Single Marker Analysis and Interval Mapping

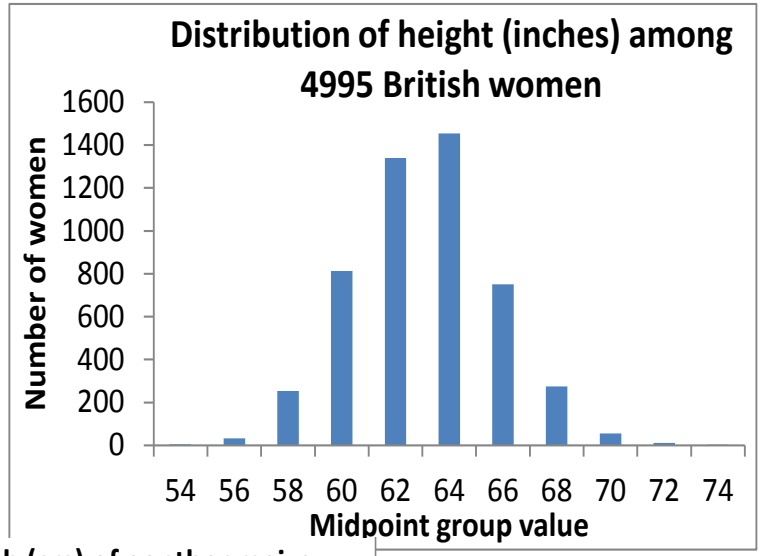
Jiankang Wang
E-mail: jkwang@cgiar.org; wangjiankang@caas.cn
Web: http://www.isbreeding.net

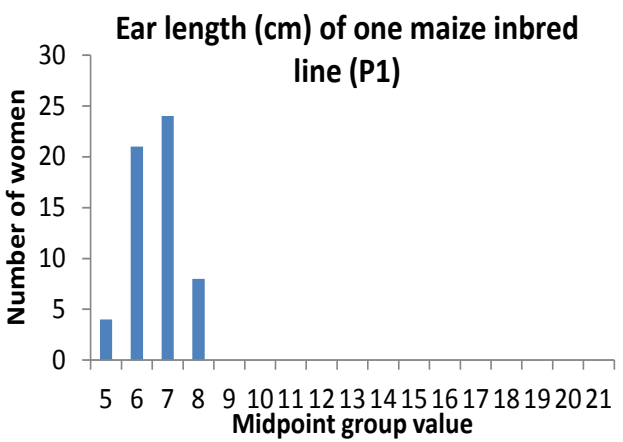
Outlines

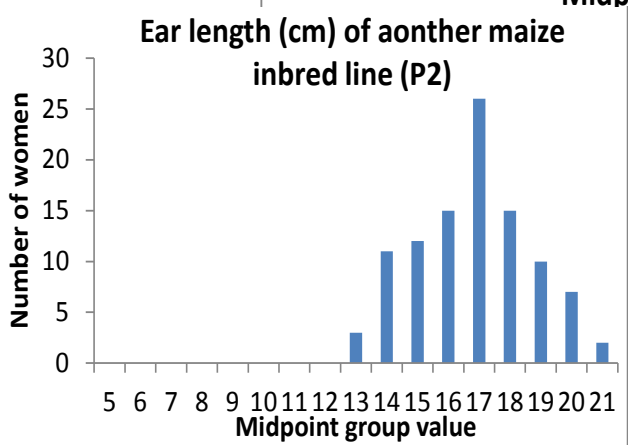
- **Quantitative Traits and QTL Mapping**
- **Single Marker Analysis**
- **The Conventional (Simple) Interval Mapping**

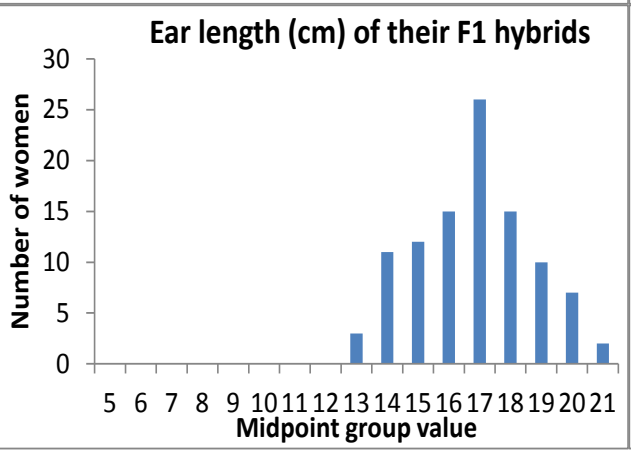
Quantitative Traits and QTL Mapping

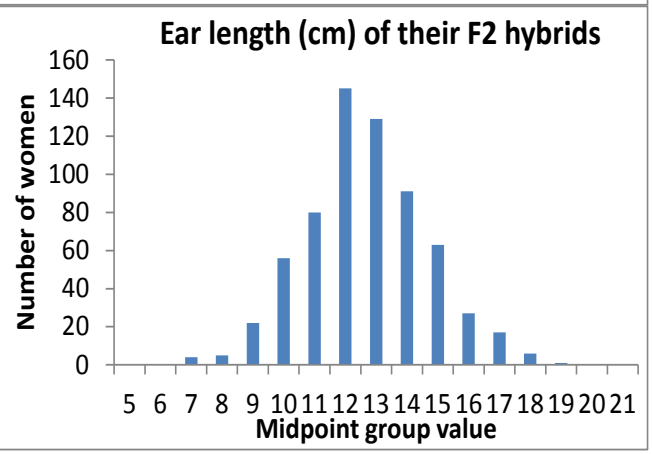
Quantitative traits in genetics











Quantitative traits

- **Continuous phenotypic variation**

- **Affected by many genes**
- **Affected by environment**
- **Epistasis**

- **Polygene (or multi-factorial) hypothesis**
- **Classical quantitative genetics**

Quantitative trait does not have to be “continuous”

- Categorical traits: traits in which the phenotype corresponds to any one of a number of discrete categories
  - Number of skin ridges forming the fingerprints
  - Number of kernels on an ear of corn
  - Number of puppies in a litter
- Threshold traits: traits that have only two, or a few, phenotypic classes, but their inheritance is determined by the effects of multiple genes acting together with the environment
  - Liability to express the trait, which is not directly observable.
  - When liability is high enough (above a “threshold”), the trait will be expressed; Otherwise, the trait is not expressed.

What is QTL Mapping?

- The procedure to map individual genetic factors with small effects on the quantitative traits, to specific chromosomal segments in the genome
- The key questions in QTL mapping studies are:
  - How many QTL are there?
  - Where are they located in the marker map?
  - How large an influence does each of them have on the trait of interest?
  - Are they interacting with each other?
  - Are they stably expressed across environments?
- **Which are the basic genetic questions.**

Dataset of QTL mapping

- Mapping populations
- Marker data of each individual in the mapping population
- Linkage map
- Phenotypic data

Example: 10 RIL of Rice
(linkage map of Chr. 5)

Marker	C263	R830	R3166	XNpb387	R569	R1553	C128	C1402	XNpb81	C246	R2953	C1447	Grain width (mm)
Position (cM)	0.0	3.5	8.5	19.5	32.0	66.6	74.1	78.6	81.8	91.9	92.7	96.8	
RIL1	0	0	0	0	0	0	0	0	0	0	0	0	2.33
RIL2	2	2	2	2	2	0	0	0	0	2	2	2	1.99
RIL3	0	2	2	2	2	2	2	2	2	2	2	2	2.24
RIL4	0	0	0	0	0	0	2	2	2	2	2	2	1.94
RIL5	0	0	0	0	0	2	2	0	0	0	0	0	2.76
RIL6	0	0	0	2	2	2	2	2	2	2	2	2	2.32
RIL7	0	0	0	0	0	0	0	0	0	0	0	0	2.32
RIL8	2	2	0	2	2	0	0	0	0	2	2	2	2.08
RIL9	0	0	0	0	2	2	0	0	0	0	0	0	2.24
RIL10	0	0	0	0	2	2	0	0	0	0	0	0	2.45

Classification of mapping populations

- Bi-parental mapping populations (linkage mapping)
  - Temporary population: F2 and BC
  - Permanent population: RIL, DH, CSSL
  - Secondary population
- Association mapping
  - Natural populations: human and animals

Overview on QTL mapping methods

- **Single marker analysis (Sax 1923; Soller et al. 1976)**
  - The single marker analysis identifies QTLs based on the difference between the mean phenotypes for different marker groups, but cannot separate the estimates of recombination fraction and QTL effect.
- **Interval mapping (IM) (Lander and Botstein 1989)**
  - IM is based on maximum likelihood parameter estimation and provides a likelihood ratio test for QTL position and effect. The major disadvantage of IM is that the estimates of locations and effects of QTLs may be biased when QTLs are linked.
- **Regression interval mapping (RIM) (Haley and Knott 1992; Martinez and Curnow 1992)**
  - RIM was proposed to approximate maximum likelihood interval mapping to save computation time at one or multiple genomic positions.
- **Composite interval mapping (CIM) (Zeng 1994)**
  - CIM combines IM with multiple marker regression analysis, which controls the effects of QTLs on other intervals or chromosomes onto the QTL that is being tested, and thus increases the precision of QTL detection.

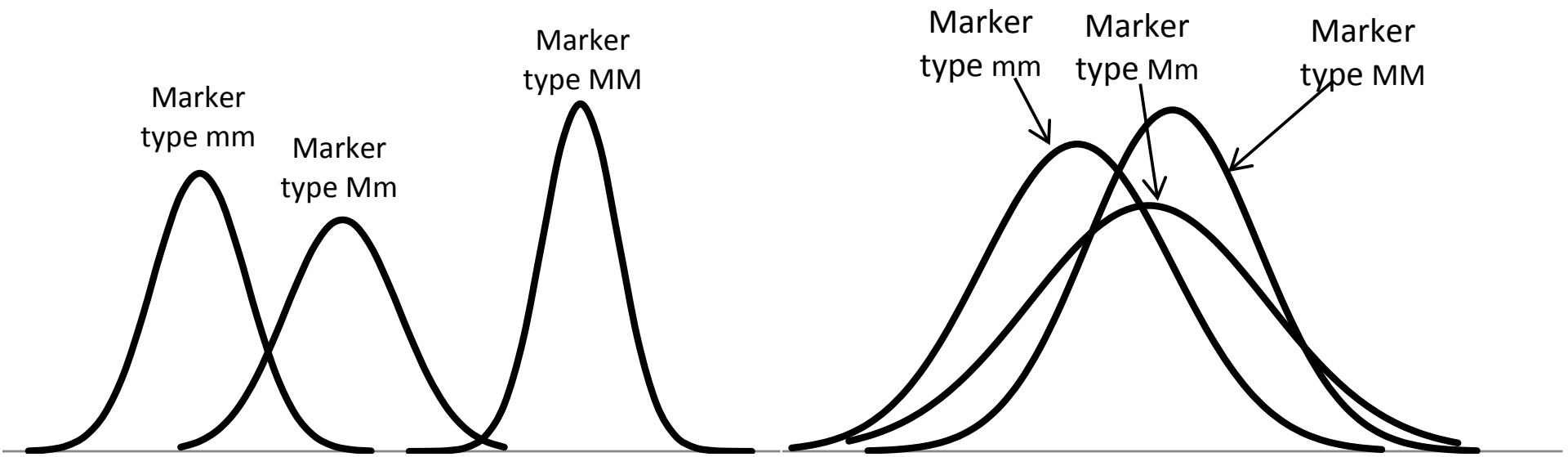
Overview on QTL mapping methods

- **Multiple interval mapping (MIM) (Kao et al. 1999)**
  - MIM is a state-of-the-art gene mapping procedure. But implementation of the multiple-QTL model is difficult, since the number of QTL defines the dimension of the model which is also an unknown parameter of interest.
- **Bayesian model (Sillanpää and Corander 2002)**
  - In any Bayesian model, a prior distribution has to be considered. Based on the prior, Bayesian statistics derives the posterior, and then conduct inference based on the posterior distribution. However, Bayesian models have not been widely used in practice, partially due to the complexity of computation and the lack of user-friendly software.
- **Inclusive Composite Interval Mapping (Li et al. 2007)**
  - In the first step, stepwise regression was applied to identify the most significant regression variables in both cases but with different probability levels of entering and removing variables. In the second step, a one-dimensional scanning or interval mapping was conducted for mapping additive and a two-dimensional scanning was conducted for mapping digenic epistasis.

Single Marker Analysis

Evidence for marker and QTL association

- Three marker types MM, Mm, and mm at one marker locus
- When marker is linked with QTL, the three marker types will have un-equal means.



Backcrosses (P1BC1 and P2BC1) of P1: MMQQ and P2: mmqq

P1BC1F1			P2BC1F1		
Genotype	Frequency	Genotypic value	Genotype	Frequency	Genotypic value
MMQQ	(1-r)/2	*m+a*	MmQq	(1-r)/2	*m+d*
MMQq	r/2	*m+d*	Mmqq	r/2	*m-a*
MmQQ	r/2	*m+a*	mmQq	r/2	*m+d*
MmQq	(1-r)/2	*m+d*	mmqq	(1-r)/2	*m-a*

Difference between the two marker types (P1BC1 as example)

- Two marker types:

$$\mu_{\mathrm{MM}} = (1-r)\mu_{\mathrm{MMQQ}} + r\mu_{\mathrm{MMQq}}$$

$$= (1-r)(m+a) + r(m+d) = m + (1-r)a + rd$$

$$\mu_{\mathrm{Mm}} = r\mu_{\mathrm{MmQQ}} + (1-r)\mu_{\mathrm{MmQq}}$$

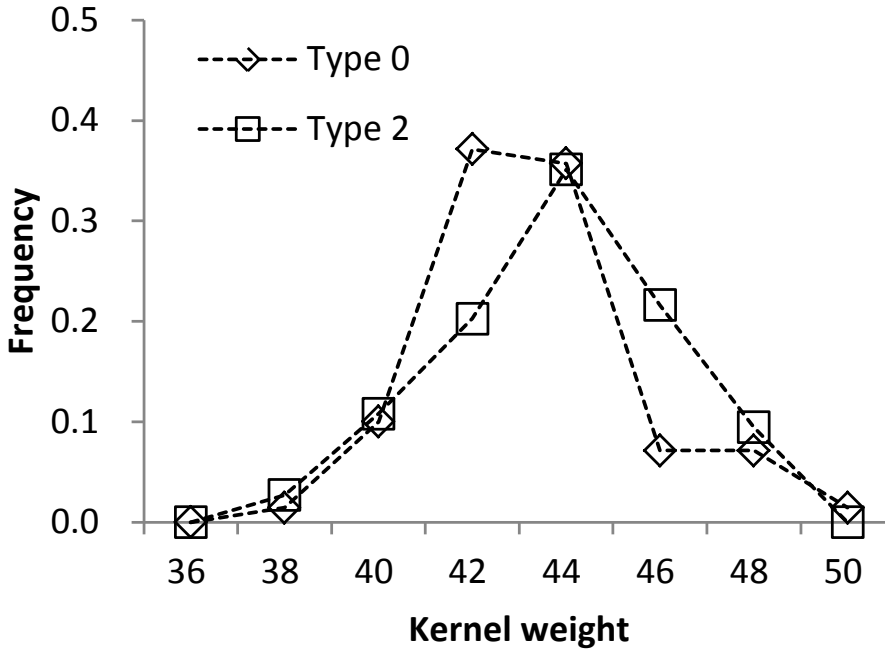
$$= r(m+a) + (1-r)(m+d) = m + ra + (1-r)d$$

- Difference in phenotype between the two types

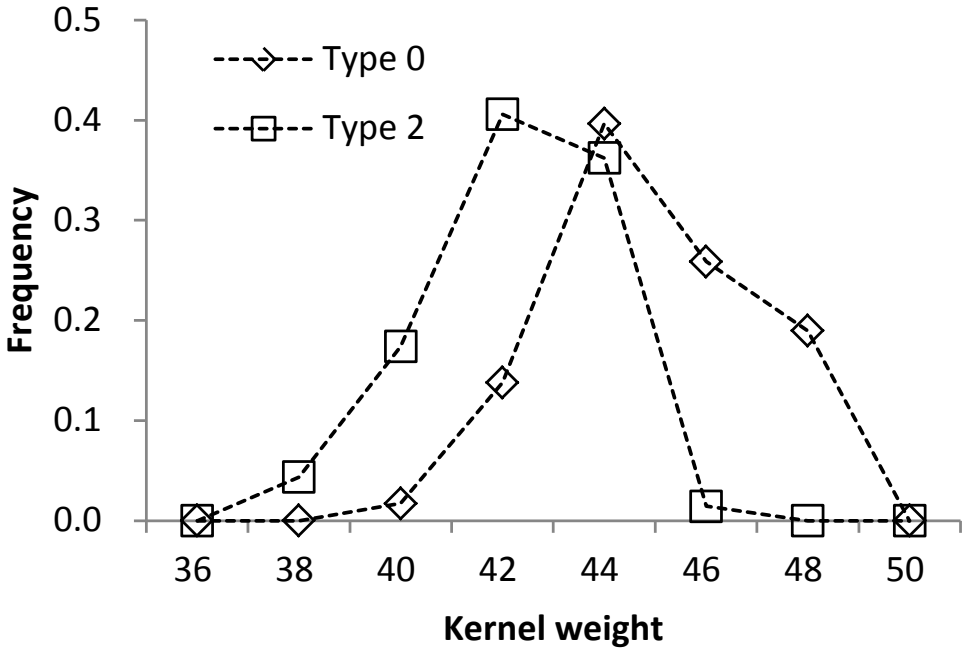
$$\mu_{\mathrm{MM}} - \mu_{\mathrm{Mm}} = (1-2r)(a-d)$$

A barley DH population

Marker locus Act8A



Marker locus Act8B

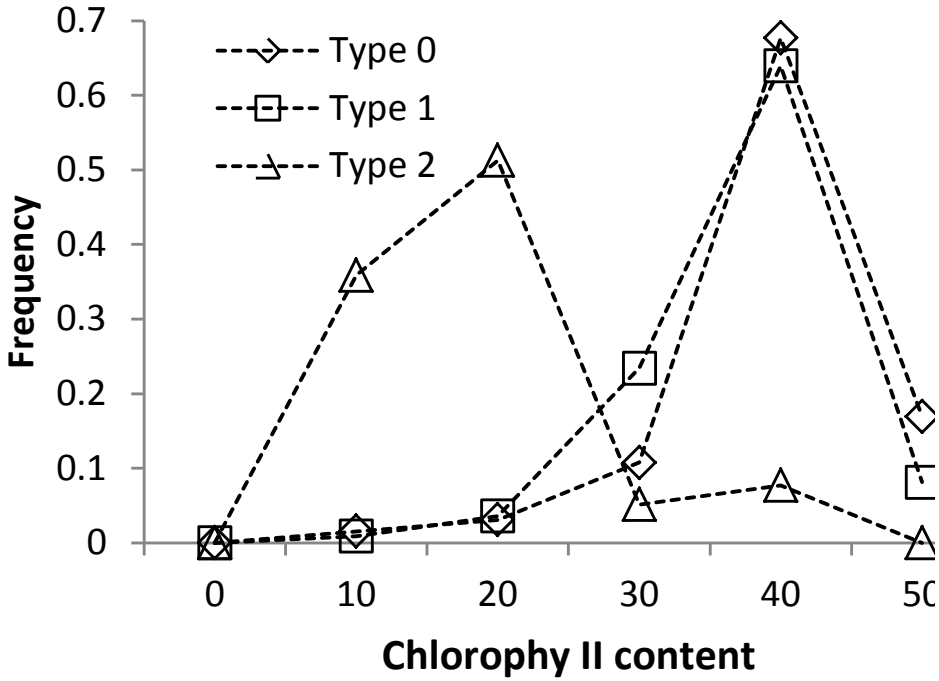


Significance test on phenotypic means of marker types

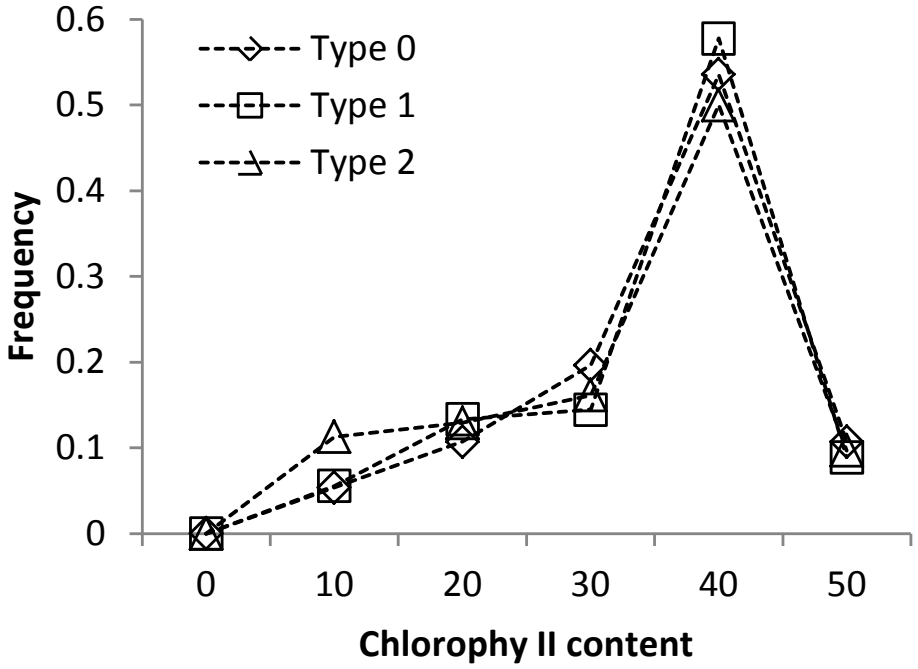
Parameter	Marker Act8A		Marker Act8B	
	Type 0	Type 2	Type 0	Type 2
Sample size	70	74	58	69
Degree of freedom	69	73	57	68
Mean	**42.23**	**42.79**	**43.89**	**41.25**
Variance	4.45	5.32	3.53	2.79
Standard error	2.11	2.31	1.88	1.67
Combined variance	4.90		3.13	
T-test	1.51 (*P*=0.1342)		8.37 (*P*=1.00E-13)	

A soybean F2 population

Marker locus *Satt339



Marker locus *Sat_033



Significance test on phenotypic means of marker types

Parameter	Marker Act8A			Marker Act8B		
	Type 0	Type 1	Type 2	Type 0	Type 1	Type 2
Sample size	65	111	39	56	90	62
Mean	**35.16**	**32.76**	**14.22**	**30.72**	**30.47**	**29.20**
Variance	47.71	40.42	65.52	92.13	97.24	133.28
Standard error	6.91	6.65	8.09	9.60	9.86	11.54
T-test of additive	15.06 (*P*=1.43E-33)			0.80 (*P*=0.4264)		
T-test of dominance	8.47 (*P*=5.89E-13)			0.35 (*P*=0.7270)		

Problems with the Single Marker Analysis

- Cannot separate QTL effect and the marker-QTL distance
- Low detection power
- Does not take the advantage of genetic linkage map

Conventional Interval Mapping

Interval mapping (IM)
(Lander and Botstein 1989; Milestone in QTL mapping methodology and applications)

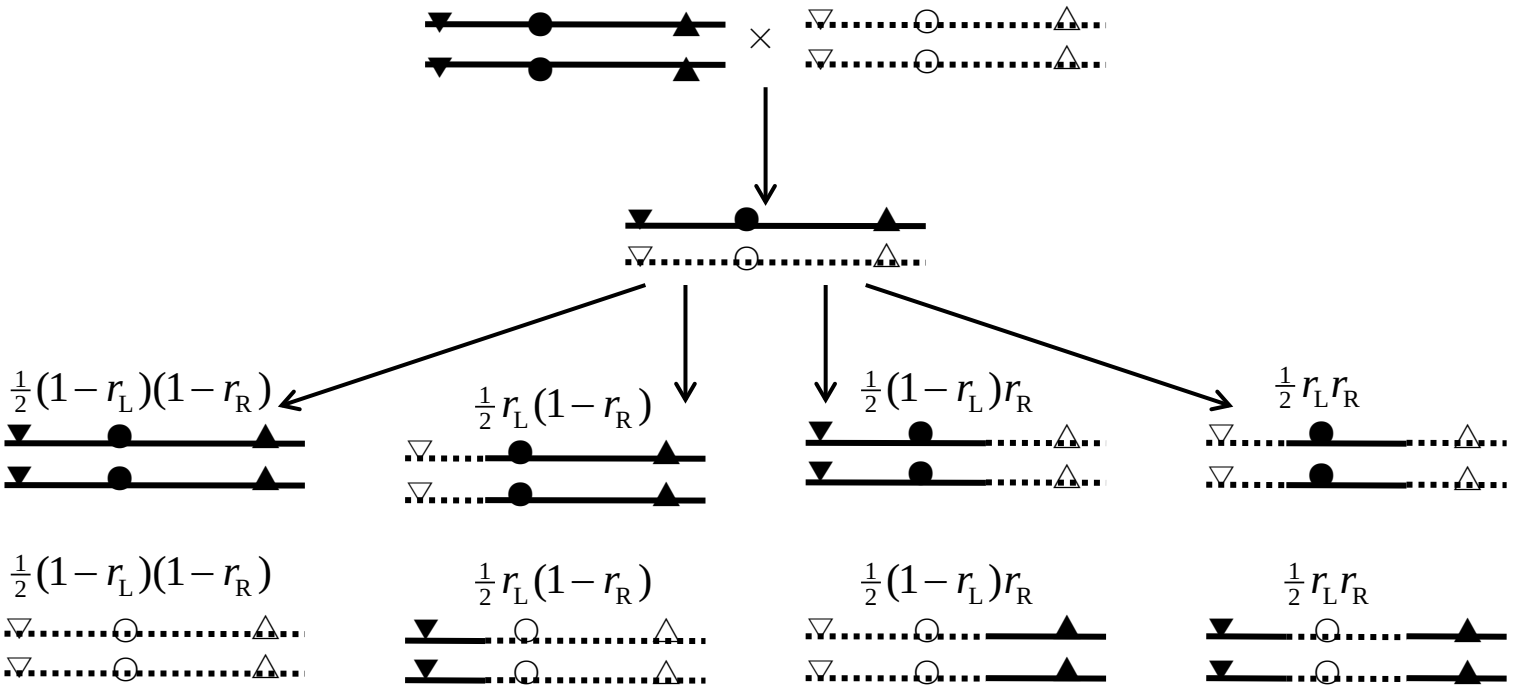
- Linear model (*j*=1，2，…，*n*)

$$y_i = b_0 + b^* x_j^* + e_j$$

- b^* represent QTL effect，x_j^* is the indicator variable (0 or 1) for QTL genotype
- Likelihood profile
- Support interval: One-LOD interval

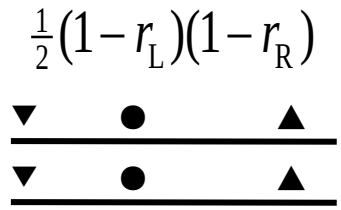
Marker types and QTL types in DH populations

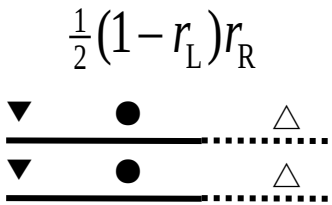
(double crossover is considered in this slide)

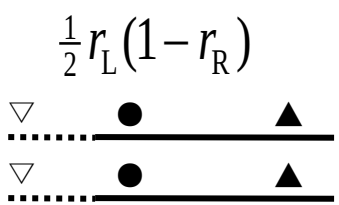


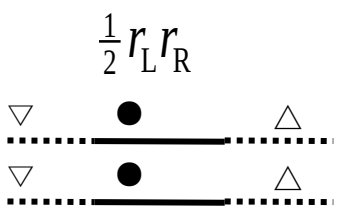
No crossover	One crossover between left marker and QTL	One crossover between QTL and right marker	Two crossovers between the two markers
$(1-r_L)(1-r_R)$	$r_L(1-r_R)$	$(1-r_L)r_R$	r_Lr_R

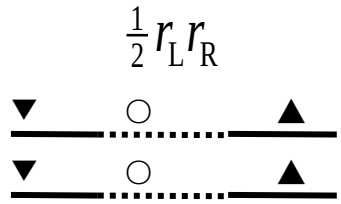
QTL types under each marker class

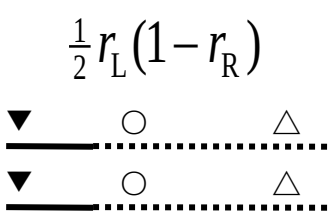


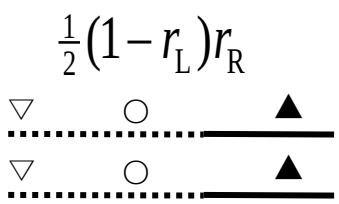


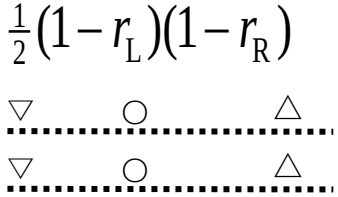












Marker class I

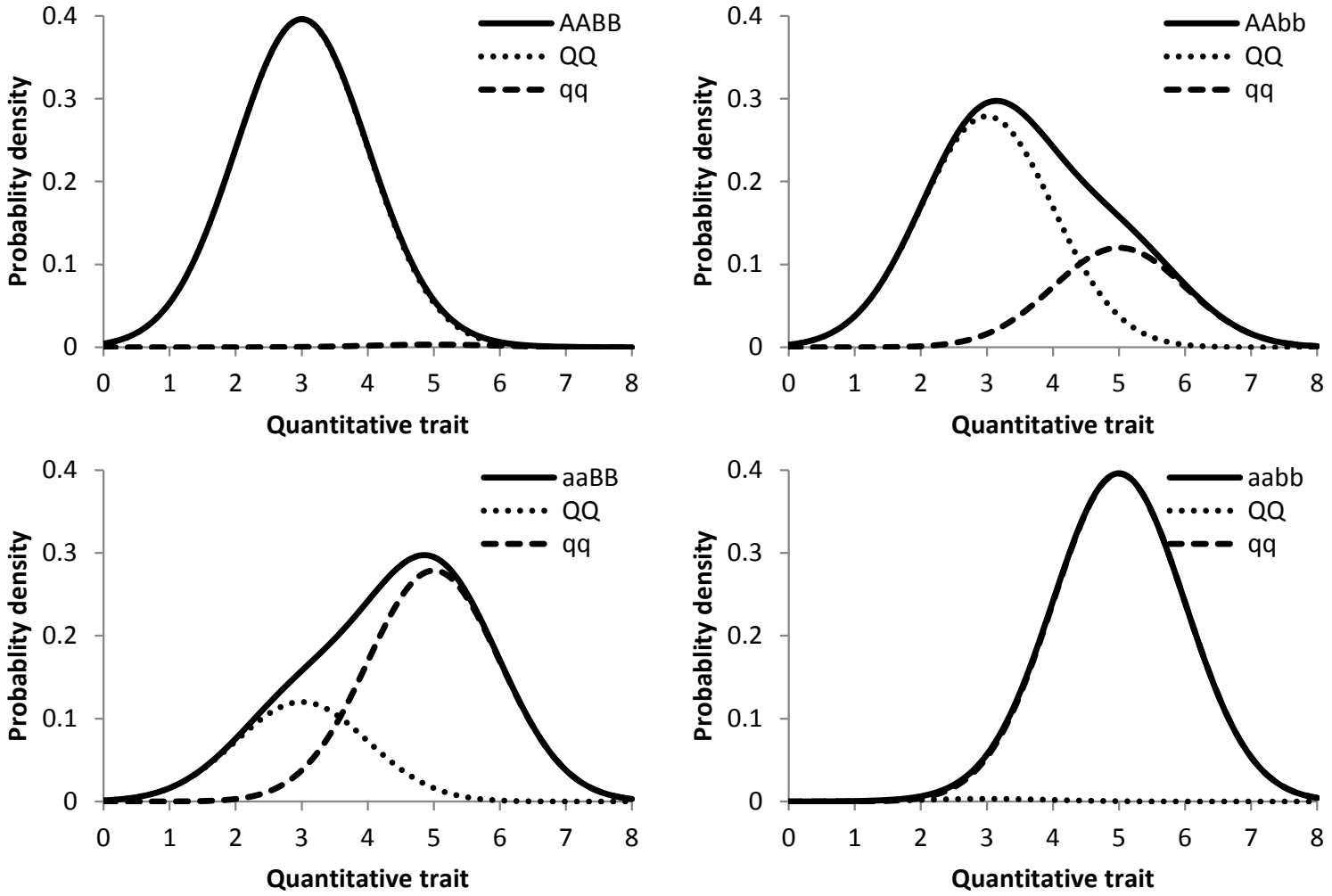
Marker class II

Marker class III

Marker class IV

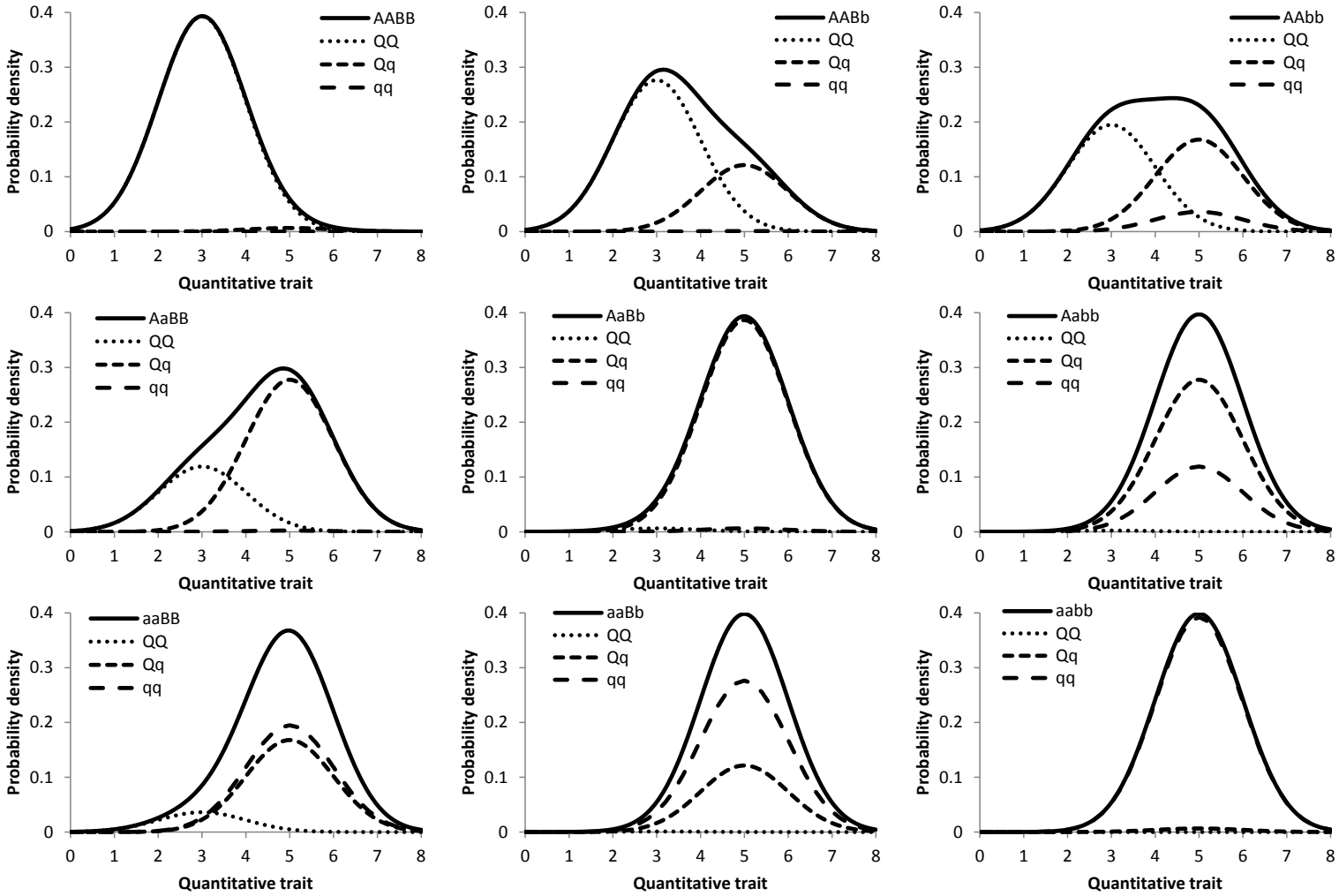
Two QTL genotypes in 4 marker classes in DH population

Proportion of QTL genotypes depends on QTL position and the marker interval



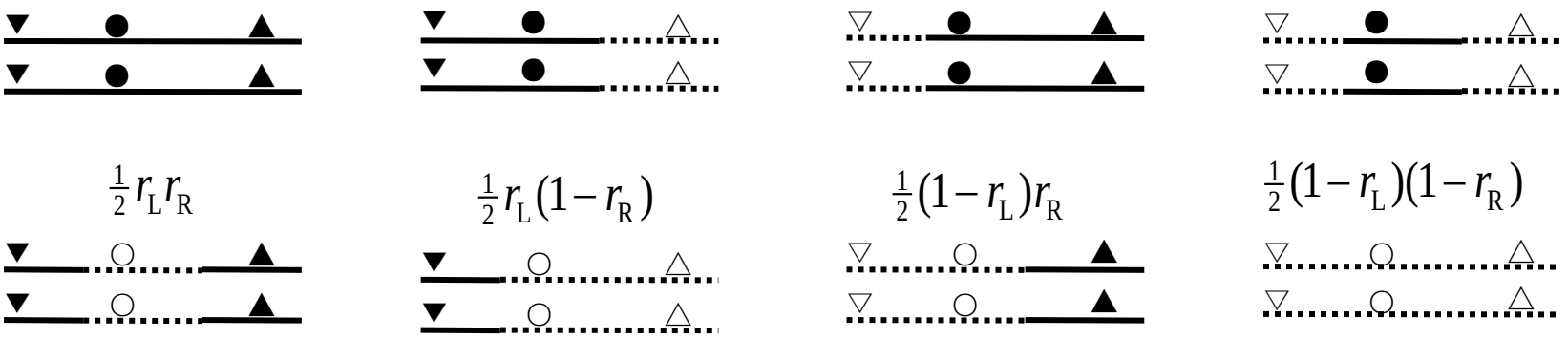
Three QTL genotypes in 9 marker classes in F2 population

Proportion of QTL genotypes depends on QTL position and the marker interval



Frequency of QTL genotypes in each marker class in DH population

$\frac{1}{2}(1-r_L)(1-r_R)$ $\frac{1}{2}(1-r_L)r_R$ $\frac{1}{2}r_L(1-r_R)$ $\frac{1}{2}r_Lr_R$



$\frac{1}{2}r_Lr_R$ $\frac{1}{2}r_L(1-r_R)$ $\frac{1}{2}(1-r_L)r_R$ $\frac{1}{2}(1-r_L)(1-r_R)$

Marker interval		Sample size	Frequency	QTL genotype	
Left	Right			QQ	qq
AA	BB	n_1	½(1-r)	½(1-r_L-r_R+r_Lr_R)	½r_Lr_R
AA	bb	n_2	½r	½(1-r_L)r_R	½r_L(1-r_R)
aa	BB	n_3	½r	½r_L(1-r_R)	½(1-r_L)r_R
aa	bb	n_4	½(1-r)	½r_Lr_R	½(1-r_L-r_R+r_Lr_R)

$$r = r_L + r_R - 2r_Lr_R$$

MLEs of means of QTL genotypes

$$Y_{ij} \sim \sum_{k=1,\cdots,q} \pi_{ik} N(\mu_k, \sigma^2)$$

$$\ln L(\mu_1, \cdots, \mu_q, \sigma^2 \mid \mathbf{Y} = \mathbf{y}) = \sum_{\substack{i=1,\cdots,m; \\ j=1,\cdots,n_i}} \ln\left(\sum_{k=1,\cdots,q} \pi_{ik} f(y_{ij} \mid \mu_k, \sigma^2) \right)$$

EM algorithm for calculating MLE

- The Expectation step, given initial values

$$w_{ijk} = \frac{\pi_{ik} f(y_{ij} \mid \mu_k^{(0)}, \sigma^{2(0)})}{\sum_{k'=1,\cdots,q} \pi_{ik'} f(y_{ij} \mid \mu_{k'}^{(0)}, \sigma^{2(0)})}$$

- w_{ijk} measures the probability of QTL genotypes of each DH line given the marker class

EM algorithm for calculating MLE

- The Maximization step, given QTL genotypes are known from w_{ijk} in E-step

$$\ln L(\mu_1, \cdots, \mu_q, \sigma^2 \mid \mathbf{X} = \mathbf{x}) = \sum_{\substack{i=1,\cdots,m;\\ j=1,\cdots,n_i}} \sum_{k=1,\cdots,q} w_{ijk} \ln f(y_{ij} \mid \mu_k^{(0)}, \sigma^{2(0)})$$

$$\mu_k^{(1)} = \frac{\sum_{\substack{i=1,\cdots,m;\\ j=1,\cdots,n_i}} w_{ijk} y_{ij}}{\sum_{\substack{i=1,\cdots,m;\\ j=1,\cdots,n_i}} w_{ijk}} \qquad \sigma^{2(1)} = \frac{\sum_{\substack{i=1,\cdots,m;\\ j=1,\cdots,n_i}} w_{ijk} (y_{ij} - \mu_k^{(1)})^2}{\sum_{\substack{i=1,\cdots,m;\\ j=1,\cdots,n_i}} w_{ijk}}$$

Test the existence of QTL

$$H_0 : \mu_1 = \cdots = \mu_q = \mu_0$$

$$H_A : \mu_1, \cdots, \mu_q \text{ are not equal}$$

Likelihood under H_0: $$L(\mu_0, \sigma_0^2 \mid \mathbf{Y} = \mathbf{y}) = \prod_{\substack{i=1,\cdots,m; \\ j=1,\cdots,n_i}} f(y_{ij} \mid \mu_0, \sigma_0^2)$$

Likelihood ratio test: $$LRT = -2 \ln \frac{\max L(H_0)}{\max L(H_A)} \sim \chi^2 (df = q-1)$$

Likelihood of odd (LOD): $$LOD = \log_{10} \left(\frac{\max L(H_A)}{\max L(H_0)} \right)$$

Estimation of genetic effects

DH populations: 1 for QQ, 2 for qq

$$\mu_1 = \mu + a \qquad \mu_2 = \mu - a$$

$$\hat{\mu} = \tfrac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2) \qquad \hat{a} = \tfrac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)$$

F2 populations, 1 for QQ, 2 for Qq, 3 for qq

$$\mu_1 = \mu + a \qquad \mu_2 = \mu + d \qquad \mu_3 = \mu - a$$

$$\hat{\mu} = \tfrac{1}{2}(\hat{\mu}_1 + \hat{\mu}_3) \quad \hat{a} = \tfrac{1}{2}(\hat{\mu}_1 - \hat{\mu}_3) \quad \hat{d} = \mu_2 - \tfrac{1}{2}(\hat{\mu}_1 + \hat{\mu}_3)$$

Contribution of a QTL

$$PVE = \frac{V_G}{V_P} \times 100\%$$

No distortion

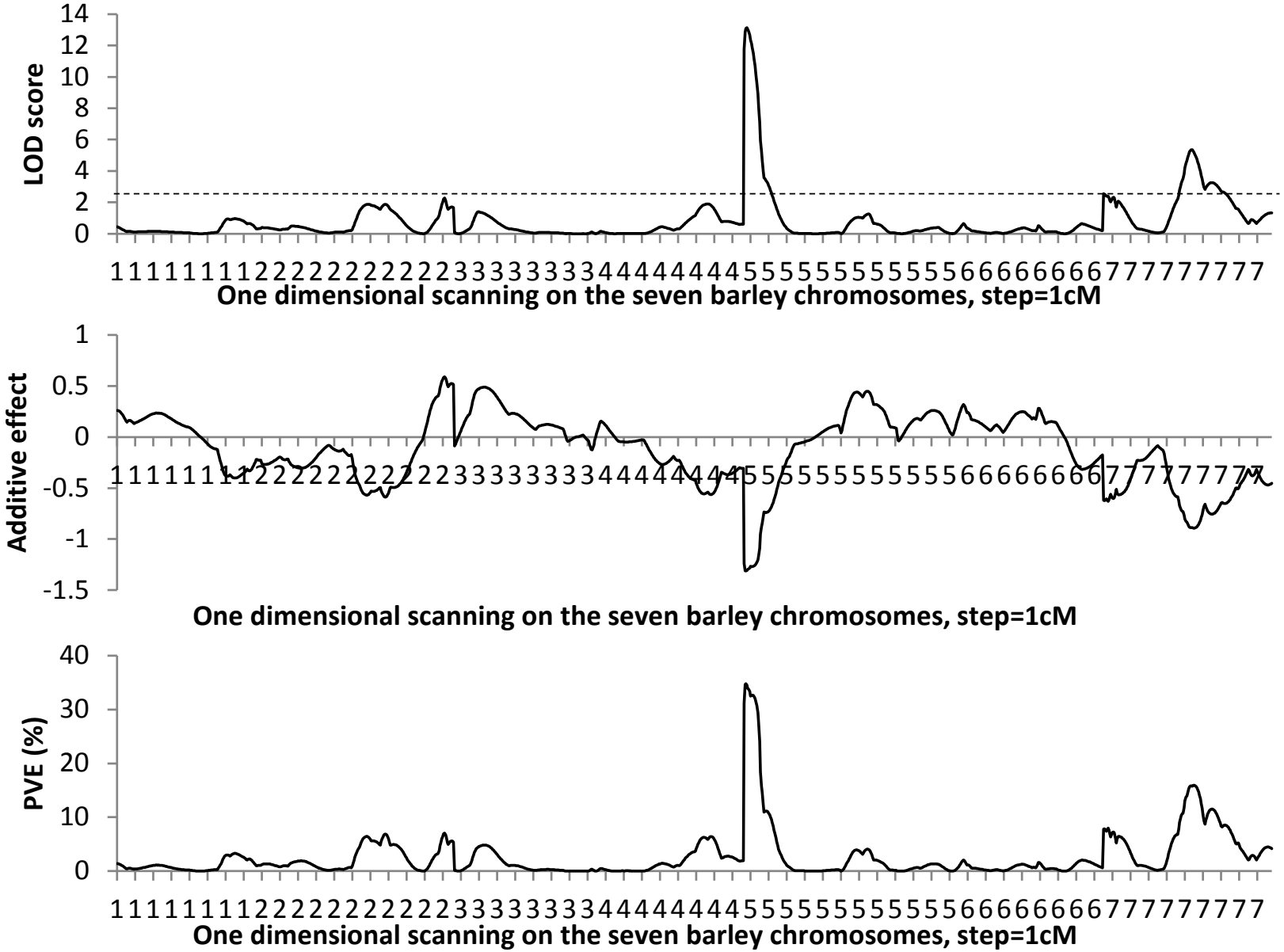
$$V_{G(DH)} = \hat{a}^2 \qquad V_{G(F2)} = \tfrac{1}{2}\hat{a}^2 + \tfrac{1}{4}\hat{d}^2$$

With distortion

$$V_{G(DH)} = 4 f_{QQ} f_{qq} a^2$$

$$V_{G(F2)} = [\, f_{QQ} + f_{qq} - (f_{QQ} - f_{qq})^2 \,]a^2 - 2 f_{Qq}(f_{QQ} - f_{qq})ad + (f_{Qq} - f_{Qq}^2)d^2$$

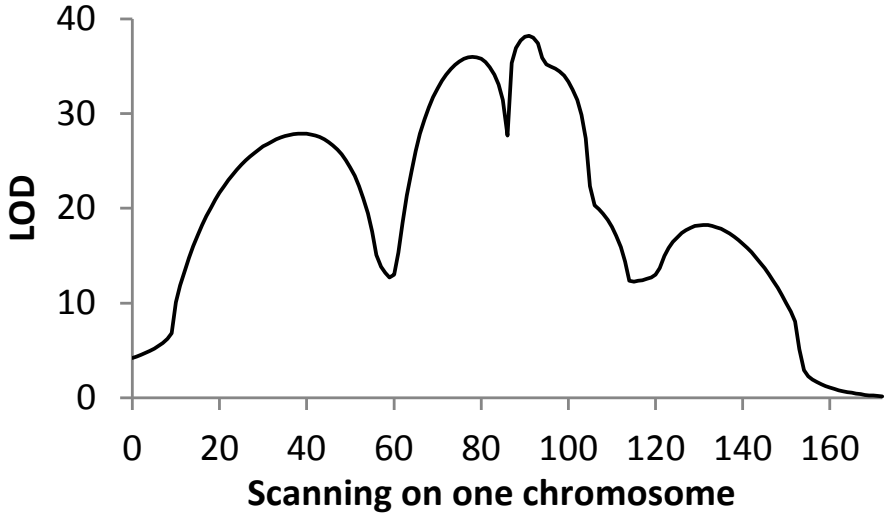
Interval mapping in a barley DH population

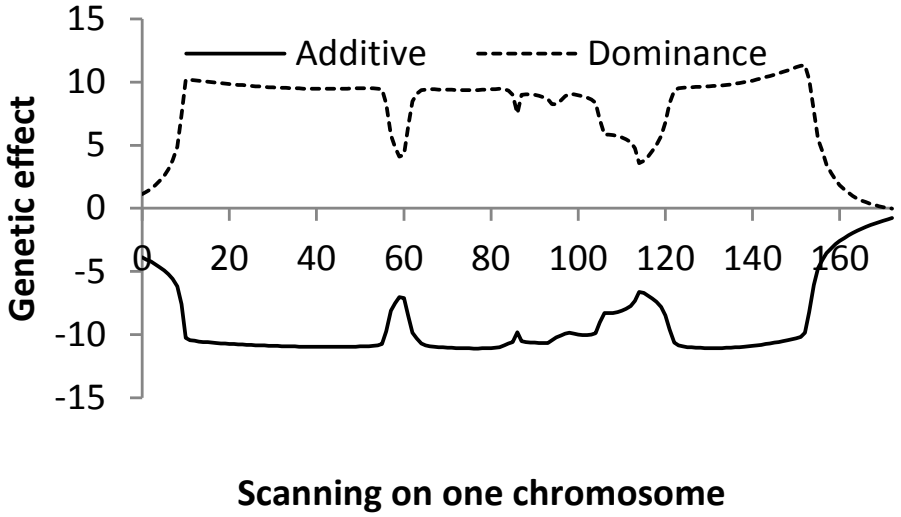


QTL identified in the barley DH population

Chromo.	Position (cM)	Left marker	Right marker	LOD	PVE (%)	Additive
5	3	ABA306B	Act88	13.15	34.55	-1.31
7	0	dRpg1	iPgd1A	2.55	7.79	-0.62
7	98	VAtp57A	MWG571D	5.36	15.77	-0.89

Interval mapping in a soybean F2 population



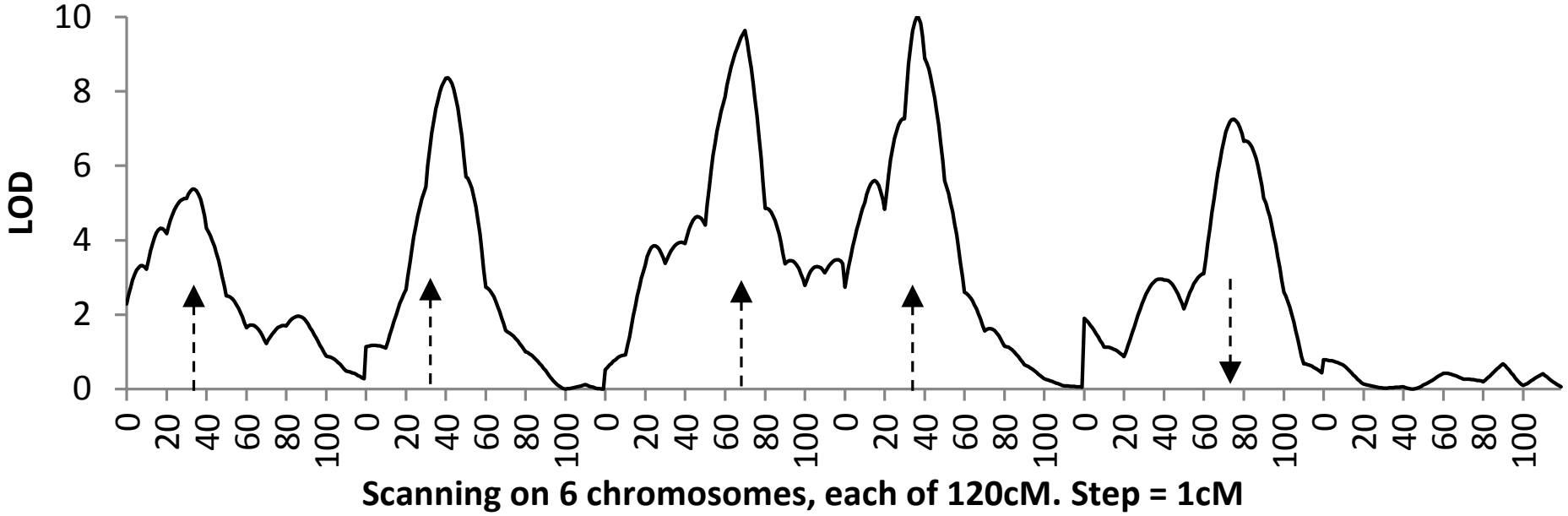


QTL identified in the soybean F2 population

Position (cM)	Left marker	Right marker	LOD	PVE (%)	Additive	Dominance
39	*Satt285	*Sat_239	27.89	73.02	-10.96	9.48
78	*Sat_239	*Satt255	36.00	68.49	-11.09	9.39
91	*Satt255	*Satt339	38.23	58.21	-10.66	8.90
131	*Satt521	*Sat_033	18.24	69.56	-11.07	9.68

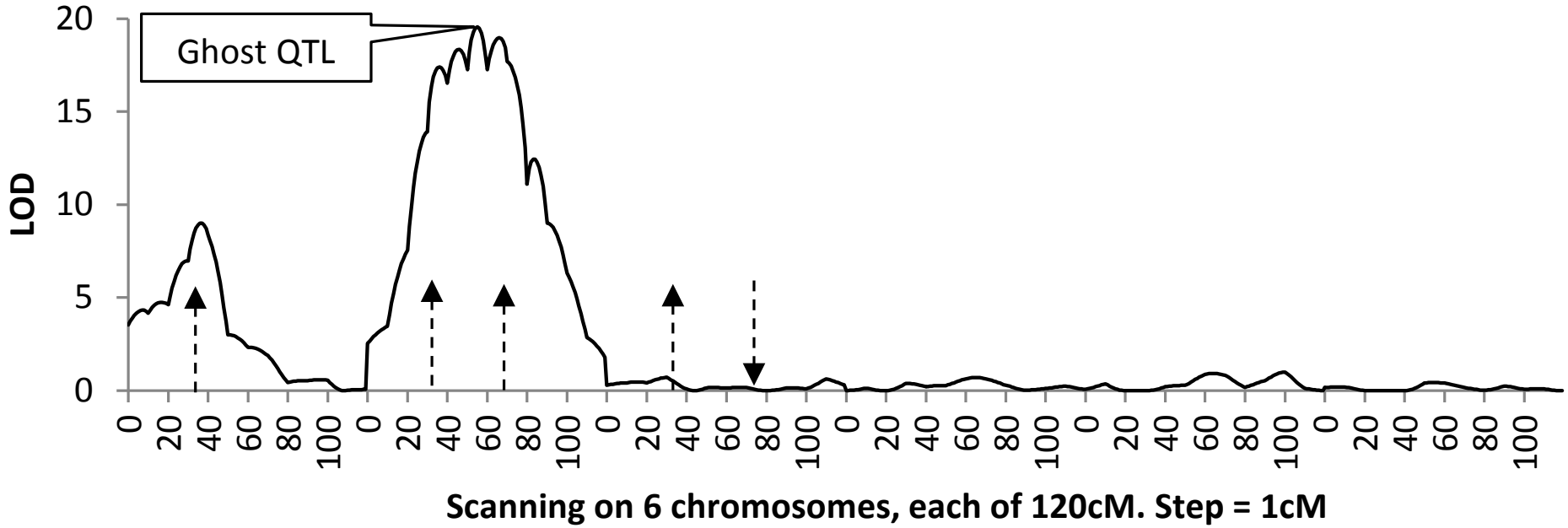
Problems with Simple Interval Mapping

- **Multiple peaks when QTLs are unlinked**



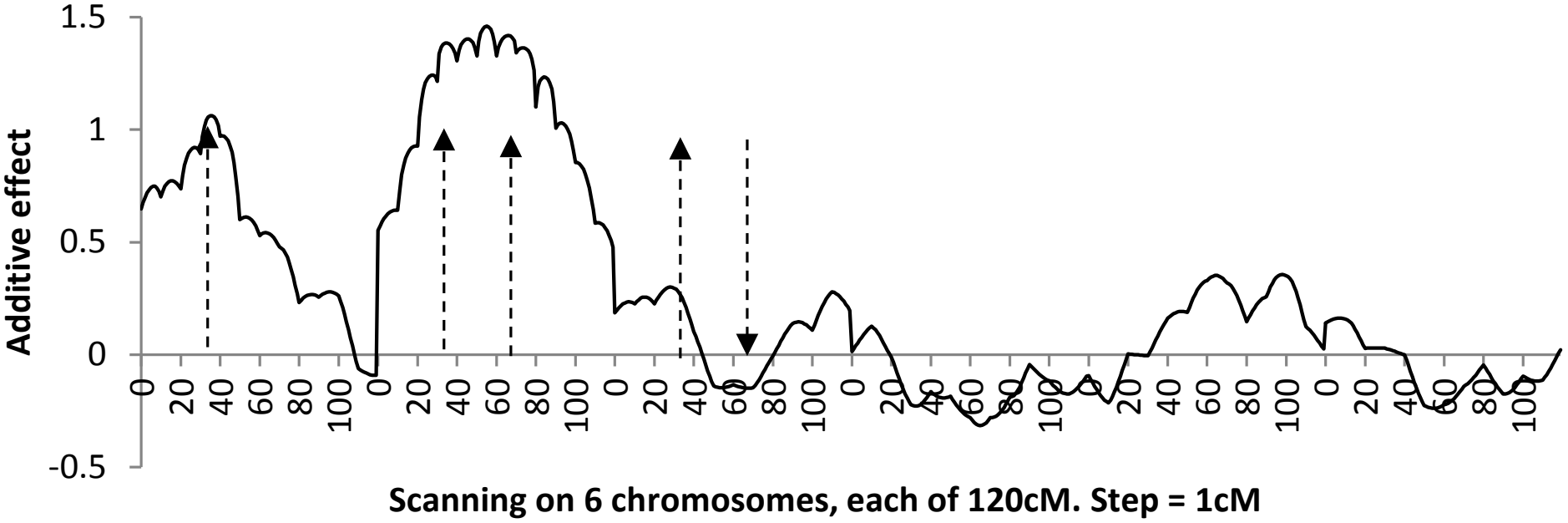
Problems with Simple Interval Mapping

- **Ghost QTL when two QTLs are linked**



Problems with Simple Interval Mapping

- **Biased estimation of QTL effects**



Exercises

There is one QTL (Q-q) located between marker locus A-a and marker locus B-b. Recombination frequencies are r_1 between A and Q, r_2 between Q and B, and r between A and B. Two parents have genotypes AAQQBB and aaqqbb.

- Assume there is no crossover interference, show that $r=r_1+r_2-2r_1r_2$
- Assume $r_1=0.1$, $r_2=0.2$.
  - Work out the frequencies of the 8 gametes of F1
  - Work out the QQ and qq frequencies in each of the 4 marker types in DH populations
  - Work out the QQ, Qq and qq frequencies in each of the 9 marker types in F2 populations