

SISG 2015

SISG Module 14: QTL Mapping

20th Summer Institute in Statistical Genetics

W UNIVERSITY *of* WASHINGTON

(This page left intentionally blank.)

Introduction to Quantitative Trait Locus (QTL) Mapping

R.W. Doerge

Zhao-Bang Zeng

Summer Institute in Statistical Genetics

1

General Schedule:

Day 1:

Session 1: Introduction, experimental design, segregation analysis

Session 2: Introduction to genetic mapping, estimating recombination

Day 2:

Session 2(cont): Introduction to genetic mapping, estimating recombination

Session 3: Introduction to QTL detection, single marker QTL analysis, linkage analysis

Session 4: Introduction to genetic mapping, map estimation exercise

Session 5: Likelihood functions for single marker analysis, interval mapping

Session 6: Computer lab I: QTL-Cartographer

Day 3:

Session 7: Permutation thresholds; example QTL analysis

Session 8: Composite interval mapping

Session 9: Multiple interval mapping

Session 10: Computer lab II: QTL-Cartographer

Session 11: Introduction to eQTL mapping

2

What is a QTL?

What are QTL?

...and why do we want to find them???

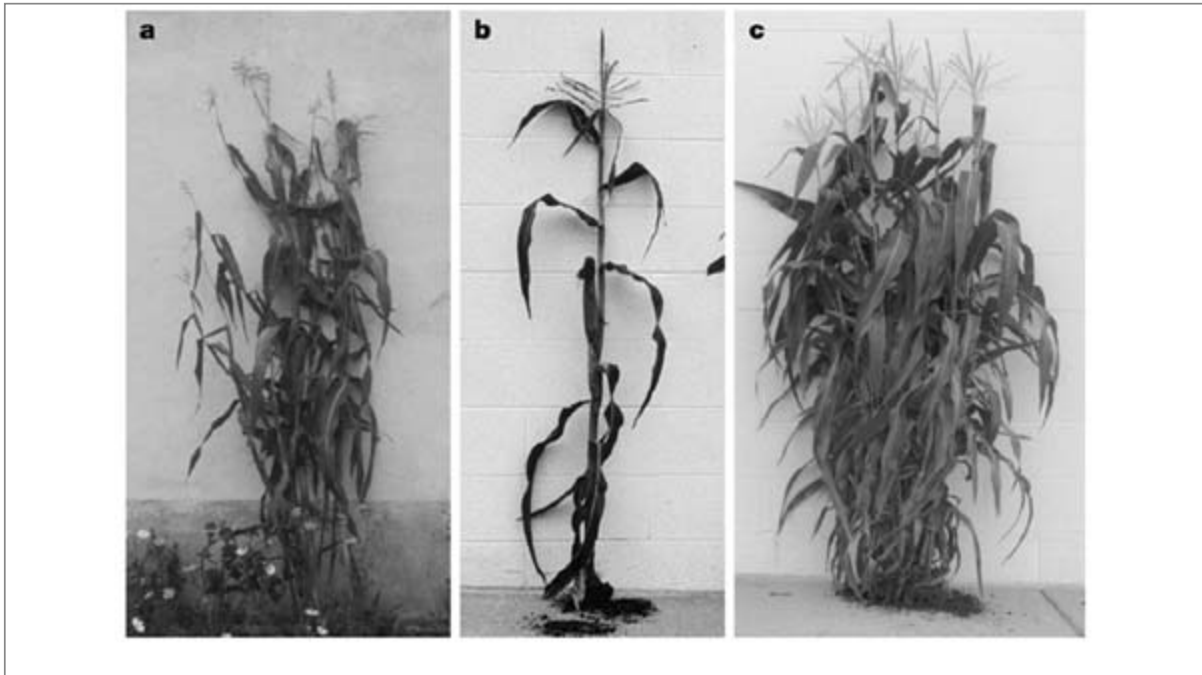
3

QTL analysis in maize . . .

- Cross: teosinte and a primitive variety of maize (F_2 population).
- Result: The chromosome 1 QTL (30% of the phenotypic variance) that affects lateral branching.
 - mapped to within 0.5 cM of a previously known major mutation, *teosinte branched1 (tb1)*
- ***This locus is the first case of a QTL that has been cloned on the basis of its map position.

4

Major Quantitative Trait Locus



5
Doebley & Stec (1991) *Genetics*

What are the data collected for QTL experiment?

- Quantitative trait values, or phenotypes, are collected on every individual in the QTL experiment.
 - height, weight, etc.
 - tens or hundreds of phenotypes collected
- Genetic marker data are collected from every individual in the QTL experiment.
 - hundreds and hundreds of markers available
- Consider one quantitative trait. Each individual i ; has data (X_{ji}, Y_i) where X is the genotype of marker j and Y is the phenotype; $j=1, \dots, m$ and $i = 1, \dots, n$.
 - assess variation in the quantitative trait
 - map quantitative trait variation/information to the genetic map provided by the genetic markers

6

Statistical genetics relies on the level of variation provided by...

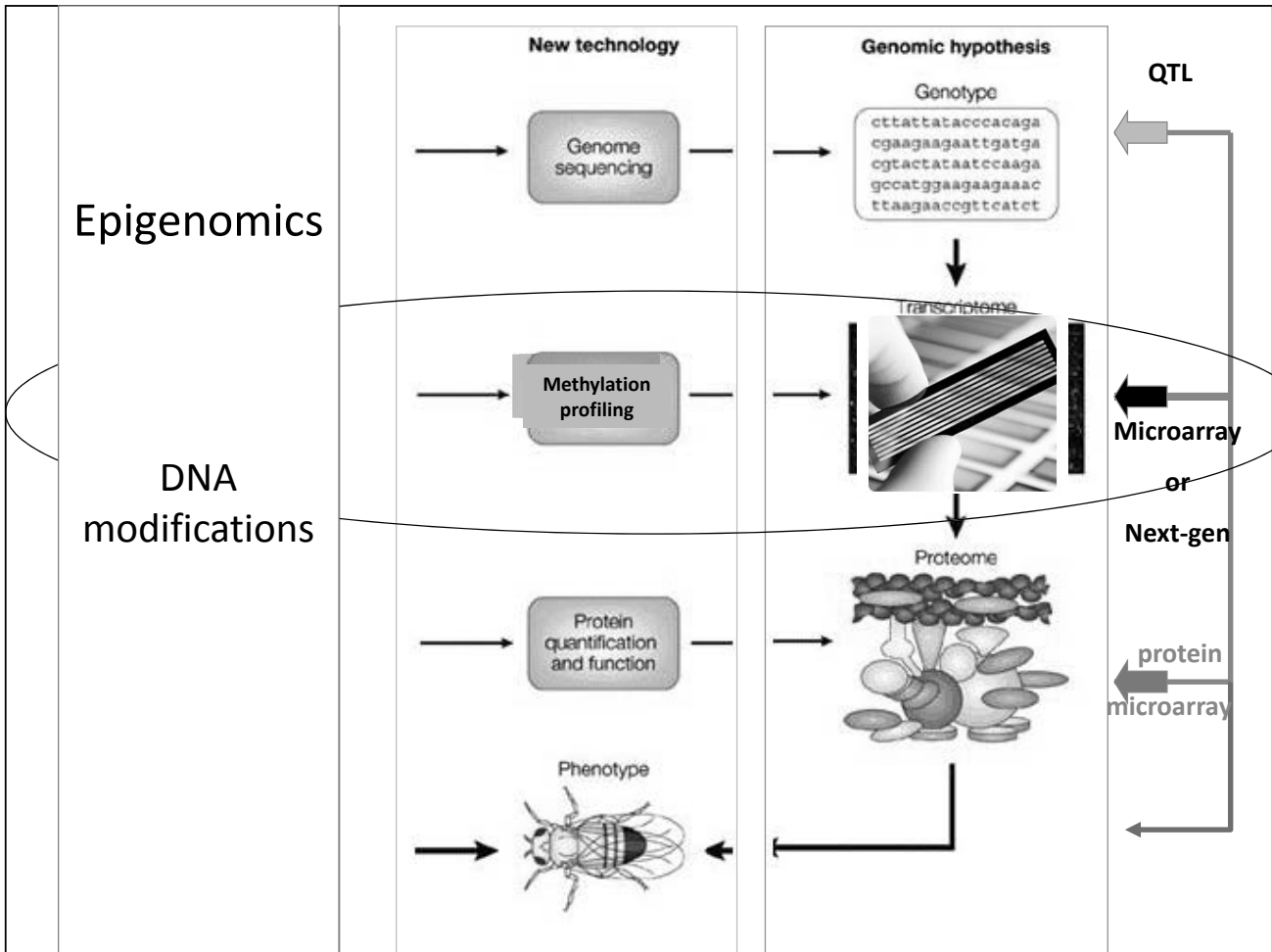
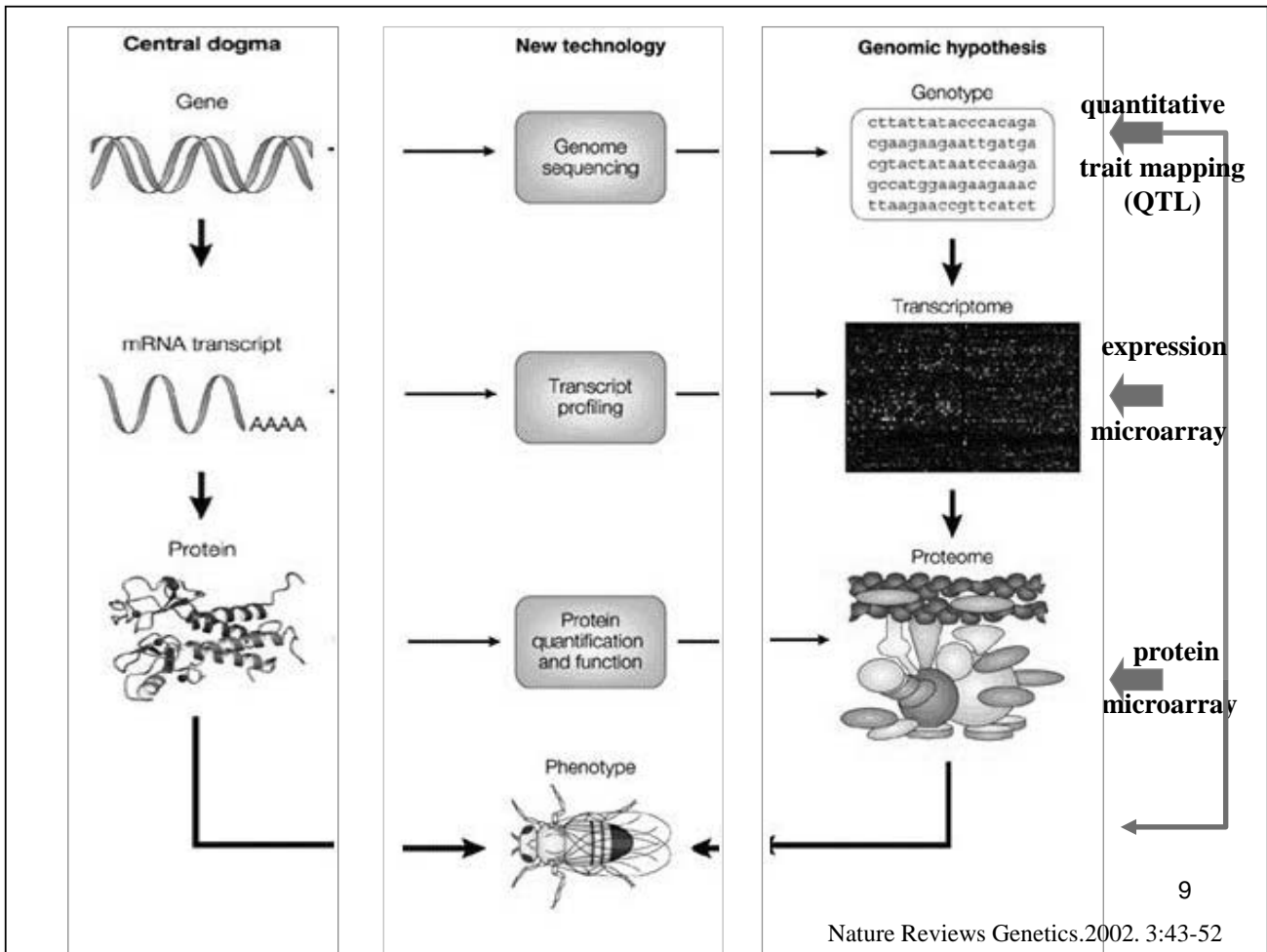
- meiosis (crossover or recombination)
 - assessment of genetic variation
 - genetic map estimation
 - detecting quantitative trait loci
 - locating quantitative trait loci
- Genetics: the basic unit of study is the gene or genetic marker, and we are interested in how these “units” are transmitted from parents to offspring.

7

Statistical genomics

- Genomics: the basic unit of study is the individual “base pairs” that make up a gene, and we are interest in how these base pairs differ between individuals.

8



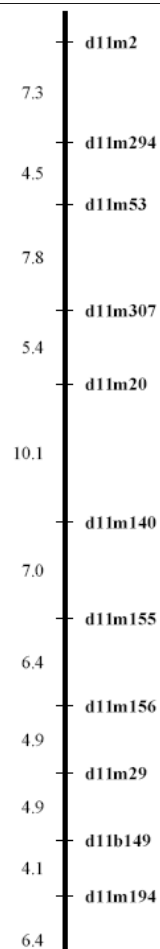
Quantitative Trait Locus or Loci (QTL): Specific regions of the genome that are associated with quantitative traits of interest.

Examples:

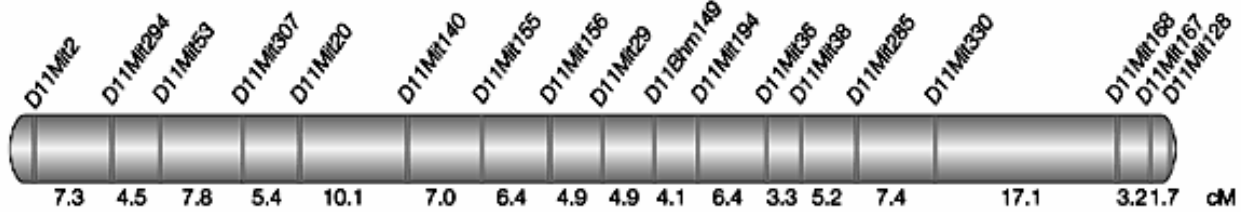


- QTL controlling grown and wood quality traits in *Eucalyptus grandis*
- QTL affecting response to short-term selection for abdominal bristle number in *Drosophila melanogaster*.
- QTL controlling susceptibility to subtypes of experimental allergic encephalomyelitis (EAE), the principal animal model of multiple sclerosis (MS).
- Honey Bee, Tomato, Rice, Sugarcane, Sorghum, Mouse, Wheat, Fern...

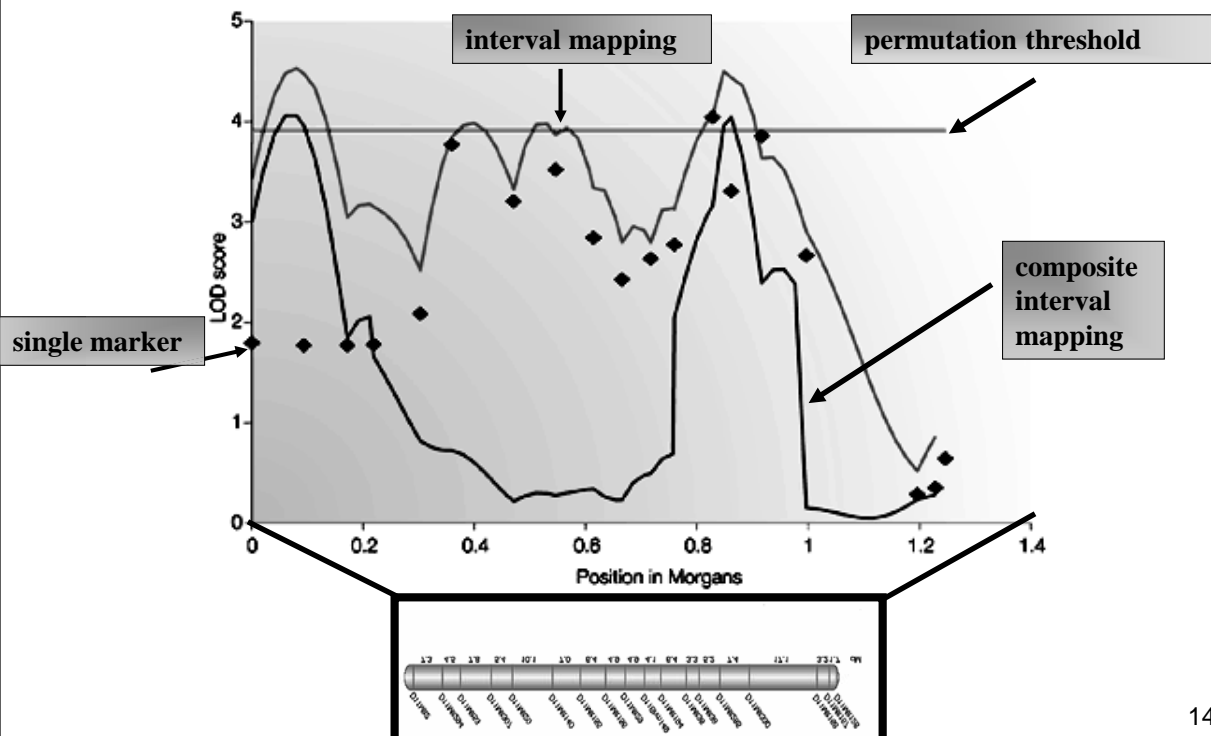
Chromosome chrom11



Estimated Genetic Map



QTL mapping methodology



➔ Why???

- Marker-based selection
- Cloning and characterization of genes.
- Connect with functional genomics?
 - expression QTL (e-QTL)

15

Main Goal:

Our main goals in this module are to:

1. examine and understand the statistical issues surrounding the search for QTL (genes)
2. understand the basic set-up and methodology for QTL mapping; introduce e-QTL
3. gain experience with QTL-Cartographer software;
4. accumulate a working knowledge of how to analyze QTL data for experimental populations
5. understand how a working knowledge of QTL analysis benefits eQTL analysis

16

Current methods for locating QTL:

- Single Marker Methods

- Interval Mapping (Lander and Botstein 1989)
 - Mapping: constructing genetic maps
 - Locating QTL: use the genetic map information to locate QTL

- Composite Interval Mapping (Jansen 1993; Zeng 1993, 1994)
 - Locating QTL: use the genetic map information to locate QTL

17

Statistical issues surrounding the search for QTL

- Hypotheses
- Distribution of Test Statistics
- Multiple tests
- Multiple QTL
- Significance levels

18

Three Basic Steps

1. Experimental design and genetic data
 - vocabulary
 - material
 - understanding the biological process that provides genetic variation
2. Building the “Genetic Map”
 - a genetic map provides the structure for the eventual location of QTL (genes)
 - need to resolve the “order” of the observable genetic markers
3. Locating QTL (genes) for the trait in which we are interested

19

Experimental Design and Data Structure

Zhao-Bang Zeng

Summer Institute in Statistical Genetics

20

QTL Mapping Data

Marker Data:

- Molecular markers: specific patterns of DNA sequences; polymorphic, abundant, neutral, co-dominant or dominant.
 - examples: RFLP, SSR, RAPD, AFLP, VNTR
- Markers data are categorical (i.e., different classifications):
 - presence or absence of a band of molecular segment.
 - the number of categories depends on mapping population and marker type
 - examples:
 - two marker types (homozygote or heterozygote) for backcross population
 - three marker types for F_2 population with co-dominant markers.
- Markers contain information about segregation at various positions of a genome in a population.

21

Quantitative trait data:

- Measurement of a phenotype.
 - examples:
 - 12 week body weight of mouse
 - grain yield of maize
 - little size of pigs
 - blood pressure
 - disease resistant score,
 - expression (traits) from microarrays...
- Continuous or discrete data.
- Quantitative trait data contain information about segregation and effects of QTL in a population.

22

Quantitative Trait Loci

Quantitative Trait Loci (QTL): the regions or genes whose variation has an effect on a trait in a population.

The statistical task of mapping QTL is to detect and estimate the association between the variation at the phenotypic level (trait data) and the variation at genetic level (marker data) in terms of number, positions, effects and interaction of QTL.

23

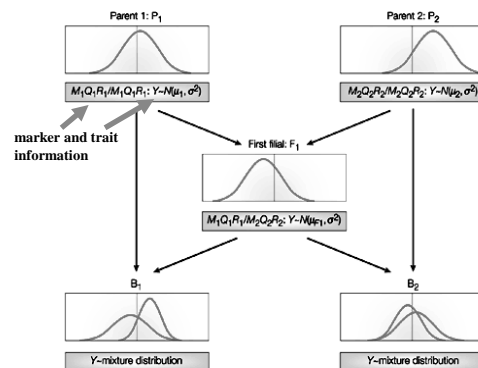
Experimental Designs

Traditional experimental designs for locating QTL start with two parental inbred lines, P_1 and P_2 , differing both in trait values and in the marker (M, N, ...) variants or alleles ($M_1, M_2, N_1, N_2, \dots$) they carry.

- in practice, markers are sought that have different variants (alleles) in the parents.

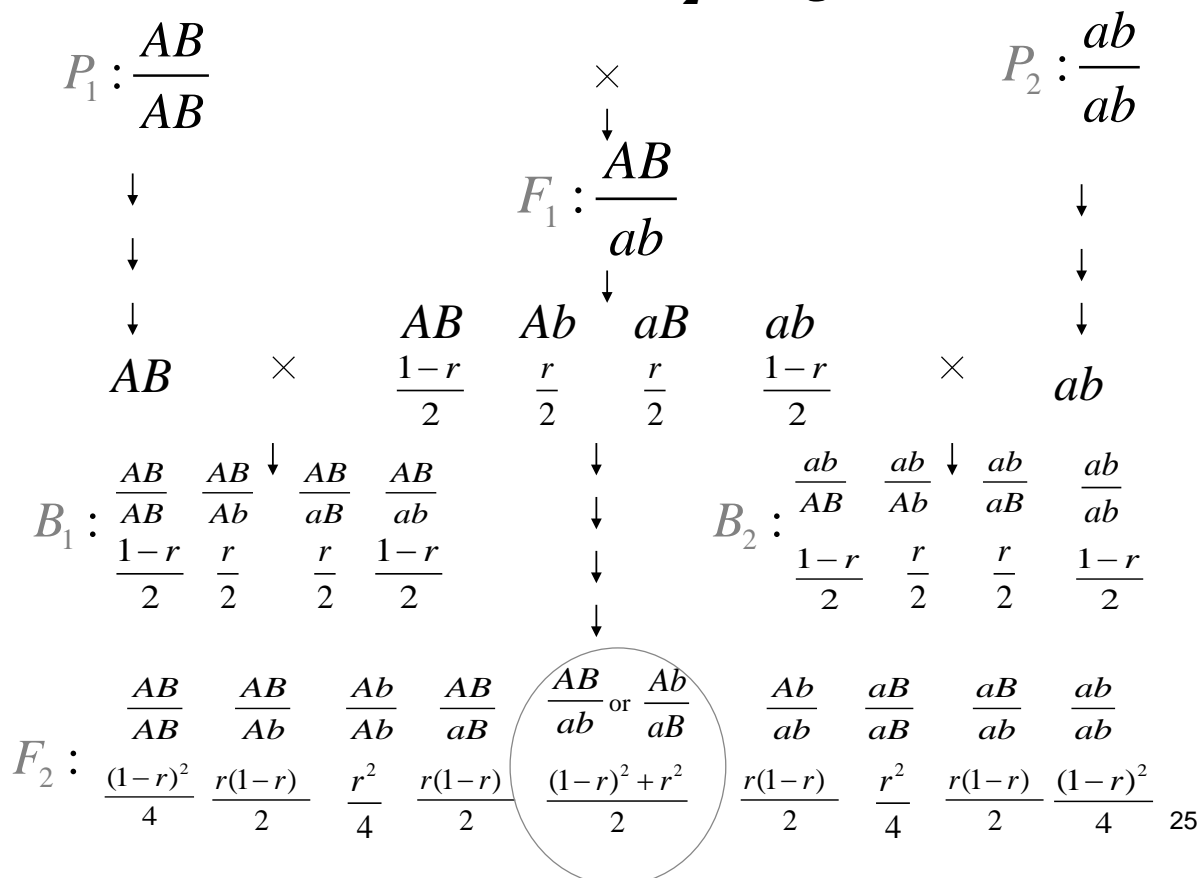
- Advantage:

- F_1 is heterozygote for all loci
 - which differ in P_1 and P_2
- maximum linkage disequilibrium
- for mapping QTL
 - this type of experimental design has the maximum power.



24

Backcross and F₂ Designs



- **Backcross (BC):**

- two genotypes at a locus
- simple to analyze

- **F₂:**

- three genotypes at a locus,
 - can estimate both additive and dominance effects
- more complex for data analysis, particularly for multiple QTL with epistasis (i.e., interaction)
- more opportunity and information to examine genetic structure or architecture of QTL
- more power than BC for QTL analysis

Other commonly used inbred line crosses

- Note: QTL-Cartographer deals with the below, and there are additional details in Lynch and Walsh (1998)
- **Advanced intercross**
 - selfing (SF_t): $F_2 = F_1 \times F_1$; $F_3 = F_2 \times F_2$; $F_4 = F_3 \times F_3$; \dots through selfing):
 - continual selfing (6+ generations) leads to recombinant inbred lines (RI lines)
 - random mating (RF_t):
 - increase recombination
 - expand the length of linkage map
 - increase the mapping resolution (estimation of QTL position)
- Doubled haploid (RI_0): similar to BC and RI in analysis
- Repeated backcross:
 - $B_{1t}, B_{2t}; B_{12}=P_1 \times B_1; B_{13}=P_1 \times B_{12}; B_{14}=P_1 \times B_{13}; \dots$
- Testcross of SF_t or RF_t to P_j
- NC design III:
 - marker genotype data on SF_t and trait phenotype data on both $SF_t \times P_1$ and $SF_t \times P_2$)

Recombinant inbred lines (RI lines):

- selfing ($RI_1 = SF_t, t > 6$)
- brother-sister mating (RI_2)
 - more mapping resolution as more recombination occurs when constructing RI lines.
- May improve the measurement of mean phenotype of a line with multiple individuals, i.e., increase heritability.
 - potentially a very big advantage for QTL analysis,
 - a big factor for power calculation and sample size requirement.

Recombinant Inbred Lines

$$P_1 : \frac{AB}{AB}$$

$$P_2 : \frac{ab}{ab}$$

$$\times$$

$$F_1 : \frac{AB}{ab} \times F_1 : \frac{AB}{ab}$$

$$F_2 :$$

$\frac{AB}{AB}$	$\frac{AB}{Ab}$	$\frac{Ab}{Ab}$	$\frac{AB}{aB}$	$\frac{AB}{ab} / \frac{Ab}{aB}$	$\frac{Ab}{ab}$	$\frac{aB}{aB}$	$\frac{aB}{ab}$	$\frac{ab}{ab}$
$\frac{(1-r)^2}{4}$	$\frac{r(1-r)}{2}$	$\frac{r^2}{4}$	$\frac{r(1-r)}{2}$	$\frac{(1-r)^2 + r^2}{2}$	$\frac{r(1-r)}{2}$	$\frac{r^2}{4}$	$\frac{r(1-r)}{2}$	$\frac{(1-r)^2}{4}$

$$F_\infty : \frac{AB}{AB} \quad \frac{Ab}{Ab} \quad \frac{aB}{aB} \quad \frac{ab}{ab}$$

$$\text{By selfing (RI}_1\text{)} \longrightarrow \frac{1}{2(1+2r)} \quad \frac{2r}{2(1+2r)} \quad \frac{2r}{2(1+2r)} \quad \frac{1}{2(1+2r)}$$

$$\text{By brother-sister mating (RI}_2\text{)} \longrightarrow \frac{1+2r}{2(1+6r)} \quad \frac{4r}{2(1+6r)} \quad \frac{4r}{2(1+6r)} \quad \frac{1+2r}{2(1+6r)} \quad 29$$

Outcross populations

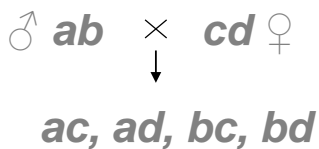
- Cross segregating populations (no inbred available):
 - similar model and analysis procedure used as inbred cross, but more complex in analysis.
 - need to estimate the probability of allelic origin for each genomic point from observed markers.
 - less powerful for QTL analysis
 - QTL alleles may not be preferentially fixed in the parental populations
 - more difficult for power calculation (more unknowns)

Half-sib families

- analyze the segregation of one parent
 - similar to backcross in model and analysis.
- less powerful for QTL detection
 - more uncontrollable variability in the other parents.
- analyze allelic effect difference in one parent, not the allelic effect difference between widely differentiated inbred lines, populations and species.
- generally the relevant heritability is low for QTL analysis.

31

Full-sib families



♀ a	a	b
c	ac	bc
d	ad	bd

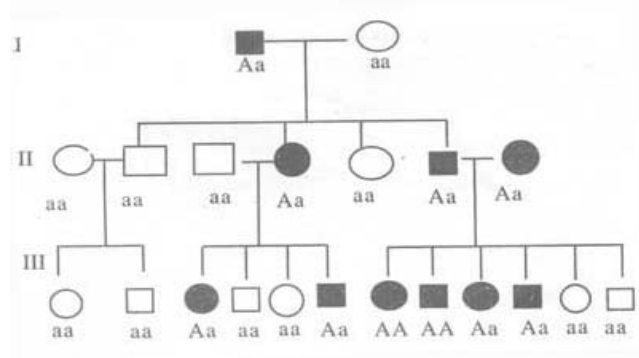
- four genotypes at a locus
 - possible to estimate allelic substitution effects for male and female parents and their interaction (dominance).
 - ♂ $\alpha = [ac + ad] - [bc + bd]$
 - ♀ $\alpha = [ac + bc] - [ad + bd]$
 - $\beta = [ac + bd] - [ad + bc]$
- doubled information for QTL analysis compared to half-sibs
 - should be more powerful.
- Note: If we use the double pseudo-backcross approach for mapping analysis, we do NOT utilize full genetic information,
 - it actually uses less than half the information available.
 - not powerful for QTL identification.
- **Power calculation depends on how the data are analyzed.**

32

Human populations

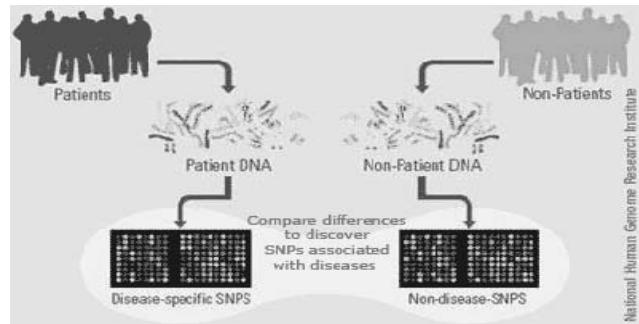
Pedigree :

- limited by sample size
- linkage analysis
 - association between markers and disease locus is due to recent genetic linkage
- suited for Mendelian diseases, not for complex diseases



Case-Control:

- large population available for study
- association analysis
 - association between markers and disease locus is due to historical genetic linkage, and restricted to short regions
- Can be used for complex diseases



33

Designs to reduce sample size and increase statistical power

- Selective genotyping
 - Bulked segregant analysis
- Progeny testing
 - Replicated progeny
 - Granddaughter design

34

Examples: Mapping Data

- Two data sets are used as examples for various analyses:
 - Mouse (Table 1)
 - Maize (Table 2)

35

- **Mouse data** (Dragani et al. 1995 Mammalian Genome 6:778-781):

- backcross population (B_1)
- 103 individuals (sample size $n = 103$)
- 181 microsatellite markers (SSR: simple sequence repeats) distributed across 20 chromosomes
 - including 14 markers on chromosome X
 - chromosome X is used here as an example
- quantitative trait is 12 week body weight (BW)
- Throughout, we use the trait data and marker data on chromosome X to illustrate the analyses of segregation, linkage, single marker analysis, interval mapping and composite interval mapping (which also uses some markers on other chromosomes).

36

Table 1: A sample of a mouse data set (backcross, n=103)

Ind.	BW	Markers (on Chromosome X)													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	50	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	54	1	1	1	1	1	1	1	1	1	1	1	1	1	0
3	49	0	1	1	1	1	1	1	1	1	1	1	1	1	1
4	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	36	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	37	0	0	0	0	0	0	1	1	1	1	1	1	1	1
8	55	1	1	1	1	1	1	1	1	1	1	1	1	1	0
9	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0
:															

1 = AA homozygote genotype; 0 = Aa heterozygote genotype

37

▪ **Maize data:**

- F₂ population.
- 171 lines (sample size n =171)
- 132 markers distributed on 10 chromosomes,
 - including 12 markers on a chromosome used as an example.
- quantitative trait is disease resistant score.
- Eventually, we will use the trait data and the marker data to illustrate segregation, linkage, single marker analysis, interval mapping and composite interval mapping (which also uses some markers on other chromosomes).

- A partial data are shown in Table 2.

38

Table 2: A sample of a maize data set (F_2 , $n=171$)

Ind.	Trait	Markers											
		1	2	3	4	5	6	7	8	9	10	11	12
1	6.25	2	1	0	0	0	0	0	1	1	1	0	0
2	3.00	1	1	1	1	1	2	2	2	2	0	0	1
3	3.00	1	2	2	2	2	1	1	1	1	2	2	2
4	4.00	1	0	0	0	0	0	0	0	0	1	2	2
5	3.00	0	0	1	1	1	1	1	1	1	1	1	1
6	3.75	1	0	0	0	0	1	1	1	1	0	0	0
7	8.25	2	2	2	1	1	0	0	0	0	1	2	2
8	2.50	0	0	0	0	0	0	0	1	1	1	1	2
9	4.25	1	0	1	1	1	1	1	0	0	1	1	1
10	4.50	0	1	1	1	1	1	1	1	0	0	0	0
	⋮												

2 = AA homozygote genotype of P1; 1 = Aa heterozygote genotype; 0 = aa homozygote genotype of P2

39

Segregation Analysis

Zhao-Bang Zeng

Summer Institute in Statistical Genetics

40

Understanding the inheritance of markers: segregation analysis

- Statistically test whether markers are segregating independently
 - no external forces acting on the population
 - random mating
 - no mutation
 - no selection

- Employ a chi-square test

41

Testing Mendelian Segregation

Backcross population: cross between **A/A** and **A/a** produces the following zygotes

	<u>A/A</u>	<u>A/a</u>
Frequency under H_0	1/2	1/2
Expected number	$n/2$	$n/2$
Observed number	n_1	n_2

A test statistic can be constructed by using χ^2 under the null hypothesis $p(A/A) = p(A/a) = 1/2$ (Mendelian Segregation).

$$\chi^2 = \sum \frac{(\text{Obs. \#} - \text{Exp. \#})^2}{\text{Exp. \#}} = \frac{(n_1 - n/2)^2}{n/2} + \frac{(n_2 - n/2)^2}{n/2} = \frac{(n_1 - n_2)^2}{n}$$

Under the null hypothesis, this statistic is chi-square distributed with 1 degree freedom. $\alpha = 0.05; \chi_1^2 = 3.84$

42

Table 3: Example of testing Mendelian segregation: Mouse data

Marker	n_1	n_2	χ^2	P value
1 Hmg1-rs13	41	62	4.282	0.038
2 DXMit57	42	61	3.505	0.061
3 Rps17-rs11	43	60	2.806	0.094
4 Rps18-rs17	42	61	3.505	0.061
5 DXMit48	43	60	2.806	0.094
6 DXNds1	44	59	2.184	0.142
7 DXMit109	45	58	1.641	0.20
8 Hmg14-rs6	49	54	0.243	0.61
9 DXMit60	50	53	0.087	0.77
10 DXMit16	50	53	0.087	0.77
11 DXMit97	50	53	0.087	0.77
12 Hmg1-rs14	51	52	0.010	0.92
13 DXMit3	56	47	0.786	0.38
14 Tpm3-rs9	49	54	0.243	0.61

43

F₂ population: A cross between **A/a** and **A/a**. The distribution of zygotes is as follows:

	<u>A/A</u>	<u>A/a</u>	<u>a/a</u>
Frequency under H ₀	1/4	1/2	1/4
Expected number	n/4	n/2	n/4
Observed number	n_1	n_2	n_3

Under the null hypothesis (Mendelian Segregation)

$$p(A/A) = p(A/a) = 1/4 \text{ and } p(a/a) = 1/4$$

$$\chi^2 = \frac{(n_1 - n/4)^2}{n/4} + \frac{(n_2 - n/2)^2}{n/2} + \frac{(n_3 - n/4)^2}{n/4} \sim \chi^2_2$$

Under the null hypothesis, this statistic is chi-square distributed with 2 degrees of freedom.

44

Table 4: Example of testing Mendelian segregation: Maize data

Marker	n_2	n_1	n_3	χ^2	P value
1	43	86	42	0.018	0.99
2	48	89	34	2.579	0.28
3	42	92	37	1.281	0.52
4	44	89	38	0.708	0.70
5	43	87	41	0.099	0.95
6	43	83	45	0.193	0.91
7	44	83	44	0.146	0.93
8	47	81	43	0.661	0.72
9	41	86	44	0.111	0.95
10	40	94	37	1.795	0.40
11	45	89	37	1.035	0.61
12	46	85	40	0.427	0.80

45

Segregation Distortion

- Deviation from Mendelian Segregation is called segregation distortion.
- Significant segregation distortion can
 - bias estimation of recombination frequency between markers
 - reduce the power to identify QTL
 - bias the estimation of QTL position and effect

46

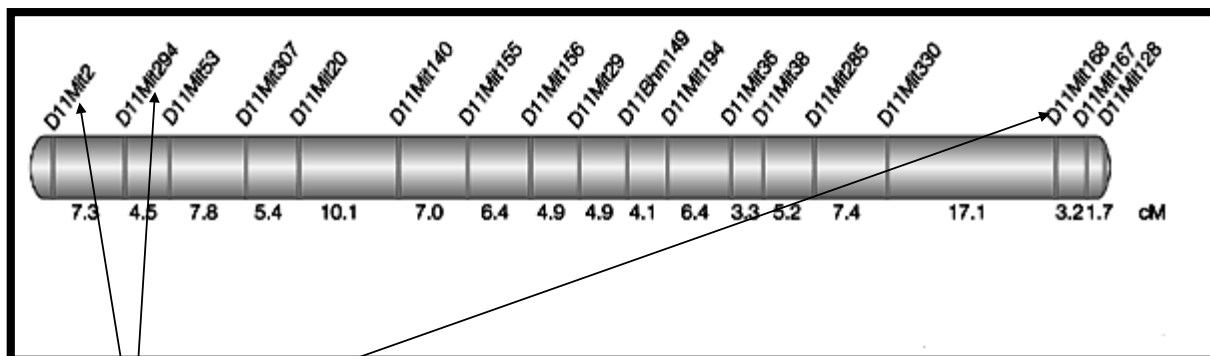
Introduction to Genetic Mapping

R.W. Doerge

Summer Institute in Statistical Genetics

47

Estimated Genetic Map (framework for QTL mapping)



***need to understand how each marker segregates, then we can estimate a genetic map

Vocabulary

- **recombination:** the transmission to progeny combinations of alleles different from those received by a parent, due to independent assortment of crossing over.
- **crossing over:** the exchange of genetic material between homologous chromosomes.
- **Morgan (unit):** a unit for expressing the relative distance between genes (or markers) on a chromosome. The distance on a genetic map between two loci for which one crossover event is expected per gamete per generation.
- **Map unit or Centimorgan (cM):** a map unit is 0.01 Morgans.
- **Interference:** the lack of independence between crossover events in different (nearby) regions.

49

Genetic Markers

Genetic markers are specific aspects of DNA. Specific patterns in the DNA. There are many ways to find these “patterns” or sequences through molecular genomic techniques.

- RFLP: restriction fragment length polymorphism (co-dominant)
- RAPD: randomly amplified polymorphic DNA (dominant)
- VNTR: variable number of tandem repeats
- AFLP, SSR (microsatellites), etc.
- SNP: single nucleotide polymorphism
- SFP: single feature polymorphism
- etc.

50

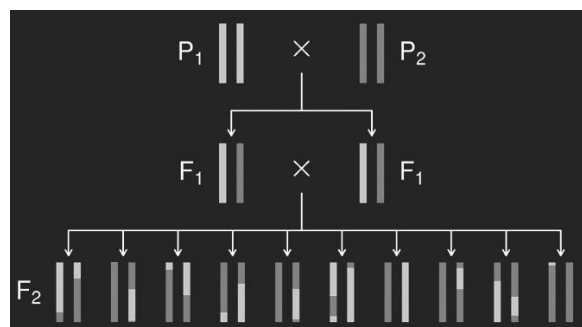
Genetic Markers, Genotypes, and Statistical Variation

- The state of specific genetic marker is called the “genotype”.
- Individuals sharing the same parents may have different genotypes for the same genetic marker.
 - these differences provide the variation we need to statistically estimate the relationship between genetic markers for the purpose of resolving their “linear” order (or genetic map) across chromosomes.

51

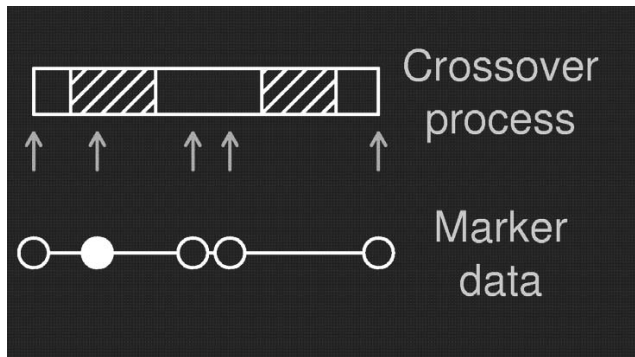
Recombination

- During the production of “gametes” an exchange of material (cross over) between pairs of chromosomes may occur.
 - eggs or sperm: each will eventually contain half the normal chromosome number of a “diploid” organism
- Occurs in the Prophase I stage of Meiosis.
- The result of meiosis is the formation of “haploid” cells containing one set of “unique chromosomes”.



52

Recombination fraction



- Recombination across an interval indicates an odd number of crossovers
- locations of crossovers are not observed.

Recombination fraction:

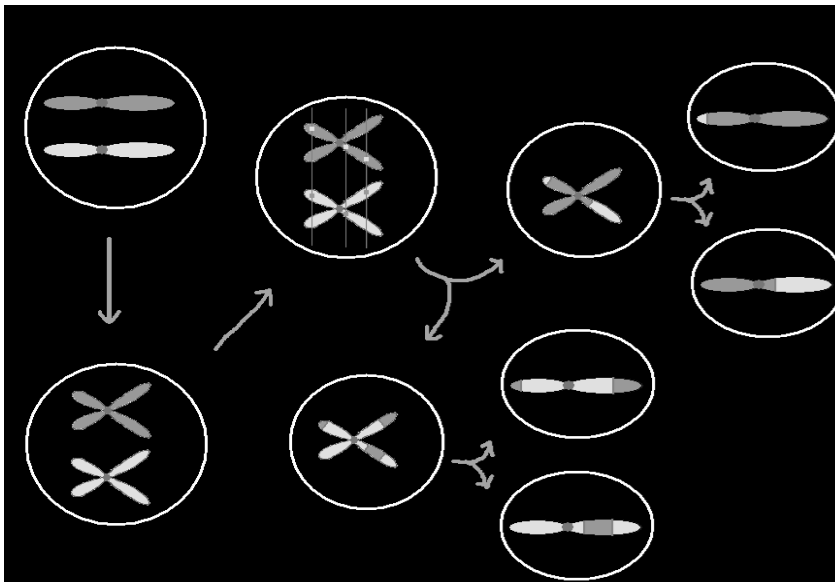
Probability of recombination in an interval = Probability of an odd number of crossovers in interval

Illustration: K. Broman

53

Connecting Genetic Material, Lab Results, and Statistics

1. DNA: long stretches of base pairs.
2. We understand how (basically) the variation occurs during meiosis.
3. The variation (recombination) can be detected using laboratory techniques (i.e., the genotypes of genetic markers are observed).



54

FACTS:

- The closer two markers are the less likely a recombination event is to occur.
- Markers that reside on different chromosomes undergo free recombination.
- Two markers that never experience a recombinant event between them are said to be “completely linked”.
 - they travel together during the meiosis process.
- If an even number of crossing over events occurs between two genetic markers, this event is undetectable.

55

Next step...

- Once each marker is tested for independent segregation (and passes), then the task becomes (linearly) ordering the markers into linkage groups or chromosomes.
 - equivalent to the traveling salesman problem in mathematics.
 - requires a measure of distance between pairs of markers
 - this distance is a function of recombination
 - A map function translates between recombination and genetic distance
 - Haldane map function
 - Kosambi map function

56

Estimating Recombination

R.W. Doerge

Summer Institute in Statistical Genetics

57

Estimating Recombination Between Two Genetic Markers

The number of (odd) crossovers (k) in an interval defined by two genetic markers has a Poisson distribution with mean θ .

$$\begin{aligned}\text{Pr}(\text{recombination}) &= \sum_k \frac{\theta^k \exp^{-\theta}}{k!} \\ &= \exp^{-\theta} \left(\frac{\theta}{1!} + \frac{\theta^3}{3!} + \dots \right) \\ &= \frac{\exp^{-\theta} (\exp^{\theta} - \exp^{-\theta})}{2} \\ &= \frac{1}{2} (1 - \exp^{-2\theta})\end{aligned}$$

- call this probability r with limits $0 \leq r \leq \frac{1}{2}$
- θ is the number of map units (M) between two markers

58

Haldane Map Function:

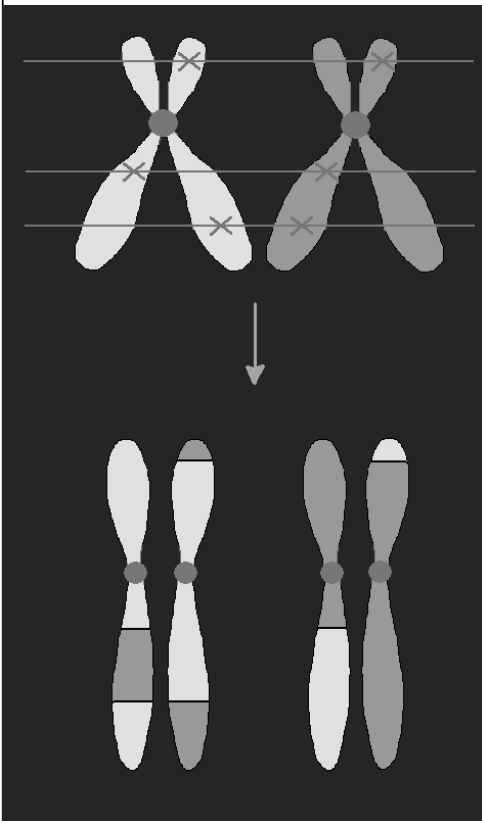
Solving the previous equation for θ gives Haldane's map function:

$$\theta = -\frac{1}{2} \ln(1 - 2r)$$

- Let $r \rightarrow 0, \theta = 0$ (completely linked)
- Let $r \rightarrow \frac{1}{2}, \theta = \infty$ (markers are unlinked)
 - markers on the same chromosome, far apart
 - markers reside on different chromosomes

59

Crossover interference



- Strand choice
 - chromatid interference
- Spacing
 - crossover interference
- Positive crossover interference:
 - crossovers tend not to occur too close together.

60

Kosambi Map Function:

If **interference** is taken into account, the Kosambi map function should be used:

$$\theta = \frac{1}{4} \ln \left(\frac{1 + 2r}{1 - 2r} \right)$$

- As two loci (or markers) become further apart, the amount of interference allowed by the Kosambi map function decreases.

61

Haldane versus Kosambi:

- Haldane map function assumes that crossover events are independent.
 - as the loci (genetic markers) become further apart, recombination increases from 0.0 to 0.50.
- Kosambi map function assumes there is *interference*
 - one crossover tends to prevent other crossovers in the same or close regions
 - for unlinked loci, interference is 0.
- When the genetic distance is small (less than 10cM), both Kosambi and Haldane map functions provide, essentially the same values.

62

Estimating Recombination from Experimental Data

- Estimate the probability of recombination between each pair of genetic markers
 - pairwise recombination estimates
- Recall: Recombination occurs in the F_1 generations, transmitted in the F_1 gametes, and is detectable in the final generation.
 - backcross, F_2
- Use a genetic map function to convert recombination (probability) to genetic distance (additive).

63

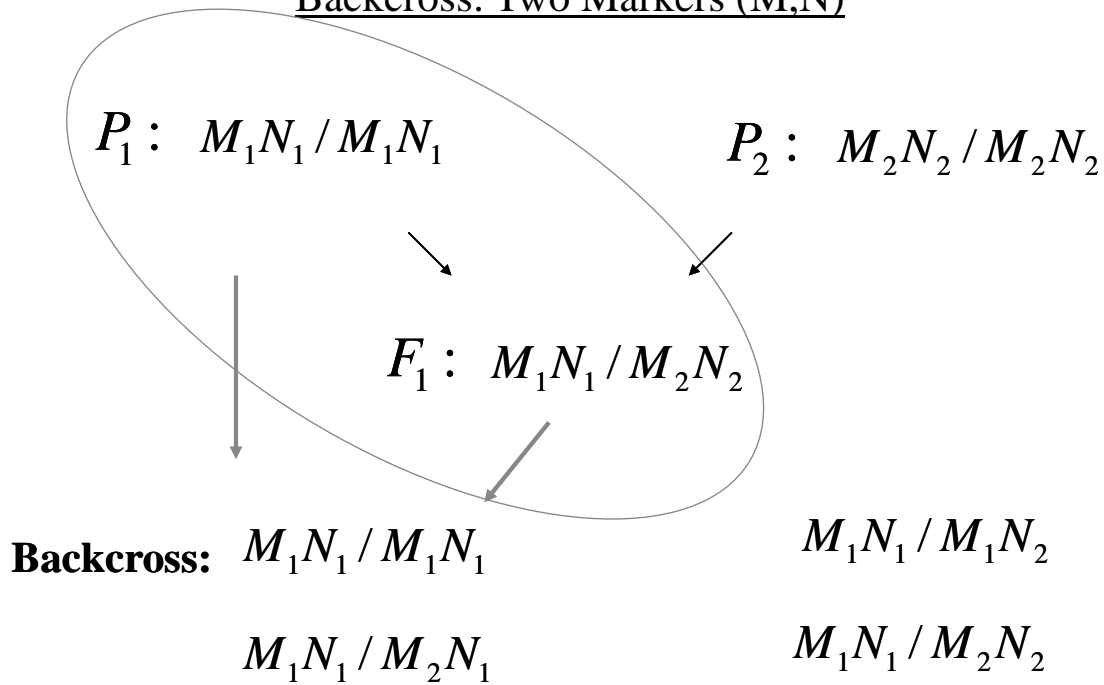
Genetic distance

- The genetic distance between two markers (in cM) is the average number of crossovers in the interval in 100 meiotic outcomes.
- Recombination rate varies by
 - organism
 - sex
 - chromosome
 - position on chromosome

64

WHAT WE SEE

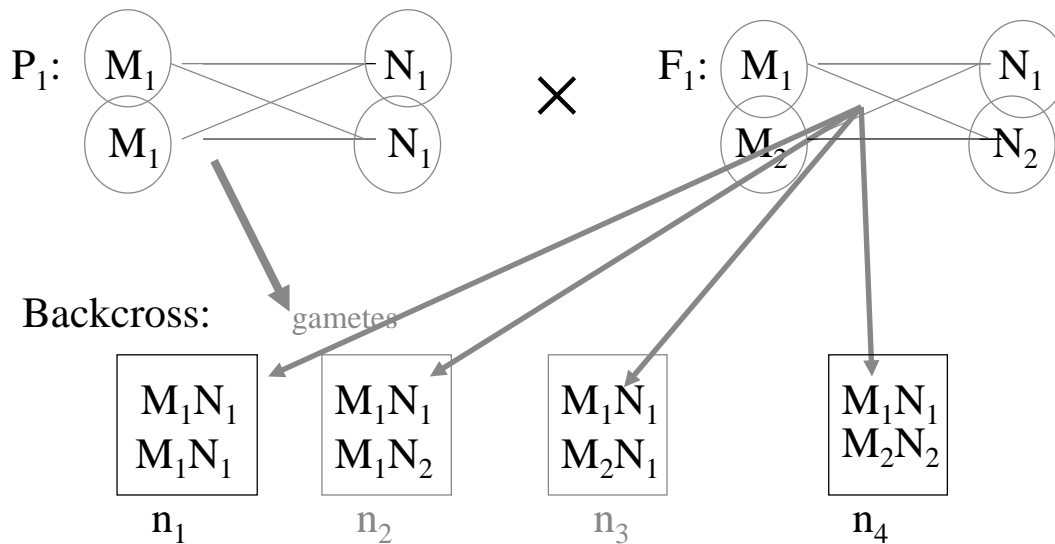
Backcross: Two Markers (M,N)



65

A View of Crossing Over...

Recall... recombination occurs in the parents, and the result is observed in the offspring.



$n_1+n_4 =$ non-recombinant class

$n_2+n_3 =$ recombinant class

66

Counting Recombinants Between Two Markers (Backcross)

- Assume we have two markers **M** and **N**, each having two alleles:
 - M_1, M_2 and N_1, N_2 .
- The possible “genotypes” of the two genetic markers are:
 - M_1/M_1 and M_1/M_2
 - N_1/N_1 and N_1/N_2

67

Or... we can derive it...

The likelihood function describing the backcross situation:

$$L(r) = Cr^{n_2+n_3} (1-r)^{n_1+n_4},$$

where C is the binomial distribution constant $n_1+n_2+n_3+n_4$ choose n_2+n_3 (i.e., $\binom{n_1+n_2+n_3+n_4}{n_2+n_3}$).

The MLE is

$$\hat{r} = \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4}$$

68

Maximizing the Likelihood:

- The likelihood function:

$$L(r) = Cr^{n_2+n_3} (1-r)^{n_1+n_4}$$

- Take the natural logarithm (or log base 10):

$$\ln L(r) = \ln C + (n_2 + n_3) \ln r + (n_1 + n_4) \ln(1-r)$$

- The first partial derivative is the slope of a function.
 - the slope will be zero at the maximum (global/local and/or minimum)
 - check the second derivative to ensure maximum

69

- The partial derivative with respect to r is:

$$\frac{\partial \ln L(r)}{\partial r} = \frac{n_2 + n_3}{r} - \frac{n_1 + n_4}{(1-r)} = 0$$

- Solve this equation for r :

$$\frac{n_2 + n_3}{r} = \frac{n_1 + n_4}{(1-r)}$$

- The MLE is

$$\hat{r} = \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4}$$

70

Example Data Set

data type backcross
100 120 1

***M1** H A A A H H H H H A A A H A A A H A H H A A A A H H H A A H A H H
H A H A A A A A H H A H A H A H A H A H H A H H A H A H H H H A H A H H A H
A H A A A A H A H A A H H H A H H H H H H A A A H A H

***M2** H A A A H H H H H A A A H A A A H A H H A A A A H H H A A H A H H
H A H A A A A A H H A H A H A H A H A H H A H H A H A H H H H A H A H H A H
A H A A A A H A H A A H H H A H H H H H H A A A H A H

.
. .
.

***M120** H H A H H H H H A H H A A H A A H H A A H H H A A H H A H H H H
H H A A A H H A H H A H H H A A A H A H A H H A A A A H H A A H H H H H H H A
H H A A H A H H H H H A H H A H A H H A A H H H A H H H

***trait**

9.5512 10.8668 11.0566 10.0179
11.1773 11.7145 9.9619 11.2285
11.9308 10.5303 9.8150 11.0253
13.3965 9.8091 11.8568 11.8308
...
11.5215 11.2149 9.4704 10.2907
12.2647 11.4211 10.2202 9.8874

Example: estimating pairwise recombination

using data example from previous slide

H = Marker is heterozygous:

marker M1: M1₁/M1₂

marker M2: M2₁/M2₂

A = Marker is homozygous

marker M1: M1₁/M1₁

marker M2: M2₁/M2₁

		A	H
		M1 ₁ /M1 ₁	M1 ₁ /M1 ₂
A	M2 ₁ /M2 ₁	<i>n</i> ₁	<i>n</i> ₂
H	M2 ₁ /M2 ₂	<i>n</i> ₃	<i>n</i> ₄

- Total number of recombinant events is $n_2 + n_3$

- $$\hat{r} = \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4}$$

Pairwise recombination between every pair of markers

- Consider two markers M and N:
- FACTS:
 - A “linkage group” is a group of markers where each marker is linked ($r < .50$) to at least one other marker.
 - If a marker is not linked to any marker in a linkage group, it does not belong in that group, and most likely belongs to some other linkage group.
 - $r_{MN} > 0$ for $M \neq N$
 - $r_{MN} = 0$ for $M = N$
 - $r_{MN} = r_{NM}$
 - Let O be a third marker, $M - N - O$; $r_{MO} \leq r_{MN} + r_{NO}$
 - recombination fractions are not additive

73

Introduction to Quantitative Trait Loci Detection: Hypotheses and Single Marker QTL Analysis

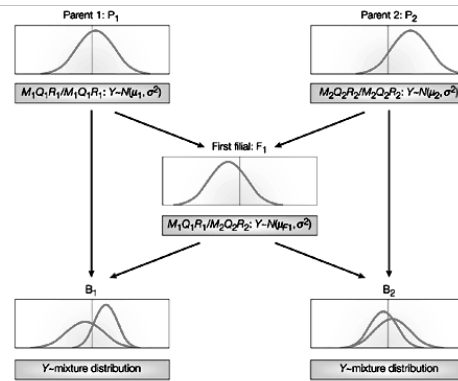
R.W. Doerge

Summer Institute in Statistical Genetics

74

Data Notation

- Assume:
 - backcross experimental design
- Many genetic Markers ...
 - consider marker M with alleles M_1 and M_2
 - every marker has 2 states:
 - homozygous: M_1/M_1
 - heterozygous: M_1/M_2
 - trait Y
- The unknown quantity is the genotype of the QTL.
 - denote the QTL by Q with alleles Q_1 and Q_2



75

Phenotype Data (backcross)

- Measured quantitative trait values can be described via a line:

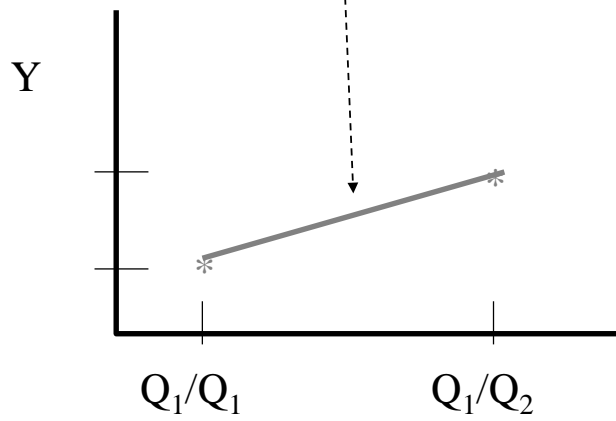
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i; \quad \varepsilon \sim N(0, \sigma^2)$$

- The trait value Y_i is related to the QTL genotype,
 - the indicator variable X_i takes the value 1 or 0 according to whether individual Y_i has QTL genotype Q_1/Q_1 ($X_i = 0$) or Q_1/Q_2 ($X_i = 1$)
- Idea: when testing for a relationship between a marker and a QTL
 - consider the two QTL genotypic classes
 - Q_1/Q_1 and Q_1/Q_2

76

... effect of the QTL

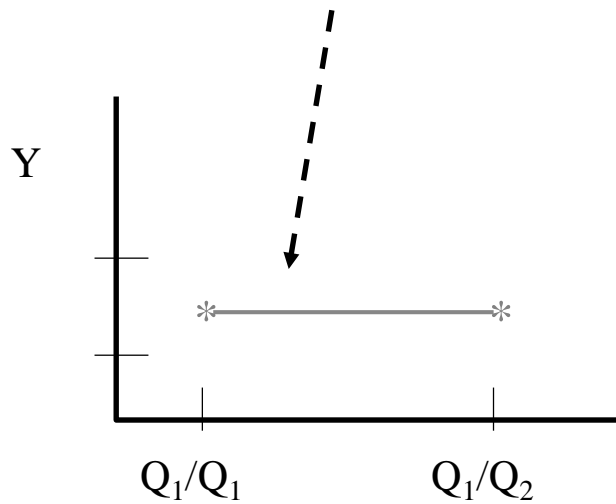
$$Y_i = \beta_0 + \beta_1 X_i$$



77

...no QTL effect

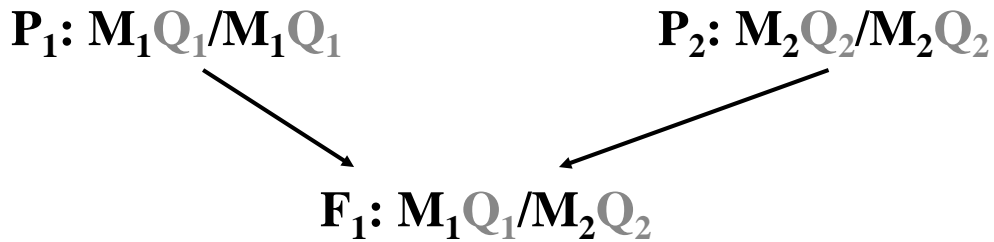
$$Y_i = \beta_0 + \beta_1 X_i$$



78

Reality: QTL genotype is unknown, marker genotypes are known... use marker information. If the marker (M) is linked to the QTL, knowing the marker is like knowing the QTL.

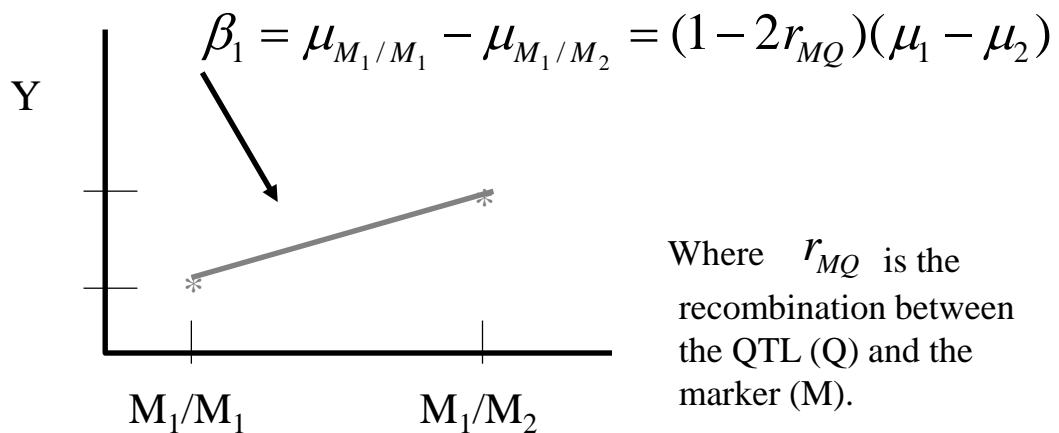
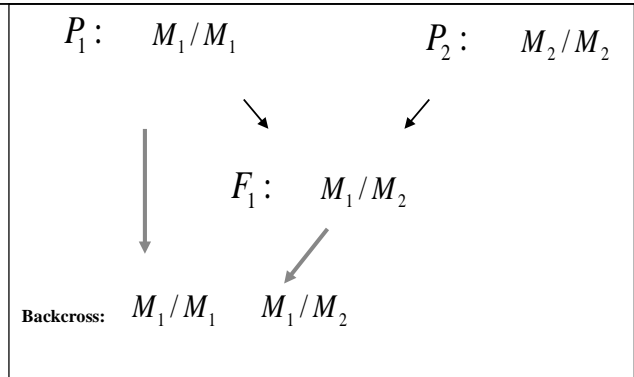
Recall:



Backcross:

...effect of QTL

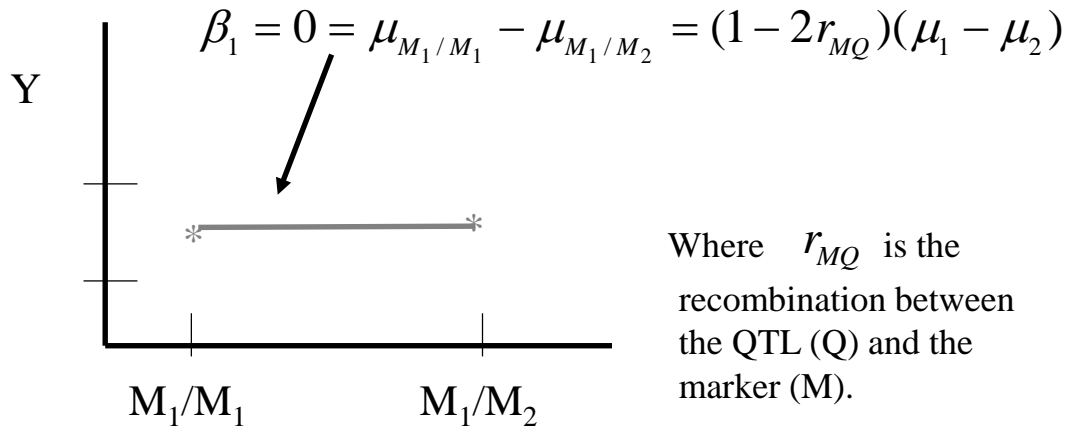
$$Y_i = \beta_0 + \beta_1 X_i$$



Where r_{MQ} is the recombination between the QTL (Q) and the marker (M).

...no QTL effect

$$Y_i = \beta_0 + \beta_1 X_i$$



81

Traditional Single Marker Methods

- t-test
- Hypotheses:

$$H_0 : \mu_{M_1/M_1} - \mu_{M_1/M_2} = 0$$

$$H_a : \mu_{M_1/M_1} - \mu_{M_1/M_2} \neq 0$$

Test

Statistic:

$$t = \frac{\bar{Y}_{M_1/M_1} - \bar{Y}_{M_1/M_2}}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{(n_1+n_2-2)}$$

where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

- $\mu_{M_1/M_1} - \mu_{M_1/M_2} = (1 - 2r_{MQ})(\mu_1 - \mu_{M_{F1}}) = \beta_1$

82

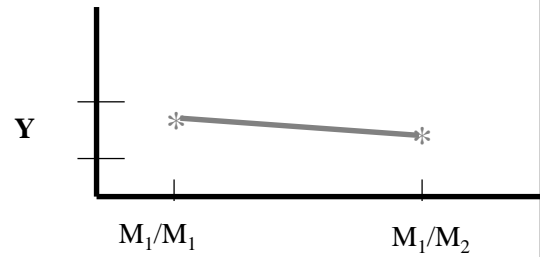
• effect of allelic substitution: t-test

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

$$t = \frac{b_1 - E[b_1]}{s_{b_1}}$$

$$= \frac{b_1 - \beta}{s_{b_1}} \sim t_{n-2} = t_{(n_1+n_2-2)}$$



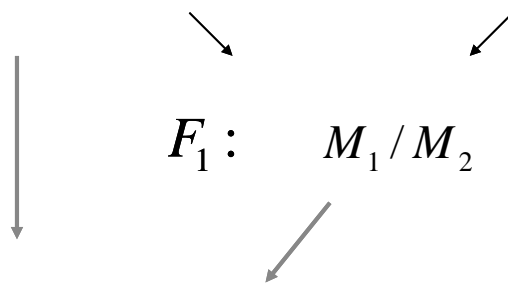
where $s_{b_1} = \sqrt{\frac{MSE}{\sum_{i=1}^n (y_i - \bar{y})^2}}$

83

WHAT WE SEE

Backcross: One Marker (M)

$P_1 : M_1 / M_1$ $P_2 : M_2 / M_2$



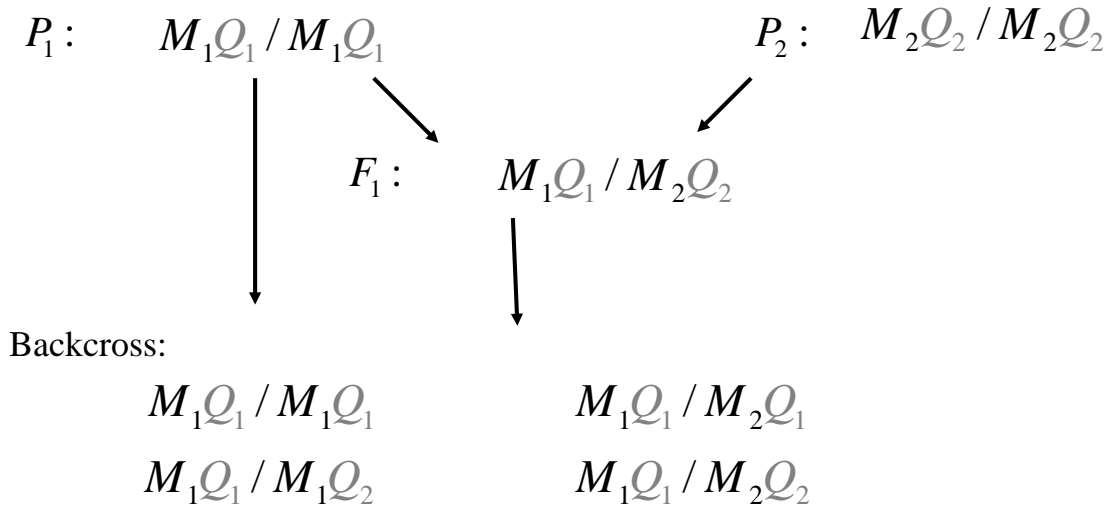
Backcross:

M_1 / M_1 M_1 / M_2

84

WHAT WE THINK

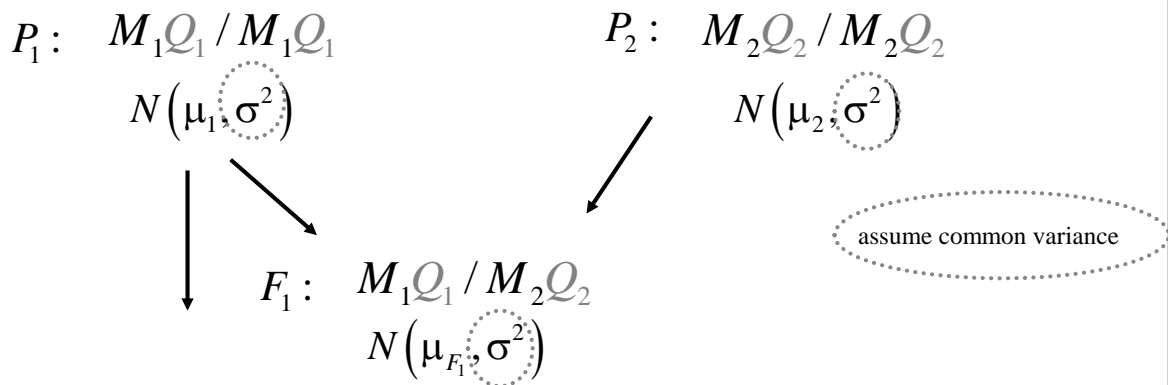
Backcross: One Marker (M)



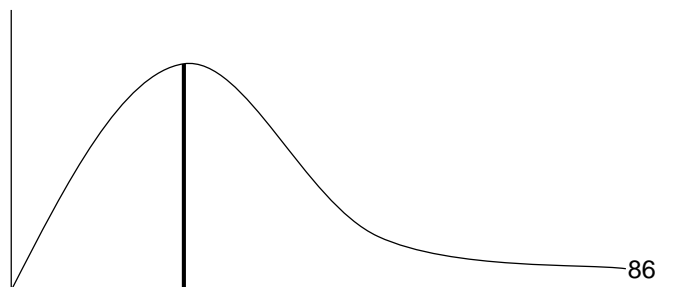
85

WHAT WE ASSUME

Backcross: One Marker (M)



**The distribution (shape) of the quantitative trait values in the backcross population follows a mixture of normal distributions within each of the known genotypic marker classes.



86

- Distribution (shape) of the quantitative trait values within each backcross genotypic marker class.

- two observable (backcross) marker genotypes
- four possible (observable) marker and QTL (unobservable) genotypes
- the distribution of the trait values:

$$M_1/M_1: \begin{cases} M_1Q_1 / M_1Q_1 : f_1 = (1-r)N(\mu_1, \sigma^2) + rN(\mu_{F_1}, \sigma^2) \\ M_1Q_1 / M_1Q_2 : \end{cases}$$

$$M_1/M_2: \begin{cases} M_1Q_1 / M_2Q_1 : f_2 = rN(\mu_1, \sigma^2) + (1-r)N(\mu_{F_1}, \sigma^2) \\ M_1Q_1 / M_2Q_2 : \end{cases}$$

87

Likelihood approach for single marker analysis (backcross):

- Discussed in more detail later...
- Obtain maximum likelihood estimates (MLEs) of $(\beta_0, \beta_1, \sigma^2)$
 - the MLEs are the values that maximize the likelihood of the observed values
 - or, the probability that the observed data would have occurred
 - write the likelihood as

$$L(\beta_0, \beta_1, \sigma^2 | Y, X, r) = \prod_1^{n_1} f_1 \prod_1^{n_2} f_2$$

- where $\beta_1 = \mu_{M_1/M_1} - \mu_{F_1} = (1 - 2r_{MQ})(\mu_1 - \mu_2)$

$$f_1 = (1-r)N(\mu_1, \sigma^2) + rN(\mu_{F_1}, \sigma^2)$$

$$f_2 = rN(\mu_1, \sigma^2) + (1-r)N(\mu_{F_1}, \sigma^2)$$

88

Summary

- Single marker analysis is a method of QTL “**detection**”, not location.
- Testing for differences in the means of the genetic marker classes actually tests whether $(1-2r_{MQ})(\mu_1 - \mu_{M_{F_1}})$ departs from zero.
- The location of the QTL and the effect of the QTL are confounded.
- Single marker analysis can be accomplished with QTL-Cartographer.

89

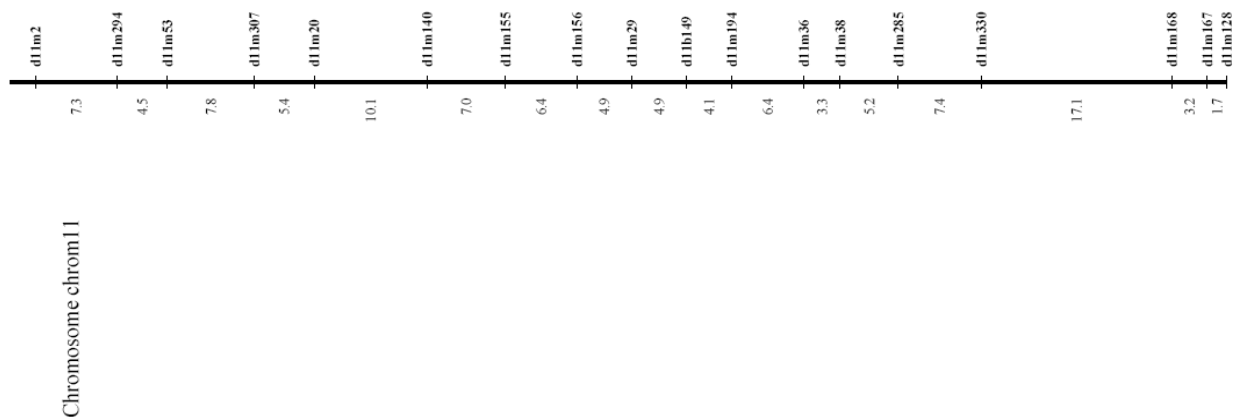
Example: single marker analysis...

Experimental details:

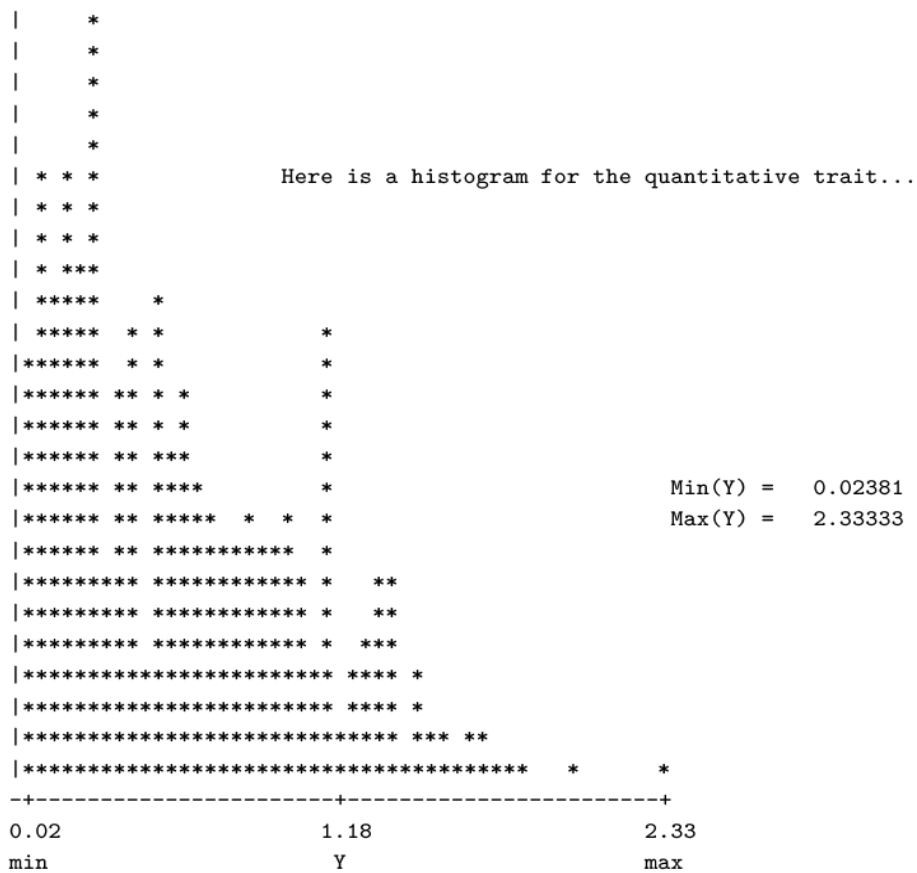
- mouse F_2
- $n = 291$
- chromosomes 11 only
- marker system
 - microsatellites ($m = 172$ genome; $m_{11} = 19$)
- mouse model ... multiple sclerosis (MS) in humans
- EAE: experimental Allergic encephalomyelitis is the principal animal model for human MS
- parental lines EAE-susceptible SJL/J and EAE-resistant B 10.S/DvTe inbred lines
- quantitative trait is severity of EAE. (Butterfield *et. al.*, 1999. Journal of Immunology. 162:(5)3096-3102).
- Analysis: Single Marker Analysis (LRmapqtl) in QTL-Cartographer.

90

Estimated Genetic Map for Chromosome 11



91



92

This output is based on the map in (qtlcart.map)

And the data in (qtlcart.cro)

Sample Size..... 291

This analysis fits the data to the simple linear regression model

$$y = b_0 + b_1 x + e$$

The results below give the estimates for b_0 , b_1 and the F statistic for each marker. The F statistic is for the hypothesis that the marker is unlinked to the quantitative trait. The column headed by PR is the probability that the trait is unlinked to the marker. Significance at the 5%, 1%, 0.1% and 0.01% levels are indicated by *, **, *** and ****, respectively. LR is $-2\log(L_0/L_1)$.

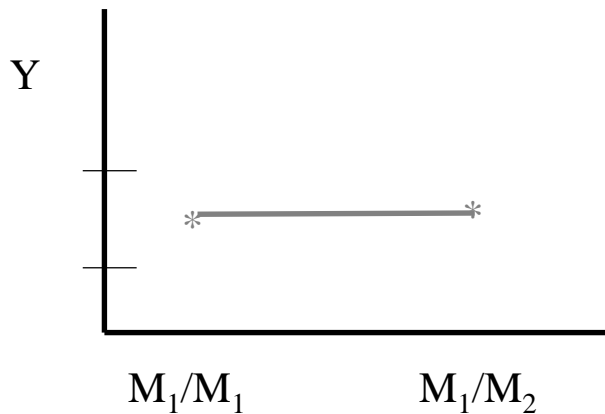
This trait is: sev, and

-t 1 is the number of trait being analyzed.

Chrom.	Marker	b_0	b_1	LR	F(1,n-2)	pr(F)
11	1	0.619	0.113	8.164	8.223	0.004 **
11	2	0.624	0.115	8.167	8.226	0.004 **
11	3	0.624	0.110	8.201	8.261	0.004 **
11	4	0.620	0.127	9.615	9.709	0.002 **
11	5	0.617	0.168	17.372	17.778	0.000 ****
11	6	0.621	0.156	14.771	15.048	0.000 ***
11	7	0.619	0.155	16.221	16.567	0.000 ****
11	8	0.620	0.144	13.101	13.308	0.000 ***
11	9	0.623	0.136	11.191	11.330	0.001 ***
11	10	0.622	0.136	12.137	12.308	0.001 ***
11	11	0.622	0.137	12.781	12.976	0.000 ***
11	12	0.620	0.165	18.627	19.104	0.000 ****

93

Chrom.	Marker	b_0	b_1	LR	F(1,n-2)	pr(F)
11	13	0.620	0.152	15.229	15.527	0.000 ***
11	14	0.618	0.164	17.762	18.189	0.000 ****
11	15	0.619	0.138	12.284	12.461	0.000 ***
11	16	0.628	0.046	1.336	1.330	0.250
11	17	0.628	0.050	1.631	1.624	0.204
11	18	0.626	0.068	2.976	2.971	0.086



Recap so far...

- Introduction to concept of QTL
- Experimental designs
- Source of data
 - genotype and phenotype
- Checking independent marker segregation
 - test for segregation distortion
- Introduction to concept of genetic map and estimating recombination
- Single marker QTL analysis

95

Linkage Analysis

Zhao-Bang Zeng

Summer Institute in Statistical Genetics

96

Linkage Analysis

$H_0: r = .50$ (markers are unlinked)

$H_1: r < .50$ (markers are linked)

Test Statistic: χ^2 test

- Question: are two markers linked?
- Need to identify recombinant and non-recombinant individuals
- Examples:
 - Backcross: 4 genotypic classes (mouse example)
 - F_2 : 9 identifiable genotypic classes (maize example)

97

Backcross population:



	AB/Ab	AB/aB	AB/AB	AB/ab
Frequency	$r/2$	$r/2$	$(1-r)/2$	$(1-r)/2$
Observed	n_2	n_3	n_1	n_4

- recombinant: $n_R = n_2 + n_3$ and non-recombinant: $n_{NR} = n_1 + n_4$
- total sample size: $n = n_1 + n_2 + n_3 + n_4 = n_R + n_{NR}$

Under the null hypothesis $r = .50$ (no linkage), the test statistic can be constructed as

$$\chi^2 = \frac{(n_{NR} - n_R)^2}{n} = \frac{(n_1 + n_4 - n_2 - n_3)^2}{n} \sim \chi_1^2$$

Recall: estimate of recombination frequency is $\hat{r} = n_R / n$

98

Markers are in map order

Example of linkage analysis: Mouse data

Markers		n_{NR}	n_R	χ^2	\hat{r}	cM(H)	cM(K)
1 Hmg1-rs13	2 DXMit57	96	7	76.903	0.068	7.3	6.8
2 DXMit57	3 Rps17-rs11	102	1	99.039	0.010	1.0	1.0
3 Rps17-rs11	4 Rps18-rs17	102	1	99.039	0.010	1.0	1.0
4 Rps18-rs17	5 DXMit48	100	3	91.350	0.029	3.0	2.9
5 DXMit48	6 DXNds1	100	3	91.350	0.029	3.0	2.9
6 DXNds1	7 DXMit109	98	5	83.971	0.049	5.1	4.9
7 DXMit109	8 Hmg14-rs6	99	4	87.621	0.039	4.0	3.9
8 Hmg14-rs6	9 DXMit60	102	1	99.039	0.010	1.0	1.0
9 DXMit60	10 DXMit16	101	2	95.155	0.019	2.0	1.9
10 DXMit16	11 DXMit97	101	2	95.155	0.019	2.0	1.9
11 DXMit97	12 Hmg1-rs14	100	3	91.350	0.029	3.0	2.9
12 Hmg1-rs14	13 DXMit3	96	7	76.903	0.068	7.3	6.8
13 DXMit3	14 Tpm3-rs9	92	11	63.699	0.107	12.0	10.8

H denotes Haldane map function; K denotes Kosambi map function

↑ 99 ↑

Mouse data: Estimated pairwise recombination frequencies

	2	3	4	5	6	7	8	9	10	11	...
1	0.07	0.08	0.09	0.12	0.15	0.19	0.23	0.24	0.26	0.26	
2		0.01	0.02	0.05	0.08	0.13	0.17	0.17	0.19	0.19	
3			0.01	0.04	0.07	0.12	0.16	0.17	0.18	0.18	
4				0.03	0.06	0.11	0.15	0.16	0.17	0.17	
5					0.03	0.08	0.12	0.13	0.15	0.17	
6						0.05	0.09	0.10	0.12	0.14	
7							0.04	0.05	0.07	0.09	
8								0.01	0.03	0.05	
9									0.02	0.04	
10										0.02	
⋮											

F₂ population:

A mating between AB/ab and AB/ab can produce ten genotypes, but only nine observable genetic classes:

$$\begin{array}{c}
 P_1 : \frac{AB}{AB} \quad \times \quad P_2 : \frac{ab}{ab} \\
 \downarrow \\
 F_1 : \frac{AB}{ab} \times F_1 : \frac{AB}{ab} \\
 \downarrow \\
 F_2 : \\
 \begin{array}{ccccccc}
 \frac{AB}{AB} & \frac{AB}{Ab} & \frac{Ab}{Ab} & \frac{AB}{aB} & \frac{AB_{or}Ab}{ab_{or}aB} & \frac{Ab}{ab} & \frac{aB}{aB} \\
 \frac{(1-r)^2}{4} & \frac{r(1-r)}{2} & \frac{r^2}{4} & \frac{r(1-r)}{2} & \frac{(1-r)^2+r^2}{2} & \frac{r(1-r)}{2} & \frac{r^2}{4} \\
 \downarrow & & & & & & \\
 F_\infty : & \frac{AB}{AB} & \frac{Ab}{Ab} & \frac{aB}{aB} & \frac{ab}{ab} & &
 \end{array}
 \end{array}$$

- The two double heterozygotes (AB/ab and Ab/aB) are generally not distinguishable.
- The specific nine unique genotype expected frequencies follow.

101

Genetic class	Code	H ₀ : r = .50	H ₁ : r < .50	Rec. Event	Observed #
AB/AB	2 2	1/16	(1-r) ² /4	0	n ₁
AB/Ab	2 1	2/16	r(1-r)/2	1	n ₂
Ab/Ab	2 0	1/16	r ² /4	2	n ₃
AB/aB	1 2	2/16	r(1-r)/2	1	n ₄
AB/ab	1 1	4/16	[(1-r) ² +r ²]/2	0	n ₅
Ab/aB	1 1			2	
Ab/ab	1 0	2/16	r(1-r)/2	1	n ₆
aB/aB	0 2	1/16	r ² /4	2	n ₇
aB/ab	0 1	2/16	r(1-r)/2	1	n ₈
ab/ab	0 0	1/16	(1-r) ² /4	0	n ₉

where $c = \frac{r^2}{[(1-r)^2+r^2]}$

102

Estimating Recombination in an F₂ population

- A little more complicated largely because of genetic class 5 (n₅).
- When estimating recombination frequency, we can utilize the genetic classification of recombination events and estimate r as:

$$r = \frac{1}{2n} \left[(n_2 + n_4 + n_6 + n_8) + 2(n_3 + n_7 + cn_5) \right] \quad (1)$$

- Recall:
 - genetic classes 2, 4, 6, 8 are the result of a single recombination event
 - genetic classes 3, 7, and 5 (with probability c) are the result of two recombination events...

103

Since the probability c is unknown, it has to be estimated. However, c is a function on recombination r (which is also unknown) :

$$c = \frac{r^2}{(1-r)^2 + r^2} \quad (2)$$

- Therefore, an analysis has to be performed in an iteratively updated loop between equations (2) and (1).
 - guess r (usually start with r = 0.25)
 - calculate c
 - stop when estimates converge
- This algorithm is called the EM algorithm (Dempster, Laird, and Rubin 1977).
 - the E-step (Expectation step; equation (2))
 - the M-step (Maximization step equation (1)).
 - after a few iterations the estimate usually converges quickly.

104

Testing linkage in an F₂ population

- Given that we can estimate recombination (r) in an F₂
- Test for linkage between pairs of genetic markers or loci
- The statistical test for linkage can be performed by LOD score (log₁₀ of odds, a likelihood ratio test statistic)
- The likelihood function is:

$$L(r) \propto \underbrace{\left[\frac{1}{4}(1-r)^2\right]^{n_1+n_9}}_{\text{no recombinant events}} \underbrace{\left[\frac{1}{2}r(1-r)\right]^{n_2+n_4+n_6+n_8}}_{\text{1 recombinant event}} \underbrace{\left[\frac{1}{4}r^2\right]^{n_3+n_7}}_{\text{2 recombinant events}} \underbrace{\left[\frac{1}{2}(1-r)^2 + \frac{1}{2}r^2\right]^{n_5}}_{\text{complicated part}}$$

105

The statistical test for linkage can be performed by LOD score (log₁₀ of odds, a likelihood ratio test statistic):

$$\text{LOD} = \log_{10} \frac{L(\hat{r})}{L(r = 1/2)}$$

with

$$L(\hat{r}) \propto \left[\frac{1}{4}(1-\hat{r})^2\right]^{n_1+n_9} \left[\frac{1}{2}\hat{r}(1-\hat{r})\right]^{n_2+n_4+n_6+n_8} \left[\frac{1}{4}\hat{r}^2\right]^{n_3+n_7} \left[\frac{1}{2}(1-\hat{r})^2 + \frac{1}{2}\hat{r}^2\right]^{n_5}$$

and

$$L(r = 1/2) \propto \left[\frac{1}{16}\right]^{n_1+n_3+n_7+n_9} \left[\frac{2}{16}\right]^{n_2+n_4+n_6+n_8} \left[\frac{4}{16}\right]^{n_5}$$

106

Maize data: Estimated pairwise recombination frequencies

	2	3	4	5	6	7	8	9	10	11	12
1	0.31	0.39	0.49	0.50	0.46	0.46	0.45	0.45	0.49	0.50	0.51
2		0.12	0.22	0.22	0.31	0.33	0.40	0.39	0.40	0.39	0.42
3			0.11	0.12	0.28	0.32	0.44	0.44	0.43	0.42	0.43
4				0.02	0.22	0.26	0.39	0.39	0.41	0.41	0.42
5					0.21	0.27	0.39	0.40	0.41	0.41	0.43
6						0.08	0.21	0.22	0.34	0.37	0.40
7							0.13	0.14	0.31	0.34	0.39
8								0.06	0.26	0.30	0.35
9									0.20	0.25	0.31
10										0.06	0.13
11											0.09

Maize data: example of linkage analysis using estimated recombination

Markers	n_{0R}	n_{1R}	n_{2R}	n_5	LOD	r	cM(H)	cM(K)	
1	2	39	77	6	49	6.08	0.31	47.7	35.8
2	3	62	37	0	72	30.87	0.12	13.1	11.7
3	4	63	33	1	74	32.08	0.11	12.3	11.0
4	5	79	4	2	86	60.60	0.02	2.4	2.4
5	6	54	58	3	56	15.92	0.21	27.0	22.2
6	7	75	22	2	72	41.55	0.08	8.6	8.0
7	8	68	40	1	62	29.52	0.13	15.2	13.4
8	9	78	19	0	74	48.92	0.06	6.1	5.7
9	10	50	58	2	61	15.73	0.20	26.1	21.5
10	11	70	19	0	82	46.53	0.06	6.1	5.8
11	12	70	26	1	74	38.86	0.09	9.4	8.6

Recall: $n_{0R} = n_1 + n_9$; $n_{1R} = n_2 + n_4 + n_6 + n_8$; $n_{2R} = n_3 + n_7$

Estimating Genetics Maps

R.W. Doerge

Summer Institute in Statistical Genetics

109

Ordering a Set of Genetic Markers

The problem is equivalent to the “Traveling Salesman Problem”.

▪ Methods:

1. Brand and Bound (Thompson, 1984)
2. Simulated Annealing (Weeks and Lange, 1987)
3. Seriation (Buetow and Chakravarti, 1987a,b)
4. Rapid Chain Delineation (RCD) (Doerge, 1993)
5. Many more...

110

- Methods 1-3 are “multipoint analysis”, meaning that they rely on the calculation of all recombinant classes, between chains of markers (not just two markers).
- Method 4 starts with pairwise recombination estimates
 - forms linkage groups and preliminary order
 - then resolves local inversions, by “permuting” all possible n-lets (i.e., triplets, quads, etc.)
 - very fast
 - RCD implemented in QTL-Cartographer

111

Motivation:

- Building a genetic map quickly.
 - some experiments have only 50 individuals, but have 1500 markers
 - 10 markers alone, provide 1,814,400 possible orderings
 - $\binom{10!}{2}$
 - it is not computationally feasible to try all possible orders of m markers

112

Example of RCD

- Assume we have four markers: M, N, T, U
- We learned how to estimate pairwise recombination estimates
- The pairwise recombination between markers is represented in the following matrix:

$$\begin{array}{c} \text{M} \\ \text{N} \\ \text{T} \\ \text{U} \end{array} \begin{pmatrix} 0 & .09 & .19 & .17 \\ & 0 & .26 & .22 \\ & & 0 & .32 \\ & & & 0 \end{pmatrix}$$

- Step 1: chain together M – N; SAR = .09
- Step 2: add U to chain (U – M – N); SAR = .26
- Step 3: add T to chain (U – M – N – T); SAR = .52

113

- Step 4: Permute overlapping (triplets) markers.
- Final order: U – N – M – T; SAR = .50
- Most of the time, markers very close together may be transposed. The permutation stage of RCD takes care of this.

114

Mapping Software

- MAPMAKER/EXP (version 3.0): Software for the calculation of genetic maps of certain experimental populations.
- JoinMap: “JoinMap provides high quality tools that allow detailed study of the experimental data and the generation of publication-ready map charts.”
- One Map (R function): includes RCD Method

115

Summary:

- Determine linkage groups, resolve order within linkage groups.
- Order across all linkage groups.
- A “genetic map” is a collection of all linkage groups.
 - if there are enough markers to cover the entire chromosome, a linkage group is then referred to as a chromosome.
- **The genetic map is the structure that we rely on to locate “quantitative trait loci” (QTL), the genomic regions affecting a trait of interest.**
 - *if your estimated genetic map is poor, then your QTL location will be poor.*

116

Map Estimation Exercise

Calculate pairwise recombination by hand, and estimate genetic map of 6 markers for increasing sample size

R.W. Doerge

Summer Institute in Statistical Genetics

117

Simulation Setting and Goal

- **Simulation Input:**
 - **Experimental design:** backcross
 - **Sample size:** $n = 25, 50, 100, 500, 1000$
 - **Marker number:** 6
 - **Recombination between markers**

- **Simulation Output:**
 - Genotype information on 6 markers

- **Goal:**
 - Estimate pairwise recombination and linear order (i.e., genetic map) by hand

118

Recall: estimating pairwise recombination

two markers *11 and *12

H = Marker is heterozygous:

marker *11: *11₁/*11₂

marker *12: *12₁/*12₂

A = Marker is homozygous

marker *11: *11₁/*11₁

marker *12: *12₁/*12₁

		A	H
		*11 ₁ /*11 ₁	*11 ₁ /*11 ₂
A	*12 ₁ /*12 ₁	<i>n</i> ₁	<i>n</i> ₂
H	*12 ₁ /*12 ₂	<i>n</i> ₃	<i>n</i> ₄

- Total number of recombinant events is $n_2 + n_3$

- $$\hat{r} = \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4}$$

119

Individuals= 25, marker number= 6.

Marker name	Genotype for each backcross individual
*11	HHAAHHHHHHHAHAAAHA AAHHAHH
*12	HHHA AHHHHHAHAAAHA AAHHAHH
*13	HHHA AHHHHHAHAAAHA AAHHAHH
*14	HAHAHAHHHHA AAAAHA AAHHAHH
*15	HHHA AHHHA AHAHA AHA AAHHAHH
*16	HAHA A AHHHA AHAHA AHA AAHHAHH

120

Individuals= initial 25 + 25 additional=50, marker number= 6.

Marker name	Genotype for each backcross individual
*11	ННАНННННННАНААААНАААННАНН...
*12	НННААНННННАНААААНАААННАНН...
*13	НННААНННННАНААААНАААННАНН...
*14	НАНАНАННННАААААНАААННАНН...
*15	НННАААНННААНАНААНААААНННН...
*16	НАНАААНННААНАНААНААНАННАН...
*11	...АНААНААНААНАААНАААНААНН
*12	...ААААНАННААНАНННАНАААНАННН
*13	...АНААНАННААНАНННААААНАННН
*14	...АНААНАННААНАННННААААНАННН
*15	...АНААНАННААНАНАНАНАНАААННН
*16	...АНААНААНААНАНАНАНАНААААНН

Marker name	Genotype for each backcross individual ($n = 100, m = 6$)
*11	ННАНННННННАНААААНАААННАНН...
*12	НННААНННННАНААААНАААННАНН...
*13	НННААНННННАНААААНАААННАНН...
*14	НАНАНАННННАААААНАААННАНН...
*15	НННАААНННААНАНААНААААНННН...
*16	НАНАААНННААНАНААНААНАННАН...
*11	...АНААНААНААНАААНАААНААНН...
*12	...ААААНАННААНАНННАНАААНАННН...
*13	...АНААНАННААНАНННААААНАННН...
*14	...АНААНАННААНАННННААААНАННН...
*15	...АНААНАННААНАНАНАНАНАААННН...
*16	...АНААНААНААНАНАНАНАНААААНН...

Marker name	Genotype for each backcross individual ($n = 100, m = 6$) cont...
*11	...HHAAAAAHHAAAAHAANHHAAAAHAA...
*12	...HAAAAAAHHA AAAHAANHHAAHHA...
*13	...HAAAAAHNHAAAAHAANHHAAHAAA...
*14	...HAAAAAHNHAAAAHAANHA AAAHAAA...
*15	...HAAAAAHHAAAAHAANHA AAAAAAAA...
*16	...HAAAAAHHAAAAHNAHNA AAAAAAAA...

*11	...AAAAAHAANHHNHANAANHHANHHNA
*12	...ANANANAANHHANHAANAANHHNNH
*13	...ANANANAANAANHHANHAANHHNNH
*14	...ANANHHAAAAAANHHNNHANHHNNH
*15	...ANANHHAAAAAANHHNNHANHHNNH
*16	...ANANHHAAAAAANHHNNHANHHNNH

123

ANSWERS:

Individuals= 25, marker number= 6.

marker name	*11	*12	*13	*14	*15	*16
*11	0.00	0.12	0.12	0.20	0.32	0.44
*12		0.00	0.00	0.16	0.20	0.32
*13			0.00	0.16	0.20	0.32
*14				0.00	0.28	0.32
*15					0.00	0.12
*16						0.00

124

ANSWERS:

Individuals= 50, marker number= 6.

marker name	*11	*12	*13	*14	*15	*16
*11	0.00	0.16	0.16	0.22	0.26	0.28
*12		0.00	0.04	0.14	0.18	0.28
*13			0.00	0.10	0.18	0.28
*14				0.00	0.24	0.30
*15					0.00	0.10
*16						0.00

125

ANSWERS:

Individuals= 100, marker number= 6.

marker name	*11	*12	*13	*14	*15	*16
*11	0.00	0.16	0.22	0.28	0.31	0.31
*12		0.00	0.08	0.18	0.23	0.27
*13			0.00	0.10	0.17	0.23
*14				0.00	0.15	0.19
*15					0.00	0.08
*16						0.00

126

ANSWERS:

Individuals= 500, marker number= 6.

marker name	*11	*12	*13	*14	*15	*16
*11	0.00	0.09	0.16	0.23	0.27	0.30
*12		0.00	0.08	0.17	0.22	0.26
*13			0.00	0.10	0.17	0.24
*14				0.00	0.11	0.18
*15					0.00	0.08
*16						0.00

127

ANSWERS:

Individuals= 1000, marker number= 6.

marker name	*11	*12	*13	*14	*15	*16
*11	0.00	0.09	0.17	0.24	0.28	0.32
*12		0.00	0.10	0.18	0.23	0.28
*13			0.00	0.10	0.18	0.25
*14				0.00	0.10	0.18
*15					0.00	0.10
*16						0.00

128

Things you can do with a genetic map...

- With a genetic map in place we can rely on the order of the genetic markers across linkage groups (chromosomes) to provide additional information to locate QTL.
 - incorporate the recombination information from the genetic map into the search for QTL
 - necessary to use genetic map function to translate between recombination and genetic distance (i.e., probability to additive distance)
 - Haldane
 - Kosambi

129

Introduction to QTL-Cartographer

- Download from:
<http://statgen.ncsu.edu/qtlcart/WQTLCart.htm>
- Modular based:
 - Simulation and analysis
- Simulation:
 - Genetic map
 - QTL
- Need to understand parameter settings and order of events
- Need to stay organized
 - Keep real data and simulated data in separated files/directories

130

Introduction to QTL Detection

single marker likelihood to interval mapping

R.W. Doerge

Summer Institute in Statistical Genetics

131

The concept...

- likelihood based QTL analysis...
- moving from single marker analysis to interval mapping
 - develop likelihood approach for single marker
- incorporate additional marker information into likelihood function
 - develop likelihood approach for two markers
 - consider two markers M and N, and the distance between them
 - each with two alleles
- QTL-Cartographer uses likelihood based approaches
 - important to understand the parameters that are being estimated

132

Recall: Single Marker Backcross QTL Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i; i = 1, \dots, n$$

- consider a backcross experiment
- the QTL genotype can be one of two states Q_1/Q_1 or Q_1/Q_2
- recall the equation for a straight line
 - β_0 is overall mean
 - β_1 is the additive effect of the QTL
 - allelic substitution at the QTL
 - $\beta_1 = \mu_{M_1/M_1} - \mu_{M_1/M_2} = (1 - 2r_{MQ})(\mu_1 - \mu_2)$
 - X_i is the genotype of the unobservable QTL
- use marker genotype as X_i ; maybe QTL and marker are linked

133

- Distribution (shape) of the quantitative trait values within each backcross genotypic marker class.

- two observable (backcross) marker genotypes
- four possible marker and QTL genotypes
- the distribution of the trait values:

$$M_1/M_1: \begin{cases} M_1Q_1 / M_1Q_1 : & f_1 = (1-r)N(\mu_1, \sigma^2) + rN(\mu_{F_1}, \sigma^2) \\ M_1Q_1 / M_1Q_2 : & \end{cases}$$

$$M_1/M_2: \begin{cases} M_1Q_1 / M_2Q_1 : & f_2 = rN(\mu_1, \sigma^2) + (1-r)N(\mu_{F_1}, \sigma^2) \\ M_1Q_1 / M_2Q_2 : & \end{cases}$$

134

Likelihood approach for single marker analysis (backcross):

- Obtain maximum likelihood estimates (MLEs) of $(\beta_0, \beta_1, \sigma^2)$
 - the MLEs are the values that maximize the likelihood of the observed values
 - or, the probability that the observed data would have occurred
 - write the likelihood as

$$L(\beta_0, \beta_1, \sigma^2 | Y, X, r) = \prod_1^{n_1} f_1 \prod_1^{n_2} f_2$$

- where

$$\beta_1 = \mu_{M_1/M_1} - \mu_{F_1} = (1 - 2r_{MQ})(\mu_1 - \mu_2)$$

$$f_1 = (1 - r)N(\mu_1, \sigma^2) + rN(\mu_{F_1}, \sigma^2)$$

$$f_2 = rN(\mu_1, \sigma^2) + (1 - r)N(\mu_{F_1}, \sigma^2)$$

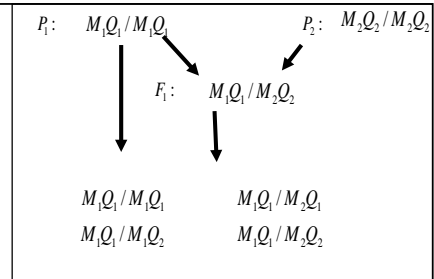
135

Same hypotheses:

- $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$
- Test statistic is:

$$LOD = \log_{10} \left[\frac{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{L(\hat{\beta}_0, 0, \hat{\sigma}^2)} \right]$$

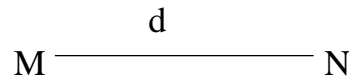
- no QTL effect (denominator), implies $\hat{\beta}_0 = \mu_1$ and $\hat{\sigma}^2 = \sigma_{B_1}^2$
- the LOD score demonstrates (statistically) how much more likely (probable) the data are if there was a QTL present as compared to the situation when there is no QTL present.



136

Extend this idea to interval mapping...

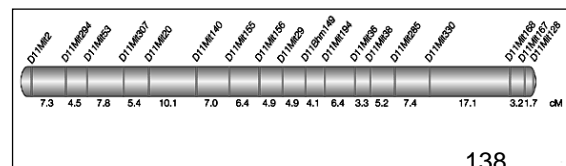
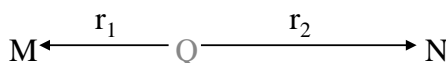
- Consider two markers M and N, each with two alleles.
- The genetic distance d (or, recombination, r) between markers M and N has been previously estimated (known)



- A map function (Haldane or Kosambi) is utilized to translate between recombination and genetic distance.
- Working in the units of genetic distance, incrementally step through the defined interval, testing the same hypotheses as before...
 - only now we need to incorporate the fact that we have information about recombination (genetic distance)
 - calculate a LOD score at each increment in the interval.

137

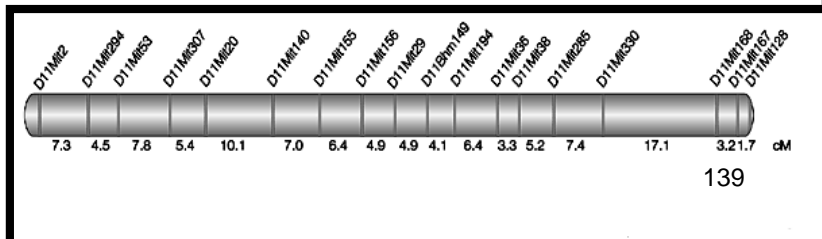
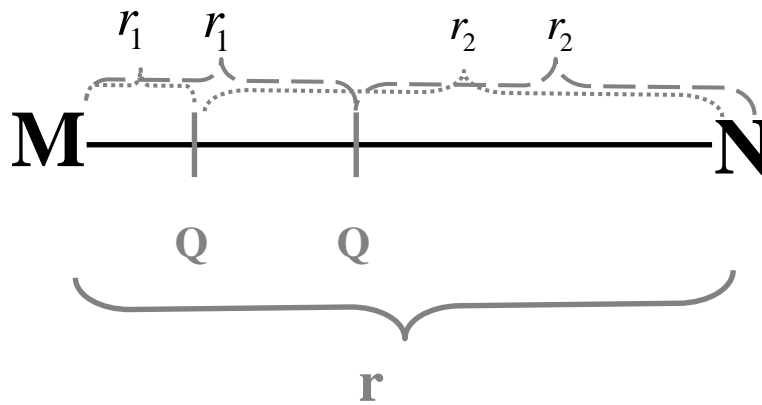
- Marker M: alleles M_1 and M_2
- Marker N: alleles N_1 and N_2
- Relationship between M and N defined by recombination r
 - the value of r is estimated and known
 - $M \xleftarrow{\quad r \quad} N$
- Use the additional information from knowing 'r' to locate QTL
- Notation:
 - r_1 is the recombination between marker M and the putative QTL
 - r_2 is the recombination between the putative QTL and marker N
 - any function of both r_1 and r_2 will be denoted as $k_i, i = 1, 2$.



138

Lander and Botstein (1989)

Locate QTL by stepping through the interval defined by M and N



WHAT WE SEE

Backcross: Two Markers (M,N)

$$P_1 : M_1N_1 / M_1N_1$$

$$P_2 : M_2N_2 / M_2N_2$$

$$F_1 : M_1N_1 / M_2N_2$$

Backcross: M_1N_1 / M_1N_1

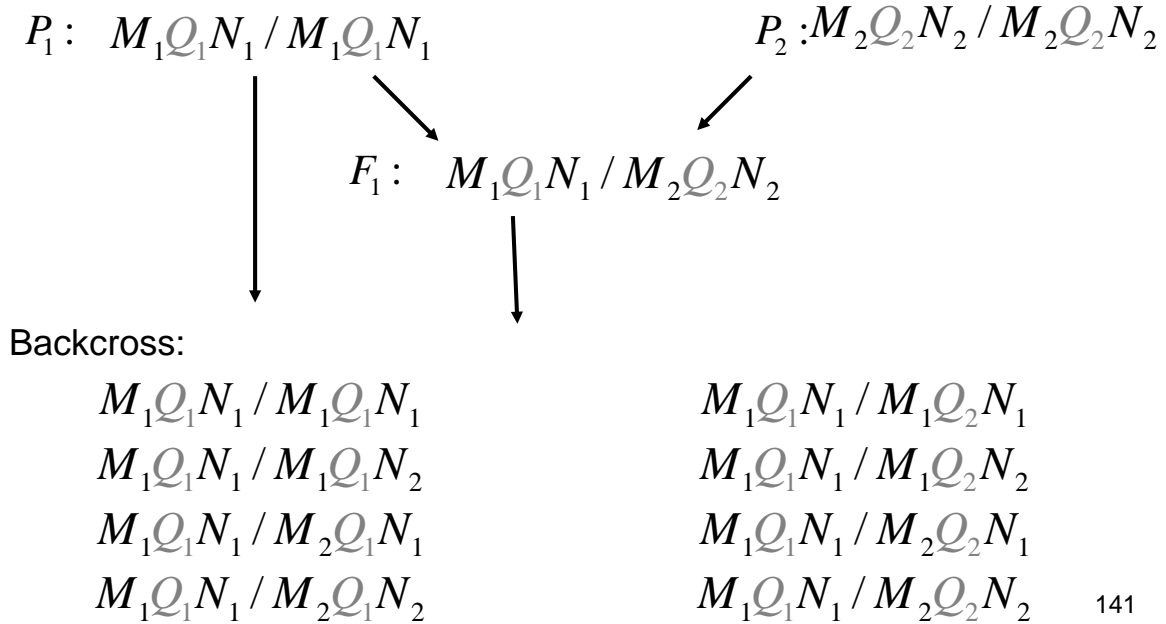
$$M_1N_1 / M_1N_2$$

$$M_1N_1 / M_2N_1$$

$$M_1N_1 / M_2N_2$$

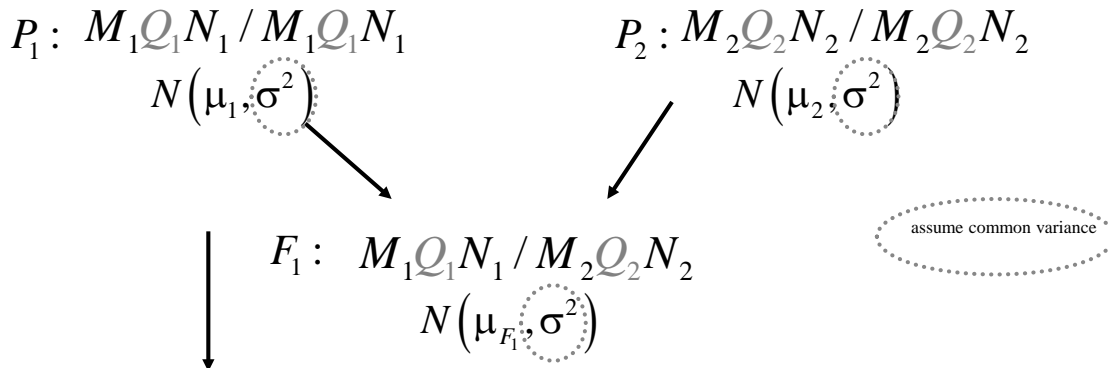
WHAT WE THINK

Backcross: Two Markers (M,N)



WHAT WE ASSUME

Backcross: Two Markers (M,N)



**The distribution (shape) of the quantitative trait values in the backcross population follows a mixture of normal distributions within each of the known genotypic marker classes.

Backcross classes for two markers, one QTL:

- Possible backcross genotypes and the distribution of the trait values (four unique mixtures of distributions):

$$\begin{array}{l} M_1Q_1N_1 / M_1Q_1N_1 : \\ M_1Q_1N_1 / M_1Q_2N_1 : \end{array} \quad f_1 = k_1N(\mu_1, \sigma^2) + (1 - k_1)N(\mu_{F_1}, \sigma^2)$$

$$\begin{array}{l} M_1Q_1N_1 / M_2Q_1N_2 : \\ M_1Q_1N_1 / M_2Q_2N_2 : \end{array} \quad f_2 = (1 - k_1)N(\mu_1, \sigma^2) + k_1N(\mu_{F_1}, \sigma^2)$$

$$\begin{array}{l} M_1Q_1N_1 / M_2Q_1N_1 : \\ M_1Q_1N_1 / M_2Q_2N_1 : \end{array} \quad f_3 = (1 - k_2)N(\mu_1, \sigma^2) + k_2N(\mu_{F_1}, \sigma^2)$$

$$\begin{array}{l} M_1Q_1N_1 / M_1Q_1N_2 : \\ M_1Q_1N_1 / M_1Q_2N_2 : \end{array} \quad f_4 = k_2N(\mu_1, \sigma^2) + (1 - k_2)N(\mu_{F_1}, \sigma^2)$$

$$k_1 = \frac{(1 - r_1)(1 - r_2)}{(1 - r_1)(1 - r_2) + r_1r_2} \quad k_2 = \frac{(1 - r_1)r_2}{(1 - r_1)r_2 + r_1(1 - r_2)}$$

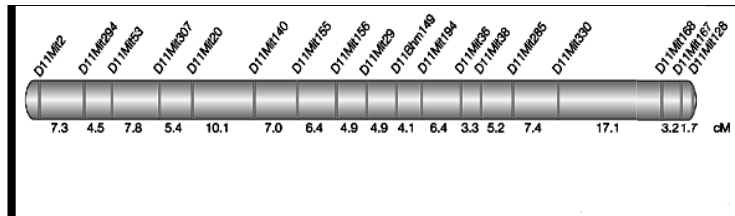
143

Hypotheses:

H_0^A : no QTL

H_0^B : QTL unlinked

H_a : QTL present and linked



Likelihood function:

$$L(\mu_1, \mu_2, \sigma^2 | X, Y, r_1, r_2) = \prod_1^{n_1} f_1 \prod_1^{n_2} f_2 \prod_1^{n_3} f_3 \prod_1^{n_4} f_4$$

$$\text{LOD} = \log_{10} \frac{L(\hat{\mu}_1, \hat{\mu}_{F_1}, \hat{\sigma}^2, r_1, r_2)}{L(\hat{\mu}_1, \hat{\mu}_{F_1}, \hat{\sigma}^2, r_1 = 0.50, r_2 = 0.50)} \quad ? \sim ? \quad 144$$

Interval Mapping

Zhao-Bang Zeng

Summer Institute in Statistical Genetics

145

Interval mapping

The idea of interval mapping is two-fold:

1. by using two markers, both position and effect of a QTL can be inferred
2. two adjacent markers (and associated genetic distance) are used to define the position in the search for QTL.

146

Interval Mapping

- The analysis is usually based a maximum likelihood analysis.
- Consider a backcross population.
- To analyze a QTL (Q) located in an interval flanked by two markers (M and N)
 - (assuming the order MQN)
- The interval mapping analysis assumes the following linear model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2) \quad (3)$$

- Y_i is the quantitative trait value
- the indicator variable X_i takes the value 1 or 0 according to whether individual Y_i has QTL genotype Q_1/Q_1 ($X_i = 0$) or Q_1/Q_2 ($X_i = 1$)

147

Note: the model is defined based on the QTL genotypes which are unobserved. However, given the marker genotypes and linkage relationship between markers and QTL, the probabilities of possible QTL genotypes can be inferred. Given a backcross design, let

$$p_{ki} = \Pr(X_i^* = k | M, N, \theta) \quad k = 0, 1.$$

which is specified as

Genotype	#	Freq	QTL genotype	
			$Q_1/Q_1 (X_i=0)$	$Q_1/Q_2 (X_i=1)$
M_1N_1 / M_1N_1	n_1	$\frac{1-r_{MN}}{2}$	$\frac{(1-r_{MQ})(1-r_{QN})}{1-r_{MN}} \approx 1$	$\frac{r_{MQ}r_{QN}}{1-r_{MN}} \approx 0$
M_1N_1 / M_1N_2	n_2	$\frac{r_{MN}}{2}$	$\frac{(1-r_{MQ})r_{QN}}{r_{MN}} \approx 1-\theta$	$\frac{r_{MQ}(1-r_{QN})}{r_{MN}} \approx \theta$
M_2N_1 / M_1N_1	n_3	$\frac{r_{MN}}{2}$	$\frac{r_{MQ}(1-r_{QN})}{r_{MN}} \approx \theta$	$\frac{(1-r_{MQ})r_{QN}}{r_{MN}} \approx 1-\theta$
M_2N_2 / M_1N_1	n_4	$\frac{1-r_{MN}}{2}$	$\frac{r_{MQ}r_{QN}}{1-r_{MN}} \approx 0$	$\frac{(1-r_{MQ})(1-r_{QN})}{1-r_{MN}} \approx 1$

where $\theta = r_{MQ} / r_{MN}$.

148

Because there are two possible QTL genotypes each of which can be true with certain probability, the distribution of the model is a mixture distribution. Thus, the likelihood function of equation (3) is usually defined as

$$\begin{aligned}
 L(\beta_0, \beta_1, \sigma^2, \theta) &= \prod_{i=1}^n \left[p_{1i} \phi\left(\frac{y_i - \beta_0 - \beta_1}{\sigma}\right) + p_{0i} \phi\left(\frac{y_i - \beta_0}{\sigma}\right) \right] \\
 &= \prod_{i=1}^{n_1} \phi\left(\frac{y_i - \beta_0 - \beta_1}{\sigma}\right) \prod_{i=1}^{n_2} \left[(1 - \theta) \phi\left(\frac{y_i - \beta_0 - \beta_1}{\sigma}\right) + \theta \phi\left(\frac{y_i - \beta_0}{\sigma}\right) \right] \\
 &\quad \prod_{i=1}^{n_3} \left[\theta \phi\left(\frac{y_i - \beta_0 - \beta_1}{\sigma}\right) + (1 - \theta) \phi\left(\frac{y_i - \beta_0}{\sigma}\right) \right] \prod_{i=1}^{n_4} \phi\left(\frac{y_i - \beta_0}{\sigma}\right)
 \end{aligned}$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp[-z^2/2]$ is the standard normal density function

$$\theta (= r_{MQ} / r_{MN})$$

149

- In this likelihood function, the parameters include:
 - β_0 the mean of the model
 - β_1 the effect of the putative QTL
 - $\theta (= r_{MQ} / r_{MN})$ the position of the putative QTL
 - σ^2 residual variance

- The data are
 - y_i the phenotypic value of a quantitative trait for each individual
 - genotypes of markers for each individual that contributes to the analysis of $p_{ki}, k = 1, 2; i = 1, \dots, n$

150

Maximum likelihood analysis and EM algorithm

The maximum likelihood analysis of a mixture model is usually performed via an EM (Expectation and Maximization) algorithm. The EM-algorithm is an iterative procedure. In each iteration, the **E-step calculates:**

$$P_i = \Pr(x_i = 1 | M, N, y_i) = \frac{\Pr(x_i = 1 | M, N) \Pr(y_i | x_i = 1)}{\Pr(y_i)}$$

$$= \frac{p_{1i} \phi([y_j - \beta_0 - \beta_1] / \sigma)}{p_{1i} \phi([y_j - \beta_0 - \beta_1] / \sigma) + p_{0i} \phi([y_i - \beta_0] / \sigma)} \quad (4)$$

151

and the M-step calculates:

$$\hat{\beta}_0 = \sum_{i=1}^n (y_i - P_i \hat{\beta}_1) / n \quad (5)$$

$$\hat{\beta}_1 = \sum_{i=1}^n (y_i - \hat{\beta}_0) P_i / \sum_{i=1}^n P_i \quad (6)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left[(y_i - \hat{\beta}_0)^2 - P_i \hat{\beta}_1^2 \right] \quad (7)$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n_2} (1 - \hat{P}_i) + \sum_{i=1}^{n_3} \hat{P}_i}{n_2 + n_3} \quad (8)$$

This process is iterated until convergence of estimates.

152

Likelihood ratio test statistic

The test statistic can be constructed using a likelihood ratio in LOD (likelihood of odds) score

$$LOD = \log_{10} \frac{L(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2)}{L(\widehat{\widehat{\beta}}_0, \beta_1 = 0, \widehat{\widehat{\sigma}}^2)} \quad (9)$$

under the hypotheses

$$H_0 : \beta_1 = 0 \quad \text{and} \quad H_1 : \beta_1 \neq 0$$

assuming that the putative QTL was located at the point θ on the genetic map, and where $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2$ are the maximum likelihood estimates of $\beta_0, \beta_1, \sigma^2$ under H_1 , and $\widehat{\widehat{\beta}}_0, \widehat{\widehat{\sigma}}^2$ are the estimates of under H_0 with β_1 constrained to zero. 153

Note: that the LOD score test is the same test as the usual likelihood ratio test

$$LR = -2 \ln \frac{L(\widehat{\widehat{\beta}}_0, \beta_1 = 0, \widehat{\widehat{\sigma}}^2)}{L(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2)}$$

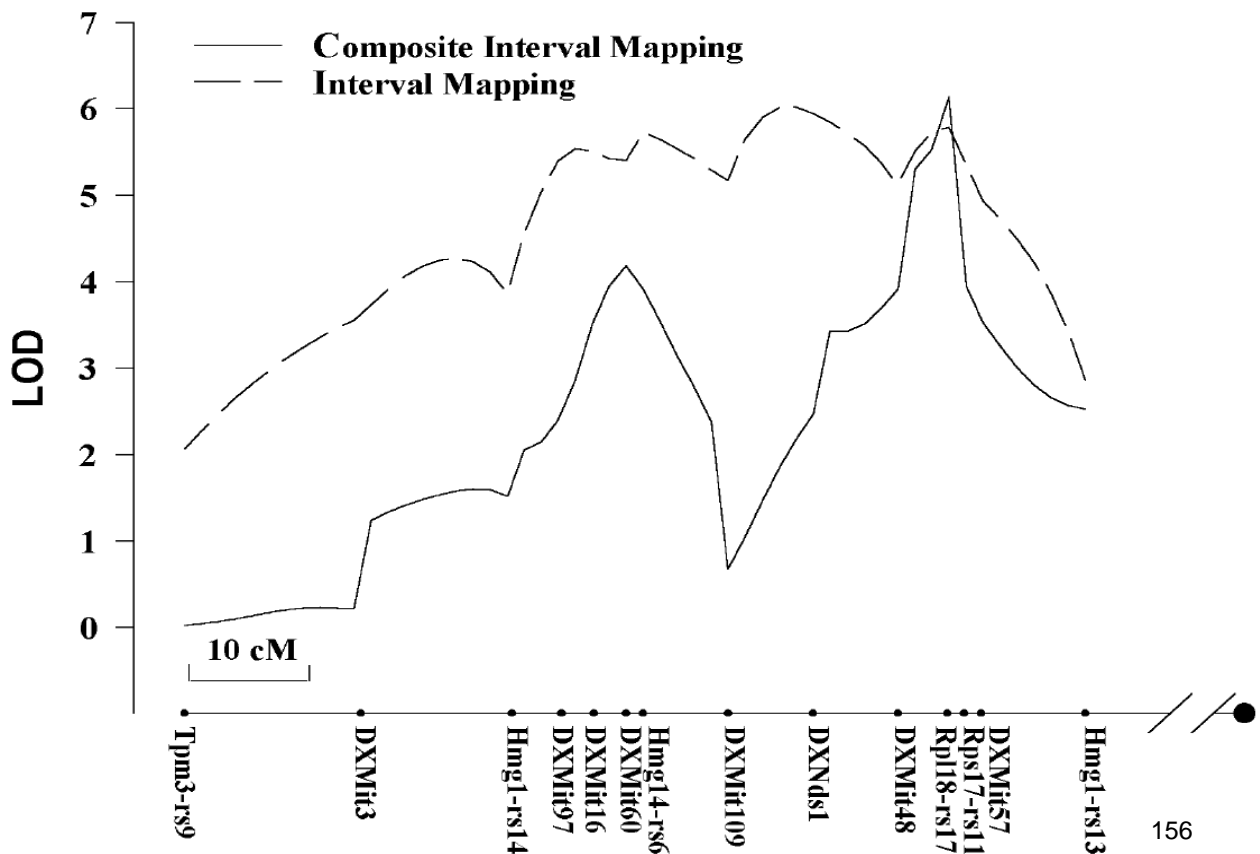
Therefore,

$$LOD = \frac{1}{2} (\log_{10} e) LR = 0.217 LR$$

Thoughts...

- Interval mapping can be performed at any position covered by markers, and thus the method creates a systematic strategy of searching for QTL.
- The amount of support or evidence for a QTL at a particular map position is often displayed graphically through the use of likelihood maps or profiles
 - plots the likelihood ratio test statistic (or a closely related quantity) as a function of map position of the putative QTL.
- If the LOD score at a region exceeds a pre-defined critical threshold, a QTL is indicated at the neighborhood of the maximum of the LOD score with the width of the neighborhood defined by one or two LOD support interval (Lander and Botstein 1989).
- By the property of the maximum likelihood analysis, the estimates of locations and effects of QTL are asymptotically unbiased if the assumption that there is at most one QTL on a chromosome is true.

155



156

Variance explained by QTL

Sometimes the magnitude of a QTL is also reported as the proportion of the variance explained by the QTL $\hat{\sigma}_{\text{explained}}^2$ and is usually estimated as

$$\hat{\sigma}_{\text{explained}}^2 = \frac{\hat{\sigma}_{\text{total}}^2 - \hat{\sigma}_{\text{reduced}}^2}{\hat{\sigma}_{\text{total}}^2}$$

where $\hat{\sigma}_{\text{total}}^2$ is an estimate of the total phenotypic variance ($\hat{\sigma}^2$ of equation (9) at the null hypothesis) and $\hat{\sigma}_{\text{reduced}}^2$ is an estimate of the residual variance of the interval mapping model ($\hat{\sigma}^2$ of equation (9) at the alternative hypothesis).

Problem: estimates of the proportion of variation explained are not additive for multiple QTL, and usually overestimate the variance explained by a QTL.

***A more appropriate way to estimate the variance explained by QTL effects will be discussed in multiple interval mapping.

157

Haley-Knott regression approximation

- A simplified approximation of the model in equation (3) was proposed by Haley and Knott (1992) and Martinez and Curnow (1992).
- Instead of treating X_i as missing data and using a mixture model via maximum likelihood for missing data analysis, the Haley-Knott approximation uses

$$p_{1i} = \Pr(x_i = 1 | M, N, \theta)$$

in the place of X_i and simplifies model (3) to

$$y_i = \beta_0 + \beta_1 p_{1i} + \varepsilon_i \quad i = 1, \dots, n \quad (10)$$

- Since this is a simple regression model, and the statistical analysis is straightforward.
- Haley and Knott (1992) and Rebai et al. (1995) have shown that this procedure gives a very good approximation to the likelihood profile for maximum likelihood interval mapping.
- Xu (1995) notes that this regression approach tends to overestimate the residual variance, and presents a correction.

158

Advantages and disadvantages

Compared with single marker analysis, interval mapping has several advantages:

1. The probable position of the QTL can be inferred by a support interval.
2. The estimated locations and effects of QTL tend to be asymptotically unbiased if there is only one segregating QTL on a chromosome.
3. The method requires fewer individuals than single marker analysis.

159

There are many problems with interval mapping:

1. The test is not an interval test
 - a test which is able to distinguish whether or not there is a QTL within a defined interval, independent of the effects of QTL that are outside a defined region.
2. Even when there is no QTL within an interval, the likelihood profile for the interval can still exceed the significance threshold if there is a QTL at some nearby location on the chromosome.
 - if there is only one QTL on a chromosome, this effect, though undesirable, may not matter because the QTL is more likely to be located at the location which shows the maximum likelihood profile
 - however, the number of QTL on a chromosome is unknown.

160

3. If there is more than one QTL on a chromosome, the test statistic at the position being tested will be affected by other QTL
 - the estimated positions and effects of “QTL” identified by interval mapping are likely to be biased.

4. It is not efficient to use only two markers at a time to test for QTL
 - the information from other markers is not utilized.

161

Permutation Thresholds for QTL Mapping

R.W. Doerge

Summer Institute in Statistical Genetics

162

Search for QTL

- Single Marker Methods:
 - t-test
 - F-test (ANOVA)
 - Regression
 - Likelihood based tests
- Interval Mapping:
 - LOD score (Lander and Botstein)
 - LRT (likelihood ratio test)
- Composite Interval Mapping:
 - LRT (Zeng, Jansen)

163

Recall: Interval Mapping...

Hypotheses:

H_0^A : no QTL

H_0^B : no QTL linked

H_a : QTL present and linked

Likelihood function:

$$L(\beta_0, \beta_1, \sigma^2 | Y, X, r_1, r_2) = \prod_1^{n_1} f_1 \prod_1^{n_2} f_2 \prod_1^{n_3} f_3 \prod_1^{n_4} f_4$$

$$LOD = \log_{10} \left[\frac{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, r_1, r_2)}{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, r_1 = r_2 = 0.5)} \right] \sim \frac{1}{2} \log_{10} e_{x_1^2}$$

164

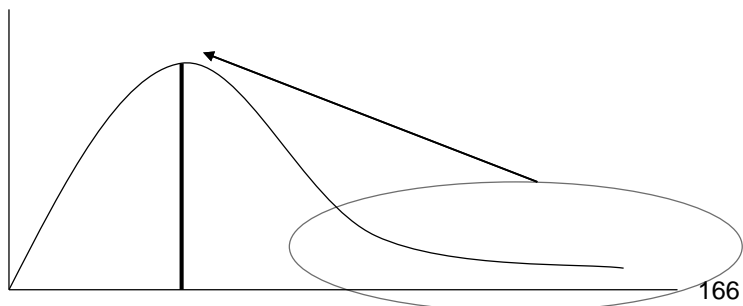
Permutation Thresholds

- Estimate the distribution of the test statistic
 - LOD or LRT
- Determine statistically significant QTL
- Empirically derived QTL thresholds
 - specific to the experiments

165

Statistical Issues

1. Distribution of trait values not always $N(\mu, \sigma^2)$.
 - mixture of distributions: $p_1N(\mu_1, \sigma^2) + p_2N(\mu_2, \sigma^2)$
2. Transformation of trait data (get rid of skewing...)
 - \log_{10} transformation (statistical fix...)
 - is it correct?



3. Smaller sample sizes
4. Statistical tests (for QTL) are not independent.
 - Bonferroni correction
 - statistical fix
 - does not address genetic issues
 - marker order and density
 - same hypothesis being tested
 - is the type I error (α) correct?
 - how do we come up with an appropriate critical value (from an unknown distribution) that reflects our Type I error?

167

Experimental Factors

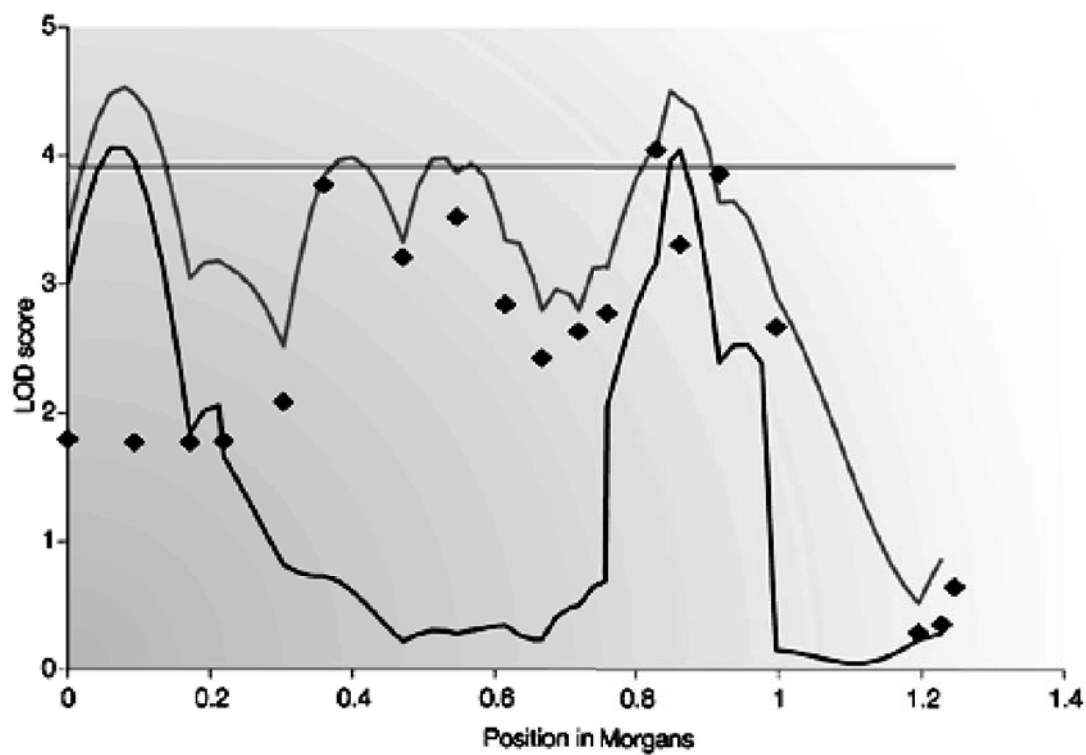
- Sample size
- Genome size
- Marker density
- Proportion of missing data
- Segregation distortion

168

Detecting QTL

- Single marker methods
 - compute test statistical at each marker
 - compare to (known?) statistical distribution
 - significant genotype-phenotype association?
- Multiple marker methods
 - order markers across genome
 - calculate test statistic at each position
 - compare to (known?) statistical distribution
 - significant genotype-phenotype association?

169



170

Permutation Theory Applied to QTL Analysis

- It is possible to derive the distribution of any test statistic under an appropriate null hypothesis by “**shuffling**” (Fisher 1935) the quantitative trait values among the individuals in the data set.
 - observations need to be “exchangeable”

- If there is a QTL effect at specific location(s) in the genome, there will be an association between the trait values and the point of analysis on the genetic map.
 - there is a phenotype-genotype association

- If there is no QTL present in the genome, or it is unlinked to the point of analysis, there is no phenotype-genotype association
 - *exactly* the situation described under the null hypothesis

171

“Shuffling Trait Data”

Original Trait Values:		
Individual	trait value	summary statistics
1	10.2	
2	11.1	
3	5.7	$n = 200$
.	.	$\bar{x} = 8.75$
.	.	$s = 2.57$
.	.	
200	9.7	

Shuffled Trait Values:		
Individual	trait value	summary statistics
1	5.7	
2	10.2	
3	9.7	$n = 200$
.	.	$\bar{x} = 8.75$
.	.	$s = 2.57$
.	.	
200	11.1	

“**Shuffling**” trait values among individuals in the data set represents the situation under the null hypothesis (randomness).

Steps for Estimating Significance Threshold Values

1. Hold the genetic map fixed.
2. “Shuffle” the trait values.
3. Analyze the ”shuffled” data set
 - t-test
 - likelihood ratio test
 - LOD score
4. Store the test statistic at each analysis point of step 3 in an “**Analysis Matrix**”.
5. Repeat steps 2-4 N times (Doerge and Churchill 1996).

173

Analysis Matrix

- Let $N = 1000$ shuffles of the original data set.
- Denote test statistic value by ts_{ij} ; $i = 1, \dots, 1000$, $j = 1, \dots, k$.
- There are k analysis points.

Shuffle Number	Analysis Points					
	1	2	3	4	...	k
1	ts_{11}	ts_{12}	ts_{13}	ts_{14}	...	ts_{1k}
2	⋮	⋮	⋮	⋮	⋮	⋮
3	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	ts_{10001}	ts_{1000k}

- Comparisonwise Threshold Values (per marker)
- Chromosomewise Threshold Values (chromosome specific)
- Experimentwise Threshold Values (experiment specific)

174

- Each analysis point is essentially being sampled from the null distribution of the test statistic.

- From the N sets of analyses, we can develop
 - comparisonwise critical values
 - experimentwise critical values.

- We use $N = 1000$ (number of permutations or “shuffles”)
 - ($\alpha = 0.05$).

175

Comparisonwise, Experimentwise, and Chromosomewise Threshold Values

- **Recall... Type I Error:** Reject the null hypothesis (no QTL) in favor of the alternative hypothesis (QTL effect) when there is really no QTL effect linked to the testing position.

- **Comparisonwise Threshold Values (per “marker”):**
 - order the N test statistics obtained at each analysis point in the map and find the $100(1-\alpha)$ percentile
 - using this critical value to define a test controls the type I error rate at that point to be α or less.

176

- **Experimentwise Threshold Values (“genome”-wise):**
 - obtain the maximum test statistic over all analysis points for each of the N analyses.
 - order these N values
 - the $100(1-\alpha)$ percentile is the estimated experimentwise critical value.
 - the experimentwise threshold value provides detection of the presence of a QTL somewhere in the genome while controlling the overall type I error rate to be α or less.

- **Chromosomewise Threshold Values:**
 - limit the scope of the analysis to one chromosome
 - treat this one chromosome as “the experiment” , and estimated the chromosomewise (“experiment”) threshold.

177

α

- Using comparisonwise thresholds of this kind increases the type I error rate over the entire genome to be much higher than α .

- The experimentwise critical value will be higher than the comparisonwise value since we are controlling the type I error rate over the entire genome.

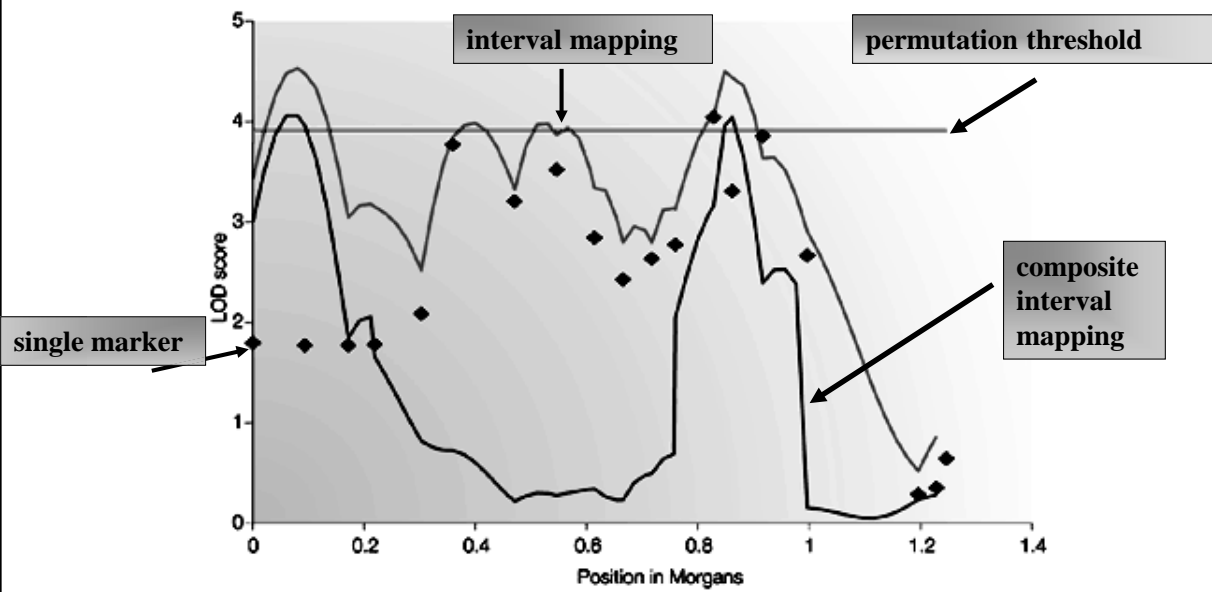
178

Summary

- specifics of the experiment affect the statistics
- test statistics for real experiments may not follow standard distribution
- empirical threshold values are specific to experiment
- excellent application for parallel computing

179

QTL mapping methodology



180

Example of QTL Mapping Experiment: single marker, interval mapping, composite interval mapping, permutation thresholds

R.W. Doerge

Summer Institute in Statistical Genetics

181

Chromosome 11 Data of Mouse: Fine Mapping Using QTL-Cartographer...

```
#
-n      291   is the sample size
-p      173   is one more than the number of markers
-cross  RF2   is the type of cross
-traits 1     is the number of traits
-Names of the traits...
  1 sev
-otraits 0   is the number of other traits
#From here, the first number is the individual
# the second is a 1 or 2 (for BC1, BC2 in Design III),
# and then come the 172 marker values and finally the trait values.
-s
  1  1
    1  1  1  0  0  0  0  0  0  0  1 -1
    1  1  1  1  1  1  0  0  0  0  1  1  1
    0  0  1 -1  1  1  2 -1  2  2  2
    1  1  1  1  1  1  1  0  0  0  1
    1  0  0  0  0  0  0
    1  1  0  0
    1  0  0  0  0  0  0  0  1  1  1  1  1  1  1
    2  1  0  0  0
    1  1  1  1  1  0  1
    1  1  1  1  1  1  1
    1  1  1  1  1  1  1  1  1  1  0  0  0  0  0  0  0  0
    1  1  1  1  1  1  1  1
    0  0  0  0  0  1  1
```

182

```

1 2 2 -1
0 0 2 2 2 2 2 1
2 2 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 -1 2 2 2
1 1 1 1 1 1
1 1 1 1 1

```

2.333330000000

2 1

```

0 0 0 0 0 1 1 1 1 2 2 2
1 0 0 0 1 1 1 2 2 2 2 2
0 0 0 -1 0 0 0 -1 1 1 1
0 0 1 1 2 2 2 2 2 1 1
2 1 1 0 0 0 0
1 1 0 1
1 1 1 1 1 1 1 1 1 0 1 1 1 1
2 2 1 1 0
1 1 1 1 1 1 1
0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2
2 2 2 2 2 1 1 1
1 1 1 1 1 1 1
2 0 0 0
2 0 0 2 2 2 2 2
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0
1 1 2 2 2 1
0 0 0 0 0

```

2.000000000000

3 1

```

0 1 0 0 0 0 0 0 0 2 2 2
0 0 0 0 0 -1 0 1 1 1 2 2 2
1 1 1 1 1 1 1 1 1 1 1

```

.
.
.
.

```
doerge@rwdstat(NEW_052302.D)% Qstats
```

```
=====
QTL Cartographer v. 1.16c, February 2002 for Unix
Copyright (C) 1996-2001 C. J. Basten, B. S. Weir and Z.-B. Zeng.
QTL Cartographer comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome to redistribute it
under certain conditions. For details see the file COPYING.
=====
```

```
=====
No.           Options                               Values:
-----
0. Continue with these parameters
1. Data Input File                          severity.cro
2. Output File                               severity.qst
3. Error File                                severity.log
4. Genetic Linkage Map File                  severity.map
5. Random Number Seed                        1022191071
-----
6. Specify Resource File                     qtlcart.rc
7. Change Filename stem                      severity
8. Change Working Directory:
9. Quit
10. Quit, but update the Resource File
=====
```

```
Please enter a number...
```

185

```
# 1022189179 -filetype Qstats.out
#
# QTL Cartographer v. 1.16c, February 2002
# This output file (severity.qst) was created by Qstats...
#
# It is 16:26:19 on Thursday, 23 May 2002
#
#
# This output is based on the map in (severity.map)
# And the data in (severity.cro)
#
#
```

```
-----
This is for -trait 1 called sev
-----
```

```
-----
Sample Size..... 291
M(1)..... 0.6284
M(2)..... 0.5992
M(3)..... 0.7038
M(4)..... 0.9424
Mean Trait Value..... 0.6284
Variance..... 0.2050
Standard Deviation..... 0.4528
Coefficient of Variation... 0.7205
Average Deviation..... 0.3764
Skw..LW(24)..... 0.0712
.....Sqrt(6/n)..... 0.1436
```

186

min Y max

```

-----
n = 291
Min(Y) = 0.02381
Q1 = 0.24109
Median = 0.53659
Q3 = 0.949495
Max(Y) = 2.33333
-----

```

Summary of missing data for trait number 1 (sev)
with 291 individuals

n(m) individuals have marker data, n(m+t) have trait and marker data

```

-----
-----

```

Chrom	Mark	Name	type	n(m)	n(m+t)	%(m+t)
11	1	d11m72	co	284	284	97.59
11	2	d11m2	co	290	290	99.66
11	3	d11m294	co	274	274	94.16
11	4	d11m53	co	285	285	97.94

-begin missmark 1

189

11	5	d11m307	co	275	275	94.50
11	6	d11m20	co	263	263	90.38
11	7	d11m140	co	278	278	95.53
11	8	d11m155	co	274	274	94.16
11	9	d11m156	co	284	284	97.59
11	10	d11m29	co	271	271	93.13
11	11	d11b149	co	281	281	96.56
11	12	d11m194	co	288	288	98.97
11	13	d11m36	co	289	289	99.31
11	14	d11m38	co	248	248	85.22
11	15	d11m285	co	285	285	97.94
11	16	d11m330	co	285	285	97.94
11	17	d11m168	co	291	291	100.00
11	18	d11m167	co	288	288	98.97
11	19	d11m128	co	289	289	99.31

-end missmark

Here is a summary of missing data for each individual.

There are 172 markers, 1 traits and 0 categorical traits.
The table below lists the raw number and the percentage of data
points for each individual.

```

-----
-----

```

Individual	Markers	Traits	Cat. Traits
------------	---------	--------	-------------

190

Number	#	%	#	%	#	%	
-----							-begin missind
1	167	97.09	1	100.00	0	0.00	
2	170	98.84	1	100.00	0	0.00	
3	166	96.51	1	100.00	0	0.00	
4	170	98.84	1	100.00	0	0.00	
5	168	97.67	1	100.00	0	0.00	
6	166	96.51	1	100.00	0	0.00	
7	172	100.00	1	100.00	0	0.00	
8	169	98.26	1	100.00	0	0.00	
9	163	94.77	1	100.00	0	0.00	
10	171	99.42	1	100.00	0	0.00	
11	171	99.42	1	100.00	0	0.00	
12	171	99.42	1	100.00	0	0.00	
13	171	99.42	1	100.00	0	0.00	
14	168	97.67	1	100.00	0	0.00	
15	170	98.84	1	100.00	0	0.00	
16	170	98.84	1	100.00	0	0.00	
17	167	97.09	1	100.00	0	0.00	
18	171	99.42	1	100.00	0	0.00	
19	168	97.67	1	100.00	0	0.00	
20	169	98.26	1	100.00	0	0.00	
21	168	97.67	1	100.00	0	0.00	
22	169	98.26	1	100.00	0	0.00	
23	168	97.67	1	100.00	0	0.00	
24	170	98.84	1	100.00	0	0.00	
25	168	97.67	1	100.00	0	0.00	
26	169	98.26	1	100.00	0	0.00	
27	169	98.26	1	100.00	0	0.00	191

.

.

.

286	161	93.60	1	100.00	0	0.00	
287	167	97.09	1	100.00	0	0.00	
288	163	94.77	1	100.00	0	0.00	
289	162	94.19	1	100.00	0	0.00	
290	156	90.70	1	100.00	0	0.00	
291	163	94.77	1	100.00	0	0.00	

-end missind

Summary of marker segregation

Chrom	Mark	Name	type	n(m)	Chi2	LR	
11	1	d11m72	co	284	4.8169	5.0935	
11	2	d11m2	co	290	6.3517	6.7668	
11	3	d11m294	co	274	3.1971	3.2782	
11	4	d11m53	co	285	1.2737	1.3088	
11	5	d11m307	co	275	6.5855	6.9654	
11	6	d11m20	co	263	3.8973	4.0959	
11	7	d11m140	co	278	5.5683	5.7788	
11	8	d11m155	co	274	2.1679	2.2048	
11	9	d11m156	co	284	5.0423	5.3047	192

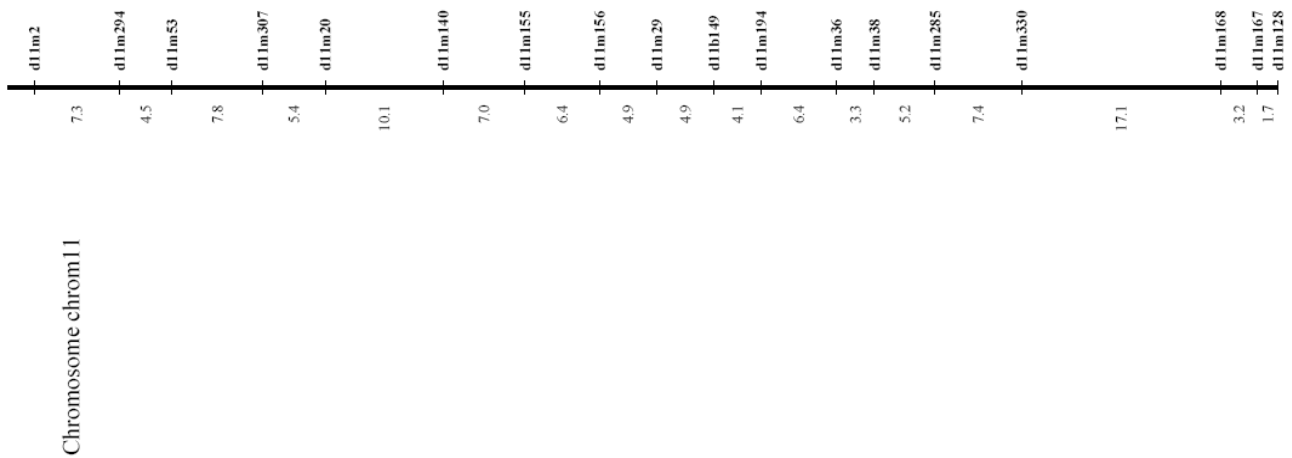
-begin segr

11	10	d11m29	co	271	3.9963	4.1107
11	11	d11b149	co	281	2.4875	2.5779
11	12	d11m194	co	288	2.2500	2.3331
11	13	d11m36	co	289	2.3356	2.4245
11	14	d11m38	co	248	3.6371	3.4865
11	15	d11m285	co	285	3.8211	4.0110
11	16	d11m330	co	285	4.3895	4.6236
11	17	d11m168	co	291	3.4124	3.4318
11	18	d11m167	co	288	1.8542	1.8702
11	19	d11m128	co	289	2.3910	2.4419

-----end | segreg

193

Estimated Genetic Map for Chromosome 11 from MAPMAKER/EXP



194

```

|      *
|      *
|      *
|      *
|      *
| * * *           Here is a histogram for the quantitative trait...
| * * *
| * * *
| * ***
| *****      *
| ***** * *           *
|***** * *           *
|***** ** * *           *
|***** ** * *           *
|***** ** ***          *
|***** ** *****      *
|***** ** ***** * * *
|***** ** ***** * *
|***** ** ***** * **
|***** ** ***** * **
|***** ** ***** * ***
|***** ** ***** **** *
|***** ** ***** **** *
|***** ** ***** **** **
|***** ** ***** * *
|***** ** ***** *
+-----+-----+-----+
0.02           1.18           2.33
min            Y            max

```

```

Min(Y) = 0.02381
Max(Y) = 2.33333

```

```
doerge@rwdstat(NEW_052302.D)% LRmapqtl
```

```

=====
      QTL Cartographer v. 1.16c, February 2002 for Unix
      Copyright (C) 1996-2001 C. J. Basten, B. S. Weir and Z.-B. Zeng.
      QTL Cartographer comes with ABSOLUTELY NO WARRANTY.
      This is free software, and you are welcome to redistribute it
      under certain conditions. For details see the file COPYING.
=====

```

```

=====
No.           Options                               Values:
-----
0. Continue with these parameters
1. Data Input File                               severity.cro
2. Output File                                   severity.lr
3. Error File                                    severity.log
4. Genetic Linkage Map File                      severity.map
5. Random Number Seed                            1022191173
6. Number of Permutations                        0
7. Trait to Analyze                              1
-----
8. Specify Resource File                         qtlcart.rc
9. Change Filename stem                          severity
10. Change Working Directory:
11. Quit
12. Quit, but update the Resource File
=====

```

Please enter a number...

This output is based on the map in (severity.map)
 And the data in (severity.cro)

Sample Size..... 291

This analysis fits the data to the simple linear regression model

$$y = b_0 + b_1 x + e$$

The results below give the estimates for b0, b1 and the F statistic for each marker. The F statistic is for the hypothesis that the marker is unlinked to the quantitative trait. The column headed by PR is the probability that the trait is unlinked to the marker. Significance at the 5%, 1%, 0.1% and 0.01% levels are indicated by *, **, *** and ****, respectively. LR is $-2\log(L_0/L_1)$.

This trait is: sev, and

-t 1 is the number of trait being analyzed.

```
-----
```

Chrom.	Marker	b0	b1	LR	F(1,n-2)	pr(F)
11	1	0.620	0.113	8.273	8.334	0.004 **
11	2	0.619	0.113	8.164	8.223	0.004 **
11	3	0.624	0.115	8.167	8.226	0.004 **
11	4	0.624	0.110	8.201	8.261	0.004 **
11	5	0.620	0.127	9.615	9.709	0.002 **
11	6	0.617	0.168	17.372	17.778	0.000 **** 197

11	7	0.621	0.156	14.771	15.048	0.000 ***
11	8	0.619	0.155	16.221	16.567	0.000 ****
11	9	0.620	0.144	13.101	13.308	0.000 ***
11	10	0.623	0.136	11.191	11.330	0.001 ***
11	11	0.622	0.136	12.137	12.308	0.001 ***
11	12	0.622	0.137	12.781	12.976	0.000 ***
11	13	0.620	0.165	18.627	19.104	0.000 ****
11	14	0.620	0.152	15.229	15.527	0.000 ***
11	15	0.618	0.164	17.762	18.189	0.000 ****
11	16	0.619	0.138	12.284	12.461	0.000 ***
11	17	0.628	0.046	1.336	1.330	0.250
11	18	0.628	0.050	1.631	1.624	0.204
11	19	0.626	0.068	2.976	2.971	0.086

```
-----
```

Here are the experimentwise significance levels for different sizes
 # Permutation significance level for alpha = 0.1 : 11.3327
 # Permutation significance level for alpha = 0.05 : 12.7709
 # Permutation significance level for alpha = 0.025 : 14.4815
 # Permutation significance level for alpha = 0.01 : 16.4431
 #end of shuffling results

doerge@rwdstat(NEW_052302.D)% Zmapqtl

```
=====
QTL Cartographer v. 1.16c, February 2002 for Unix
Copyright (C) 1996-2001 C. J. Basten, B. S. Weir and Z.-B. Zeng.
QTL Cartographer comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome to redistribute it
under certain conditions. For details see the file COPYING.
=====
```

```
=====
No.           Options                               Values:
-----
0. Continue with these parameters
1. Input File                               severity.cro
2. Output File                              severity.z
3. Error File                               severity.log
4. Genetic Linkage Map File                 severity.map
5. LRmapqtl Results file (Models 4&5)      severity.lr
6. SRmapqtl Results file (Model 6)         severity.sr
7. Random Number Seed                      1022191318
8. Model [1-6], 3=>IM                       3
9. Trait to analyze                         1
10. Chromosome to analyze (0=>all)         11
11. Walking speed in cM                     2.000000
12. Number of Background Parameters (Model 6) 5
13. Window Size in cM (Models 5&6)         10.000000
14. Number of Permutations                  0
15. Number of Bootstraps                    0
-----
```

199

- ```
16. Specify Resource File qtlcart.rc
17. Change Filename stem severity
18. Change Working Directory:
19. Quit
20. Quit, but update the Resource File
=====
```

Please enter a number...

200

```

1022185969 -filetype Zmapqtl.out
#
QTL Cartographer v. 1.16c, February 2002
This output file (severity.z) was created by Zmapqtl...
#
It is 15:32:49 on Thursday, 23 May 2002
#
#
#The position is from the left telomere on the chromosome
-window 10.00 Window size for models 5 and 6
-background 5 Background parameters in model 6
-Model 3 Model number
-trait 1 Analyzed trait [sev]
-cross RF2 Cross
Test Site * Like. Ratio Test Statistics * Additive * Dominance * Misc. HT
c m position H0:H3 H1:H3 H2:H3 H1:a H3:a H2:d
-s
11 1 0.0001 15.8142936 11.0045756 6.7958738 0.1820874 0.1003515 -0.1589014
11 1 0.0201 17.9706722 13.7760958 6.8457844 0.1983380 0.1043741 -0.1849326
.
.
.

```

201

**Composite Interval Mapping Model 1: Use all markers as cofactors.**

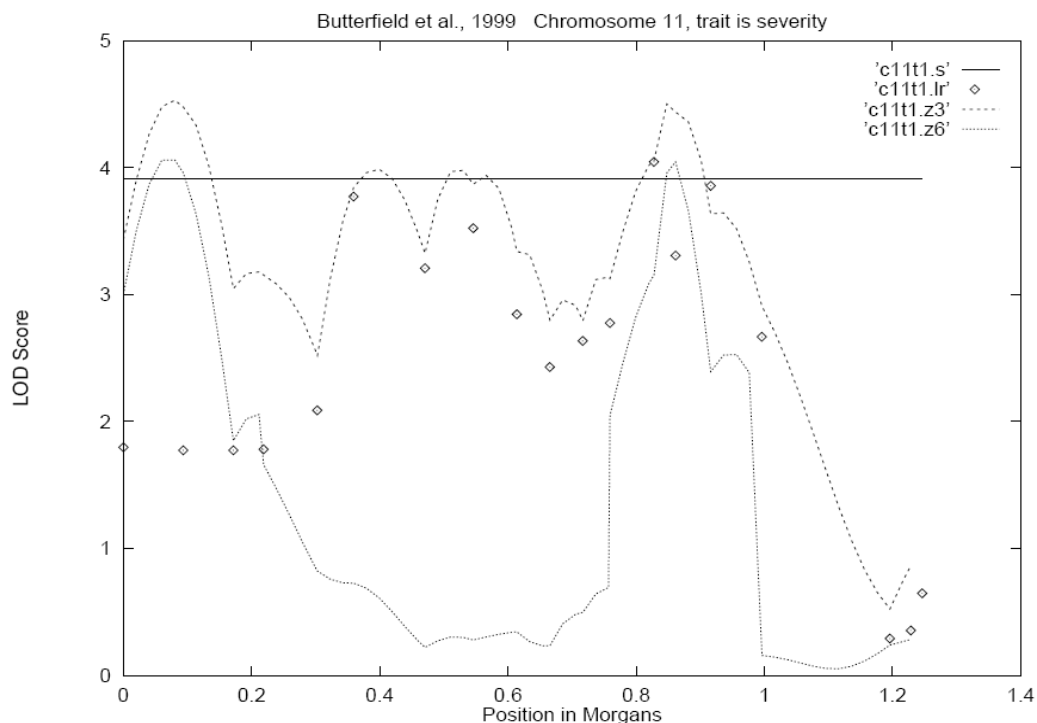
```

1022186131 -filetype Zmapqtl.out
#
QTL Cartographer v. 1.16c, February 2002
This output file (severity.z) was created by Zmapqtl...
#
It is 15:35:31 on Thursday, 23 May 2002
#
#
#The position is from the left telomere on the chromosome
-window 10.00 Window size for models 5 and 6
-background 5 Background parameters in model 6
-Model 1 Model number
-trait 1 Analyzed trait [sev]
-cross RF2 Cross
Test Site * Like. Ratio Test Statistics * Additive * Dominance * Misc. HT
c m position H0:H3 H1:H3 H2:H3 H1:a H3:a H2:d
-s
11 1 0.0001 3.8879597 7.4665236 0.3065377 0.1389767 0.0298707 -0.0918275
11 1 0.0201 5.5459852 11.7529063 0.0779620 0.1835906 0.0174144 -0.1175236
11 1 0.0401 7.4420245 17.1386366 0.0001878 0.2387654 0.0016548 -0.1398737
.
.
.

```

202

**Single marker, interval mapping, composite interval mapping, and permutation thresholds...**



203

## Interval Mapping Simulation Exercise

- Backcross
  - $n=400$ ;  $m=230$
  - some missing data
  - 5 quantitative traits
  - 5 QTL
    - Chromosome 1, 2, 3 each have one QTL
    - Chromosome 4 has no QTL
    - Chromosome 5 has two QTL
    - All QTL are independent

204



# Simulation set-up

- 5 chromosomes
- 5 quantitative traits
- 5 independent QTL

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i; \quad \varepsilon \sim N(0, \sigma^2)$$

$$\beta_0 = 100$$

| chrom \ trait               | 1<br>1 QTL   | 2<br>1 QTL    | 3<br>1 QTL   | 4<br>no QTL | 5<br>2 QTL               |
|-----------------------------|--------------|---------------|--------------|-------------|--------------------------|
| trait 1<br>$\sigma^2 = 40$  | p30<br>a=5.0 | p195<br>a=2.5 | p36<br>a=1.0 | -           | p103, p151<br>a=2.5, 2.5 |
| trait 2<br>$\sigma^2 = 20$  | ↓            | ↓             | ↓            | ↓           | ↓                        |
| trait 3<br>$\sigma^2 = 10$  | ↓            | ↓             | ↓            | ↓           | ↓                        |
| trait 4<br>$\sigma^2 = 5$   | ↓            | ↓             | ↓            | ↓           | ↓                        |
| trait 5<br>$\sigma^2 = 2.5$ | ↓            | ↓             | ↓            | ↓           | ↓                        |

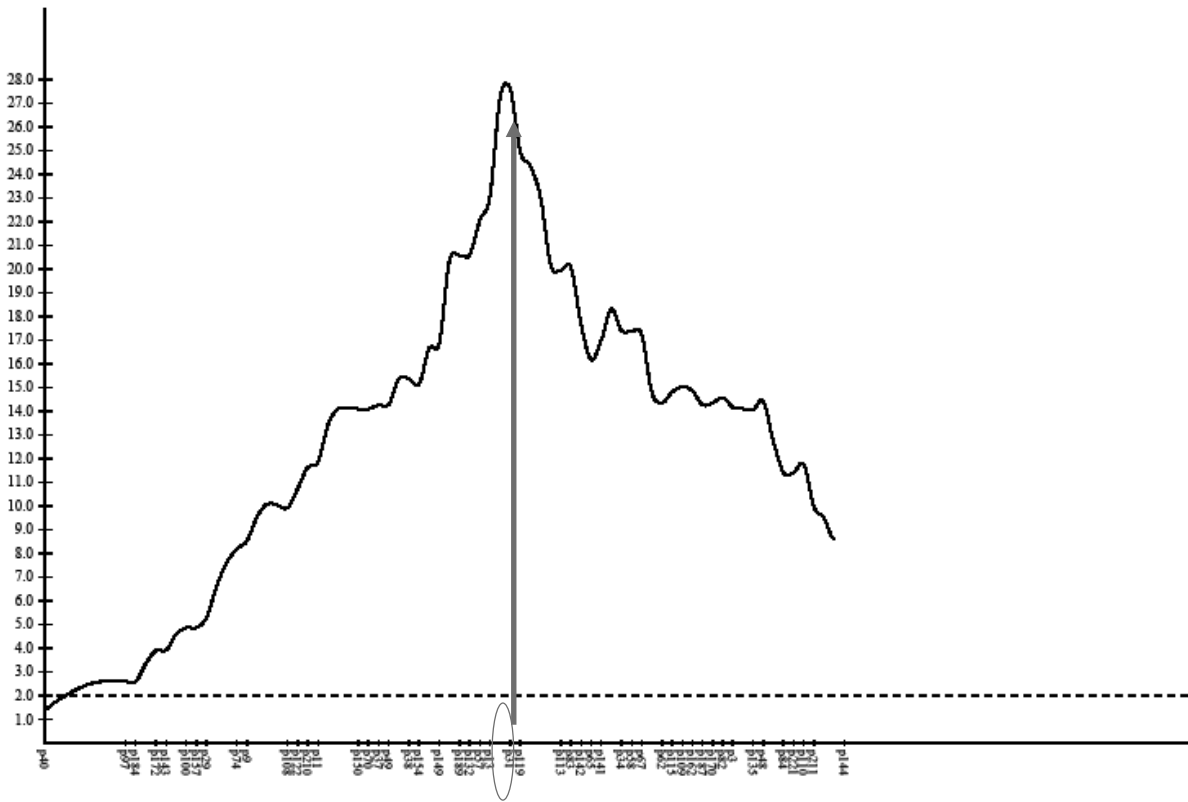
data type f2 backcross

400 230 5 0

|     |      |   |   |   |   |   |   |      |
|-----|------|---|---|---|---|---|---|------|
| *p1 | H    | A | A | A | A | A | A | A    |
|     | A    | A | A | H | H | A | A | A    |
|     | ...  | A | H | H | H | H | H | A    |
|     | H    | A | H | A | A | H | H | A    |
|     | A    |   |   |   |   |   |   |      |
| *p2 | A    | H | H | H | A | H | A | H    |
|     | A    | A | A | A | H | H | A | A    |
|     | -    | A | A | H | A | H | A | A    |
|     | A    | H | A | H | A | A | A | H    |
|     | H... | H | H | H | A | H | H |      |
| *p3 | H    | - | H | H | A | A | H | A    |
|     | A    | H | A | H | A | H | A | H    |
|     | A    | H | A | A | A | A | A | H    |
|     | H    | H | H | H | H | H | H | -    |
|     | A    | A | H | - | H | H | A | A    |
|     | H    | A | H | A | H | A | - | H... |

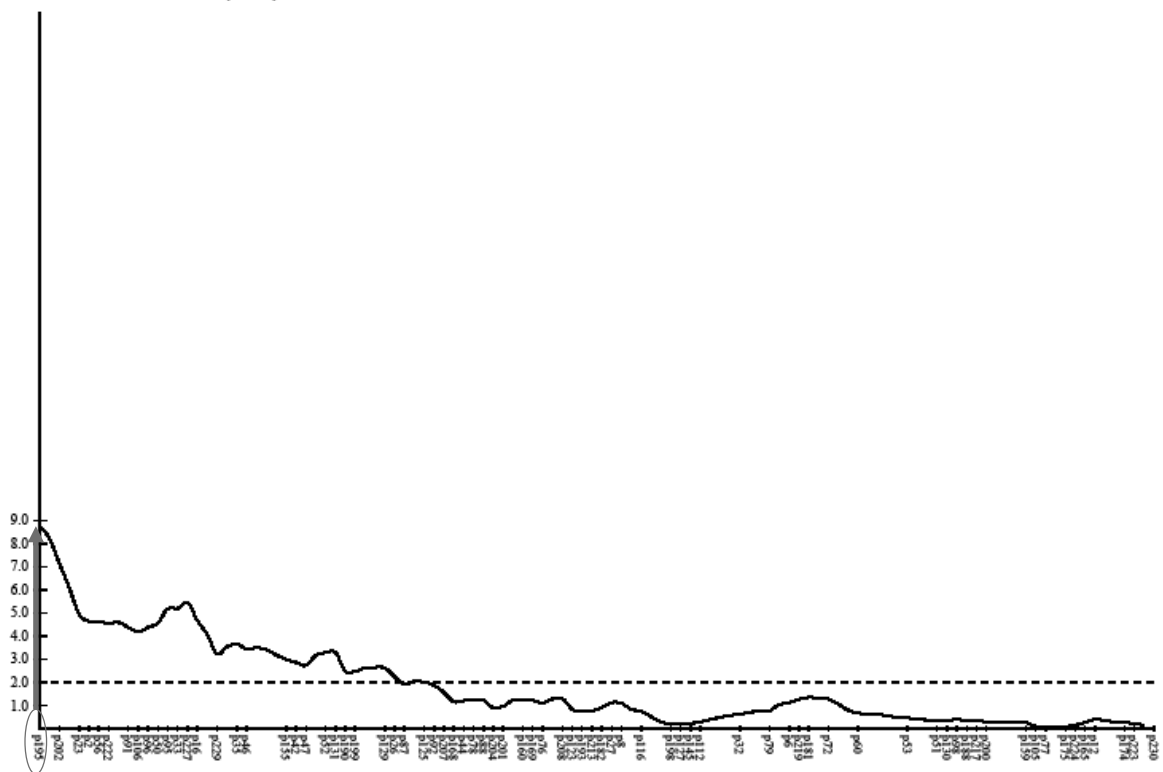
...

LOD score - Trait 5 (sim5)



207

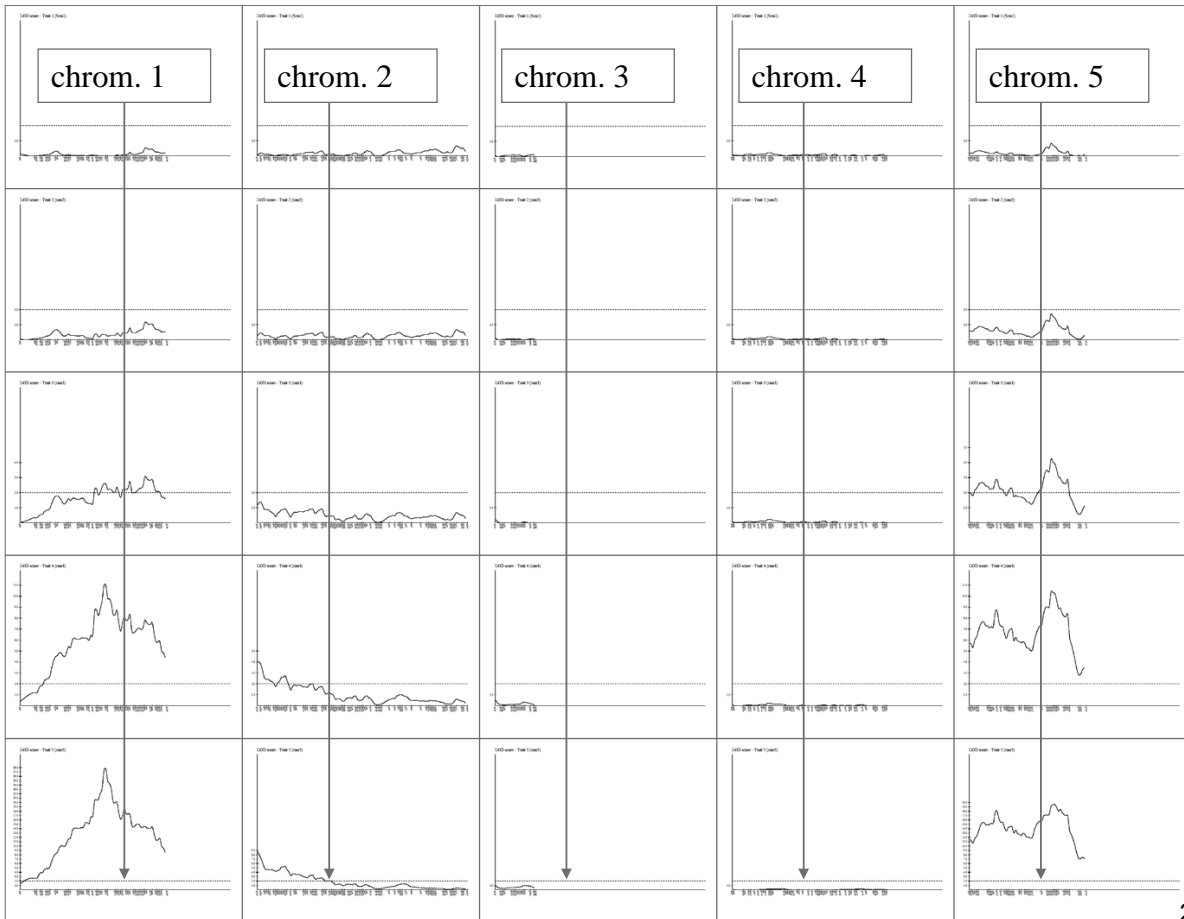
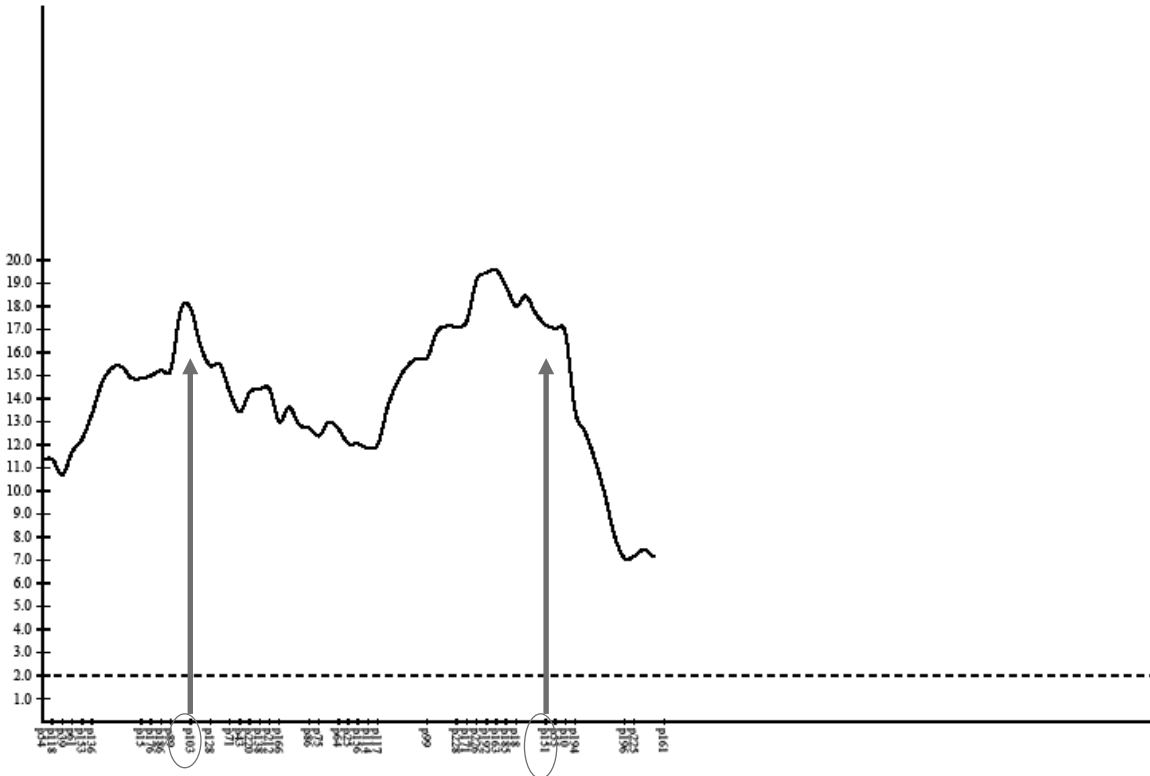
LOD score - Trait 5 (sim5)



208



LOD score - Trait 5 (sim5)





## Composite Interval Mapping

Zhao-Bang Zeng

*Summer Institute in Statistical Genetics*

## Motivation:

- Problem with interval mapping:
  - searching for a single QTL while there may be multiple QTL in the genome
  - the search for a QTL can be complicated and confounded by multiple QTL.
  
- Solution: Think about multiple QTL.

215

- Extension from one QTL to multiple QTL
  - explicitly model two or multiple linked QTL
    - multi-dimensional problem
      - gets complicated (see multiple interval mapping).

**Goal:** Test for QTL in an interval with a statistic that is independent of effects from other QTL along the chromosome.

- Improved precision and efficiency of mapping multiple QTL.

**Idea:** Employ interval mapping to scan for single QTL while using other markers as surrogates to absorb linked QTL effects.

- each interval test is independent
  - assumes no interference!
- thus, for an interval, such an approach tests and estimates only QTL in the interval.

216

- This approach simplifies searching for multiple QTL from a multiple dimensional search problem to one dimensional scan.

### Questions:

- How are marker cofactors selected?
- What considerations should be taken into account?

217

## Composite interval mapping

Composite interval mapping (CIM) is an extension of interval mapping using selected markers that are fitted in the model as cofactors and used to control the genetic variation of other possibly linked or unlinked QTL. The model is

$$y_i = \mu + b^* x_j^* + \sum_k b_k x_{jk} + e_j = b^* x_j^* + X_j B + e_j \quad (11)$$

where  $x_j^*$  refers to the putative QTL and  $x_{jk}$  refers to those markers selected for genetic background control. Appropriate selection of markers as cofactors is important for the analysis.

The likelihood function is

$$L(b^*, B, \sigma^2) = \prod_{j=1}^n \left[ p_{j1} \phi \left( \frac{y_j - b^* - X_j B}{\sigma} \right) + p_{j0} \phi \left( \frac{y_j - X_j B}{\sigma} \right) \right]$$

where  $p_{j1} = \Pr(x_j^* = 1)$  and  $p_{j0} = \Pr(x_j^* = 0)$  218

The likelihood ratio test statistic is

$$LOD = \log_{10} \frac{L(\hat{b}^*, \hat{B}, \hat{\sigma}^2)}{L(b^* = 0, \hat{B}, \hat{\sigma}^2)}$$

219

## Maximum likelihood analysis and the EM algorithm

The maximum likelihood analysis of a mixture model is usually via an EM (Expectation-Maximization) algorithm. In each iteration

▪ **E-step calculates:**

$$P_j = \frac{p_{j1} \phi\left(\frac{[y_j - b^* - X_j B]}{\sigma}\right)}{p_{j1} \phi\left(\frac{[y_j - b^* - X_j B]}{\sigma}\right) + p_{j0} \phi\left(\frac{[y_j - X_j B]}{\sigma}\right)}$$

▪ **M-step calculates:**

Note: vector notation

$$\begin{cases} b^* = (Y - XB)' P (1' P)^{-1} \\ B = (X' X)^{-1} X' (Y - P b^*) \\ \sigma^2 = \frac{1}{n} [(Y - XB)' (Y - XB) - 1' P b^{*2}] \end{cases}$$

▪ This process iterates until the estimates converge.

220



## Motivation: use of cofactors

- Cofactors are used to block the effects (outside the testing position) of other possible QTL along the chromosome.
- Consider three points (either markers or QTL)  $a$ ,  $b$ , and  $c$  on a chromosome

-----  $a$  -----  $b$  -----  $c$  -----

- Let  $r_{ab}$ ,  $r_{bc}$ ,  $r_{ac}$  be recombination frequencies between  $a$  and  $b$ ,  $b$  and  $c$ , and between  $a$  and  $c$ . For backcross and  $F_2$  populations, the correlation coefficients are

$$\gamma_{ab} = 1 - 2r_{ab}; \quad \gamma_{bc} = 1 - 2r_{bc}; \quad \gamma_{ac} = 1 - 2r_{ac}$$

- Assuming no crossing-over interference

$$r_{ac} = r_{ab}(1 - r_{bc}) + (1 - r_{ab})r_{bc} \Rightarrow (1 - 2r_{ac}) = (1 - 2r_{ab})(1 - 2r_{bc})$$

- That is,  $\gamma_{ac} = \gamma_{ab}\gamma_{bc}$

221

However, the correlation coefficient between  $a$  and  $c$  conditional on  $b$  (i.e., the partial correlation coefficient) is

$$\gamma_{ac \cdot b} = \frac{(\gamma_{ac} - \gamma_{ab}\gamma_{bc})}{\sqrt{(1 - \gamma_{ab}^2)(1 - \gamma_{bc}^2)}} = 0$$

This means for

-----  $q_1$  -----  $a$  -----  $b$  -----  $c$  -----  $q_2$  -----

$$\gamma_{bq_1 \cdot a} = 0 \quad \gamma_{bq_2 \cdot c} = 0$$

- Therefore, conditional on markers  $a$  and  $c$ , a test on the effect of  $b$  on a trait is *unaffected* by  $q_1$  and  $q_2$  despite the fact there may be linkage.
- This is the basis of composite interval mapping.

222

- The statistical power for the test of  $b$  (if it is a QTL) is affected by cofactors  $a$  and  $c$ , since the conditional test depends on the number of recombinants between  $a$  and  $b$  and between  $b$  and  $c$ .
- The closer the distance between  $a$  and  $b$  and between  $b$  and  $c$  becomes, there is less of a chance of recombination in the sample, and less statistical power for testing  $b$  conditional on  $a$  and  $c$ .
- Unlinked markers selected as cofactors (because they are likely to be close to other QTL) can potentially reduce the residual variance of the model, and thus increase the statistical power to search and test for QTL.

223

## Marker Selection

**Question:** Which markers should be added into the model?

- The answer to this question depends on the (unknown) number and (unknown) positions of underlying QTL.
- Too few selected markers may not achieve the purpose of reducing the most residual genetic variation
- Too many selected markers may reduce the power of the analysis.

224

# QTL-Cartographer

<http://statgen.ncsu.edu/qtlcart/cartographer.html>

**Zmapqtl (model 6):** module in QTL-Cartographer: a two parameter procedure

- $n_p$  = number of markers as cofactors
  - supplied by user or selected via stepwise regression by **SRmapqtl**.
- $w_s$  = width of testing window
  - blocks out a region of the genome on either side of the markers flanking the test site (supplied by user).

## Three step procedure:

1. **Cofactor step:** select  $n_p$  markers that are significantly associated with trait using (forward or backward) stepwise regression.
2. **Window marker step:** For each interval, the algorithm automatically picks 2 markers as a testing window, at least  $W_s$ cM beyond the testing interval (one for each direction).
3. **Mapping step:** Map QTL for the interval with window markers and a subset of markers outside the testing window as cofactors.

225

## Rules of thumb for Composite Interval Mapping:

- $n_p$  can be chosen from the results of the stepwise regression analysis (*SRmapqtl*) using *F*-to-enter (forward) or *F*-to-drop (backward) statistic with a specified significance level  $\alpha = 0.01$ .
- $w_s$  should be as large as possible when there is no indication of other linked QTL
  - otherwise,  $w_s$  can be gradually decreased as long as the test statistic for a putative QTL is significant.

226

**Example:** Interval mapping and composite interval mapping on chromosome X of the mouse data.

- experiment design: backcross
- $m=181$  microsatellite markers (SSR, simple sequence repeats)
- $n=103$  individuals.
- 20 chromosomes
  - this analysis using only 14 markers in chromosome X
  - The quantitative trait is 12 week body weight.
- **Composite Interval Mapping:**
  - The boundary markers  $x^L$  and  $x^R$  are chosen to be the closest markers which are at least 10cM away from the testing interval.
  - Besides  $x^L$  and  $x^R$ , 20 other linked or unlinked markers are also selected as cofactors (from stepwise regression) to absorb the effects of other QTL.

227

**Interval mapping analysis of (mouse) chromosome X:**

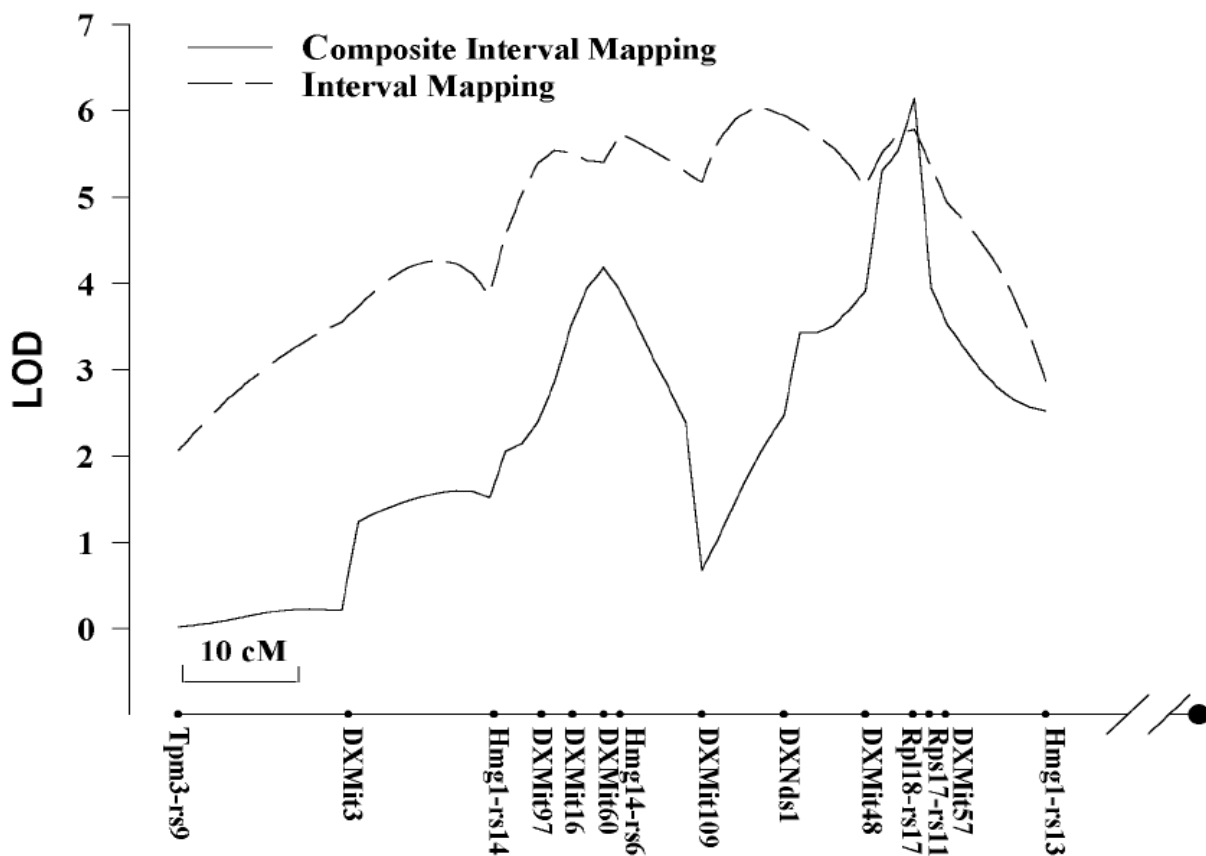
- The analysis from interval mapping indicates the existence of QTL on chromosome X
  - the LOD score is significant for a wide region
    - the arbitrary threshold is 3.3 for a backcross
  - not all significant peaks can be interpreted as QTL because of
    - linkage effects
    - the “ghost” gene phenomenon
    - statistical sampling effects.
  - the fact that a very wide region shows significant and comparable effects may suggest multiple QTL.

228

## Composite interval mapping analysis of (mouse) chromosome X:

- The boundary markers  $x^L$  and  $x^R$  are chosen to be the closest markers that are at least 10cM away from the testing interval.
- 20 additional linked or unlinked markers are also selected as cofactors in the analysis to absorb the effects of other QTL
  - markers are selected using stepwise regression
  - model 6 is employed via *Zmapqtl*
- The LOD score from this analysis reveals two distinct major peaks.
  - suggesting that there are at least two body weight QTL
    - one named *Bw1* is mapped near marker *Rp18-rs11*
    - the other, *Bw2*, mapped near *DXMIT60*
      - (Dragani et al. 1995 Mammalian Genome 6:778-781).
    - together the two QTL explain 25% of the phenotypic variance in the mapping population. In this case, the
  - Composite interval mapping achieved much better resolution in mapping QTL than interval mapping

229



230

## **Some limitations of composite interval mapping**

### **... motivation for multiple interval mapping**

- The analysis can be affected by an uneven distribution of markers in the genome
  - the test statistic in a marker-rich region may not be comparable to a test statistic in a marker-poor region
- It is difficult to estimate the joint contribution of multiple linked QTL to the phenotypic variance
- CIM is not directly extendible to the analysis of epistasis
- The use of tightly linked markers (as cofactors) can reduce the statistical power to detect a QTL

231

## **Multiple Interval Mapping**

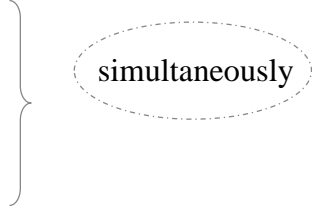
**Zhao-Bang Zeng**

*Summer Institute in Statistical Genetics*

232

## Multiple interval mapping

Multiple interval mapping (MIM) is a multiple QTL method that combines QTL mapping analysis with the analysis of genetic architecture of quantitative traits through a search algorithm that searches for

- number
  - positions
  - effects
  - interaction of significant QTL
- 
- simultaneously

The basic idea is to implement a multiple QTL model and use a search method to search for number and positions of multiple QTL.

233

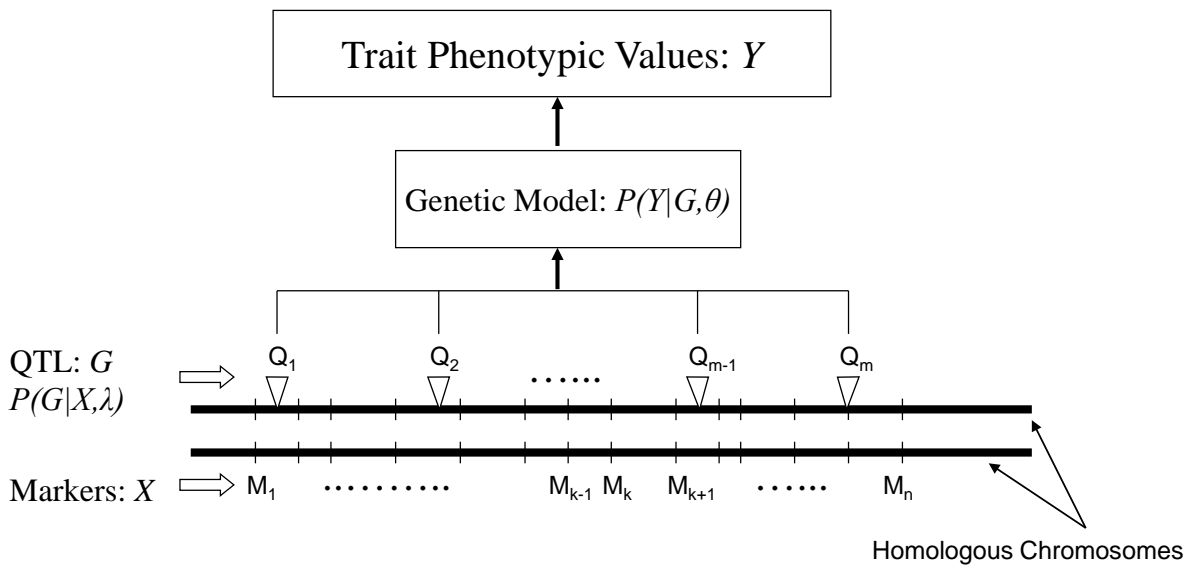
## Multiple Interval Mapping

MIM consists of four components:

- 1. An evaluation procedure** designed to analyze the likelihood of the data given a genetic model (number, positions, and epistasis of QTL).
- 2. A search strategy** optimized to select the best genetic model (among those sampled) in the parameter space.
- 3. An estimation procedure** for all parameters of the genetic architecture of the quantitative traits (number, positions, effects and epistasis of QTL; genetic variances and covariances explained by QTL effects) simultaneously given the selected genetic model.
- 4. A prediction procedure** to estimate or predict the genotypic values of individuals based the selected genetic model and estimated genetic parameter values for marker-assisted selection.

234

# QTL Mapping



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

Infer the relationship between genotypes and phenotypes

235

## MIM Model

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

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

where

- $y_i$  is the phenotypic value of individual  $i$
- $i$  indexes individuals of the sample;  $i=1, \dots, n$
- $\mu$  is the mean of the model
- $\alpha_r$  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
  - is unobserved, but can be inferred from marker data in sense of probability;

Continued...

236



Continued...

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

- $\beta_{rs}$  is the epistatic effect between putative QTL  $r$  and  $s$
- $r \neq s \in (1, \dots, m)$  denotes a subset of QTL pairs that each shows a significant epistatic effect
  - avoids 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 the residual effect of the model assumed to be  $N(0, \sigma^2)$

237

## Likelihood

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

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

The term in square braces is the weighted sum of a series of normal density functions, one for each of  $2^m$  possible multiple-QTL genotypes

- $p_{ij}$  is the probability of each multilocus genotype conditional on marker data;
- $E$  is a vector of QTL parameters ( $\alpha$ 's and  $\beta$ 's)
- $D_j$  is a vector specifying the configuration of  $x^*$ 's associated with each  $\alpha$  and  $\beta$  for the  $j^{\text{th}}$  QTL genotype;
- $\phi(y | \mu, \sigma^2)$  denotes a normal density function for  $y$  with mean  $\mu$  and variance  $\sigma^2$

238

## EM algorithm

**E-Step:**

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

**M-step:**

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

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

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

239

## Posterior genotype probability $\pi_{ij}$

Probability of QTL genotype given markers:

$$\Pr(\text{genotype} | \text{markers}) = \Pr(g | M) = p_{ij}$$

Conditional density of phenotypic given genotype:

$$\Pr(\text{phenotype} | \text{genotype}) = \Pr(y | g) = \phi(y_i | \mu + D_j E, \sigma^2)$$

Probability of QTL genotype conditional on markers and phenotype:

$$\begin{aligned} \Pr(\text{genotype} | \text{markers, phenotype}) &= \pi_{ij} \\ &= \Pr(g | M, y) = \frac{\Pr(g | M) \Pr(y | g)}{\sum_g \Pr(g | M) \Pr(y | g)} \end{aligned}$$

240

## Dealing with many QTL

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

### ***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  $\delta$  if needed); (3) normalize probs:  $\sum_j p_{ij} = 1$ .

Number of “significant” components  $\sim 10$ -100, depending on marker density, number and position of QTL. It has negligible loss of likelihood evaluation as compared to no selection. 241

## **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 (i.e., drop them);
  2. sum of “significant”  $p_{ij} > 0.95$  (adjust  $\delta$  if needed);
  3. normalize “significant” probabilities:  $\sum_j p_{ij} = 1$ .
- Number of “significant” components  $\sim 10$ -100, depending on marker density, number and position of QTL.
  - It has negligible loss of likelihood evaluation as compared to no selection.

## Conditional likelihood ratio test

Test for each QTL effect ( $E_r$ ) conditional on other QTL effects:

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

Proceed 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 on how many QTL to include?
- How do we select best genetic model?
  - number, positions, gene action, epistasis

243

## Model selection (*function in MIM QTL-Cartographer*)

1. Initial model (*New Model*): Use an automatic stepwise selection procedure, CIM, or stepwise marker selection.
2. Search for new QTL
  - *Refine Model* => *Search for New QTL* => *Search for QTL*
  - scan the genome to determine the best position of new QTL based on the criterion selected.
3. Search for QTL epistasis
  - *Refine Model* => *Search for New QTL* => *Search for Epistasis*
  - search for epistatic effects among QTL identified based on the selected criterion.
4. Re-evaluation
  - *Refine Model* => *Testing for Existing QTL*
  - re-evaluate the significance of each QTL effect currently fitted in the model based on the selected criterion.
    - this procedure can remove non-significant effects from the model.

244

## 5. Optimize QTL positions

- *Refine Model => Optimizing QTL Position*
- Optimize QTL position estimates in the current model.
- QTL position is optimized one by one in a sequential order

6. Return to step 2 and repeat the process as needed.

7. Selection criterion: Currently implemented are BIC and AIC.

245

## Challenges of searching for multiple QTL

- High, unknown dimension:
  - complicated, difficult.
- Search on whole genome,
  - not just markers
- Numerous peaks & valleys in likelihood “landscape”;
  - danger of selecting a local peak for from maximum.
- Appropriate criteria for model selection?
- Appropriate strategies to search for epistatic QTL?
- **Questions:**
  - Global (genomewide) search for multiple QTL
  - Genetic architecture: multiple components.

246

## Model selection and stopping rule

- 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  $2\log(\log n)$  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.

247

## Estimating the variance explained by QTL

Variance explained by QTL effect  $E_r$  can be estimated as

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

Covariance explained by QTL effect  $E_r$  and  $E_s$  is

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

Thus, the total genetic variance explained by QTL is

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

248

## Estimation of genotypic values

- The genotypic value of an individual can be estimated as:

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

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

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

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

- These estimates can be used for marker-assisted selection.

249

## Procedure for MIM analysis in QTL-Cartographer

- After opening Windows QTL-Cartographer
  - upload a data set
- Open **MIM** module:
  - choose *New Model* to select an initial model.
    - the default search procedure is pretty good
    - there are also a few other procedures implemented
  - choose *Refine Model* => *Optimizing QTL Position*.
  - choose *Refine Model* => *Search for New QTL* => *Search for QTL*
    - to look for more potential QTL.
  - choose *Refine Model* => *Search for New QTL* => *Search for Epistasis*
    - to look for QTL epistasis
      - note: given the identification of QTL, the criterion for searching QTL epistasis can be more relaxed
      - recommend: select **AIC** in the box of **Criteria for MIM Model Selection**).

250

- choose *Refine Model* => *Testing for Existing QTL*
  - to see whether the selected QTL effects are still significant based on the selected criterion.
- choose *Refine Model* => *MIM Model Summary* => *Graphic Result File*
  - to calculate and display the likelihood profile for each QTL.
- choose *Refine Model* => *MIM Model Summary* => *Model Summary File*
  - to show the MIM output result file.
    - information includes: position, likelihood ratio and effect of each QTL, epistatic effects of QTL, partition of the variance explained by QTL (main and interaction effects), estimates of genotypic value of individuals based on the model.
- There are also many other interactive functions in **MIM** module.

251

## More on epistasis: Why study epistasis?

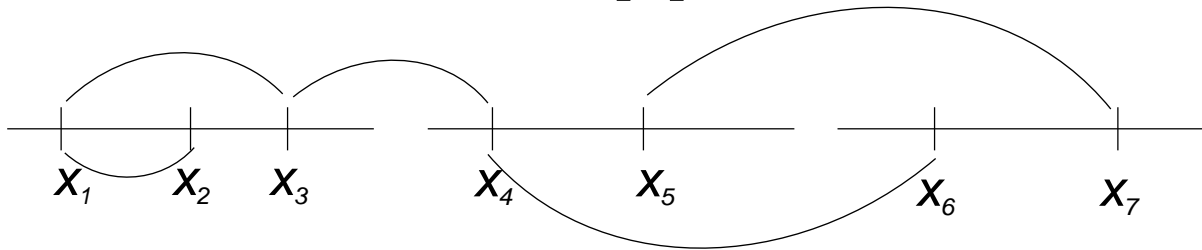
- Experiments show that most QTL have mainly additive (or marginal, or main) effects
- It is difficult to find significant QTL epistasis:
  - Small sample size
  - Multiple dimensional genome search: low statistical power
- In reality, search for QTL has always been biased for main-effect QTL, and not much effort has been put for searching for epistatic QTL
- Still, **QTL epistasis is ubiquitous biologically**

**Thus, it is important to be able to identify epistatic QTL for the purpose of our understanding of genetic complexity and also for the completeness of statistical inference of genotype and phenotype relationship**

252



## Statistical setting for the problem (for backcross population):



$$y_k = \mu + \sum_i \alpha_i x_{ik} + \sum_{i < j} \beta_{ij} x_{ik} x_{jk} + e_k \quad \text{for } k = 1, 2, \dots, n$$

with

$$x_{ik} = \begin{cases} \frac{1}{2} & \text{for } Q_i Q_i \\ -\frac{1}{2} & \text{for } Q_i q_i \end{cases}$$

$$p(x_{ik} = \frac{1}{2}) = \frac{1}{2}$$

253

### Questions:

- How to find each individual QTL ( $Q$ ), particularly those  $Q_i$  with weak  $\alpha_i$  but strong  $\beta_{ij}$ ?
- How to avoid false positive or incorrect identification of epistatic QTL?
- Also how to increase the statistical power of identifying epistatic QTL?

#### **Terminology:**

- **Main-effect QTL:** those QTL that have strong main effects ( $\alpha$ ), and may or may not have strong epistatic effects ( $\beta$ )
- **Epistatic QTL:** those QTL that have strong epistatic interactions ( $\beta$ ) (with other QTL), and may or may not have strong main effects ( $\alpha$ )

254

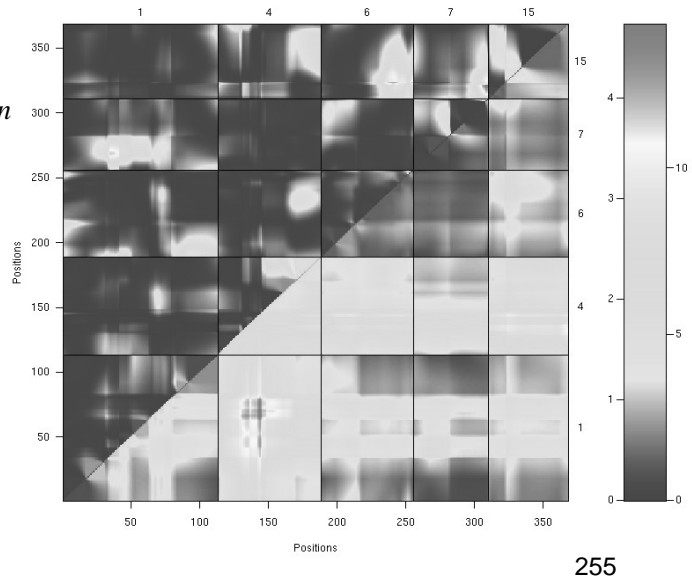
## A popular approach (simple, but ...)

Perform a 2D genome-scan to search for the best pair of  $x_1$  and  $x_2$  based on the statistical test of  $\alpha_1$ ,  $\alpha_2$  and  $\beta_{12}$ , and interpret the result on the face value.

### Statistical model:

$$y_k = \mu + \alpha_1 x_{1k} + \alpha_2 x_{2k} + \beta_{12} x_{1k} x_{2k} + e_k; k = 1, 2, \dots, n$$

- test for all combinations of genome positions for  $x_1$  and  $x_2$
- the upper-triangle shows the statistical test for  $\alpha_1$  and  $\alpha_2$
- the lower-triangle shows the statistical test for  $\beta_{12}$



## Problems... if using this simple 2D genome-scan for searching and interpreting epistatic QTL

- Search for epistatic QTL based on the 2D pattern can be misleading very easily
  - Due to complex linkage and epistatic structure of multiple QTL
- Low statistical power for this 2D genome-scan
  - Genetic variation due to other QTL effects is not fitted in the model, thus remains in the residual

### Problems: Potential bias and low statistical power

- This is a multiple QTL problem (not, a two-QTL problem) and needs a multiple QTL solution
- The challenge becomes... how to design a better analysis approach for a multiple epistatic QTL problem?

257

### **Another approach:**

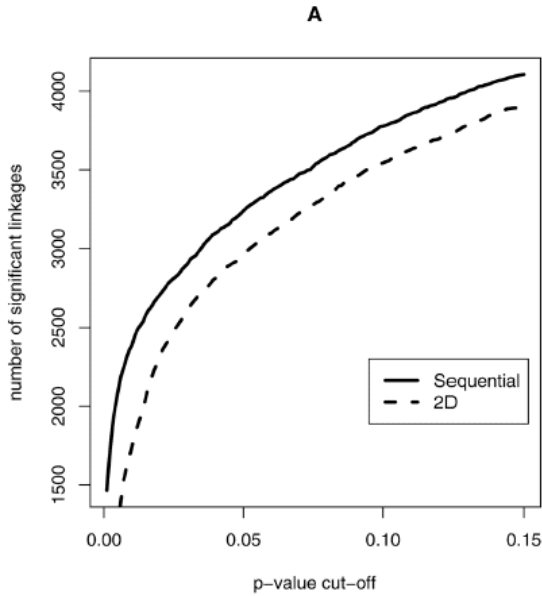
Extend the 2D search to multiple-D search for multiple epistatic QTL

- Potential problems:
  - unknown dimension in the search
  - need to assess and control statistical noise in a multiple dimensional search
    - multiple-D search is not necessarily powerful statistically
  - computational burden

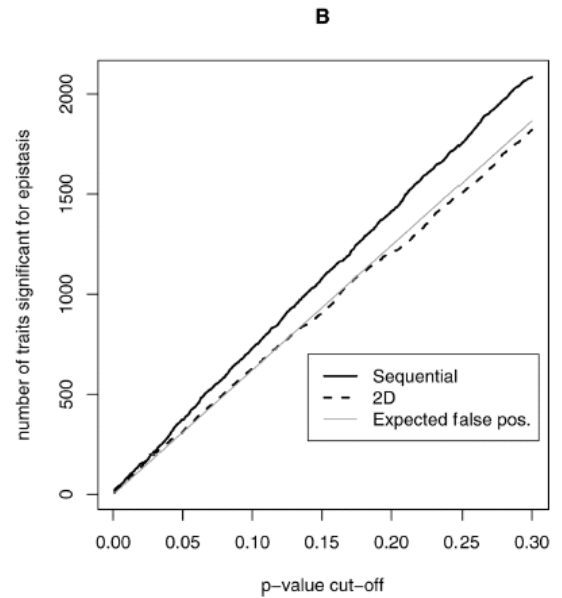
258

# A Power Comparison between 2D Locus Pair Search and Sequential Search (Storey et al. 2005)

Statistical power to detect 2 eQTL



Statistical power to detect 2 eQTL epistasis



259

## An important genetic property

$Cov(x_i, x_j) \neq 0$  if linkage  $\Rightarrow$  LD

$Cov(x_i x_j, x_k x_l) \neq 0$  if linkage  $\Rightarrow$  LD

... but,

$Cov(x_i, x_i x_j) = 0$ ;

$Cov(x_i, x_j x_k) = 0$  even with linkage

$$\Rightarrow V = \begin{bmatrix} \text{cov}(x_i, x_j) & 0 \\ 0 & \text{cov}(x_i x_j, x_k x_l) \end{bmatrix}$$

No covariance

$$y_k = \mu + \sum_i \alpha_i x_{ik} + \sum_{i < j} \beta_{ij} x_{ik} x_{jk} + e_k$$

**Implication:** The search for QTL with main effects can proceed separately from the search for QTL with epistasis without bias.

260

## Our solution: A three-stage search strategy

1. First search for main-effect QTL
  - either sequentially or other approach, each step uses 1-D genome scan),
  - then test for epistasis of identified QTL => identifying  $x_i$ 's
2. Search for epistatic QTL that interact with main-effect QTL
  - each 1-D genome scan => identifying  $x_i x_j$ 's
3. Search for additional epistatic QTL pairs
  - each 2-D genome scan => identifying  $x_i x_j$ 's with weak  $\alpha_i$  and  $\alpha_j$ , but significant  $\beta_{ij}$

$$y_k = \mu + \sum_i \alpha_i x_{ik} + \sum_{i < j} \beta_{ij} x_{ik} x_{jk} + e_k$$

261

## Justification and advantages:

1. Most QTL effects are due to “main effect” QTL that explain most genetic variance, thus need to be searched and fitted in the model first (before the subsequent analysis).
2. The search for main effect QTL *does not bias* the search for epistatic QTL.
3. After the main effect QTL are mapped and fitted in the model, further search for epistatic QTL has more statistical power.

### Advantages:

Minimize the bias and increase the statistical power

262

## How it works ...

### MIM Model and Likelihood

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

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

where  $x_{ik}$  is unobserved QTL genotype with known conditional probability from genetic markers and  $\varepsilon_i \sim N(0, \sigma^2)$ . Likelihood:

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

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

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.

263

- Since the method is based on a number of sequential searches (for main or epistatic QTL), and in each step of the search we test for significance of the searched QTL effects (main or epistatic effects). We need to figure out a way to compute the null distribution of the searched test statistic in each step efficiently.
- In each step we are testing the hypothesis (conditional on the other parameters):  $H_0: \beta=0$  vs.  $H_1: \beta \neq 0$ .

Let  $D$  be a set of indices representing the set of models examined. For a given model (represented by  $d \in D$ ), assume that the model has  $c$  parameters, split into two groups,  $\theta = (\eta, \beta) = (\theta_1, \dots, \theta_{c-1}, \beta, \mu, \sigma^2)$ , where  $\beta$  is the parameter to be tested for significance and  $\eta = (\theta_1, \dots, \theta_{c-1}, \mu, \sigma^2)$  is considered the vector of nuisance parameters. In the following,  $\tilde{\eta}$  denotes the maximum likelihood estimator of  $\eta$  for the reduced model with  $\beta = 0$ .

264

## How to assess the relevant threshold in the search process

### 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  $\eta$  and want to test for adding the  $m^{th}$  QTL with parameter  $\beta$ .

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

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

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

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

265

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

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

266

## Resampling with Score Statistic

An efficient way to simulate conditional null distribution

1. Generate  $G_i, i = 1, 2, \dots, n$  from  $N(0, 1)$ .
2. Calculate  $U^*(d) = \sum_{i=1}^n \hat{U}_i(d)G_i$ ,  $W^*(d) = U^{*'}(d)\hat{V}^{-1}U^*(d)$ , and  $S^* = \max_d W^*(d)$ .
3. Repeat step 1 and 2 for  $N$  times to find  $S_k^*$  for  $k = 1, \dots, N$ .
4. Compute the  $100(1 - \alpha)^{th}$  percentile of  $\{S_k^* : k = 1, \dots, N\}$  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.

267

Note that the  $\hat{U}_i(d)$  and  $\hat{V}(d)$  used in the resampling calculations are based on the original data and are evaluated once and used repeatedly in step 2; only the  $G_i$ 's are changed in each resample. Since it does not involve refitting the model in each iteration, the proposed method is computationally much more efficient than the permutation method.

268



# Implementation in QTL-Cartographer

## MIM procedures

- 1 MIM forward search procedure (a pre-model selection): This is an automatic QTL search procedure that is intended for generating an initial MIM model for further analysis only. QTL is searched sequentially based on its main effect and added into the model subject to a score-statistic test with a genome-wide threshold. Upon detecting a new QTL, interaction effects of the new QTL with the previously identified QTL are tested and added into the model using a score-statistic test with a point-wise threshold.
- 2 Optimizing QTL positions with or without interaction effects: This procedure optimizes position estimate of each QTL in turn. When the option with interaction effects is chosen, both the main effect of the QTL and interaction effects with other QTL are used for optimizing the estimate of position. Otherwise, only the main effect of the QTL is used for optimizing the QTL position.

269

### 3 Search for New QTL:

- (a) QTL with main effects: This procedure searches for a new QTL based on a score-statistic test on the main effect with a genome-wide threshold.
- (b) QTL with interaction effects:
  - i Search for interaction effects among identified QTL: This procedure searches for significant interaction effects among identified QTL using a point-wise threshold.
  - ii Search for new QTL that have significant interaction effects with identified QTL: This procedure searches for a new QTL based on its interaction effect with an identified QTL. A one-dimensional genome scan is performed to search for the best position for a new QTL that has an interaction effect with any identified QTL. This interaction effect is tested by a score-statistic test subject to a threshold that takes into account the search space (genome-wide with multiple identified QTL).
  - iii Search for new QTL in pair that have significant interaction effects: This procedure performs a two-dimensional genome scan that searches for the best positions for a pair of new QTL that have

270

#### 4 Testing QTL effects:

- (a) Testing QTL main effects: This procedure tests for significance the main effects of QTL in the current MIM model. If a QTL main effect is not significant at a genome-wise significant level, the QTL will be eliminated from the model. The procedure can be used only for those QTL that do not have interaction effects with other QTL to avoid to eliminate QTL that have weak main effects but significant interaction effects with other QTL.
- (b) Testing QTL interaction effects: This procedure tests for significance the interaction effects of QTL in the current MIM model. If an interaction effect is not significant at a point-wise significant level, the interaction effect will be eliminated from the model.

5 Estimating QTL effects: This procedure produces test statistics (log-likelihood ratio test statistic and score statistic) and empirical p-value of the score statistic for each QTL effect in the current MIM model. 271

6 Producing summary output: This procedure produces a comprehensive report of information of the current MIM model in two output files. One output file includes estimates of QTL number, positions, main and interaction effects,  $R^2$  value of the model (an estimate of the broad-sense heritability) and partition of the  $R^2$  value into the variance components due to individual QTL main and interaction effects and the covariance components due to a pair of QTL effects (due to linkage disequilibrium), Equation (19) of Zeng et al. (1999). It also includes estimates of QTL genotypes and genotypic values of the trait for each individual, Equation (14) of Zeng et al. (1999). The other output file includes information for generating the log-likelihood profile of each QTL in the MIM model in graphic which displays automatically. The log-likelihood profile for each QTL utilizes the combined information of the main effect of that QTL and interaction effects of the QTL with other QTL in the model.

For practical data analysis, we recommend use these procedures in the following way.

- Procedure 1 can be used to search for an initial model.
- Procedure 2 can be used next to optimize QTL position estimates. Procedure 2 can be used repeatedly when the model structure is changed by adding or removing a QTL during the model fitting process.
- Before searching for new QTL, current QTL effects should be checked by first using procedure 4(b), then 4(a).
- Procedure 3(a) can be used to search for new QTL based on main effects. This procedure can be used for multiple times in conjunction with procedure 2 until not new QTL based on main effects can be found.
- Procedure 3(b)(i) can be used to search for significant interaction effects among identified QTL.
- Then procedure 3(b)(ii) can be used to search for additional QTL that have significant interaction effects with the other identified QTL. If an additional QTL is identified, procedure 2, 4(b) and 4(a) can be used to optimize the model and check the model again.

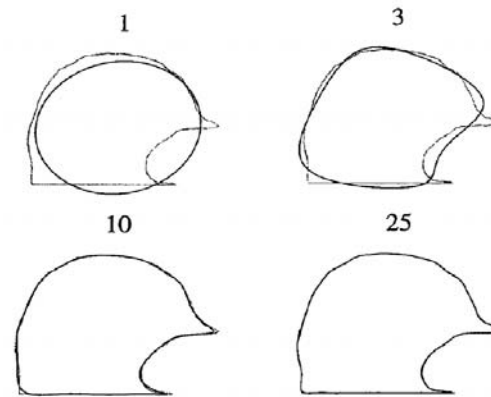
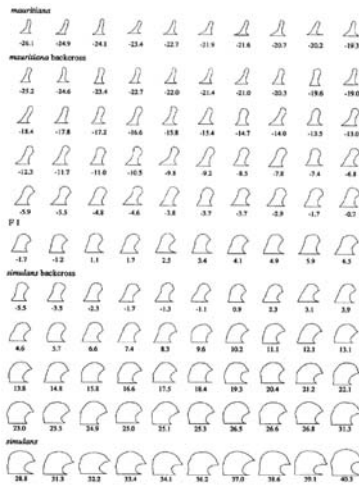
273

- Only after other procedures have been used repeatedly, i.e. QTL that have significant main effects or have significant interaction effects with the main effect QTL have already been identified and fitted in the model, should procedure 3(b)(iii) be used to search for additional new QTL that have significant interaction effects only. Procedure 3(b)(iii) should be used only in the last stage to minimize the risk of mapping epistatic QTL in wrong positions due to other unaccounted linked QTL effects. This point cannot be overemphasized enough.
- Procedure 6 can be used to generate a report for a MIM model.

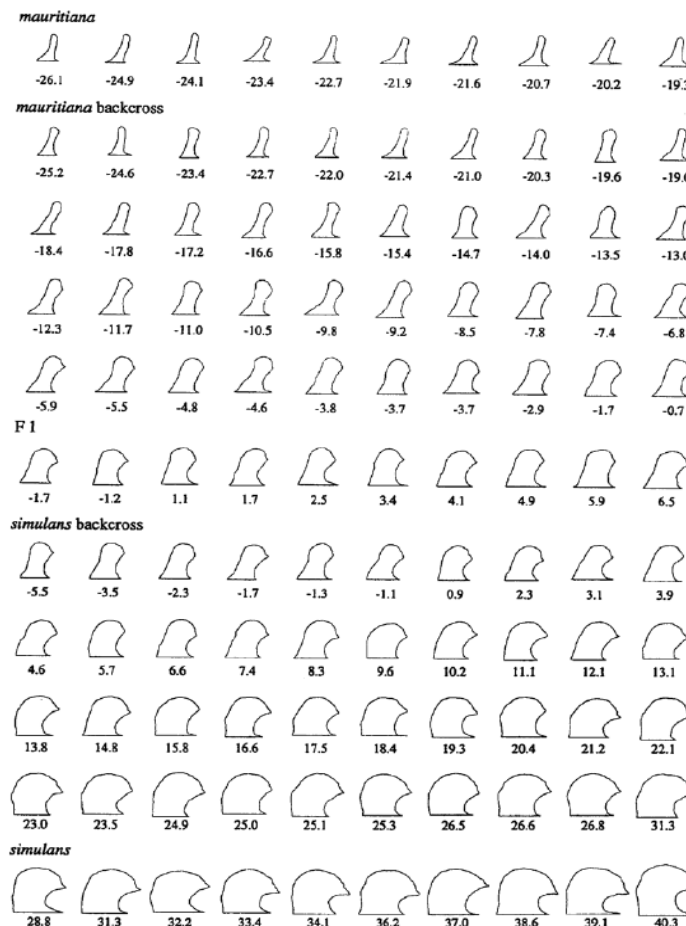
274

**Example 1 (MIM):** Genetic architecture of a morphological shape difference between two *Drosophila* species:

- **Population:** two backcrosses between *Drosophila simulans* and *D. mauritiana*, each having two independent samples of sizes 200 and 300.
  - total sample size about 1000.
- **Trait:** morphology of the posterior lobe of the male genital arch
  - analyzed as the first principal component in an elliptical Fourier analysis.



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



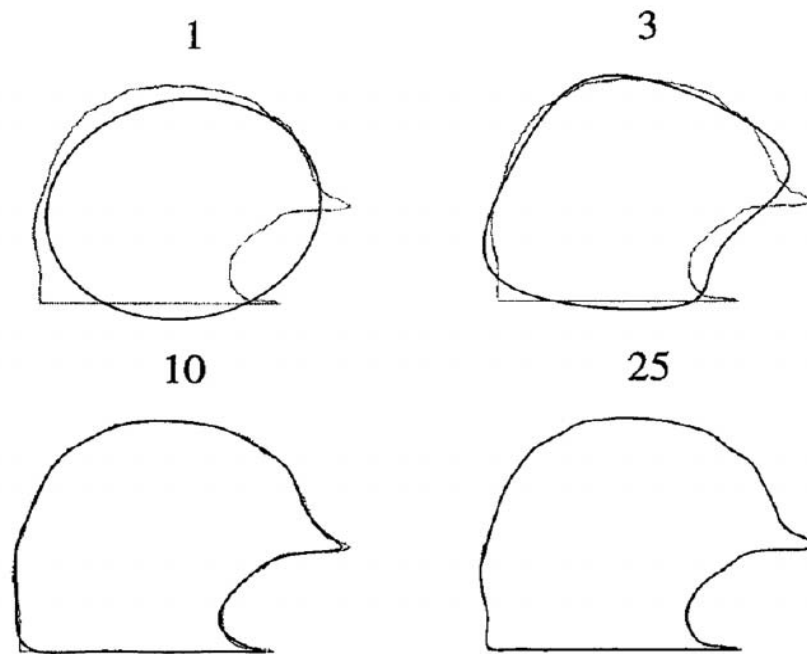
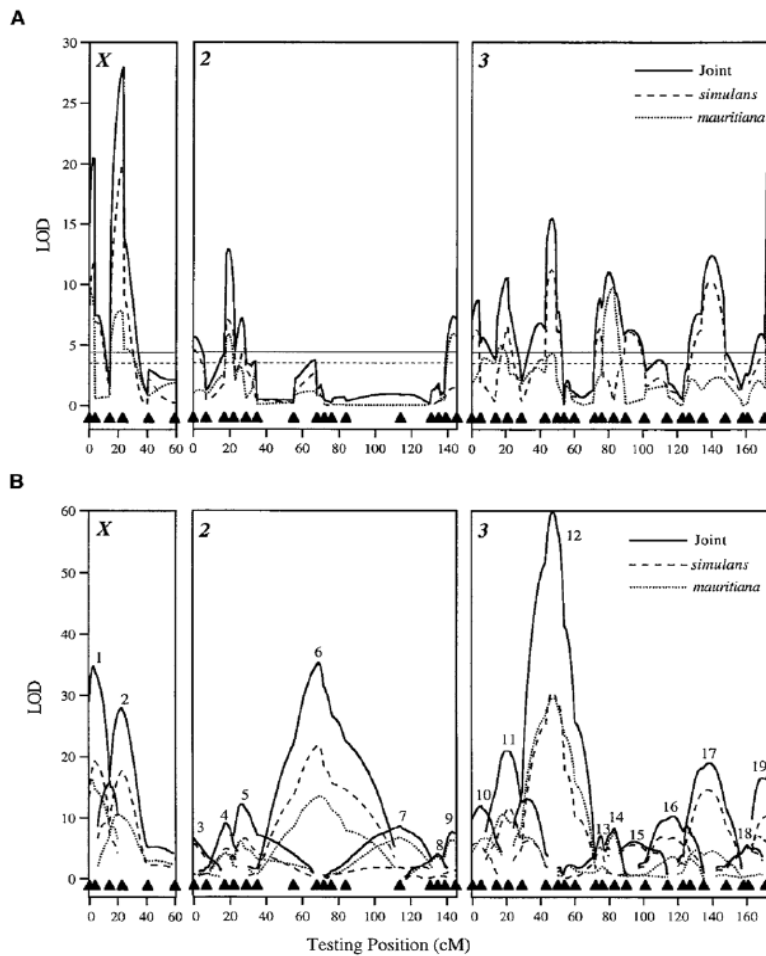


FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

277

### Results:

- There is an overall good agreement between CIM and MIM.
  - MIM identified more QTL
    - 19 in total
  - the test statistics under MIM are higher than those under CIM
  - MIM tends to have more power.
  
- There is a good agreement between the sums of individual QTL effects and the observed parental line differences
  - demonstrating the power of MIM in partitioning parental differences.
  
- Most QTL effects are additive.
- Dominant effects of QTL are substantial,
  - but marginal as compared to additive effects.
  
- There are some epistatic effects in the backcross to *D. mauritiana*.
  - overall, epistasis does not seem to be very significant for the trait
  
- QTL effects together explain
  - 93.2% of the total variance in the backcross to *D. simulans*
  - 91.6% in the backcross to *D. mauritiana*.
- There is a good predictive power of the model in the cross-validation analysis.<sup>278</sup>



279

Estimates of QTL positions, effects and variance components by MIM on

| QTL   | Posi<br>(Chro:cM) | PC1 in the two <i>Drosophila</i> backcrosses |                                    |                  |                |                                                 |                                                 |
|-------|-------------------|----------------------------------------------|------------------------------------|------------------|----------------|-------------------------------------------------|-------------------------------------------------|
|       |                   | BM (%)<br>( $a + d$ ) <sup>a</sup>           | BS (%)<br>( $a - d$ ) <sup>b</sup> | $a^c$            | $d^c$          | BM (%)<br>$\hat{\sigma}_r^2 / \hat{\sigma}_p^2$ | BS (%)<br>$\hat{\sigma}_r^2 / \hat{\sigma}_p^2$ |
| 1     | X:3               | 8.4                                          | - <sup>d</sup>                     | 3.8 <sup>e</sup> | - <sup>f</sup> | 4.4                                             | 4.4                                             |
| 2     | X:23              | 8.3                                          | - <sup>d</sup>                     | 3.7 <sup>e</sup> | - <sup>f</sup> | 4.5                                             | 3.0                                             |
| 3     | II:0              | -0.6                                         | 4.3                                | 2.1              | -2.6           | 0.1                                             | 2.8                                             |
| 4     | II:17             | 5.1                                          | 6.5                                | 5.9              | -1.2           | 3.8                                             | 5.9                                             |
| 5     | II:27             | 9.0                                          | 7.0                                | 7.9              | 0.3            | 6.7                                             | 5.7                                             |
| 6     | II:69             | 4.6                                          | 7.9                                | 6.4              | -2.2           | 3.3                                             | 5.0                                             |
| 7     | II:114            | 4.7                                          | 2.4                                | 3.5              | 0.8            | 2.5                                             | 0.9                                             |
| 8     | II:135            | -2.6                                         | 0.3                                | -1.0             | -1.4           | -0.7                                            | 0.3                                             |
| 9     | II:143            | 5.9                                          | 3.1                                | 4.4              | 1.0            | 3.2                                             | 0.9                                             |
| 10    | III:5             | 5.0                                          | 5.1                                | 5.0              | -0.5           | 4.5                                             | 3.5                                             |
| 11    | III:21            | 8.0                                          | 7.7                                | 7.8              | -0.5           | 7.7                                             | 6.8                                             |
| 12    | III:47            | 10.2                                         | 12.3                               | 11.4             | -2.0           | 12.7                                            | 11.6                                            |
| 13    | III:75            | 0.7                                          | 8.4                                | 4.9              | -4.3           | 0.7                                             | 9.1                                             |
| 14    | III:83            | 12.4                                         | -1.2                               | 5.0              | 6.3            | 14.9                                            | -0.3                                            |
| 15    | III:94            | 1.7                                          | 7.0                                | 4.6              | -3.0           | 2.6                                             | 7.6                                             |
| 16    | III:117           | 4.4                                          | 5.6                                | 5.1              | -1.1           | 4.3                                             | 6.4                                             |
| 17    | III:139           | 4.8                                          | 8.3                                | 6.8              | -4.7           | 4.2                                             | 8.9                                             |
| 18    | III:160           | 1.6                                          | 7.1                                | 4.6              | -3.2           | 1.3                                             | 5.5                                             |
| 19    | III:172           | 7.5                                          | 7.2                                | 7.3              | -0.5           | 4.4                                             | 5.2                                             |
| Total |                   | 99.1                                         | 99.0                               | 99.2             | -18.8          | 85.1                                            | 93.2                                            |

<sup>a</sup> As percentages of the phenotypic difference between F<sub>1</sub> and *D. mauritiana*.

<sup>b</sup> As percentages of the phenotypic difference between *D. simulans* and F<sub>1</sub>.

<sup>c</sup> As percentages of half the difference between *D. simulans* and *D. mauritiana*.

<sup>d</sup> QTL in chromosome X does not contribute to the observed difference.

<sup>e</sup> Only half of the additive effect contributes to the observed difference.

<sup>f</sup> There is no dominance effect for QTL in chromosome X.

280

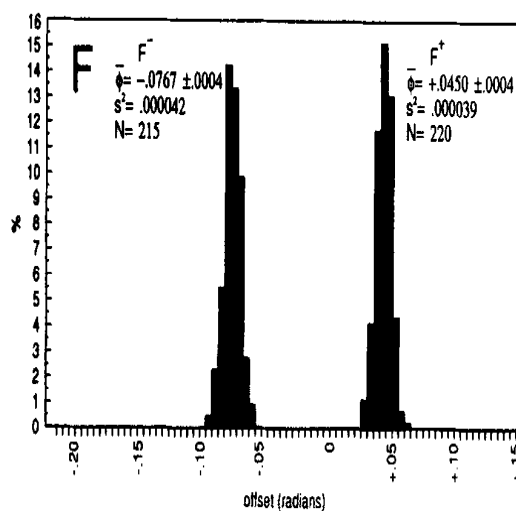
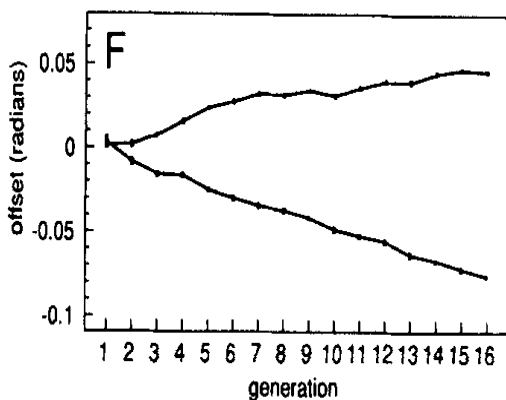
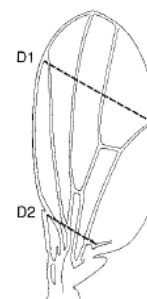
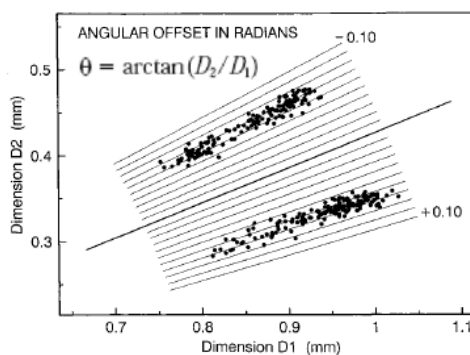
## Estimates of QTL epistatic effects and variance components in *D. mauritiana* backcross

| QTL 1 | QTL 2 | LOD  | Epis. Effect | $\hat{\sigma}_r^2/\hat{\sigma}_p^2$ (%) |
|-------|-------|------|--------------|-----------------------------------------|
| 3     | 12    | 3.29 | 0.89         | 2.2                                     |
| 8     | 15    | 3.44 | 1.48         | 1.0                                     |
| 1     | 17    | 7.32 | 2.01         | 0.8                                     |
| 3     | 17    | 3.01 | 1.17         | 0.8                                     |
| 6     | 17    | 4.22 | 1.41         | 0.6                                     |
| 12    | 17    | 7.57 | 2.00         | 1.1                                     |
| Total |       |      |              | 6.5                                     |

281

Study genetic basis of selection response on wing size in *D. melanogaster*

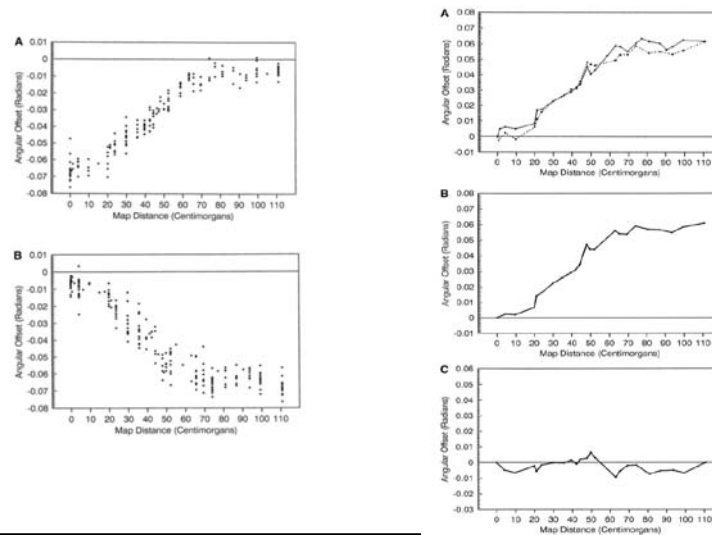
Weber et al. (*Genetics* 1999, 2001)



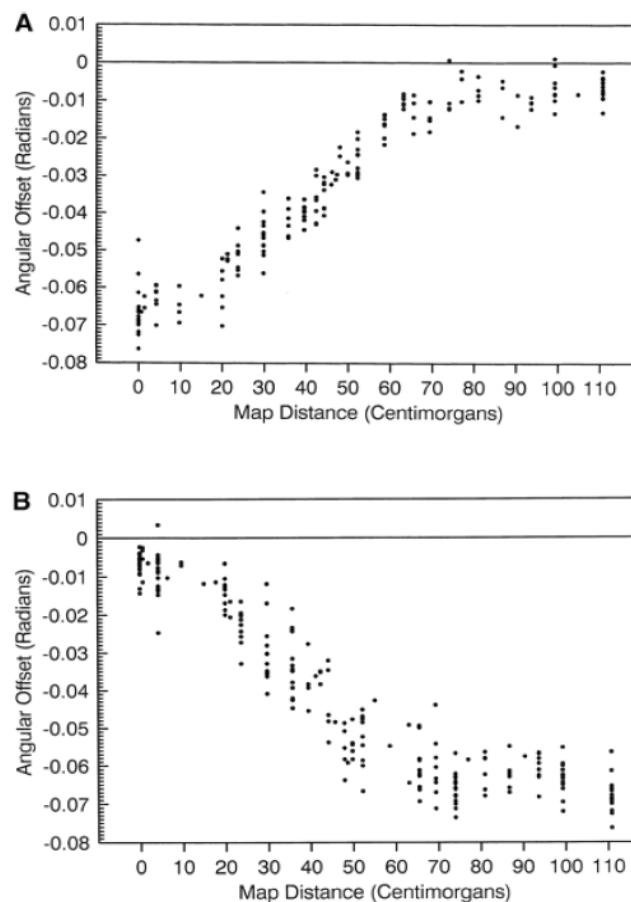
282

## Example 2 (MIM): Genetic architecture of wing size of *Drosophila melanogaster* on chromosome 3

- **Population:** 519 recombinant inbred lines (RILs) originating from a cross between high and low selected lines on wing size.
  - only QTL on chromosome 3 are segregating in the population
  - other chromosomes are identical for all RIL
- **Trait:** wing size measured in radians in an allometric analysis.

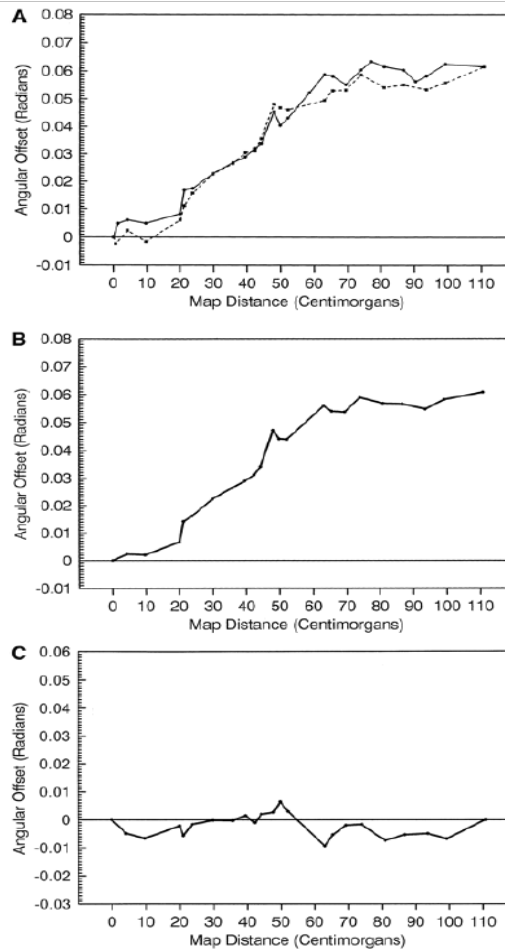


283



284

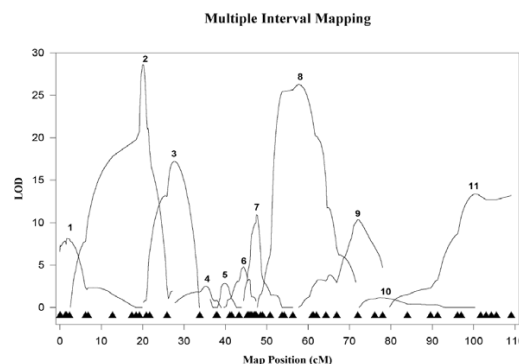




285

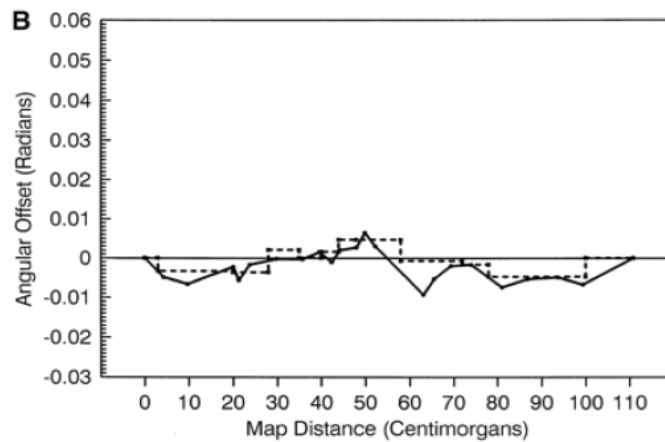
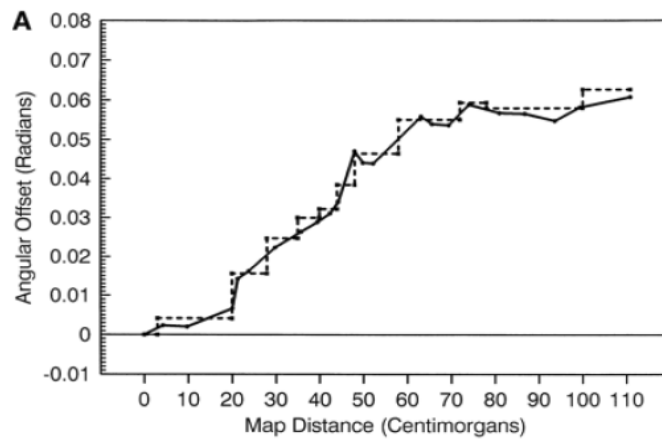
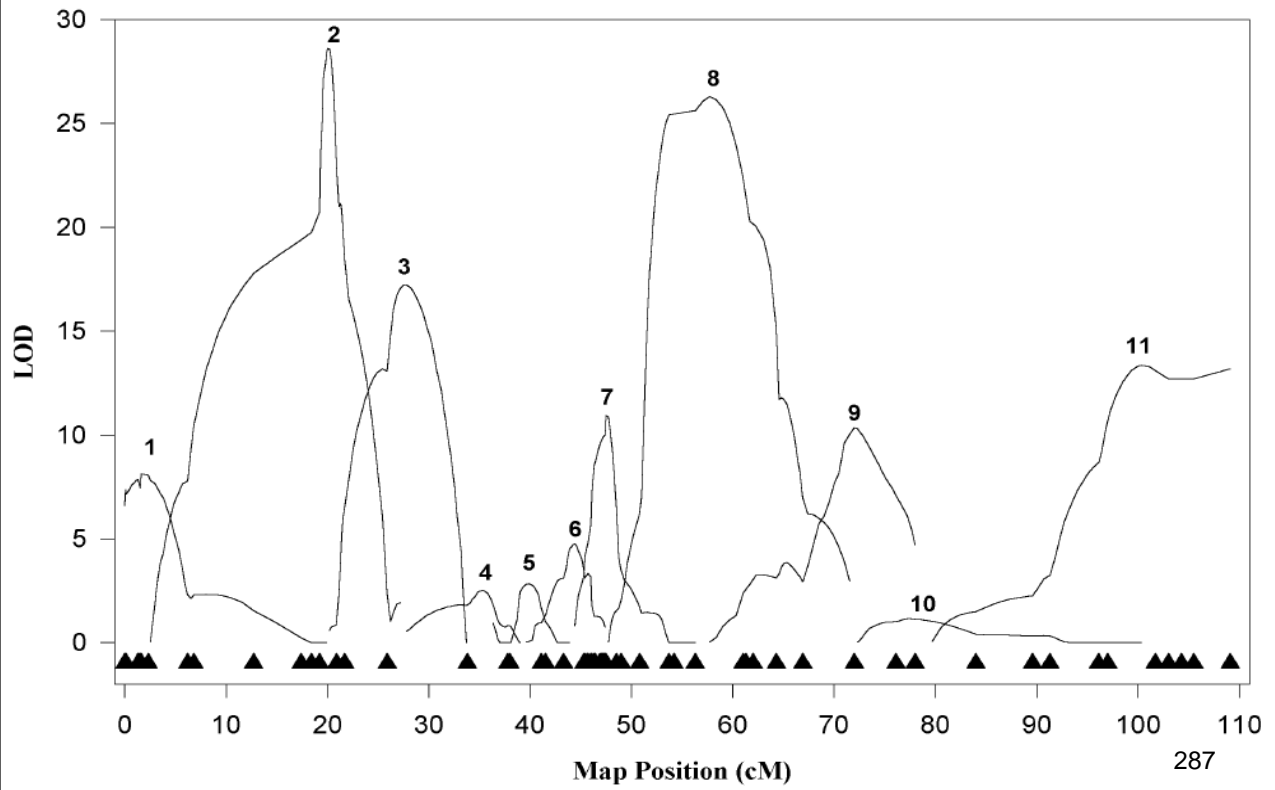
■ **Results:**

- 11 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, 11 additive and 9 additive by additive QTL effects
- 96% of the total variance in the population explained



286

# Multiple Interval Mapping



### Estimates of QTL positions and effects

| QTL   | Posi (cM) | LOD  | Effect | Effect % |
|-------|-----------|------|--------|----------|
| 1     | 3         | 8.1  | 0.41   | 6.7      |
| 2     | 20        | 28.6 | 1.15   | 18.9     |
| 3     | 28        | 17.2 | 0.91   | 14.9     |
| 4     | 35        | 2.5  | 0.53   | 8.7      |
| 5     | 40        | 2.8  | 0.22   | 3.6      |
| 6     | 44        | 4.8  | 0.62   | 10.3     |
| 7     | 48        | 11.0 | 0.81   | 13.3     |
| 8     | 58        | 26.3 | 0.86   | 14.1     |
| 9     | 72        | 10.3 | 0.43   | 7.0      |
| 10    | 78        | 1.2  | -0.14  | -2.3     |
| 11    | 100       | 13.3 | 0.47   | 7.7      |
| Total |           |      |        | 102.9    |

289

### Estimates of QTL epistatic effects

| QTL pair | LOD  | Effect |
|----------|------|--------|
| (1&2)    | 1.53 | -0.33  |
| (2&4)    | 1.73 | -0.36  |
| (3&8)    | 2.85 | -0.81  |
| (3&9)    | 7.65 | 1.38   |
| (4&5)    | 1.47 | -0.56  |
| (5&8)    | 2.85 | 0.88   |
| (5&9)    | 5.60 | -1.29  |
| (6&10)   | 1.14 | 0.31   |
| (8&11)   | 4.08 | -0.47  |

290

**Estimated variances and covariances of QTL main effects in ratio of total phenotypic variance**

|       | 1      | 2      | 3      | 4      | 5     | 6      | 7      | 8      | 9      | 10     | 11     | Sum    |
|-------|--------|--------|--------|--------|-------|--------|--------|--------|--------|--------|--------|--------|
| 1     | 0.009  | 0.018  | 0.009  | 0.003  | 0.001 | 0.001  | 0.000  | -0.002 | -0.002 | 0.001  | -0.004 | 0.034  |
| 2     | 0.018  | 0.075  | 0.044  | 0.018  | 0.006 | 0.011  | 0.011  | 0.006  | -0.002 | 0.001  | -0.009 | 0.179  |
| 3     | 0.009  | 0.044  | 0.047  | 0.020  | 0.007 | 0.014  | 0.016  | 0.011  | 0.002  | 0.000  | -0.005 | 0.163  |
| 4     | 0.003  | 0.018  | 0.020  | 0.016  | 0.005 | 0.012  | 0.013  | 0.010  | 0.003  | 0.000  | -0.002 | 0.097  |
| 5     | 0.001  | 0.006  | 0.007  | 0.005  | 0.003 | 0.006  | 0.007  | 0.006  | 0.002  | 0.000  | 0.000  | 0.042  |
| 6     | 0.001  | 0.011  | 0.014  | 0.012  | 0.006 | 0.022  | 0.025  | 0.020  | 0.007  | -0.001 | 0.000  | 0.117  |
| 7     | 0.000  | 0.011  | 0.016  | 0.013  | 0.007 | 0.025  | 0.037  | 0.028  | 0.009  | -0.002 | 0.000  | 0.145  |
| 8     | -0.002 | 0.006  | 0.011  | 0.010  | 0.006 | 0.020  | 0.028  | 0.042  | 0.015  | -0.004 | 0.005  | 0.136  |
| 9     | -0.002 | -0.002 | 0.002  | 0.003  | 0.002 | 0.007  | 0.009  | 0.015  | 0.011  | -0.003 | 0.005  | 0.045  |
| 10    | 0.001  | 0.001  | 0.000  | 0.000  | 0.000 | -0.001 | -0.002 | -0.004 | -0.003 | 0.001  | -0.002 | -0.010 |
| 11    | -0.004 | -0.009 | -0.005 | -0.002 | 0.000 | 0.000  | 0.000  | 0.005  | 0.005  | -0.002 | 0.012  | -0.001 |
| Total |        |        |        |        |       |        |        |        |        |        |        | 0.947  |

Sum of variance components:  $.009+.075+.047+.016+.003+.022+.037+.042+.011+.001+.012=0.275$

291

**Estimated variances and covariances of QTL epistatic effects in ratio of total phenotypic variance**

|       | 1-2    | 2-4    | 3-8    | 3-9    | 4-5    | 5-8    | 5-9    | 6-10   | 8-11   | Sum    |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 1-2   | 0.001  | 0.000  | 0.000  | 0.001  | 0.000  | 0.000  | -0.001 | 0.000  | 0.000  | 0.001  |
| 2-4   | 0.000  | 0.001  | 0.000  | 0.000  | 0.000  | 0.001  | -0.001 | 0.000  | 0.000  | 0.001  |
| 3-8   | 0.000  | 0.000  | 0.009  | -0.011 | 0.001  | -0.005 | 0.004  | 0.000  | -0.001 | -0.004 |
| 3-9   | 0.001  | 0.000  | -0.011 | 0.027  | -0.001 | 0.005  | -0.014 | 0.002  | 0.001  | 0.009  |
| 4-5   | 0.000  | 0.000  | 0.001  | -0.001 | 0.002  | 0.000  | -0.001 | 0.000  | 0.000  | 0.001  |
| 5-8   | 0.000  | 0.001  | -0.005 | 0.005  | 0.000  | 0.008  | -0.008 | 0.001  | 0.001  | 0.004  |
| 5-9   | -0.001 | -0.001 | 0.004  | -0.014 | -0.001 | -0.008 | 0.021  | -0.003 | 0.000  | -0.003 |
| 6-10  | 0.000  | 0.000  | 0.000  | 0.002  | 0.000  | 0.001  | -0.003 | 0.001  | -0.001 | 0.001  |
| 8-11  | 0.000  | 0.000  | -0.001 | 0.001  | 0.000  | 0.001  | 0.000  | -0.001 | 0.003  | 0.002  |
| Total |        |        |        |        |        |        |        |        |        | 0.012  |

Sum of variance components:  $.001+.001+.009+.027+.002+.008+.021+.001+.003+.003=0.073$

292

## Pattern of genetic variance partition in recombination inbred lines (RILs) of selected populations

|                             | Additive Variance | Epistatic Variance |
|-----------------------------|-------------------|--------------------|
| RI lines (Strong LD)        | 0.947             | 0.012              |
| At equilibrium (without LD) | 0.275             | 0.073              |

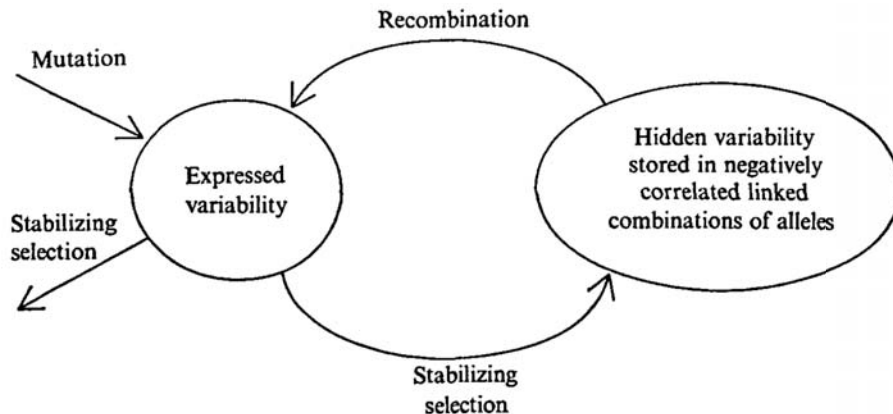


Fig. 1. Flow diagram of genetic variability. See text for explanation.

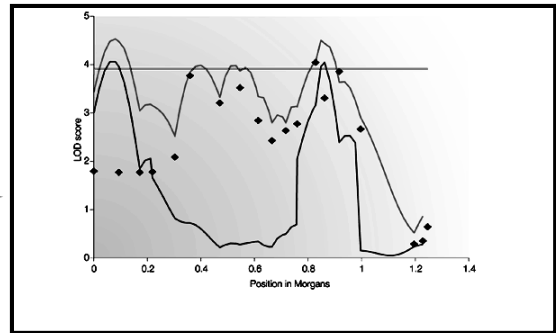
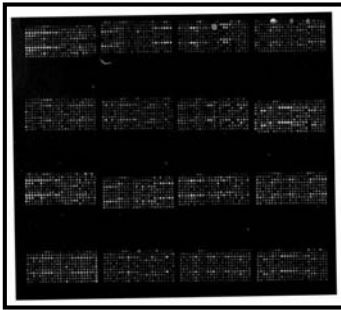
Lande R. (1975) The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet Res.* 26(3):221-35 293

## Advantages of multiple interval mapping

- More efficient and precise in the identification of QTL
- Helps to identify patterns and individual elements of QTL epistasis
- Provides appropriate estimation of individual QTL effects, variance and covariance contribution
- Improves the efficiency of marker-assisted selection,
  - particularly when the information of epistasis is used for MAS
- Multiple interval mapping helps bring QTL mapping, the study of genetic architecture, and marker assisted selection together
- Composite Interval Mapping (CIM) and Multiple Interval Mapping (MIM)
  - applied for expression QTL analysis
  - eQTL

# Connecting QTL Analysis and Microarrays expression QTL (eQTL) analysis

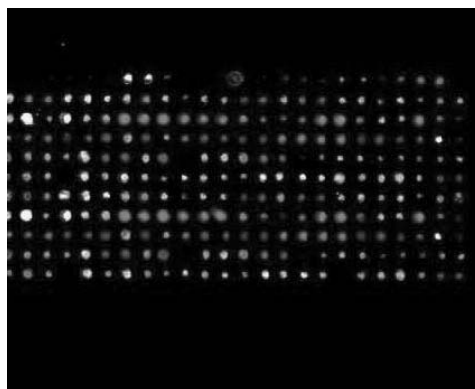
R.W. Doerge



*Summer Institute in Statistical Genetics*

295

...enter array technology  
example... Affymetrix Gene Chips



296

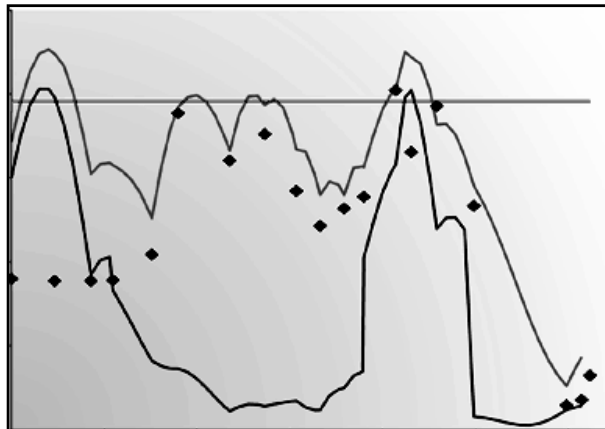
## The first use of microarray technology: Differential (Gene) Expression

- when, where, and in what quantity each gene is expressed
  - compare expression under different conditions
    - (protein-coding) genes direct the synthesis of protein
  - many features simultaneously
  
- *DNA*  $\Rightarrow$  *mRNA*  $\Rightarrow$  *Protein*

297

## The next use of microarray technology/(next gen) Associate gene expression variation to a genetic map

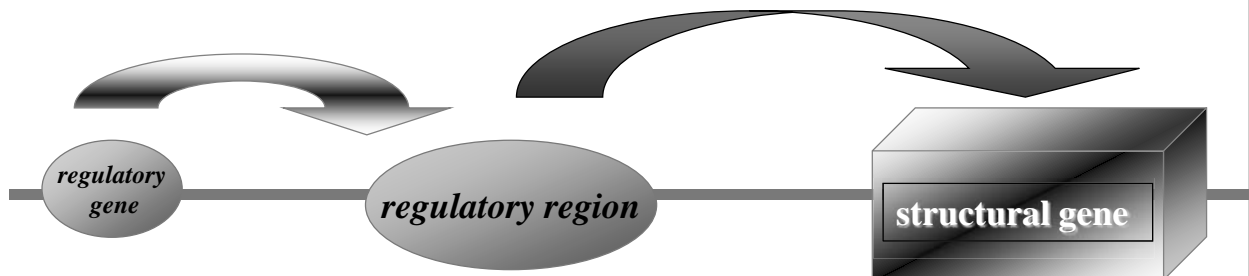
- Use QTL methodology with gene expression data
- Works best for fully sequenced organism
  - map order known
    - Yeast
    - Arabidopsis
- Requires array(s) for each individual
  - each gene's "expression" treated as quantitative trait



298

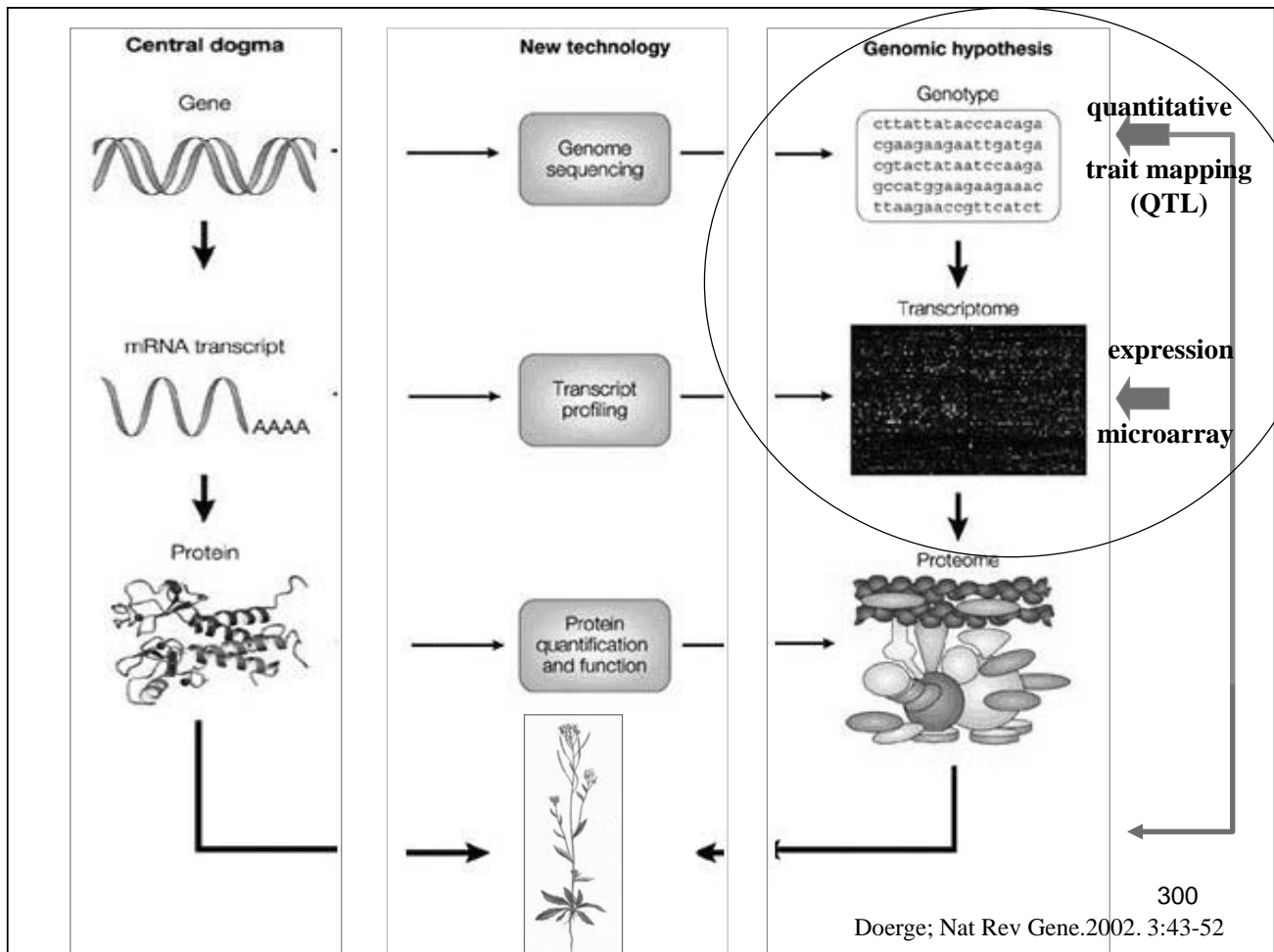
## Old methods for new ideas...

- Quantitative dissection of natural (expression) variation is likely to reveal different aspects of regulatory networks



- This combined approach benefits directly from existing QTL methodology and microarray methods

299



300



Summary of selected eQTL studies based on model species since 2005. IM, CIM and MOM represent interval mapping, composite interval mapping and mixture over marker, respectively.

| Reference             | Brem & Kruglyak                    | Brem et al.                            | Bystrykh et al.                        | Chesler et al.            |
|-----------------------|------------------------------------|----------------------------------------|----------------------------------------|---------------------------|
| Population            | 112 $F_1$ yeast haploids           | 112 $F_1$ yeast haploids               | 30 RI mice                             | 35 RI mice                |
| Genotyping            | 2,957 SFPs                         | 2,957 SFPs                             | 779 SNPs                               | 779 SNPs                  |
| Microarrays           | cDNA spotted                       | cDNA spotted                           | oligonucleotide                        | oligonucleotide           |
| #etraits              | 5,727                              | 6,216                                  | 12,422                                 | 12,422                    |
| eQTL identification   | marker-based<br>Wilcoxon & t-tests | marker-based<br>two-locus linkage scan | interval-based<br>simple-regression IM | interval-based<br>IM, CIM |
| #permutations if used | 10-100                             | 5                                      | 100-1000                               | 100-1000                  |

| Reference             | Schadt et al.        | Hubner et al.                          | Lan et al.           | Kendziorski et al.        |
|-----------------------|----------------------|----------------------------------------|----------------------|---------------------------|
| Population            | 111 $F_2$ mice       | 30 RI rats                             | 60 $F_2$ mice        | 60 $F_2$ mice             |
| Genotyping            | microsatellites      | 1,011 autosomal markers                | 194 microsatellites  | 194 microsatellites       |
| Microarrays           | oligonucleotide      | oligonucleotide                        | oligonucleotide      | oligonucleotide           |
| #etraits              | 23,574               | 15,923                                 | 45,037               | 45,037                    |
| eQTL identification   | interval-based<br>IM | interval-based<br>simple-regression IM | interval-based<br>IM | marker-based<br>MOM model |
| #permutations if used |                      | >1000                                  | 5                    |                           |

301

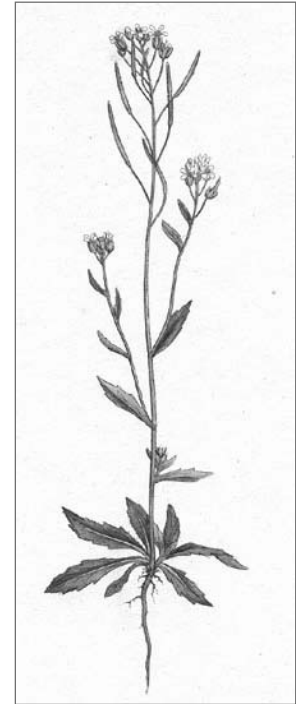
## Molecular dissection of complex traits

- A combined QTL and microarray approach
  - quantitative genetic framework
  - microarray technology
  - statistical methodology
  - take advantage of Expression Level Polymorphism: per-gene expression level differences between genotypes.
  
- Differentiate between
  - *cis*-
  - *trans*-
  
- Molecularly dissect complex expression (e-)traits
  - one gene at a time
  - networks of genes

302

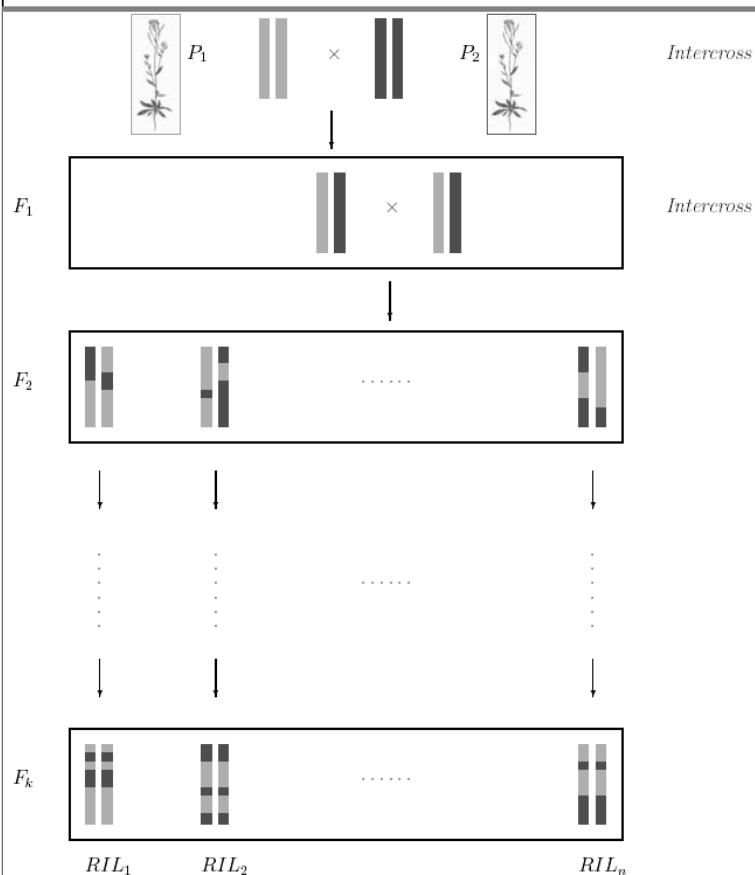
## Example 1 (eQTL): Composite Interval Mapping eQTL analysis

- Population: Arabidopsis Bay-0 x Sha
- RNA from 211 RILs assayed on 844 Affymetrix ATH1
  - 'whole genome' GeneChips (~ 22,810 genes)
- 2 treatments: salicylic acid (SA) and silwet (control)
- 28 hours post-treatment
- 2 biological reps per treatment per RIL



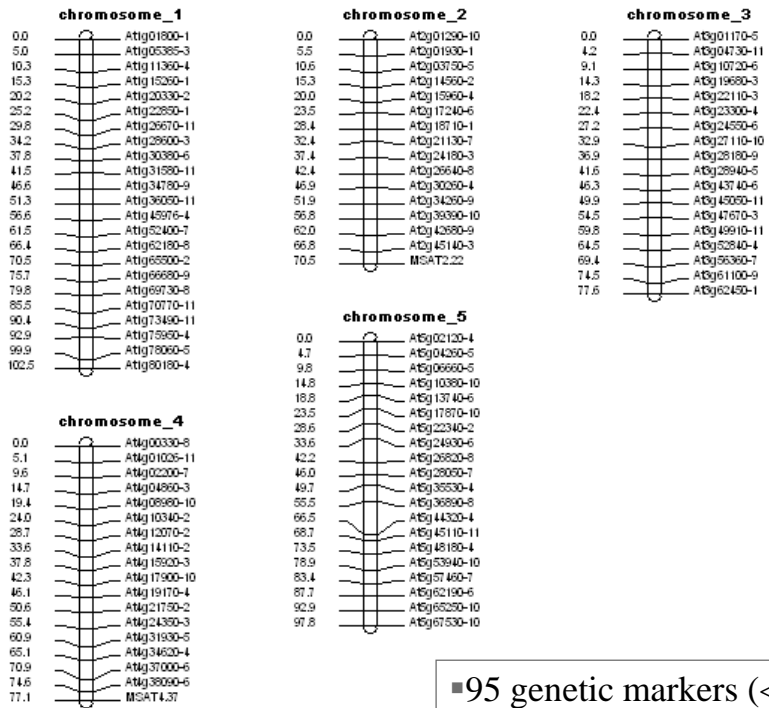
303

## Arabidopsis Bay-0 x Sha: Recombinant Inbred (RI) Line Population



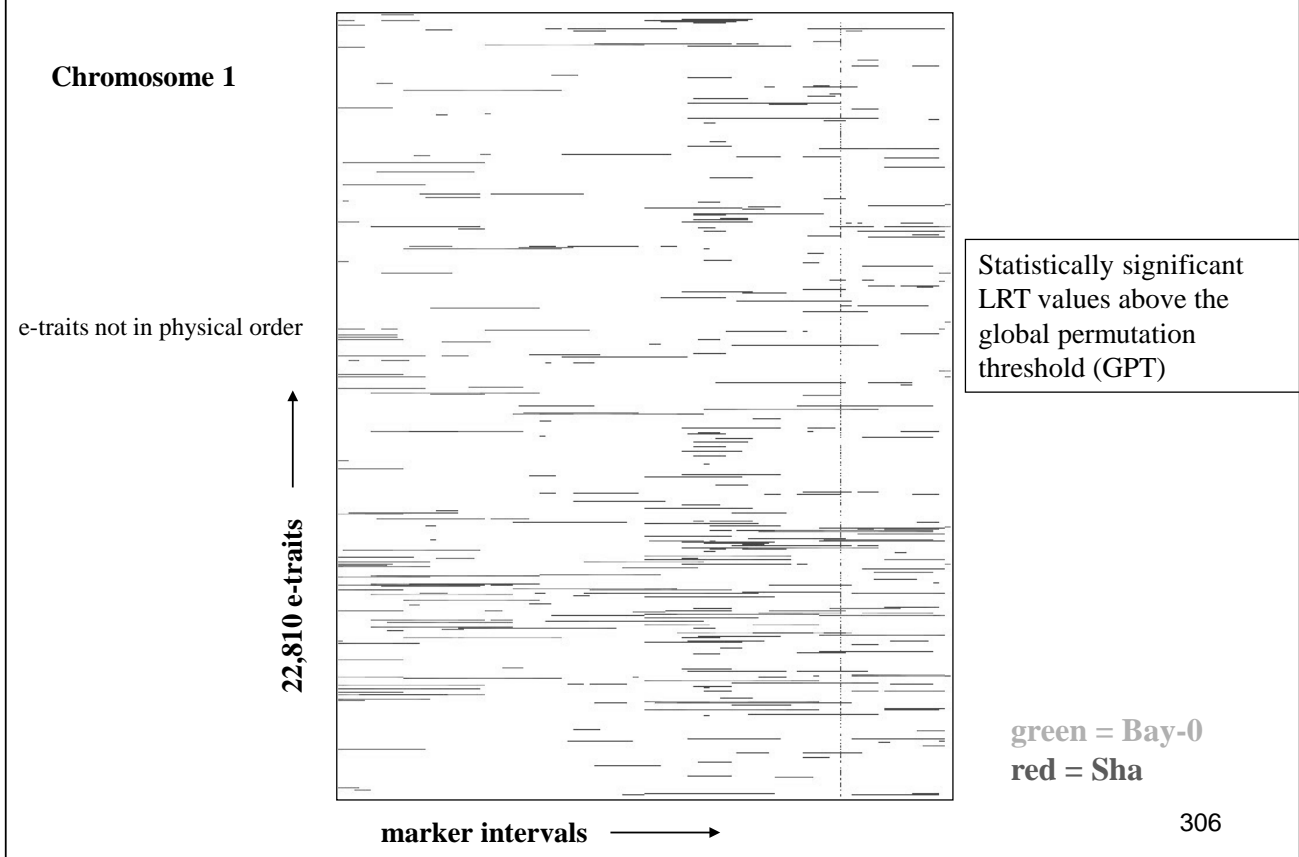
- There is allelic variation
  - in a segregating population, or
  - among genetically distinct individuals.
- Partition the variation in gene expression into:
  - genetic sources
  - non-genetic sources
  - interaction between genetic and non-genetic components
  - technical

304



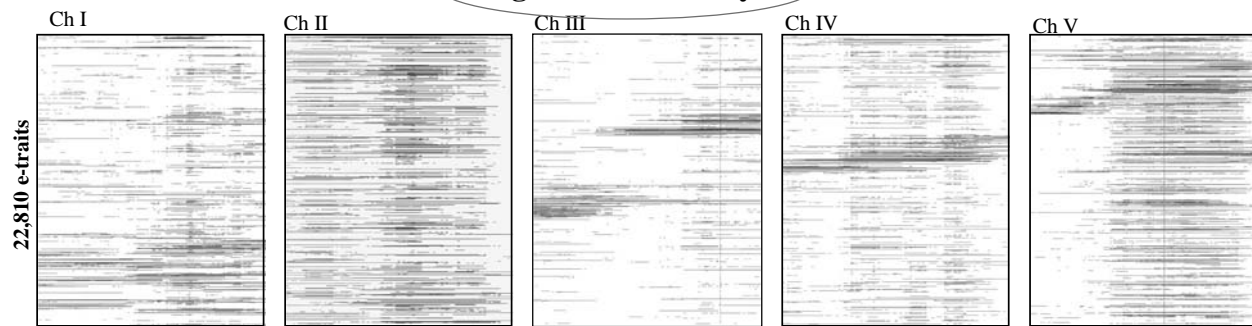
- 95 genetic markers (<5cM framework map)
  - 92 SFP markers: Single Feature Polymorphisms: difference in hybridization signal between the two genotypes (per probe).
  - 3 microsatellites
- 305

## Composite Interval Mapping Results (Control)

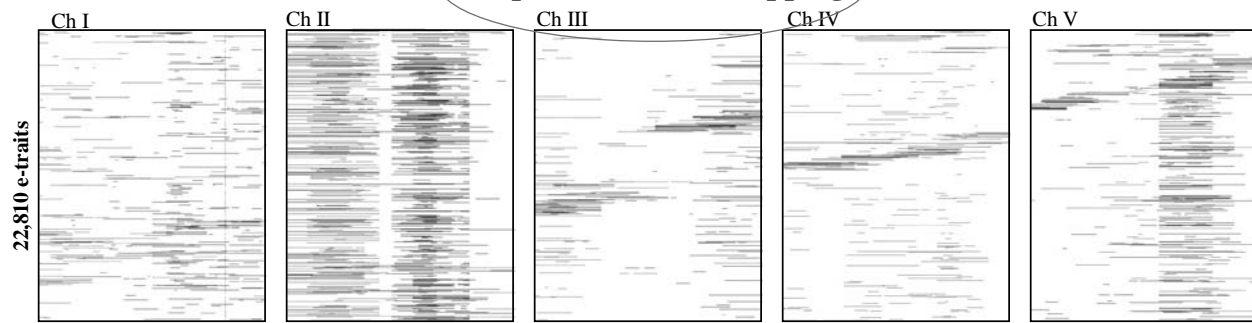


# Single Marker Analysis vs. Composite Interval Mapping (Control)

## Single Marker Analysis

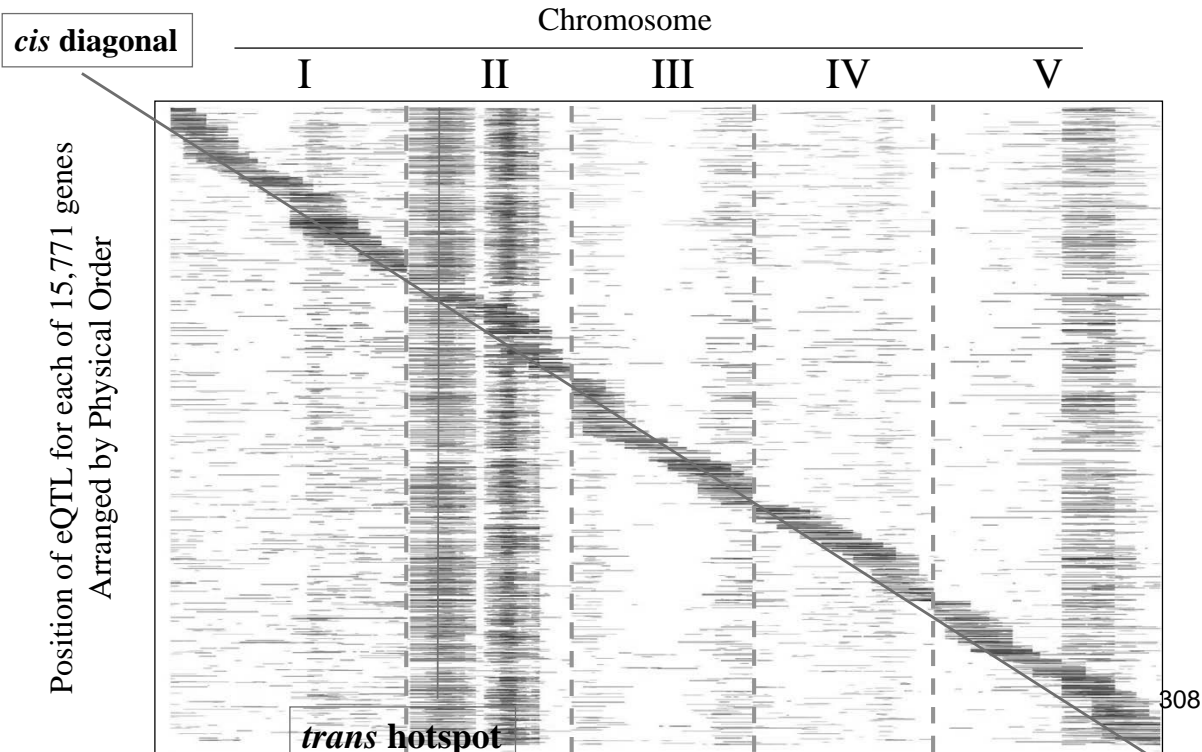


## Composite Interval Mapping



e-traits not in physical order

# eQTL (Control) Variation: 75% of 22,746 genes have at least 1 eQTL

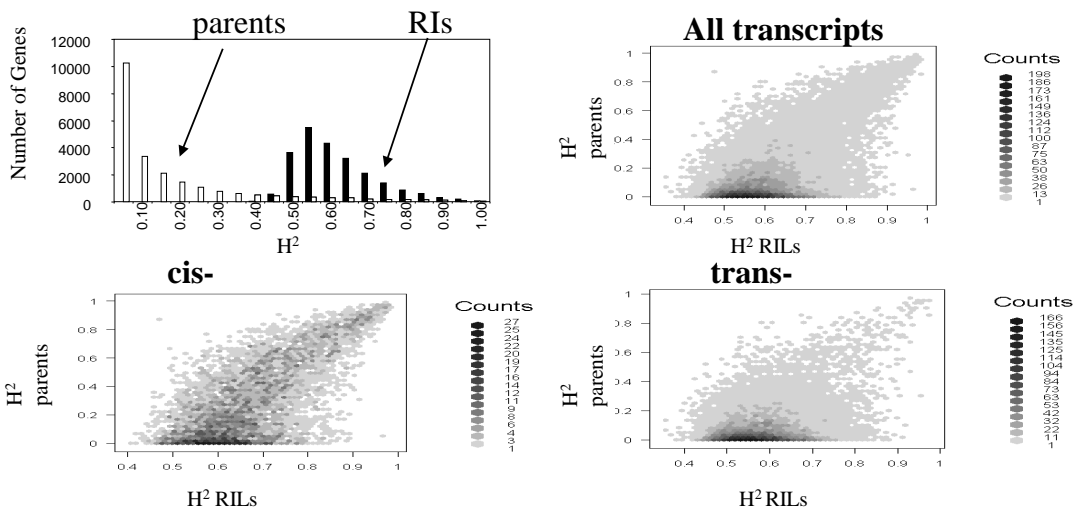
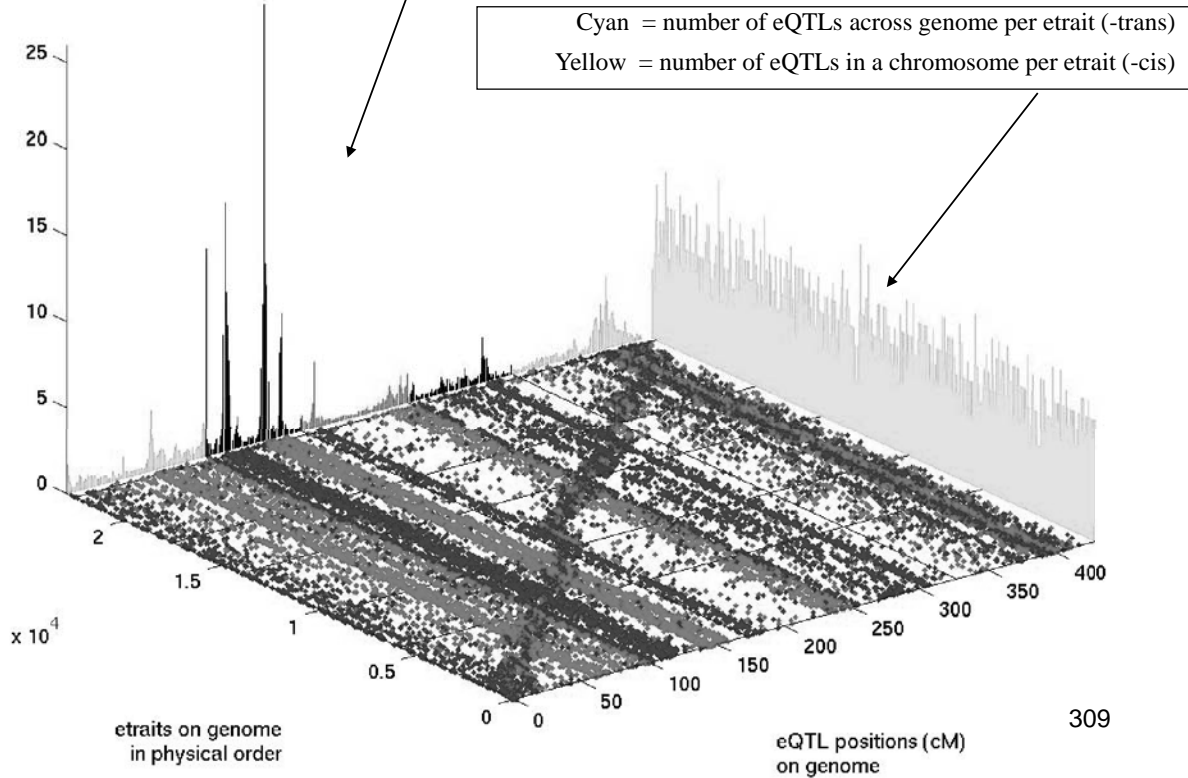


number of etraits having eQTL at each testing locus  
(scale: x 100, e.g., 20 on z-axis means 2000)

Control: CIM

Sha: red

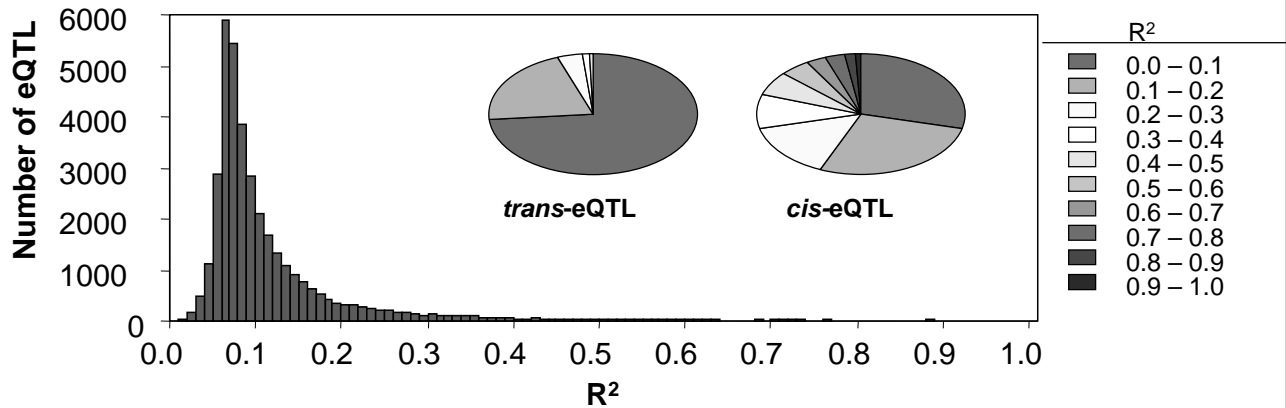
Bay: blue



**Transcript expression heritabilities (control) in RILs versus Bay and Sha parents.**

- Histogram: estimated broad-sense heritability ( $H^2$ ) in RILs and parents.
- $H^2 = V_{\text{genetic}} / V_{\text{phenotypic}}$
- Composite Interval Mapping

**Treatment = Control  
Composite Interval Mapping**



**Distribution of percent phenotypic effect ( $R^2$ ) for all eQTLs.**

Histogram: distribution of  $R^2$  values for all 36,904 eQTLs:  $\max(R^2) = 0.97$

Pie Charts:  $R^2$  distributions for eQTLs that are

*trans*- (31,777 total *trans*-eQTL)

*cis*- (5127 total *cis*-eQTL) to the gene's physical position.

311

**Multiple Interval Mapping  
eQTL Analysis**

**Zhao-Bang Zeng**

*Summer Institute in Statistical Genetics*

312

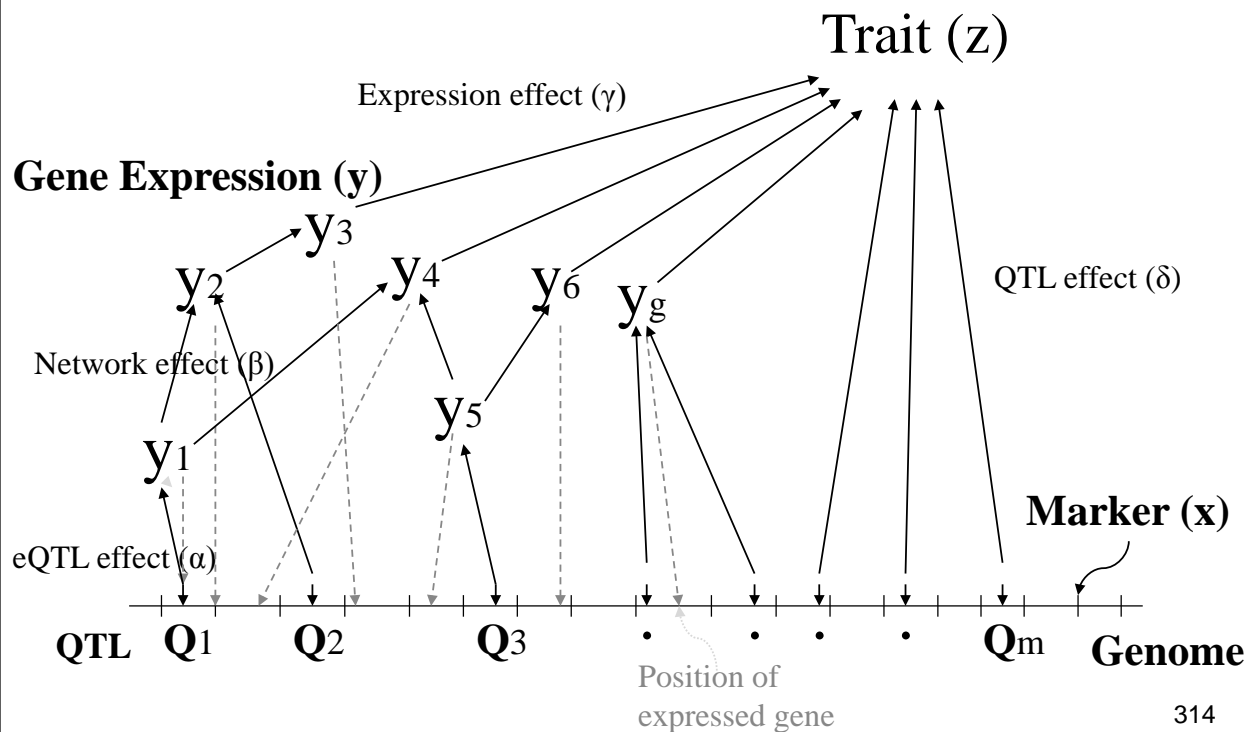
# Goals and Issues

## 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., false discovery rate (FDR)
  - epistasis of eQTL
  - multiple trait analysis
- 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
      - prioritize and suggest candidate causal genes for some eQTL.
- Moving toward network and pathway analysis

313

## Genetic Effect Network



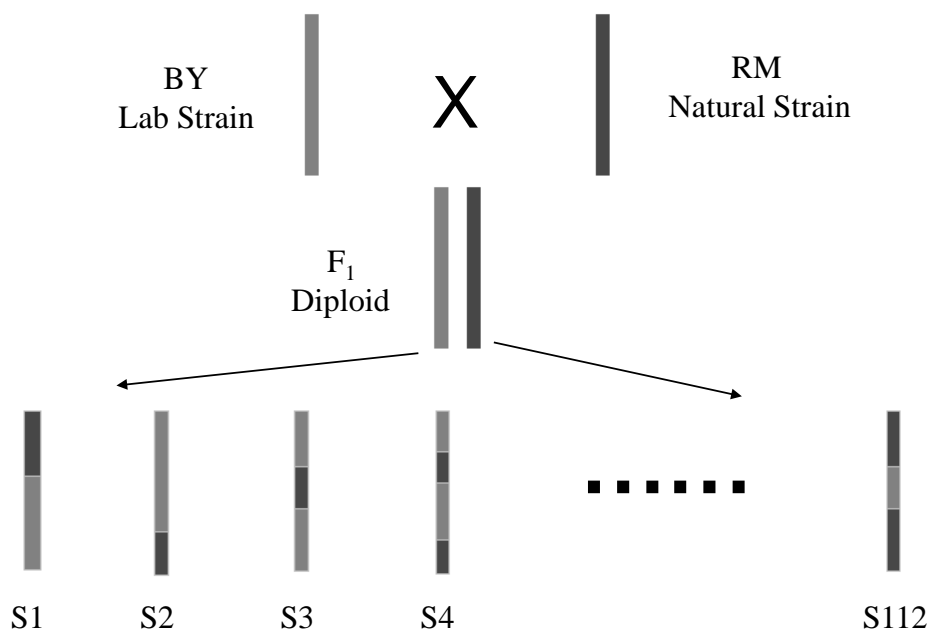
314

## Example 2 (eQTL): Multiple Interval Mapping eQTL analysis

- **Population:** BY (lab strain) x RM (natural strain)
  - n=112 F<sub>1</sub> segregants.
- **Markers:** m=3312 using yeast oligoarrays
- **Gene expression traits:**
  - F<sub>1</sub> individuals 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,
  - BY RNA was the reference for all experiments
- **Dye swap:** Two hybridizations were carried out for each sample,
  - hybridization 1: sample labeled with Cy3 (green) and reference with labeled with Cy5 (red)
  - hybridization 2: sample labeled with Cy5 (red) and reference with labeled with Cy3 (green)
  - for each gene, the two log ratios were averaged.

315

## An eQTL study on a yeast hybrid segregant population



316



## Yeast experiment data structure

|             | <u>Markers</u> |   |   |   |   |   |   |   |   |    |    | <u>Expressions</u> |      |   |   |   |   |   |   |   |   |   |    |    |       |      |
|-------------|----------------|---|---|---|---|---|---|---|---|----|----|--------------------|------|---|---|---|---|---|---|---|---|---|----|----|-------|------|
| 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 |
| <b>BY</b>   | 1              | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1  | 1  | .....              | 1    | X | X | X | X | X | X | X | X | X | X  | X  | ..... | X    |
| <b>RM</b>   | 0              | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  | 0  | .....              | 0    | X | X | X | X | X | X | X | X | X | X  | X  | ..... | X    |
| <b>S1</b>   | 1              | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1  | 1  | .....              | 0    | X | X | X | X | X | X | X | X | X | X  | X  | ..... | X    |
| <b>S2</b>   | 0              | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0  | 0  | .....              | 1    | X | X | X | X | X | X | X | X | X | X  | X  | ..... | X    |
| <b>S3</b>   | 0              | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1  | 0  | .....              | 0    | X | X | X | X | X | X | X | X | X | X  | X  | ..... | X    |
| ....        |                |   |   |   |   |   |   |   |   |    |    |                    |      |   |   |   |   |   |   |   |   |   |    |    |       |      |
| ....        |                |   |   |   |   |   |   |   |   |    |    |                    |      |   |   |   |   |   |   |   |   |   |    |    |       |      |
| <b>S112</b> | 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  $i = 1, 2, \dots, 112+2$

$X_{ij}$  (marker  $j$  on individual  $i$ )  $j = 1, 2, \dots, 3312$

$Y_{ik}$  (expression trait  $k$  for individual  $i$ )  $k = 1, 2, \dots, 6215$

317

## Multiple interval mapping for eQTL analysis

- **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

318

## MIM for eQTL analysis

- The significance threshold is first determined by a permutation test with a controlled type I error rate for the genome scan
  - 95 percentile of test statistic in a genome scan under the null
- The threshold is then 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.

Churchill and Doerge (1994); Doerge and Churchill (1996); Storey et al. (2005); Zou and Zeng (2006)<sup>319</sup>

## The role of threshold in MIM-eQTL

- In the later cycles of the genome scans, the search is restricted within the parameter space where the chance of detecting a strong association is high
  - focus is on those traits that have shown significant QTL in the previous cycles
- The threshold serves as a stopping rule for deciding how many QTL are found for each trait.

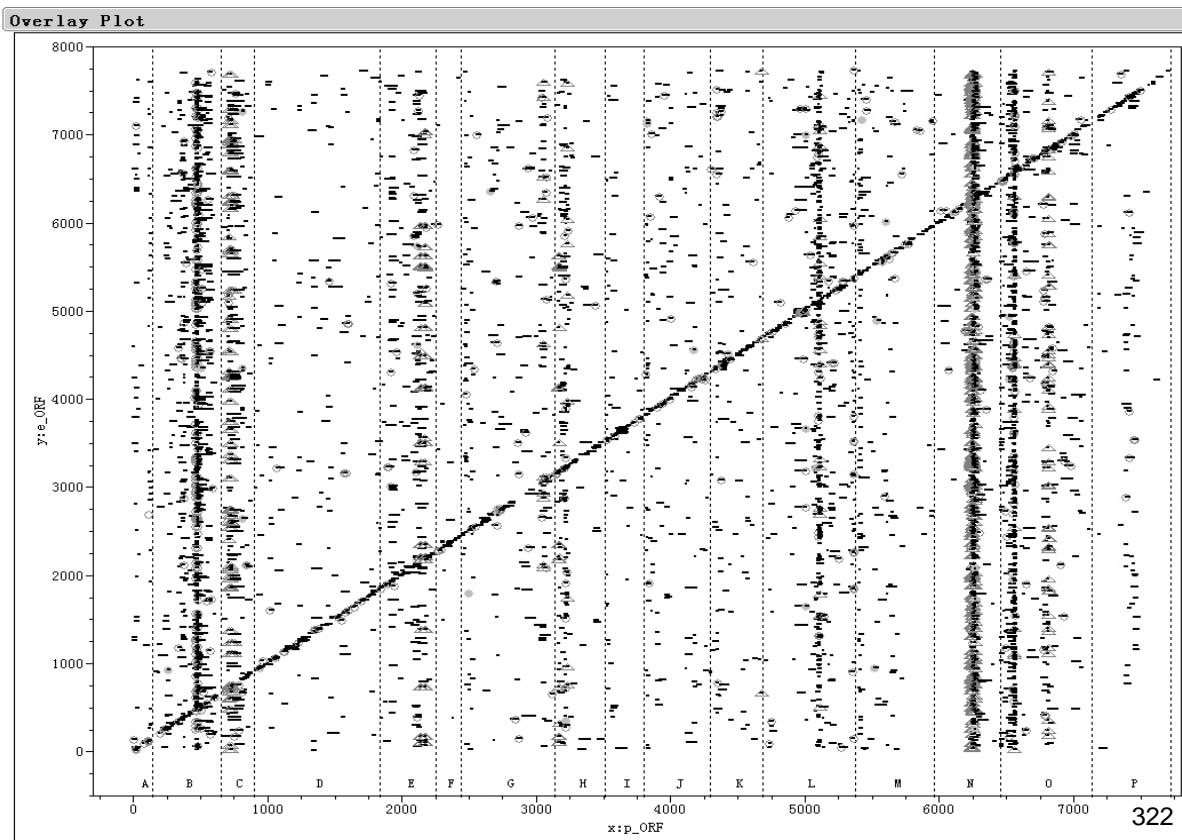
## Sequential genome scan using MIM

| Cycle | # Scanned <sup>1</sup> | # Retained <sup>2</sup> | #Claimed <sup>3</sup> |
|-------|------------------------|-------------------------|-----------------------|
| 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 traits in each cycle
  2. # of traits in the initial genome scans using the 10% genome-wide type I error rate
  3. # of traits 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 at about 8%

321

## Re-analysis of Brem & Kruglyak (2005)



## Summary and thoughts...

- Transcript variation, when measured across a segregating population, can be used to map *cis*- and *trans*- effects.
  - identify hot spots
  - use hotspots to reduce dimension?
    - use markers from hotspots as co-factors
- There are differences in eQTL activity between environments/conditions
  - differing *cis*- and *trans*-acting effects
  - some shared

323

## Inclusion:

### Day 1:

Session 1: Introduction, experimental design, segregation analysis

Session 2: Introduction to genetic mapping, estimating recombination

### Day 2:

Session 2(cont): Introduction to genetic mapping, estimating recombination

Session 3: Introduction to QTL detection, single marker QTL analysis, linkage analysis

Session 4: Introduction to genetic mapping, map estimation exercise

Session 5: Likelihood functions for single marker analysis, interval mapping

Session 6: Computer lab I: QTL-Cartographer

### Day 3:

Session 7: Permutation thresholds; example QTL analysis

Session 8: Composite interval mapping

Session 9: Multiple interval mapping

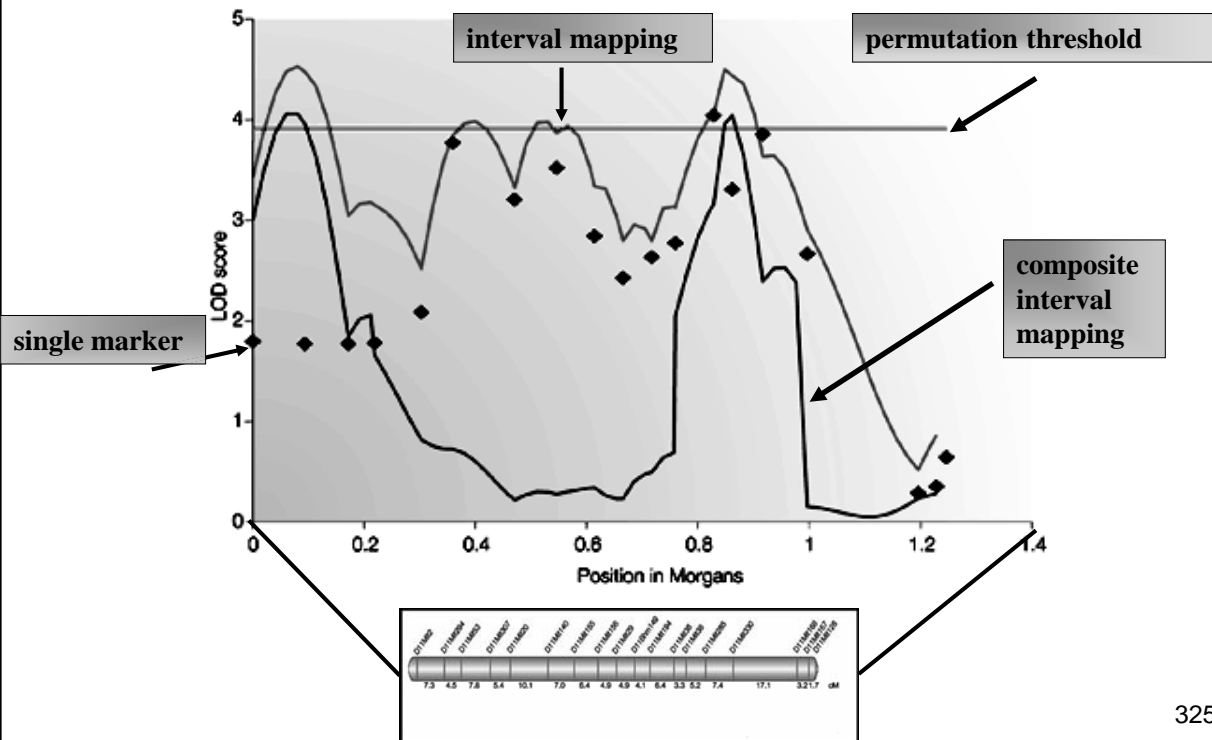
Session 10: Computer lab II: QTL-Cartographer

Session 11: Introduction to eQTL mapping

**Thank you...**

324

# QTL mapping methodology



Fisher 1935; Thoday 1961; Lander and Botstein 1989; Zeng 1994; Churchill & Doerge 1994

# References and Cited Literature

Summer Institute in Statistical Genetics

Introduction to Quantitative Trait Locus Mapping

Doerge and Zeng

- Andersson, L (2001) Genetic dissection of phenotypic diversity in farm animals. *Nature Reviews Genetics* **2**, 130-138. **Review.**
- Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102:1572–1577. **eQTL.**
- Brem, R. B., Storey, J. D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436:701–703. **eQTL.**
- Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in Budding Yeast. *Science*, 296:752–755. **eQTL.**
- Buetow, K H (1987) Multipoint gene mapping using seriation. II Analysis of simulated and empirical data. *American Journal of Human Genetics* **41**, 189-201. **Multiple locus ordering.**
- Buetow, K H and Chakravarti, A (1987) Multipoint gene mapping using seriation. I General methods. *American Journal of Human Genetics* **41**, 180-188. **Multiple locus ordering.**
- Butterfield, R J, Blankenhorn, E P, Roper, R J, Zachary, J F, Doerge, R W, Sudweeks, J, Rose, J and Teuscher, C (1999) Genetic analysis of disease subtypes and sexual dimorphisms in mouse experimental allergic encephalomyelitis (EAE): relapsing/remitting and monophasic relapsing/nonrelapsing EAE are immunogenetically distinct. *Journal of Immunology* **162**, 3096-3102. **Doerge: Mouse QTL Example.**
- Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M. T., Wiltshire, T., Su, A. I., Vellenga, E., Wang, J., Manly, K. F., Lu, L., Chesler, E. J., Alberts, R., Jansen, R. C., Williams, R. W., Cooke, M. P., and de Haan, G. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. *Nature Genetics*, 37:225–232. **eQTL.**
- Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A., Threadgi, D. W., Manly, K. F., and Williams, R.W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37:233–242. **eQTL.**
- Churchill, G A and Doerge, R W (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971. **Permutation thresholds.**
- Darvasi, A and Soller, M (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics* **85**, 353-359. **Selective genotyping to increase power.**
- Dempster, A. P., Laird, N. M. Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39, 1-38. **Original EM-Algorithm paper.**
- Doebley, J, Stec A, and Gustus C. (1991) Genetic analysis of the morphological differences between Maize and Teosinte *Genetics*. 129(1):285–295. **First QTL example.**

- Dodds, K G, Montgomery, G W and Tate, M L (1993) Testing for linkage between a marker locus and a major gene locus in half-sib families. *Journal of Heredity* **84**, 43-48. **Two-point analysis in half-sib families.**
- Doerge, R.W. (1996) Constructing genetic maps by rapid chain delineation. Journal of Quantitative Trait Loci. Volume 2. Article 6. **RCD algorithm.**
- Doerge, R.W. and Churchill, G.A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285-294. **Conditional Permutational Thresholds Used for QTL and eQTL Mapping.**
- Doerge, R W and Rebai, A (1996) Significance thresholds for QTL interval mapping tests. *Heredity* **76**, 459-464.
- Doerge, R W, Zeng, Z B and Weir, B S (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **12**, 195-219. **Review.**
- Doerge, R.W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* **3**:43-52. **QTL and eQTL Review.**
- Dragani, T A, Zeng, Z B, Canzian, F, Gariboldi, M, Ghilarducci, M T, Manenti, G and Pierotti, M A (1995) Mapping of body weight loci on mouse chromosome X. *Mammalian Genome* **6**, 778-781. **Zeng: Composite interval mapping example.**
- Elsen, J M, Mangin, B, Goffinet, B and Chevalet, C (1994) Optimal structure of protocol designs for building genetic linkage maps in livestock. *Theoretical and Applied Genetics* **88**, 129-134. **Experimental designs for mapping genes in livestock.**
- Elston, R C and Stewart, J (1971) A general model for the analysis of pedigree data. *Human Heredity* **21**, 523-542. **An algorithm to calculate likelihoods on pedigrees. Of use in linkage mapping in general pedigrees.**
- Fisher, R.A. 1935. *The Design of Experiments*, Ed. 3. Oliver & Boyd Ltd., London. **Permutation methods.**
- Goffinet, B and Mangin, B (1998) Comparing methods to detect more than one QTL on a chromosome. *Theoretical and Applied Genetics* **96**, 628-633.
- Grattapaglia, D, Bertolucci, F L G, Penchel, R and Sederoff, R R (1996) Genetic mapping of quantitative trait loci controlling growth and wood quality traits in eucalyptus grandis using a maternal half-sib family and RAPD markers. *Genetics* **144**, 1205-1214. **QTL example.**
- Hackett, C A (1997) Model diagnostics for fitting QTL models to trait and marker data by interval mapping. *Heredity* **79**, 319-328.
- Haldane, J.B.S. 1919. The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*. **8**:299-309. **Map Function.**
- Haley, C S and Knott, S A (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315-324. **Interval mapping using regression methods.**
- Haley, C S, Knott, S A and Elsen, J M (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195-1207. **Interval mapping in outbred lines using regression methods.**
- Hetzl, D J S (1991) The use of reference families for genome mapping in domestic livestock. In: Schook LB, Lewin HA, McLaren DG (eds) *Gene-mapping techniques and applications*. Marcel Dekker, New York, pp 51-64. **Experimental designs for mapping genes in livestock.**

- Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Muller, A., Cook, S. A., Kurtz, T. W., Whittaker, J., Pravenec, M., and Aitman, T. J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, 37:243–253. **eQTL.**
- Jansen, R C (1993) Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205-211. **Composite interval mapping.**
- Jansen, R C and Stam, P (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447-1455. **Composite interval mapping.**
- Jayakar, S D (1970) On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. *Biometrics* **26**, 451-464. **First consideration of QTL mapping in outbred populations.**
- Jiang, C and Zeng, Z B (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**, 47-58. **Composite Interval Mapping.**
- Kao, C H (2000) On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics* **156**, 855-865.
- Kao, C H and Zeng, Z B (1997) General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653-665.
- Kao, C H, Zeng, Z B and Teasdale, R D (1999) Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216. **Multiple interval mapping.**
- Kearsey, M J and Hyne, V (1994) QTL analysis - a simple marker-regression approach. *Theoretical and Applied Genetics* **89**, 698-702. **Multiple regression.**
- Kendziorski, C., Chen, M., Yuan, M., Lan, H., and Attie, A. D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, 62:19–27. **eQTL.**
- Knott, S A, Elsen, J M and Haley, C S (1996) Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**, 71-80. **Interval mapping in half-sib families, using regression.**
- Knott, S A and Haley, C S (1992) Aspects of maximum likelihood methods for mapping of quantitative trait loci in line crosses. *Genetical Research* **60**, 139-151. **Interval mapping - hypotheses, power, precision.**
- Knott, S A and Haley, C S (1992) Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* **132**, 1211-1222.
- Kosambi, D. D. (1944) The estimation of map distances from recombination values. *Annals of Eugenics*, 12:172–175. **Map Function.**
- Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, C., and Attie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2:e6. **eQTL.**
- Lande R. (1975) The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet Res.* 26(3):221-35.
- Lander, E S and Botstein, D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199. **Genome-wide interval mapping using maximum likelihood.**



- Lander, E S and Botstein, D (1994) Corrigendum. *Genetics* **36**, 705.
- Lander, E S and Green, P (1987) Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Science* **84**, 2363-2367. **Software for linkage mapping in general pedigrees (CRI-MAP).**
- Lander, E S, Green, P, Abrahamson, J, Barlow, A, Daly, M J, Lincoln, S E and Newburg, L (1987) MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174-181. **Software for linkage and QTL mapping in crosses between inbred lines.**
- Lander, E S and Kruglyak, L (1995) Genetic dissection fo complex traits--guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241-247. **QTL Thresholds.**
- Lange, K, Weeks, D and M, B (1988) Programs for pedigree analysis: MENDEL, FISHER and dGENE. *Genetic Epidemiology* **5**, 471-472. **Software for linkage mapping in general pedigrees.**
- Lathrop, G M, Lalouel, J M, Julier, C and Ott, J (1984) Strategies for multilocus analysis in humans. *Proceedings of the National Academy of Science*, 3443-3446. **Software for linkage mapping in general pedigrees.**
- Liu, J, Mercer, J M, Stam, L F, Gibson, G C, Zeng, Z B and Laurie, C C (1996) Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*. *Genetics* **142**. **Zeng: QTL example.**
- Long, A D, Mullaney, S L, Mackay, T F and Langley, C H (1996) Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in *Drosophila melanogaster*. *Genetics* **144**, 1497-1510. **Zeng : QTL example.**
- Lynch, M and Walsh, B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.
- Mackay, T F C (2001) Quantitative trait loci in *Drosophila*. *Nature Reviews Genetics* **2**, 11-20. **Review.**
- Mangin, B and Goffinet, B (1997) Comparison of several confidence intervals for QTL location. *Heredity* **78**, 345-335.
- Mangin, B, Goffinet, B and Rebai, A (1994) Constructing confidence interval for QTL location. *Genetics* **138**, 1301-1308.
- Martinez, O and Curnow, R N (1992) Estimation the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480-488. **Interval mapping using regression methods.**
- Nettleton, D and Doerge, R W (2000) Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics* **56**, 52-8. **Permutation Thresholds.**
- Nielsen, D, Cockett, N E and Georges, M (1995) Mapping markers and quantitative traits in large half-sib pedigrees. *Proceedings of the Western Section of the American Society of Animal Science* **46**, 205-208.
- Ott, J (1991) *Analysis of human genetic linkage*. Johns Hopkins University Press, Baltimore, pp 302. **A comprehensive treatment of linkage mapping with particular reference to human studies.**
- Rebai, A, Goffinet, B and Mangin, B (1994) Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **138**, 235-240. **Permutation Thresholds.**
- Rebaï, A, Goffinet, B and Mangin, B (1995) Comparing power of different methods for QTL detection. *Biometrics* **51**, 87-99.
- Sax, K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8**, 552-560. **First QTL experiment.**

- Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusk, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37:710–717. **eQTL.**
- Schadt, E. E., Li, C., Su, C., and Wong, W. (2000). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 80:192–202. **eQTL.**
- Schadt, E. E., Monks, S. A., Drake, T. A., Lusk, A. J., Che, N., Collnayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse, and man. *Nature*, 422:297–302. **eQTL.**
- Soller, M, Brody, T and Genizi, A (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* 47, 35-59. **Single marker QTL analysis (t-test).**
- Spelman, R J, Coppieters, W, Karim, L, Van Arendonk, J A M and Bovenhuis, H (1996) Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics* 144, 1799-1807. **QTL example.**
- Stam, P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *The Plant Journal* 3: 739-744. **JoinMap: Mapping software.**
- Stam, P (1995) JoinMap 2.0 deals with all types of plant mapping populations. *Plant Genome III Abstracts*, World Wide Web site: [www.intl-pag.org](http://www.intl-pag.org). **JoinMap: Mapping software.**
- Storey, J. D., Akey, J. M., and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, 3:e267. **eQTL.**
- Thoday, J M (1961) Location of polygenes. *Nature* 191, 368-370. **QTL mapping using flanking markers.**
- Thompson, E A (1984) Information gain in joint linkage analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* 1, 31-49. **Multiple locus ordering and likelihoods.**
- van der Beek, S and van Arendonk, J A M (1993) Criteria to optimize designs for detection and estimation of linkage between marker loci from segregating populations containing several families. *Theoretical and Applied Genetics* 86, 269-280. **Experimental designs for mapping genes in livestock.**
- Visscher, P M, Thompson, R and Haley, C S (1996) Confidence intervals in QTL mapping by bootstrapping. *Genetics* 143, 1013-1020. **Bootstrap Thresholds.**
- Wang S., C. J. Basten, and Z.-B. Zeng (2007). Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC. <http://statgen.ncsu.edu/qtlcart/WQTLCart.htm>. **QTL-Cartographer: QTL Mapping Software.**
- Weber, K., R. Eisman, S. Higgins, L. Kuhl, A. Patty, J. Sparks and Z. B. Zeng (1999) An analysis of polygenes affecting wing shape on chromosome three in *Drosophila melanogaster*. *Genetics* 153: 773–786.
- Weber, K., R. Eisman, S. Higgins, L. Morey, A. Patty, M. Tausek and Z. B. Zeng (2001) An analysis of polygenes affecting wing shape on chromosome 2 in *Drosophila melanogaster*. *Genetics* 159: 1045–1057.
- Weeks, D E and Lange, K (1987) Preliminary ranking procedures for multilocus ordering. *Genomics* 1, 236-242. **Multiple locus ordering.**

- Weller, J I (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627-640. **Single marker QTL analysis (maximum likelihood).**
- Weller, J I, Kashi, Y and Soller, M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science* **73**, 2525-2537. **Experimental designs for mapping genes in livestock.**
- West, M. A. L., van Leeuwen, H., Kozik, A., Kliebenstein, D. J., Doerge, R. W., St. Clair, D. A., and Michelmore, R. W. (2006). High-density haplotyping with microarray-based expression and single feature polymorphism markers in *Arabidopsis*. *Genome Research*, 16:787–795. **Doerge: eQTL example.**
- West, M. A. L., Kim, K., Kliebenstein, D. J., van Leeuwen, H., Michelmore, R. W., Doerge, R. W., and St. Clair, D. A. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript level variation in *Arabidopsis*. *Genetics*, 175:1441–1450. **Doerge: eQTL example.**
- Wright, A J and Mowers, R P (1994) Multiple regression for molecular-marker, quantitative trait data from large F2 populations. *Theoretical and Applied Genetics* **89**, 305-312. **Multiple regression.**
- Wu, W R and Li, W M (1994) A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. *Theoretical and Applied Genetics* **89**, 535-539. **Multiple regression.**
- Zeng, Z B (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Science* **90**, 10972-10976. **Composite interval mapping.**
- Zeng, Z B (1994) Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-1468. **Composite interval mapping.**
- Zeng, Z B, Kao, C H and Basten, C J (1999) Estimating the genetic architecture of quantitative traits. *Genetical Research* **74**, 279-289. **Multiple interval mapping.**
- Zeng, Z B, Liu, J J, Stam, L F, Kao, C H, Mercer, J M and Laurie, C C (2000) Genetic architecture of a morphological shape difference between two drosophila species. *Genetics* **154**, 299-310. **Multiple interval mapping.**
- Zou et al. (2004) *Genetics* 168:2307-2316.
- Zou, W. and Z.-B. Zeng (2008) Multiple interval mapping for gene expression QTL analysis. *Genetica* (submitted). **eQTL.**