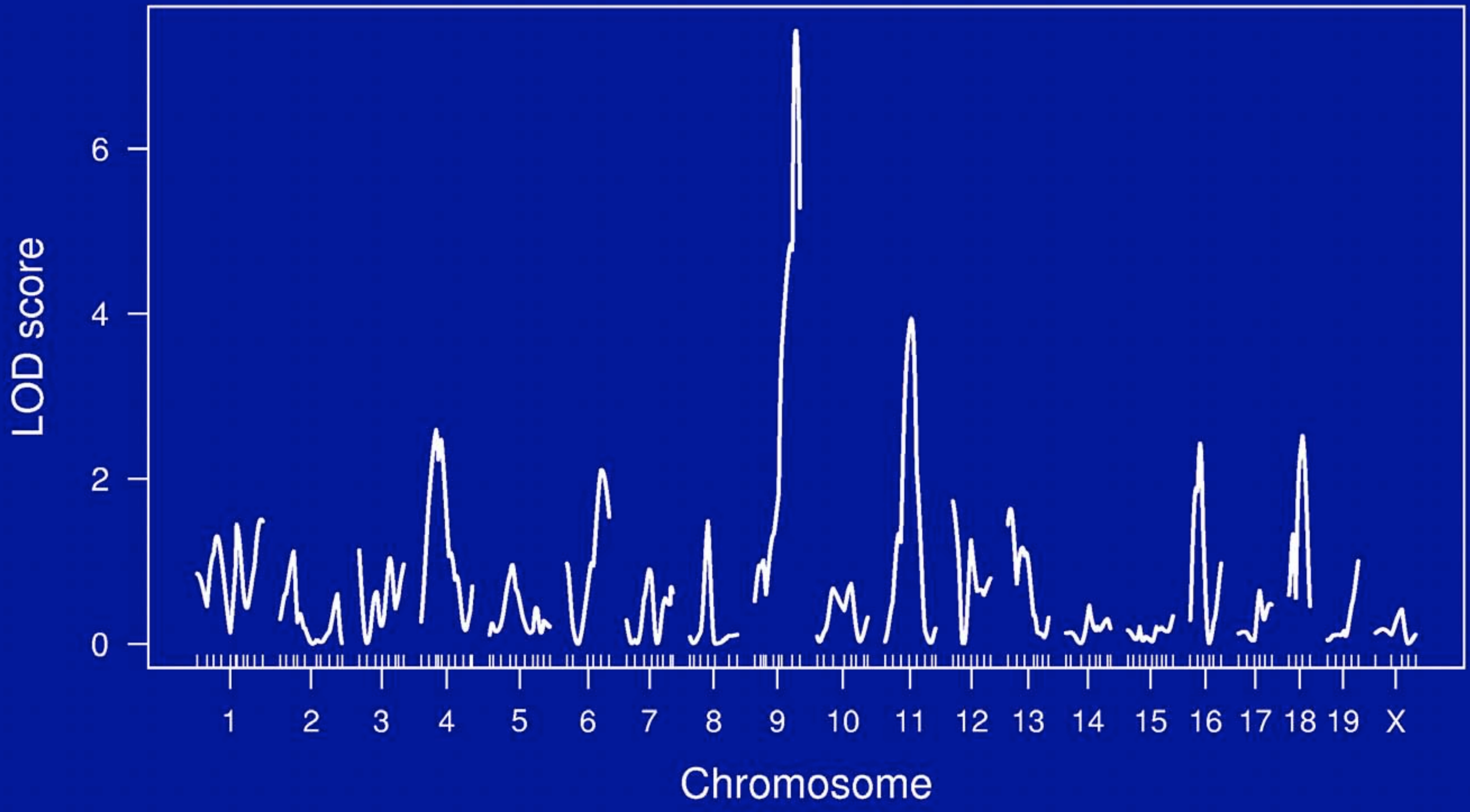
Lander and Botstein 1989

- Imagine that there is a single QTL, at position z.
- Let q_i = genotype of mouse i at the QTL, and assume

  $y_i \mid q_i \sim \text{normal}(\,\mu(q_i), \sigma\,)$
- We won’t know q_i, but we can calculate

  $p_{ig} = \Pr(q_i = g \mid \text{marker data})$
- 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.

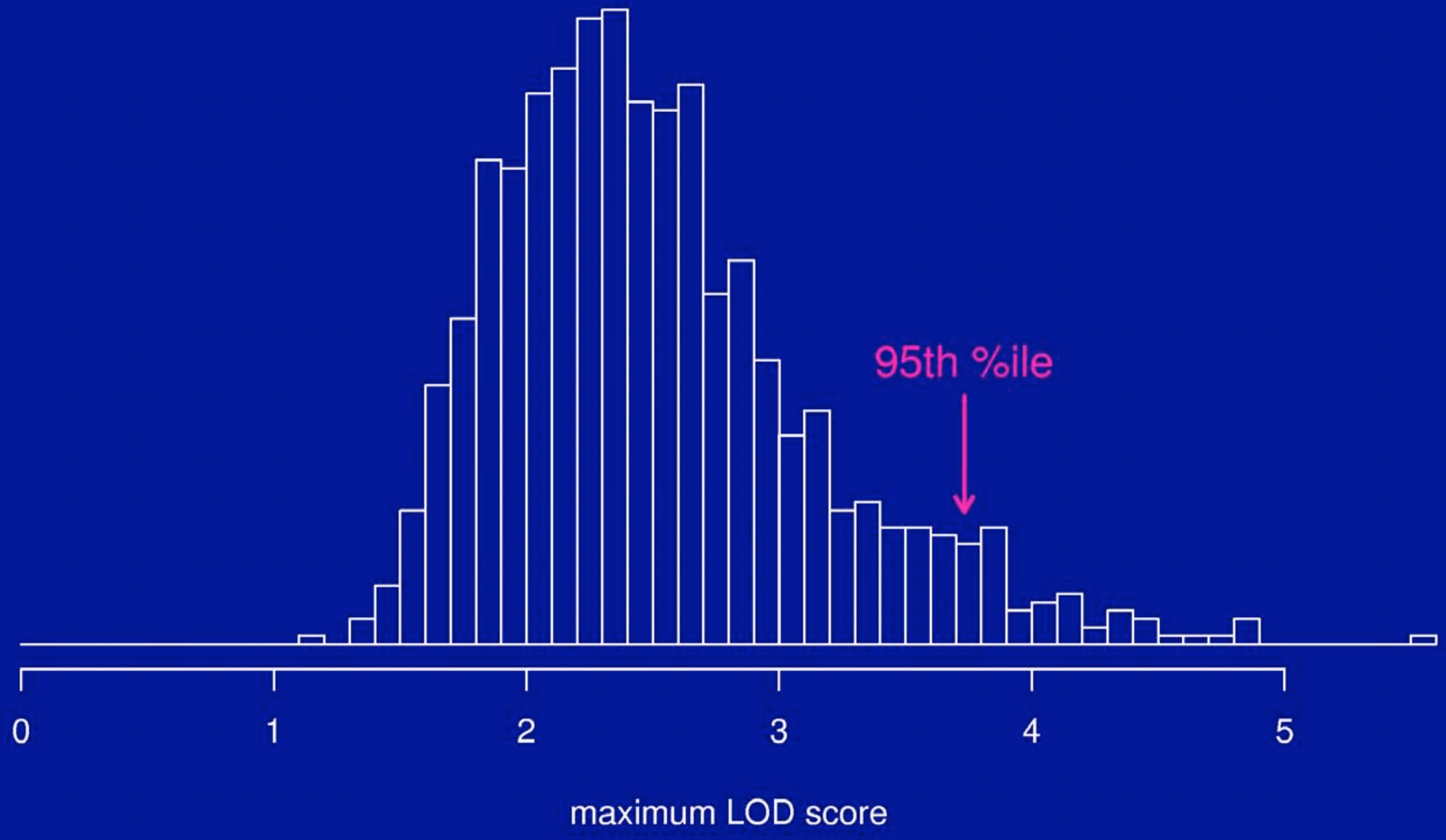
LOD curves


LOD score
0
2
4
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

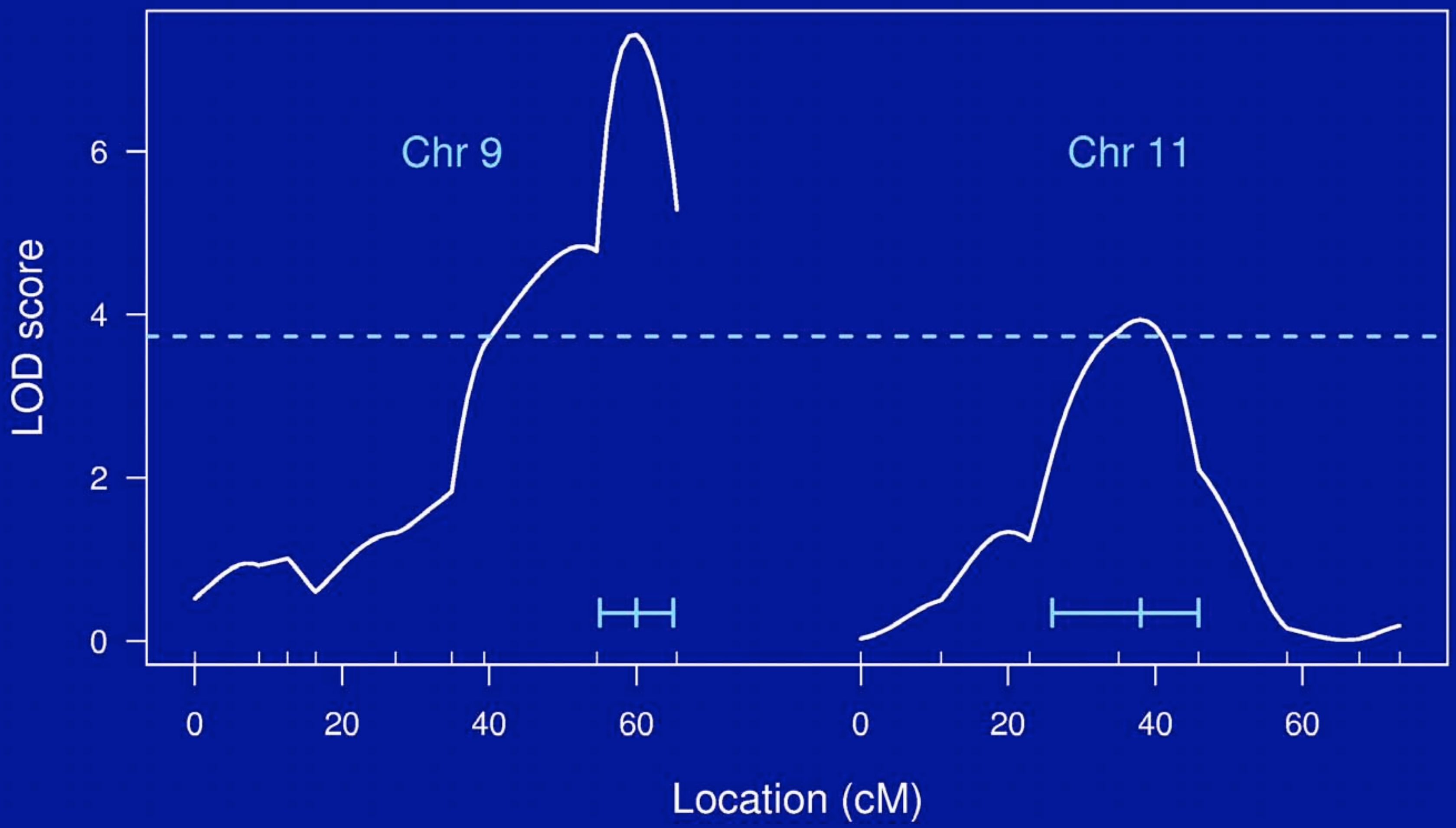
LOD thresholds

- 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).

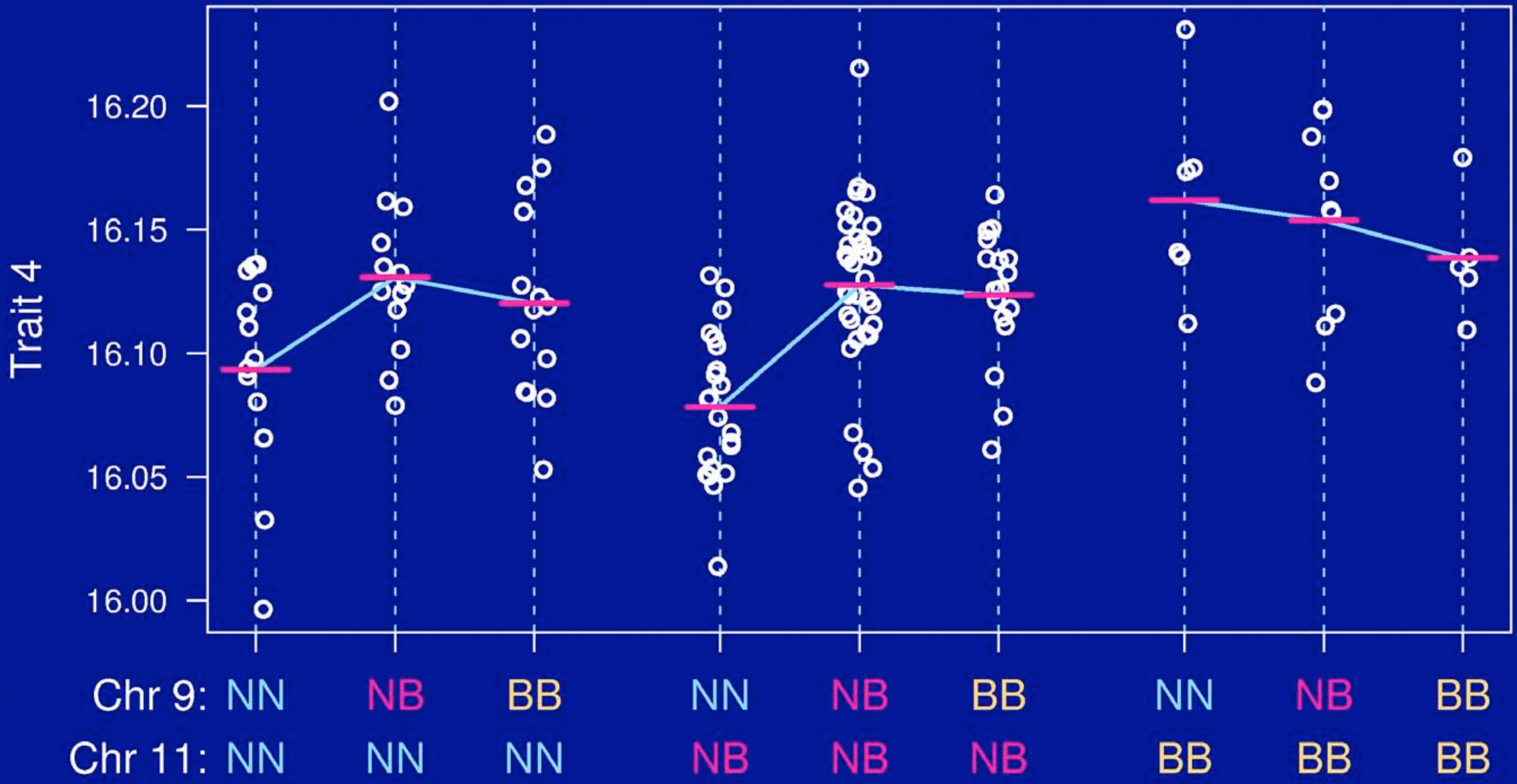
Permutation distribution



Chr 9 and 11



Epistasis



Trait 4
16.20
16.15
16.10
16.05
16.00
Chr 9: NN NB BB NN NB BB NN NB BB
Chr 11: NN NN NN NB NB NB BB BB BB

Going after multiple QTLs

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

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.
  - $\mathrm{BIC}_{\square}(\square) = \log L(\square) + (\square/2)\,|\square|\,\log n$

Miss important loci □ include extraneous loci.

Special features

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

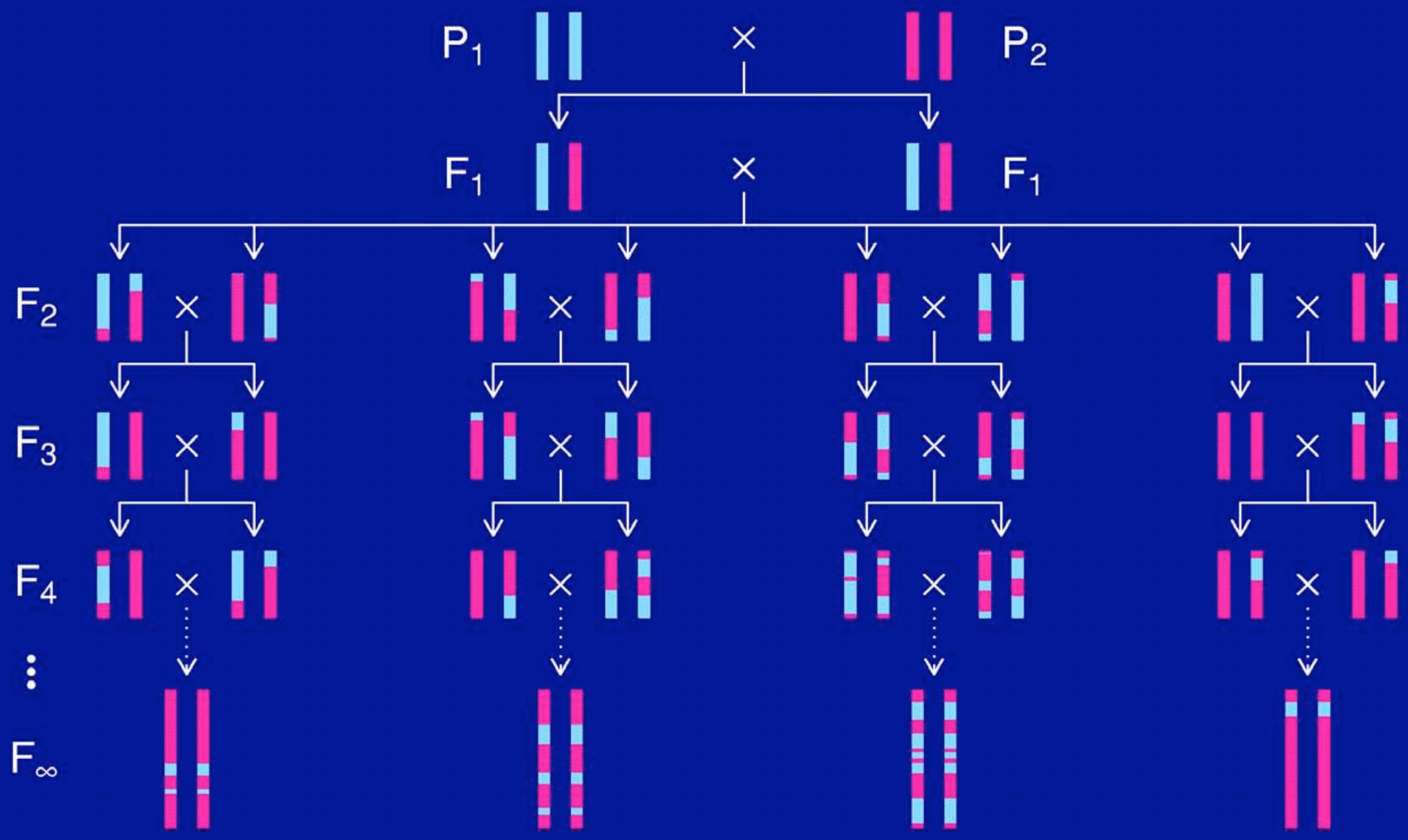
Opportunities for improvements

- 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.

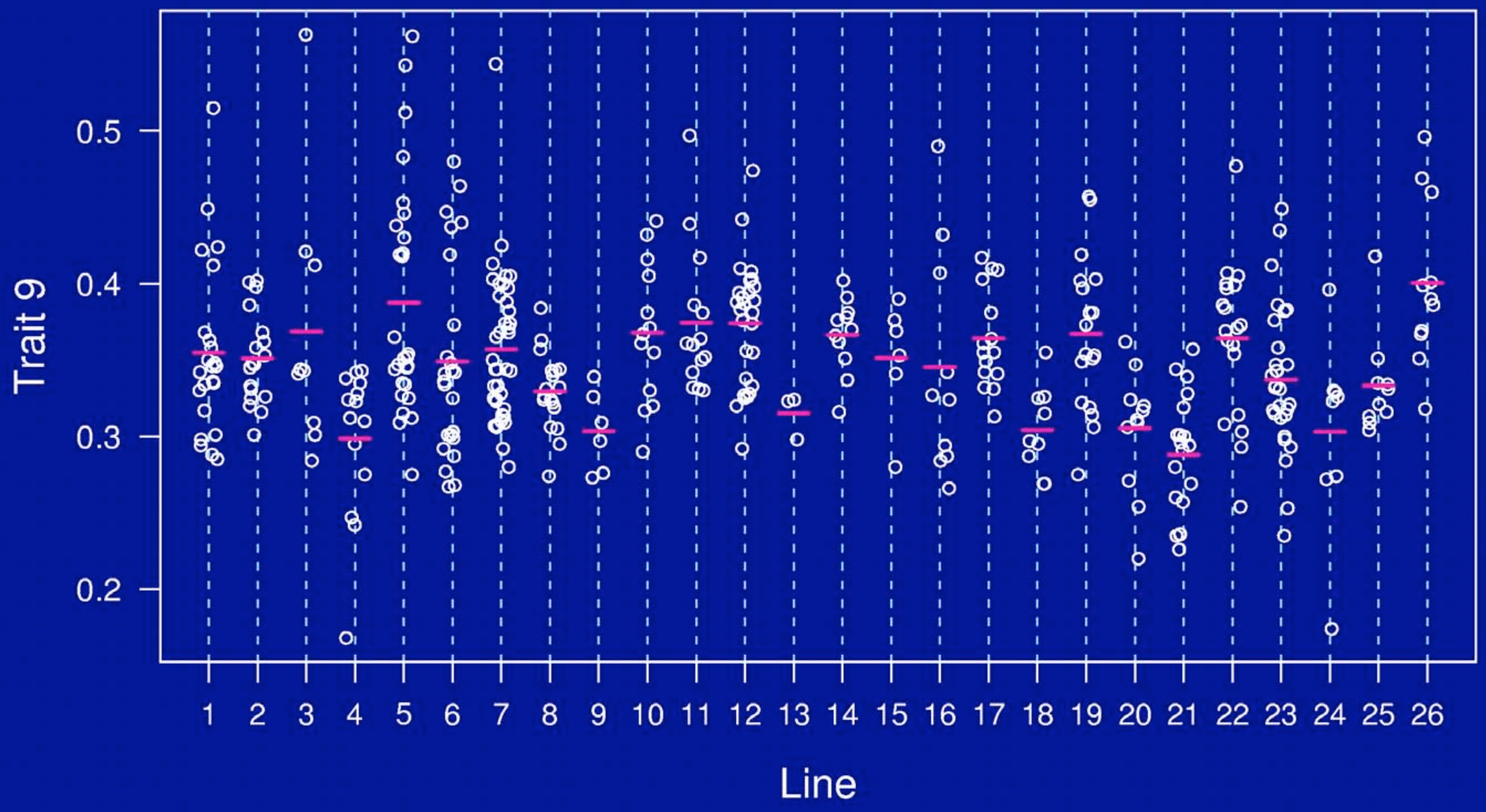
□ Design a set of inbred mouse strains.

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

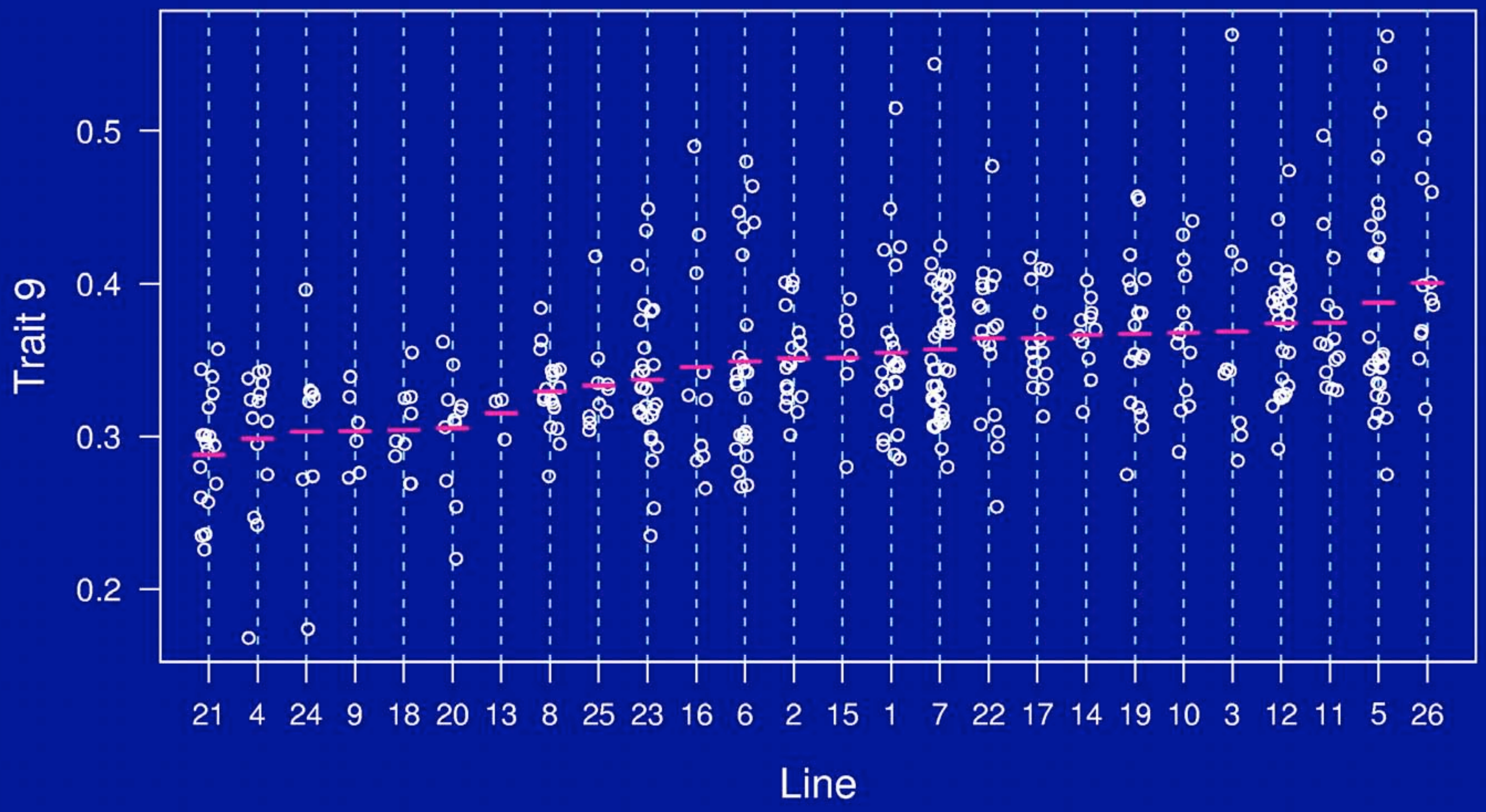
Recombinant inbred lines



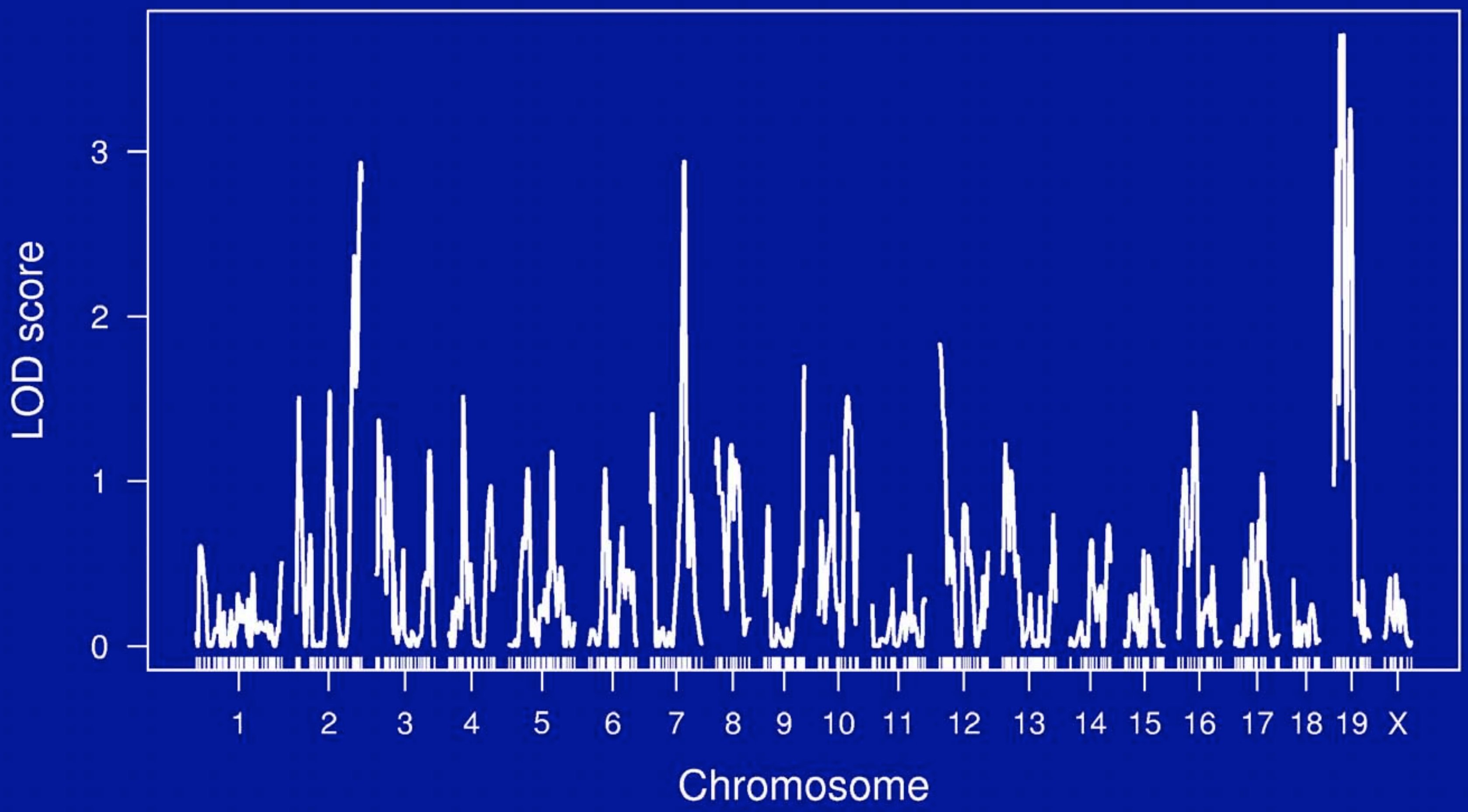
AXB/BXA panel



AXB/BXA panel

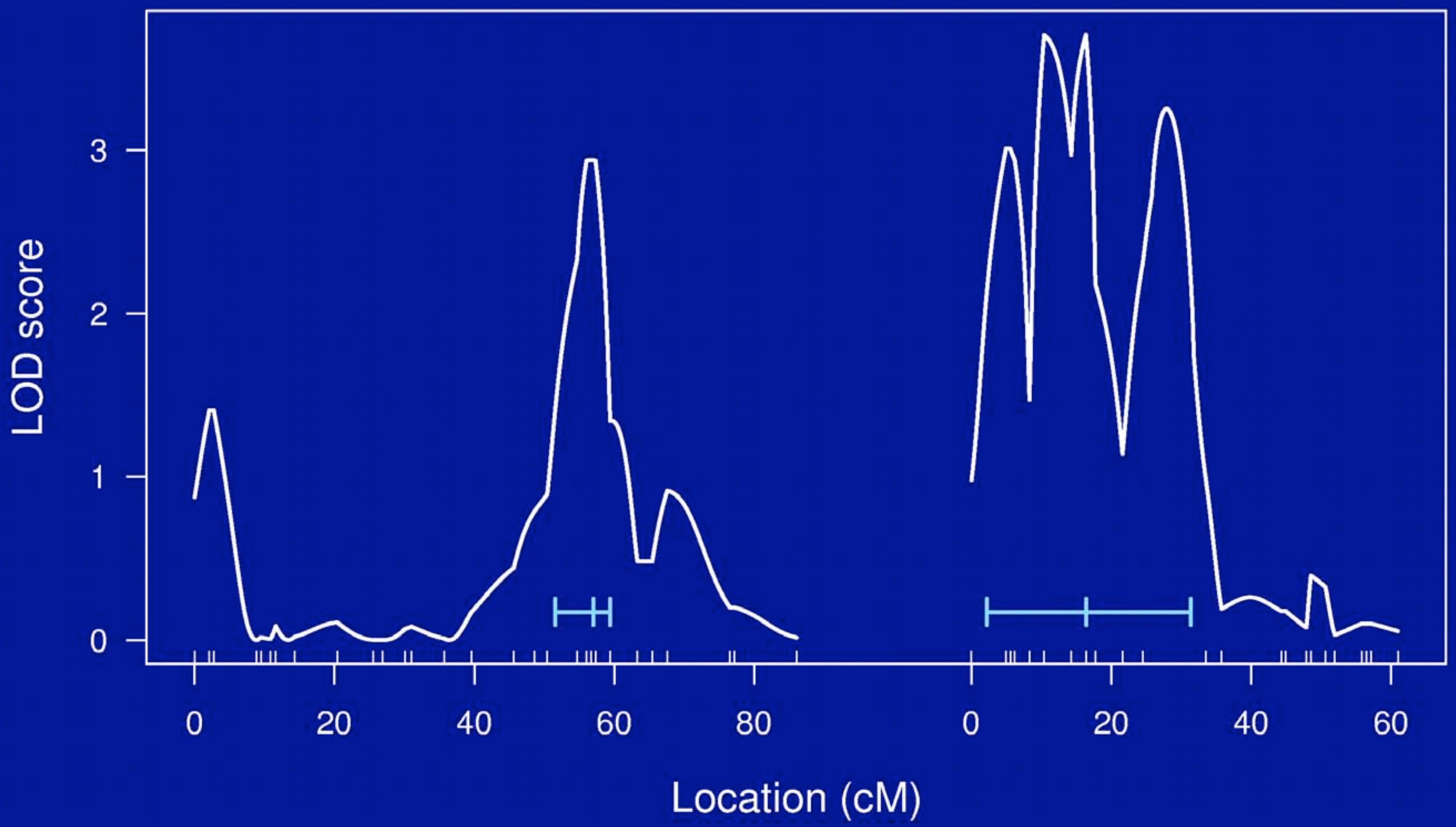


LOD curves

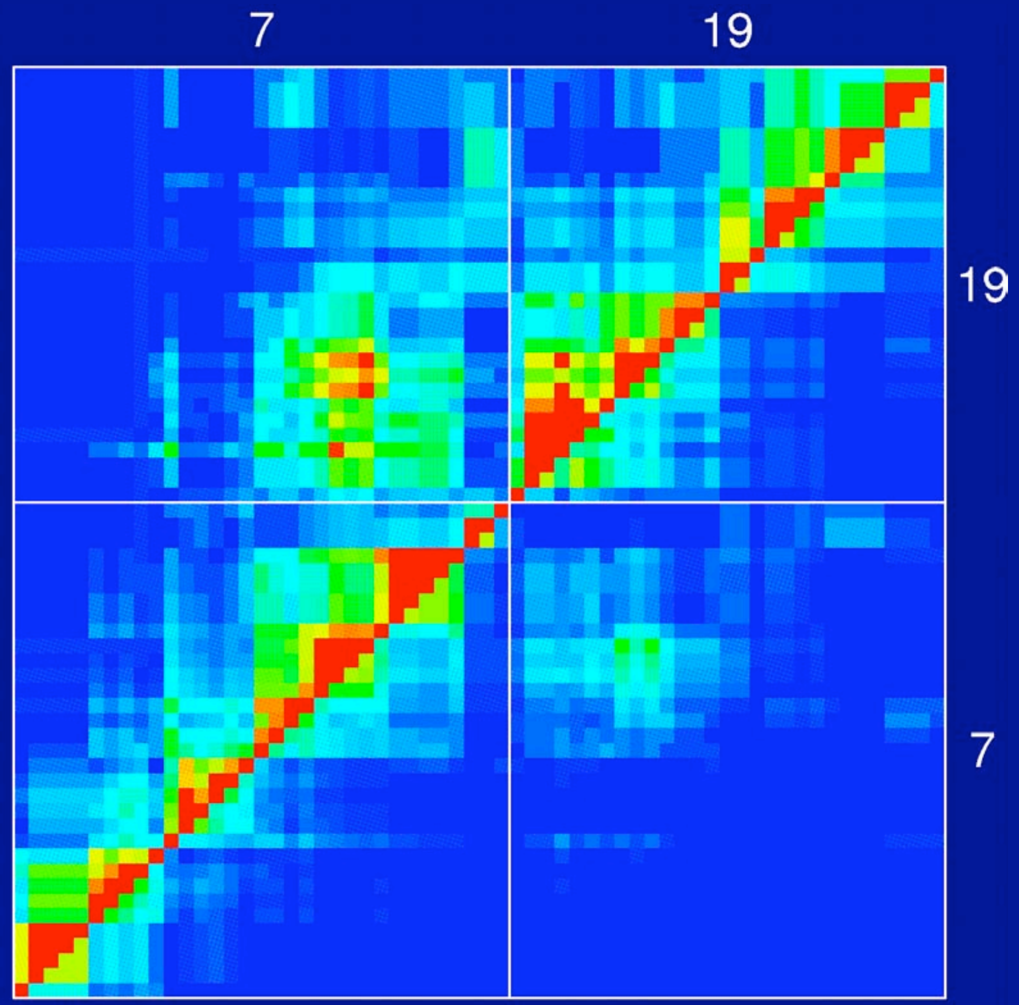


LOD score
0
1
2
3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

Chr 7 and 19



Recombination fractions



RI lines

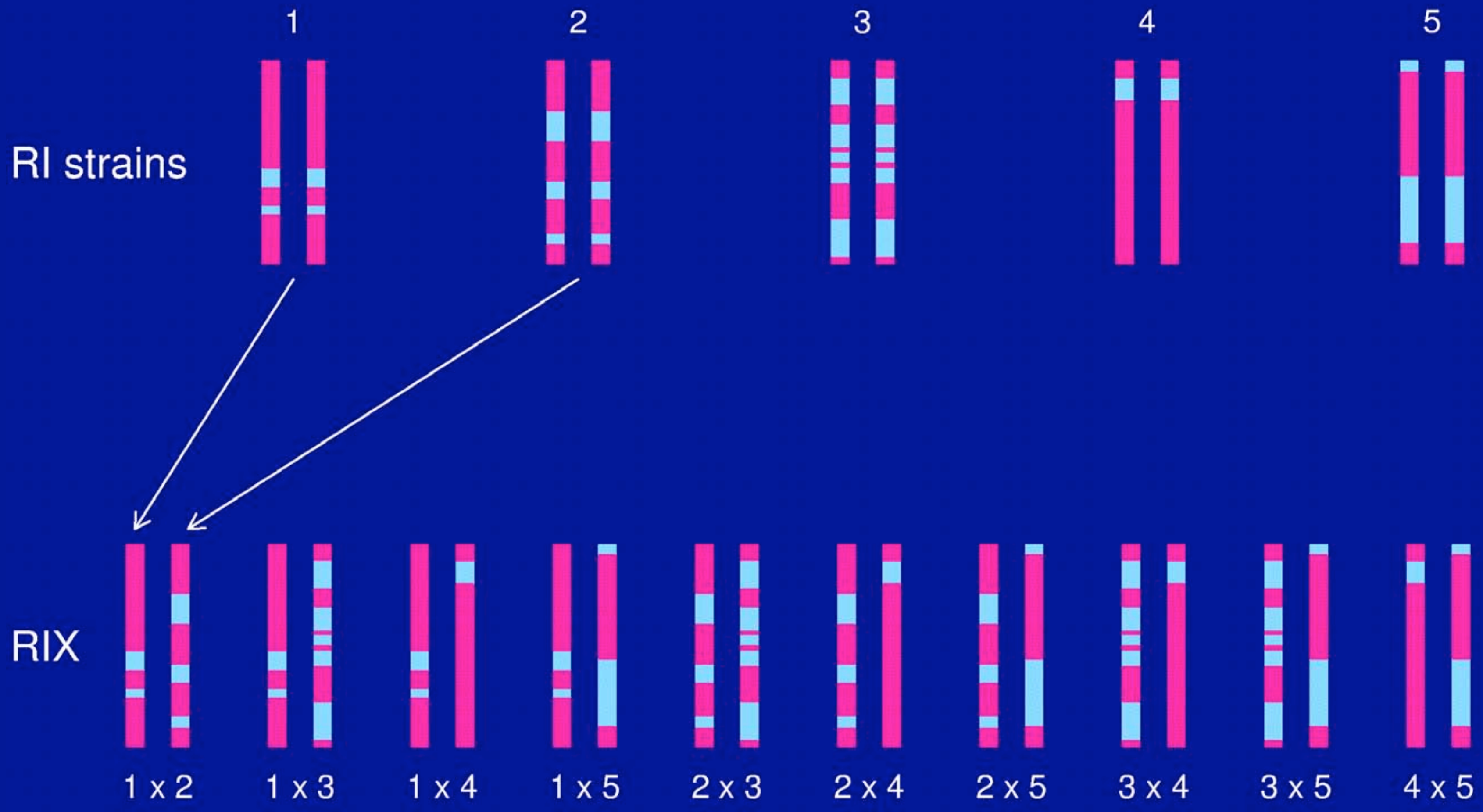
Advantages

- 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.

The RIX design



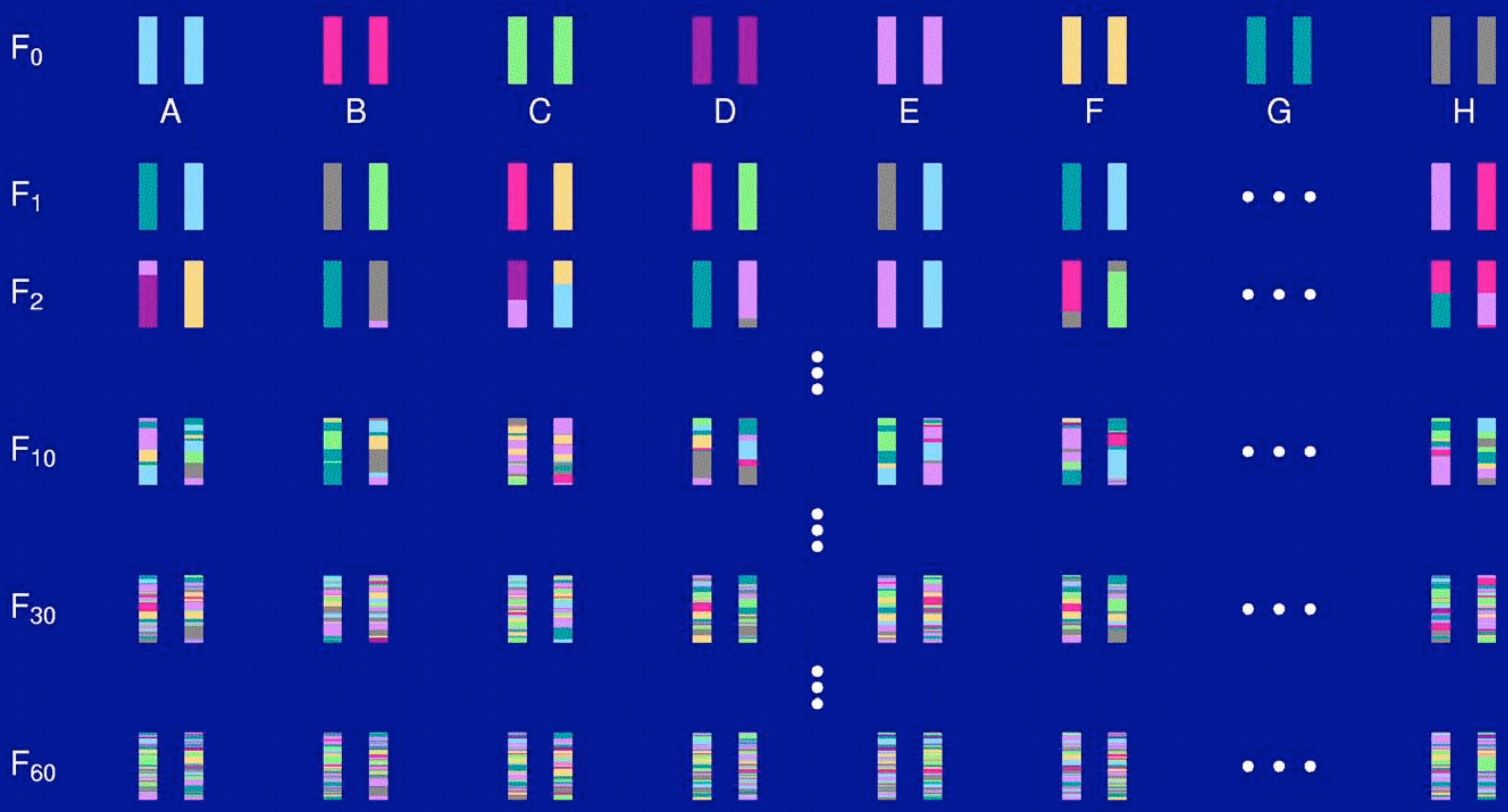
Heterogeneous stock

McClearn et al. (1970)

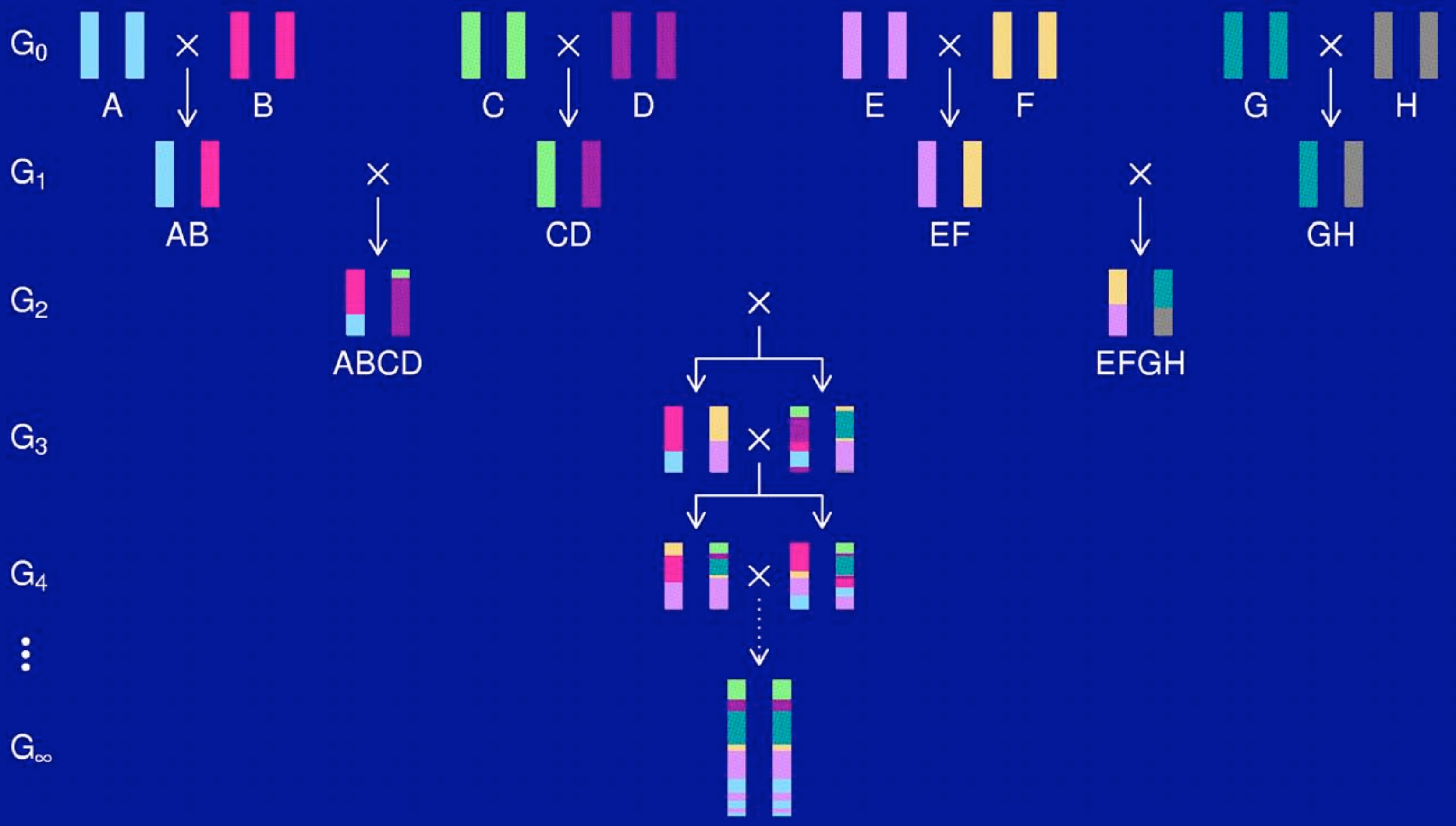
Mott et al. (2000); Mott and Flint (2002)

- 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.

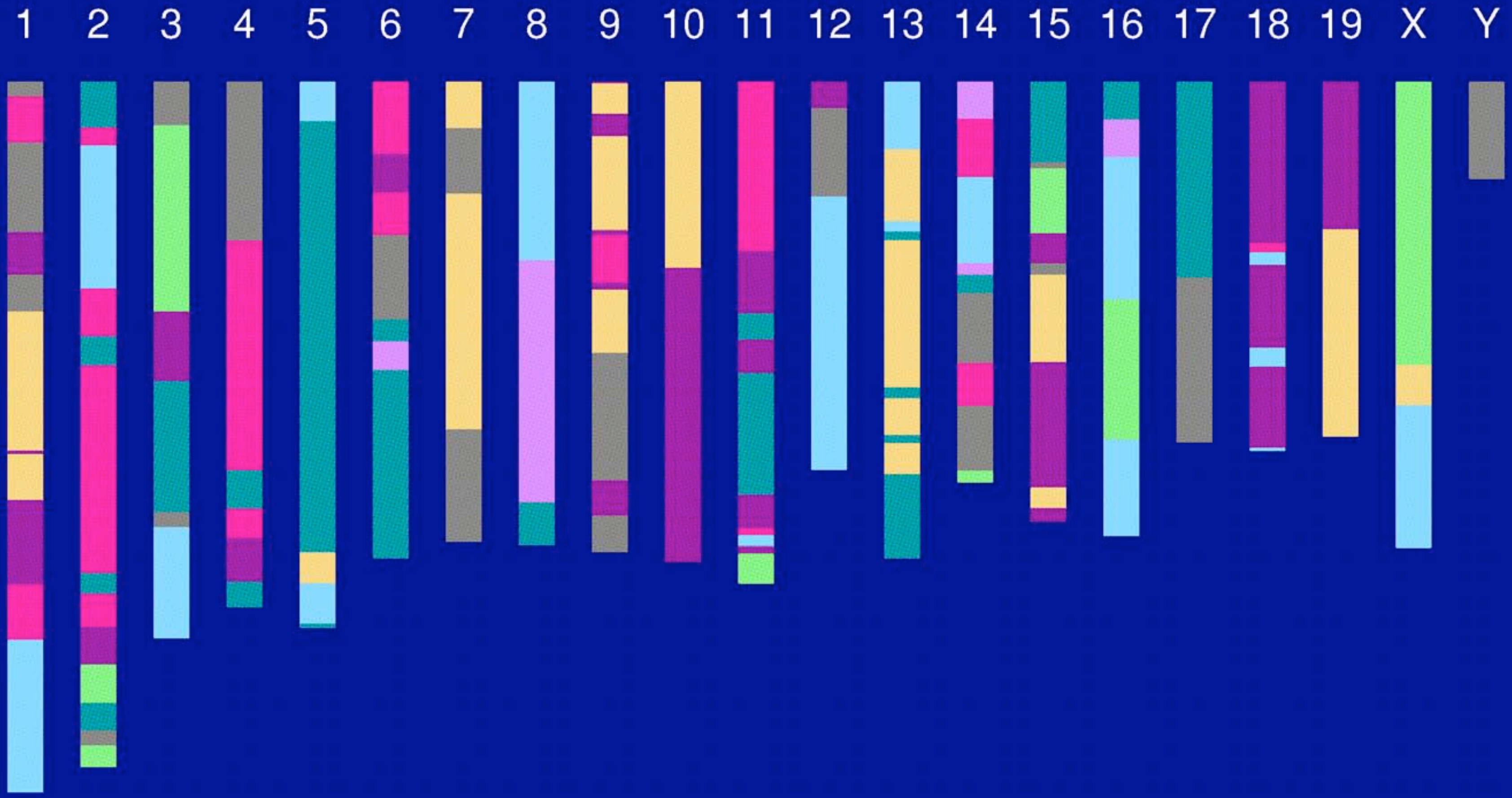
Heterogeneous stock



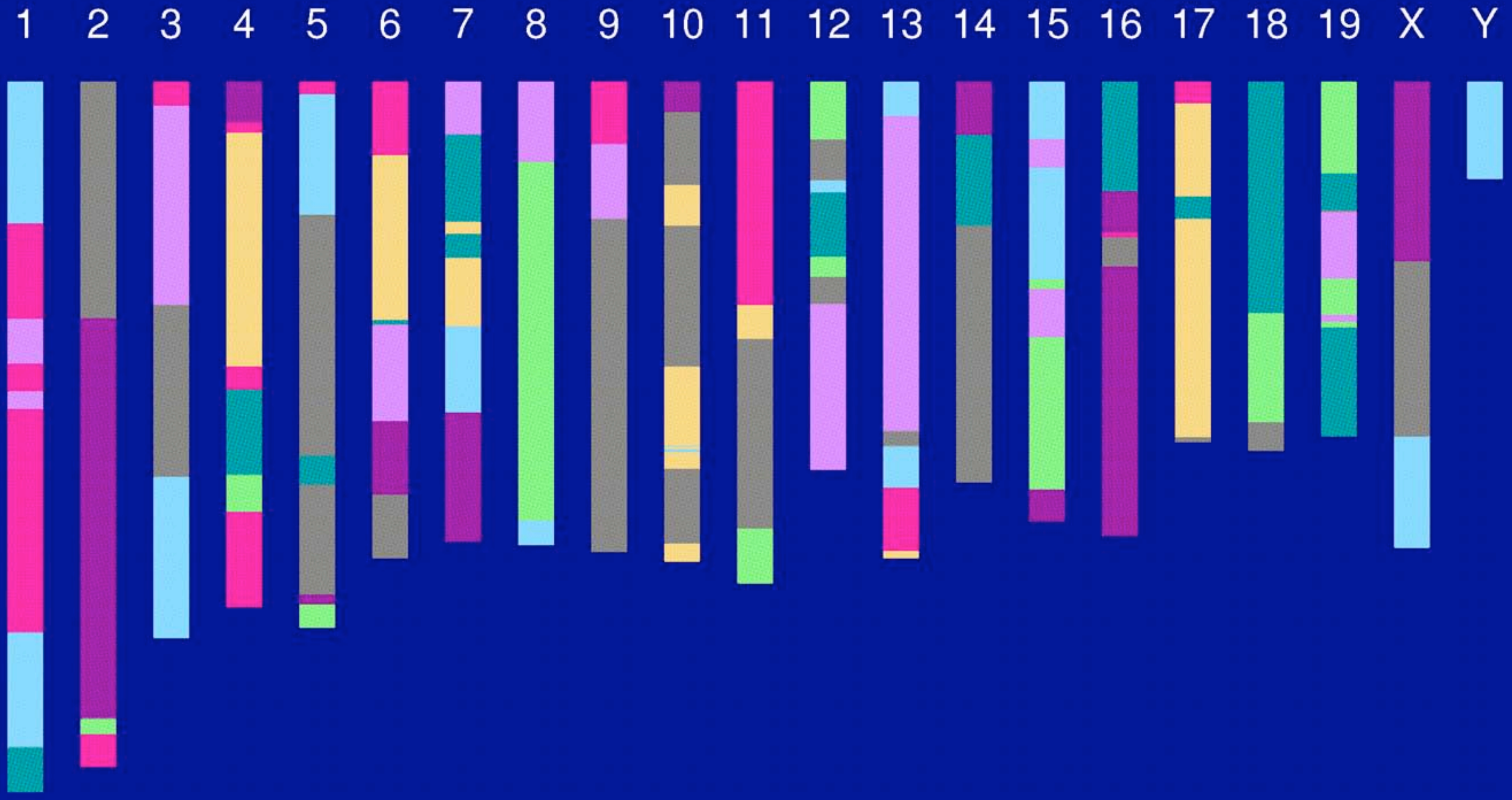
The “Collaborative Cross”



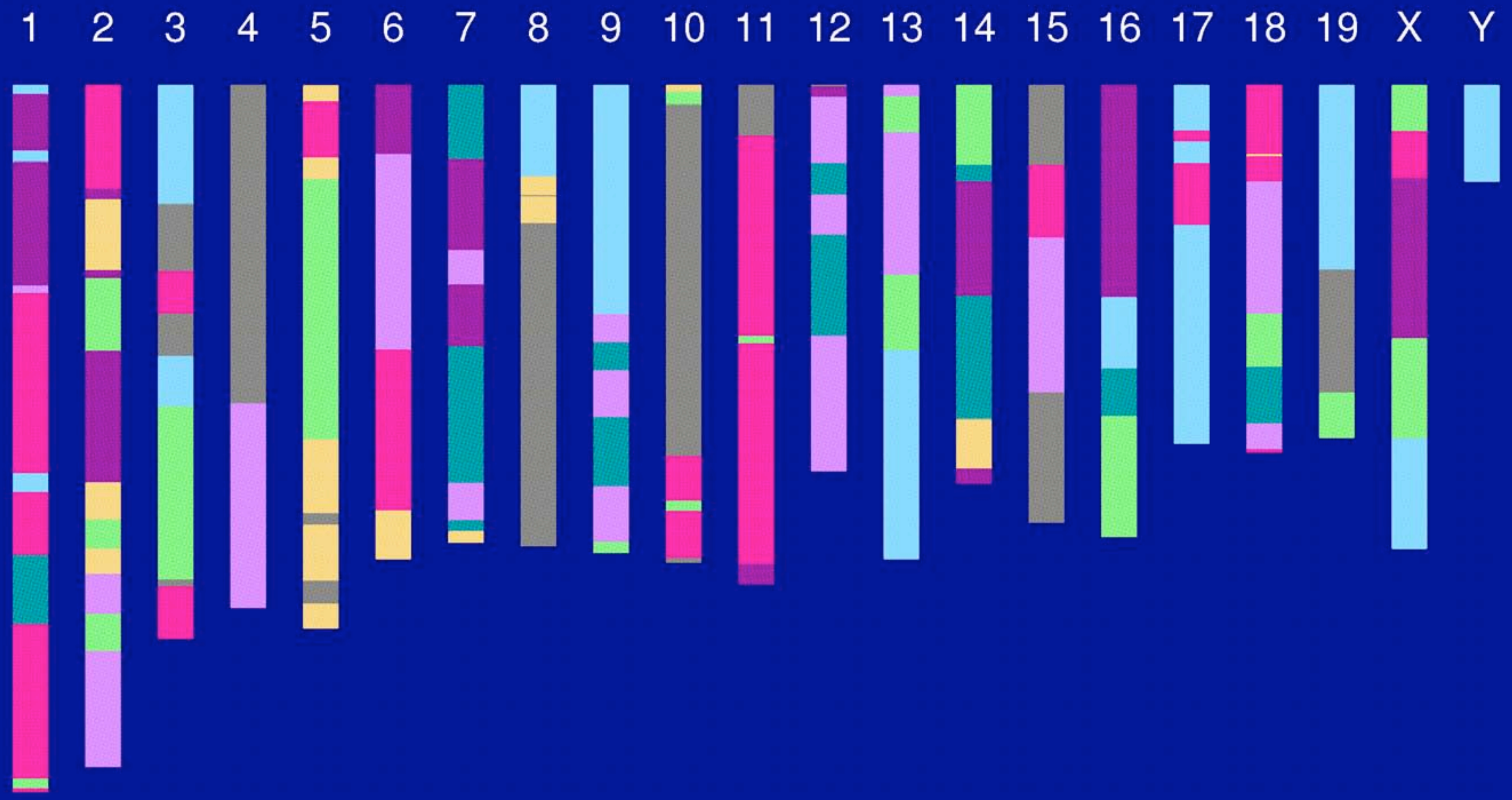
Genome of an 8-way RI



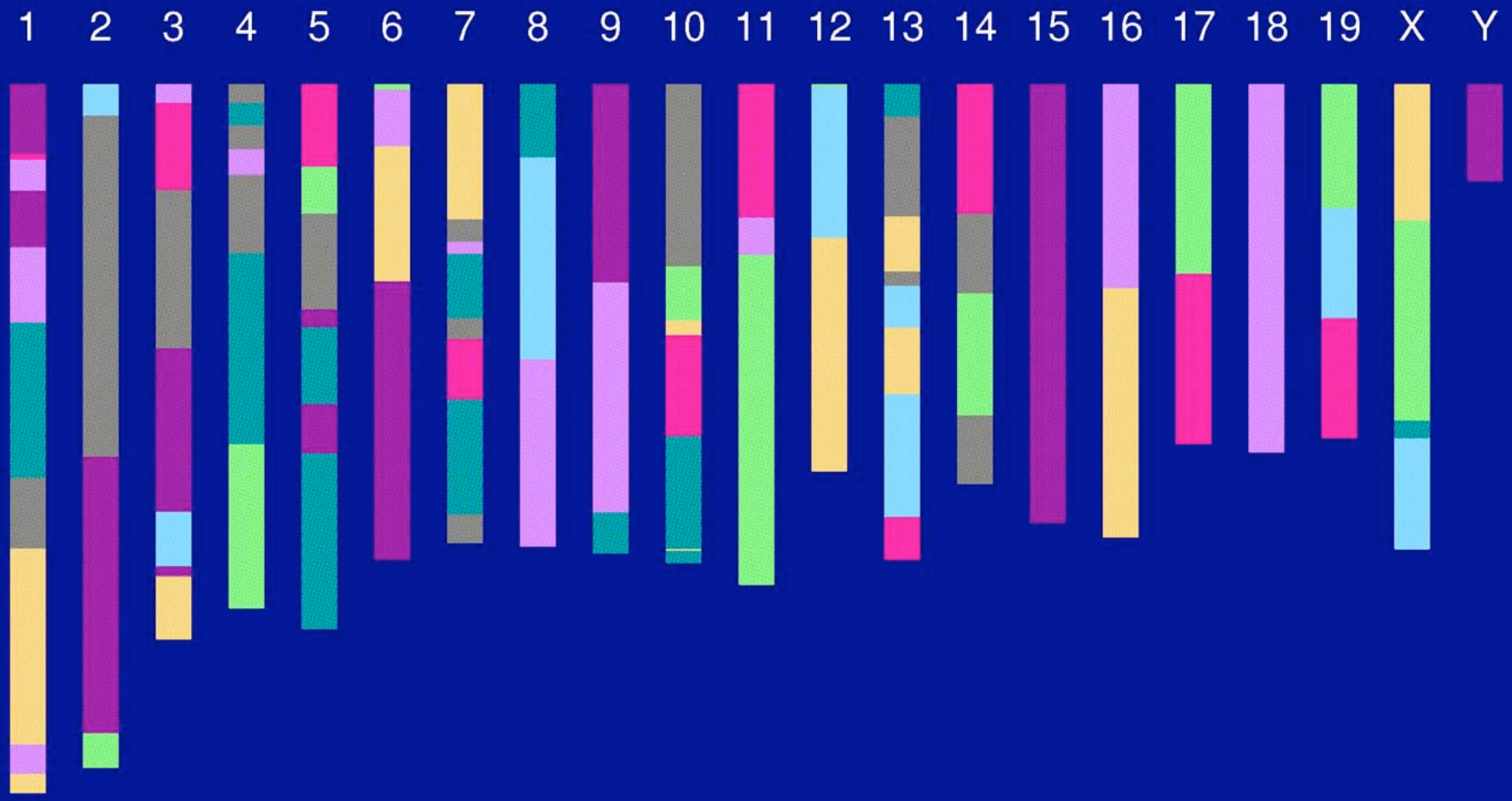
Genome of an 8-way RI



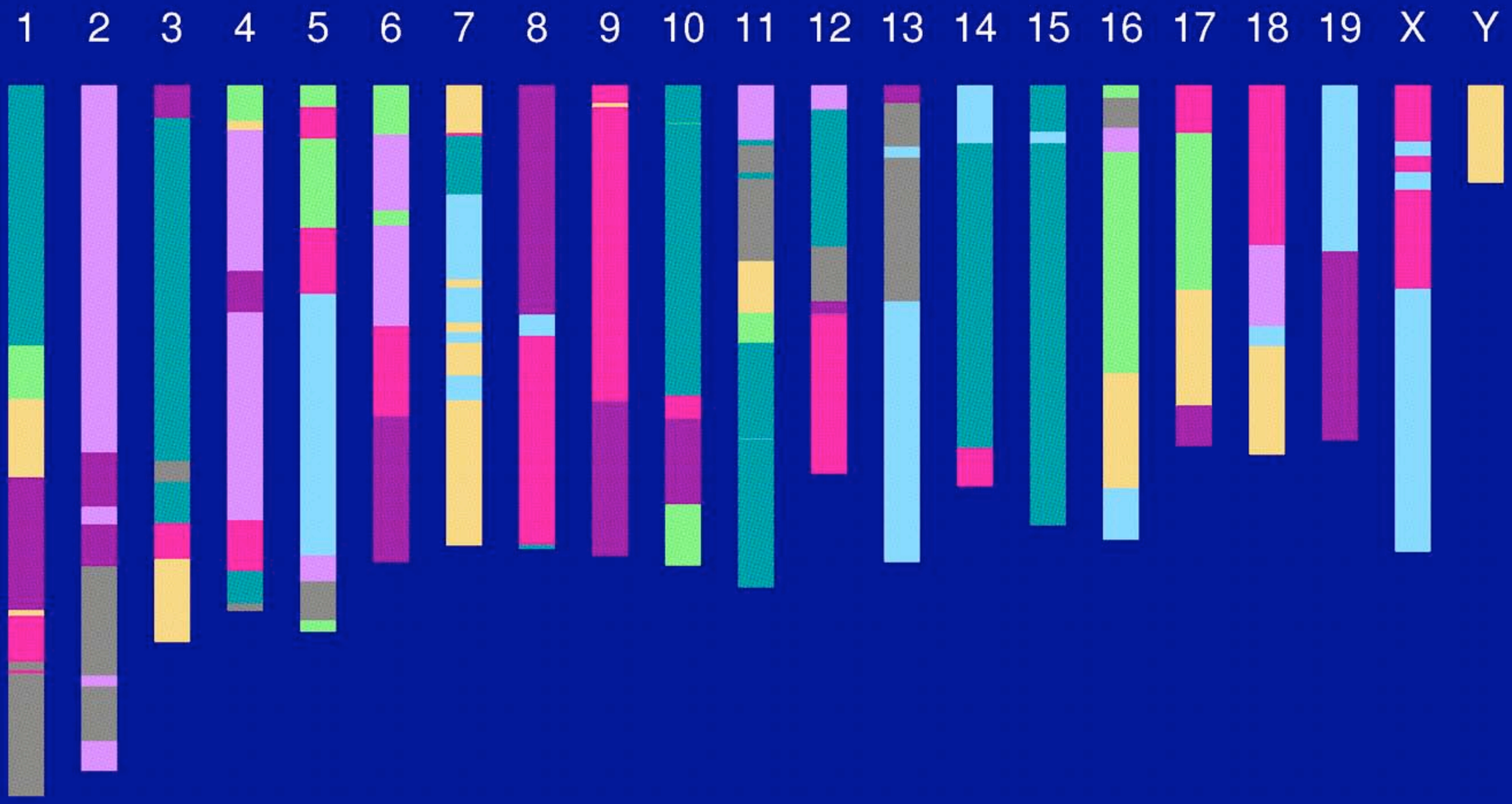
Genome of an 8-way RI



Genome of an 8-way RI



Genome of an 8-way RI



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.

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).

Acknowledgments

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