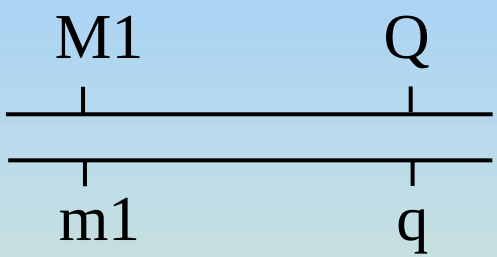
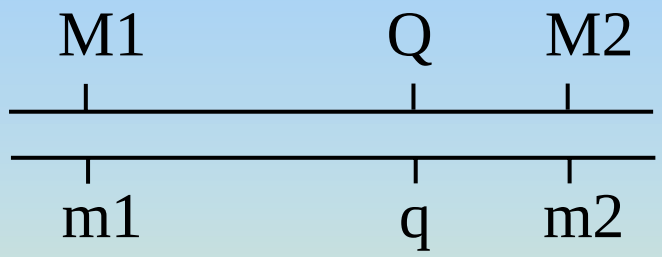
Interval mapping of QTL

- Single markers do not allow to distinguish between distance and size of QTL-effect
- Marker brackets do so, and they also provide more power

Single vs multiple markers



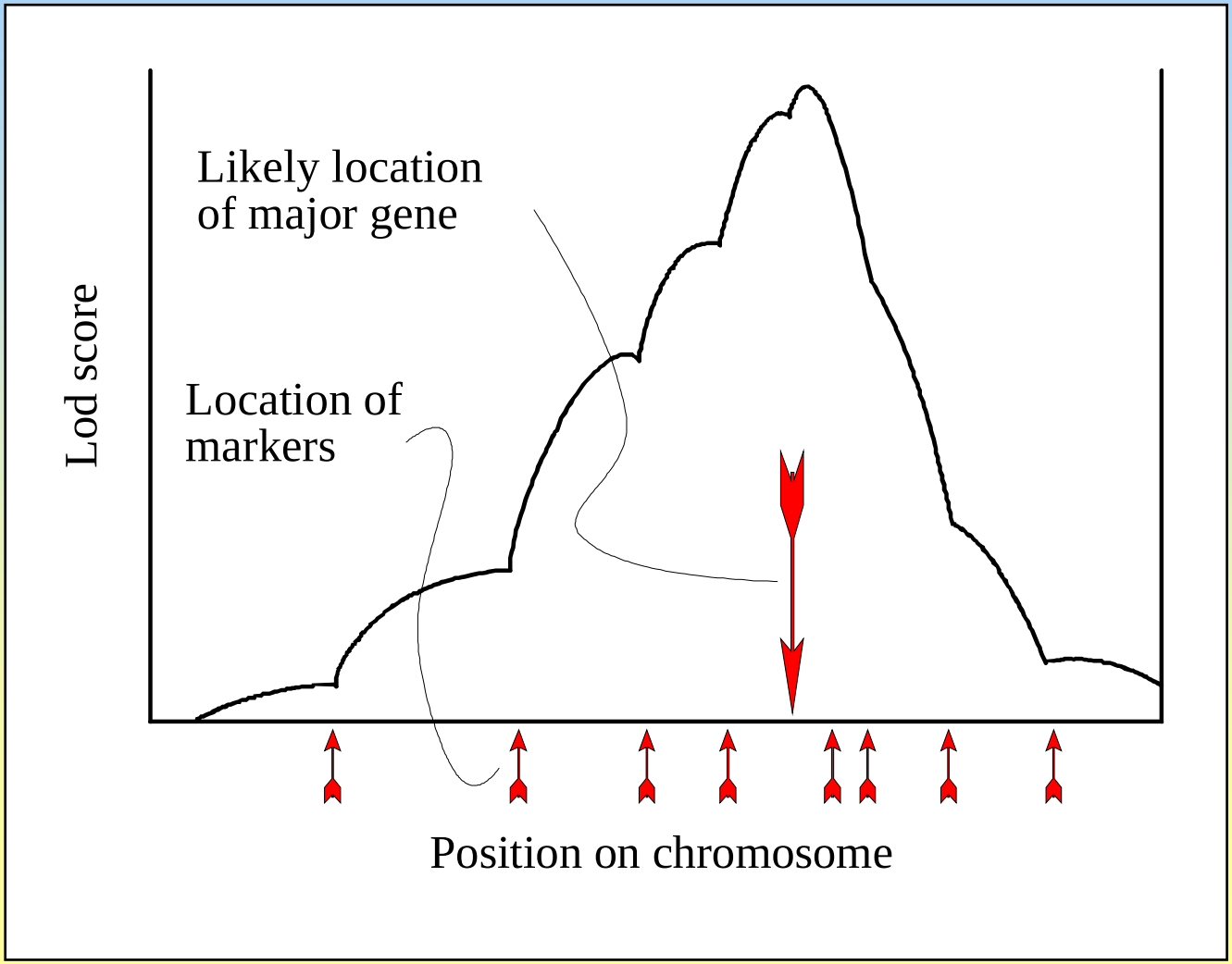
Single markers: not possible to distinguish between QTL effect and QTL position



Two (or more) markers: a lot less confounding between QTL effect and QTL position

Proper mapping of a QTL requires the use of multiple marker genotypes

QTL detection with markers



Gamete probabilities with two markers

Parental genotype M1 Q M2
m1 q m2

Recombination M1 – Q = r1
M2 – Q = r2
M1 – M2 = r12

Possible gametes			Recombination?	Gamete probability
M1	Q	M2	No	(1-r1)(1-r2) /2
M1	q	M2	Double: M1-q, q-M2	r1.r2 /2
M1	Q	m2	yes: Q-m2	(1-r1)r2 /2
M1	q	m2	yes: M1-q	r1(1-r2) /2
m1	Q	M2	yes: m1-Q	r1(1-r2) /2
m1	q	M2	yes: q-M2	(1-r1)r2 /2
m1	Q	m2	double: m1-Q, Q-m2	r1.r2 /2
m1	q	m2	no	(1-r1)(1-r2) /2

Sum = 1

Difference between marker genotypes

Marker alleles obtained from sire	QTL allele obtained from sire	Frequency	Expected mean of progeny
M1M2	**Q**	(1-r1)(1-r2)/2	$\mu + \alpha$
M1M2	q	r1.r2/2	μ
M1m2	Q	(1-r1)r2/2	$\mu + \alpha$
M1m2	q	r1(1-r2)/2	μ
m1M2	Q	r1(1-r2)/2	$\mu + \alpha$
m1M2	q	(1-r1)r2	μ
m1m2	Q	r1.r2/2	$\mu + \alpha$
m1m2	q	(1-r1)(1-r2)/2	μ

Marker genotype means

M1M2	$\frac{\frac{1}{2}(1-r1)(1-r2)(\mu+\alpha)+\frac{1}{2}r1.r2.\mu}{\frac{1}{2}(1-r12)}$	=	$\mu + (1-\frac{r1r2}{1-r12})\alpha$
M1m2	$\frac{\frac{1}{2}(1-r1).r2.(\mu+\alpha)+\frac{1}{2}r1(1-r2)\mu}{\frac{1}{2}r12}$	=	$\mu + \frac{r2-r1r2}{r12}\alpha$
m1M2	$\frac{\frac{1}{2}r1(1-r2)(\mu+\alpha)+\frac{1}{2}(1-r1).r2.\mu}{\frac{1}{2}r12}$	=	$\mu + \frac{r1-r1r2}{r12}\alpha$
m1m2	$\frac{\frac{1}{2}r1.r2(\mu+\alpha)+\frac{1}{2}(1-r1)(1-r2)\mu}{\frac{1}{2}(1-r12)}$	=	$\mu + \frac{r1r2}{1-r12}\alpha$

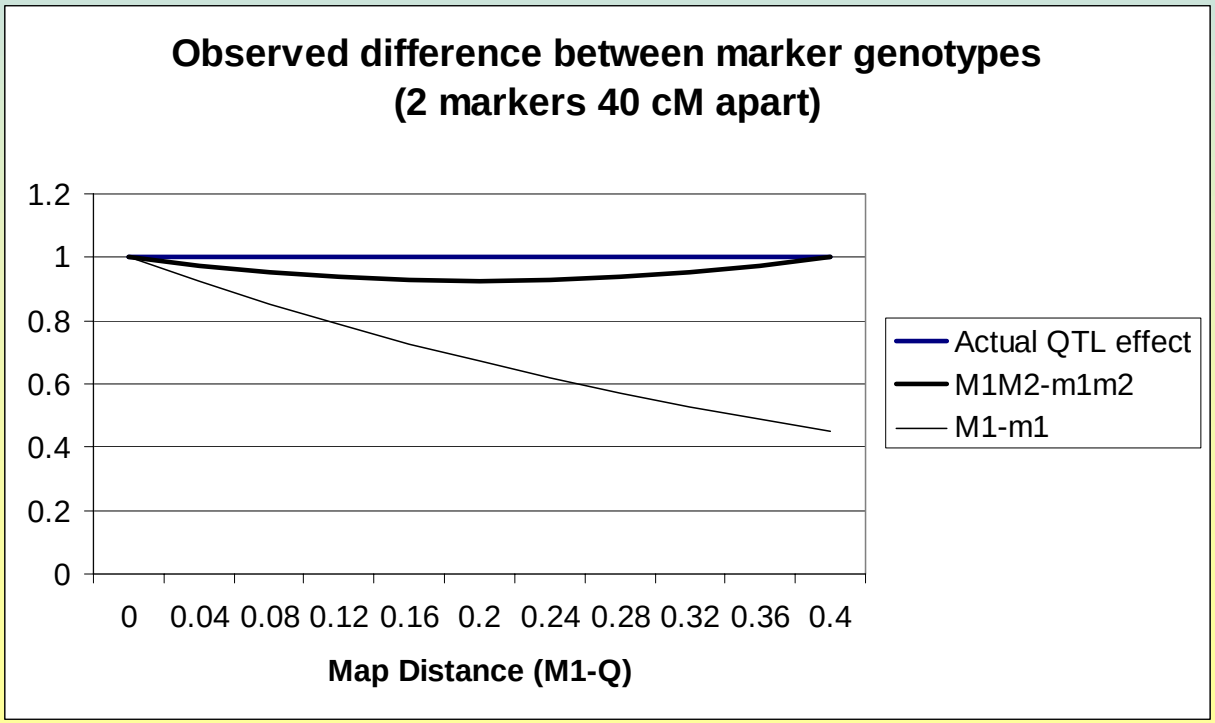
These means can be used to estimate μ and α for a given map position of QTL

This leads to a QTL mapping method (later).

Marker genotype differences

$$M_1 - m_1 = (1 - 2r_1)\alpha$$

$$M_1M_2 - m_1m_2 = \left[\mu + \left(1 - \frac{r_1 r_2}{1 - r_{12}}\right)\alpha\right] - \left[\mu + \frac{r_1 r_2}{1 - r_{12}}\alpha\right] = \left(1 - \frac{2r_1 r_2}{1 - r_{12}}\right)\alpha$$



Principle of QTL mapping

To test if there is a QTL at a specific location:

For each progeny, use markers to get the probability of each QTL genotype …

Then regress progeny phenotype on these probabilities to see if there is an association.

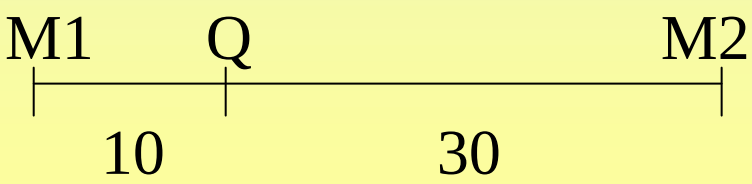
Difference between marker genotypes

Marker alleles obtained from sire	QTL allele obtained from sire	Frequency	Expected mean of progeny
M1M2	**Q**	(1-r1)(1-r2)/2	$\mu + \alpha$
M1M2	q	r1.r2/2	μ
M1m2	Q	(1-r1)r2/2	$\mu + \alpha$
M1m2	q	r1(1-r2)/2	μ
m1M2	Q	r1(1-r2)/2	$\mu + \alpha$
m1M2	q	(1-r1)r2	μ
m1m2	Q	r1.r2/2	$\mu + \alpha$
m1m2	q	(1-r1)(1-r2)/2	μ

Probability of marker haplotypes

Markertype	(M1M2)	P(M1QM2)	**P(Q|M1M2)**
M1M2	0.362	0.352	0.972
M1m2	0.138	0.103	0.745
m1M2	0.138	0.035	0.255
m1m2	0.362	0.010	0.028

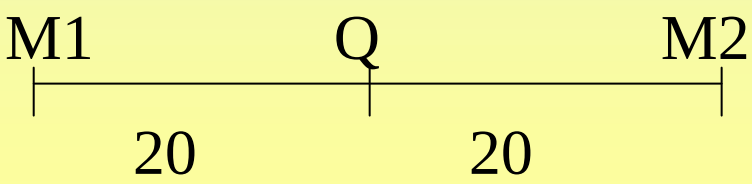
conditional probability of having Q, given paternal marker haplotype



Probability of marker haplotypes

Markertype (M1M2)	P(M1QM2)	**P(Q|M1M2)**
M1M2	0.362	0.349	0.963
M1m2	0.138	0.069	0.500
m1M2	0.138	0.069	0.500
m1m2	0.362	0.014	0.037

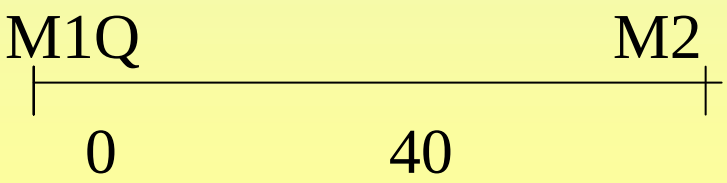
conditional probability of having Q, given paternal marker haplotype



Probability of marker haplotypes

Markertype	(M1M2)	P(M1QM2)	**P(Q|M1M2)**
M1M2	0.362	0.362	1.000
M1m2	0.138	0.138	1.000
m1M2	0.138	0.000	0.000
m1m2	0.362	0.000	0.000

conditional probability of having Q, given paternal marker haplotype



Fitting the Goodness of Fit of a certain position to the data: M1-Q = 0.1

mean	p(Q)	y-hat	y
1.0000	0.9718	50.4321	50.9813
1.0000	0.9718	50.4321	49.9813
1.0000	0.7451	50.3446	50.7500
1.0000	0.7451	50.3446	49.7500
1.0000	0.2549	50.1554	50.7500
1.0000	0.2549	50.1554	49.7500
1.0000	0.0282	50.0679	50.5187
1.0000	0.0282	50.0679	49.5187

dM1-Q	SST	SSE	LR
0.1	2.2139	2.0455	0.6331

-

Fitting the Goodness of Fit of a certain position to the data: M1-Q = 0

mean	p(Q)	y-hat	y
1.0000	1.0000	50.3656	50.9813
1.0000	1.0000	50.3656	49.9813
1.0000	1.0000	50.3656	50.7500
1.0000	1.0000	50.3656	49.7500
1.0000	0	50.1344	50.7500
1.0000	0	50.1344	49.7500
1.0000	0	50.1344	50.5187
1.0000	0	50.1344	49.5187

dM1-Q	SST	SSE	LR
0	2.2139	2.1070	0.3961

Interval mapping

Compare the likelihoods at different locations for the QTL

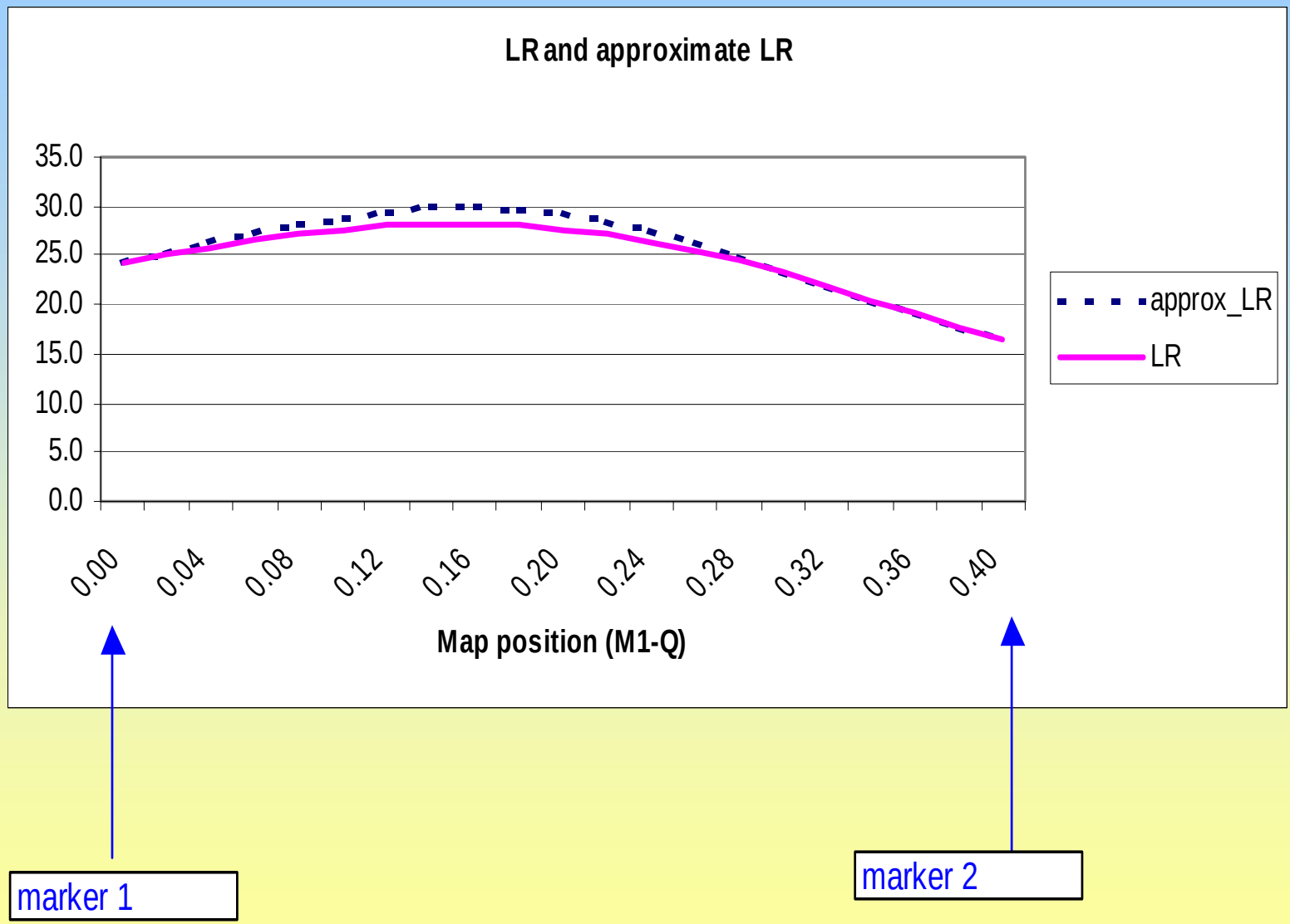
$$LR = -2\ln\frac{Max\_Likelihood(reduced\ \mathrm{mod}\ el)}{Max\_Likelihood(full\ \mathrm{mod}\ el)}$$

Full model : $y = mu + Q_r + e$

Reduced model $y = mu + e$

LOD score (Lander and Bostein, 1989)

$$-\log_{10}\frac{Max\_Likelihood(reduced\ \mathrm{mod}\ el)}{Max\_Likelihood(full\ \mathrm{mod}\ el)} = \frac{LR(c)}{4.61}$$


LR and approximate LR
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.32
0.36
0.40
Map position (M1-Q)
approx_LR
LR
marker 1
marker 2

Interval mapping using regression

Regression model

$$y = \mu + \alpha.x + e$$

where y = the observed phenotype

x = probability (Q|marker genotype)

giving $SSE = \Sigma(y- \mu_0 -\hat{a}.x)^2$

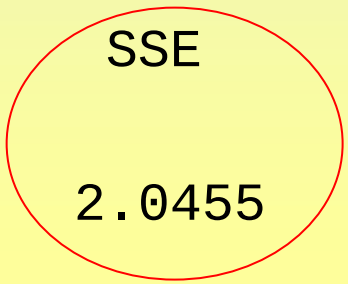
reduced model $y = \mu_0 + e$

giving $SST = \Sigma(y- \mu_0)^2$

Fitting the Goodness of Fit of a certain position to the data: M1-Q = 0.1

1.0000	0.9718	50.4321	50.9813
1.0000	0.9718	50.4321	49.9813
1.0000	0.7451	50.3446	50.7500
1.0000	0.7451	50.3446	49.7500
1.0000	0.2549	50.1554	50.7500
1.0000	0.2549	50.1554	49.7500
1.0000	0.0282	50.0679	50.5187
1.0000	0.0282	50.0679	49.5187

dM1-Q	SST	SSE	LR
0.1	2.2139	2.0455	0.6331



Approximate LR test statistic

$$LR = n \ln(SST/SSE)$$

if other fixed effects than QTL:

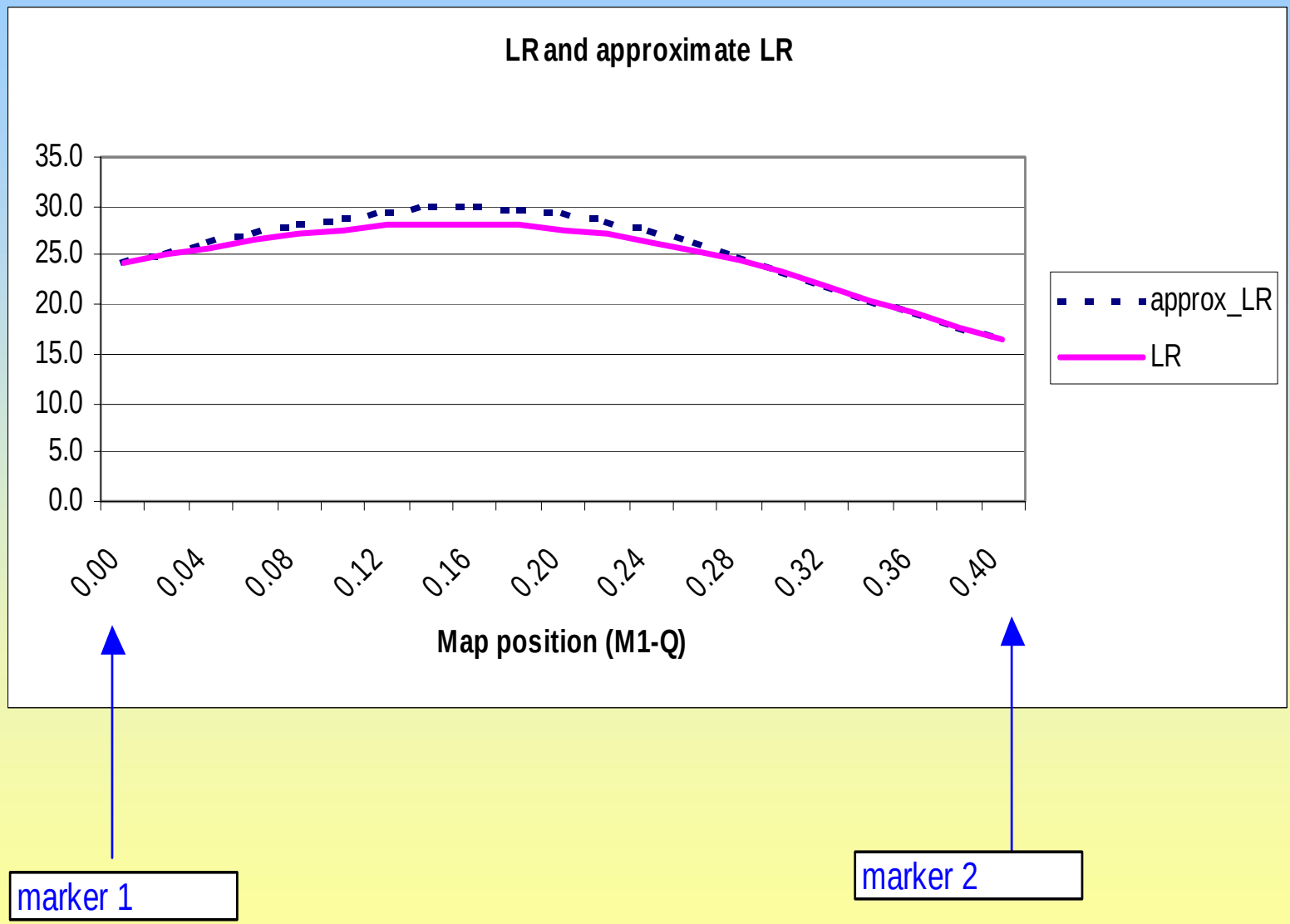
$$LR = n \ln(SSE_{reduced} / SSE_{full})$$

Fitting the Goodness of Fit of a certain position to the data: M1-Q = 0.1

1.0000	0.9718	50.4321	50.9813
1.0000	0.9718	50.4321	49.9813
1.0000	0.7451	50.3446	50.7500
1.0000	0.7451	50.3446	49.7500
1.0000	0.2549	50.1554	50.7500
1.0000	0.2549	50.1554	49.7500
1.0000	0.0282	50.0679	50.5187
1.0000	0.0282	50.0679	49.5187

dM1-Q	SST	SSE	L
0.1	2.2139	2.0455	0.57

approx LR = 8*ln(SST/SSE)= 0.63


LR and approximate LR
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.32
0.36
0.40
Map position (M1-Q)
approx_LR
LR
marker 1
marker 2

Methods for QTL analysis

- Regression
- Maximum Likelihood
- Other methods GRM-method

  MCMC
- Hypothesis testing
  - permutation tests
  - bootstrapping

Regression on marker genotypes

Regression on single markers

$y = \mu + b.MG_1 + e$

number of marker classes 2,3,……..

Quick and simple, use F-statistic

Not most powerful

Regression on multiple markers

$y = \mu + b_1.MG_1 + b_2.MG_2 + \ldots\ldots\ldots + b_nMG_n$

multiple regression (stepwise)

does not estimate exact location

more power than single trait marker

Regression on QTL probability

$y = \mu + \alpha.x + e$ as before

where y is the observed phenotype

x P(Q|Markers,recomb)

used in interval mapping, i.e. stepwise through interval

Haley-Knott regression useful in F2/BC design

$y = \mu + \alpha.x_1 + \beta.x_2 + e$

where y is the observed phenotype

$x_1 = P(QQ|M_i) - P(qq|M_i)$

$x_2 = P(Qq|M_i)$

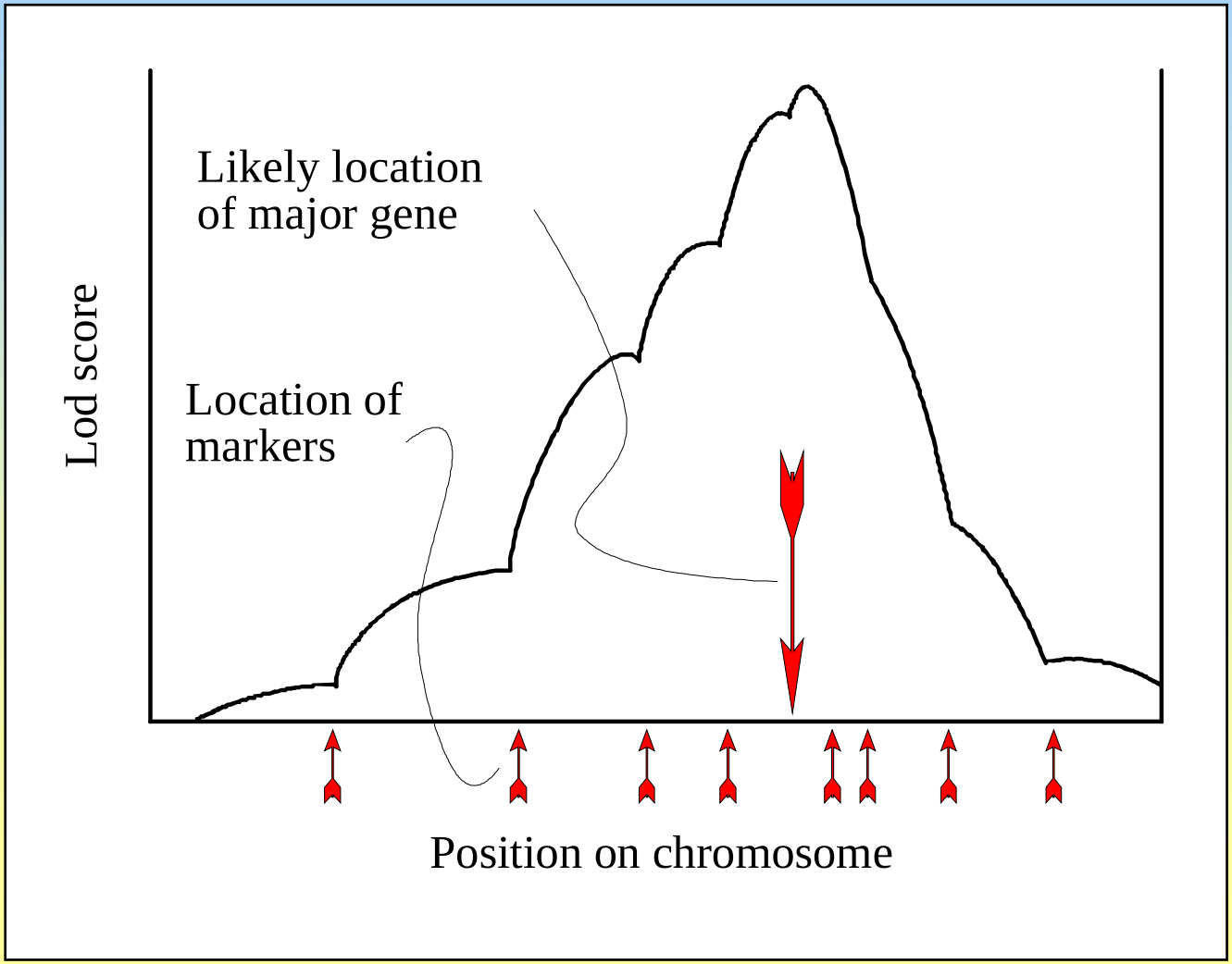
Test statistics with regression analysis

Approximate LR = $n \ln(SSE_{reduced} / SSE_{full})$

F-test = MSQ/MSE

approx.LR = $n \log_e[1+(df_1/df_2)F]$

QTL detection with markers



Regression on flanking markers

Whittaker et al. (1996)

$$y = \mu + \beta_1.x_L + \beta_2.x_R + e$$

$$= \mu + \alpha\lambda.x_L + \alpha\rho.x_R + e$$

Now $\lambda = P(Q|X_L = M1M1, X_R = m2m2)$

$\rho = P(Q|X_L = m1m1, X_R = M2M2)$.

α = **effect of Q.**

x_L **and** x_R **refer to left and right marker, and have values –1, 0 and 1**

do not need to evaluate each position of interval

Regression on flanking markers

Whittaker et al. (1996)

$$y = \mu + \alpha\lambda.x_L + \alpha\rho.x_R + e$$

From the regression coefficients: $\beta_1 = \alpha\lambda$, and $\beta_2 = \alpha\rho$, it was shown (Whittaker et al., 1996) that location and QTL effect can be estimated:

$$r_1 = 0.5\left[1 - \sqrt{1 - \frac{4\beta_2\theta(1-\theta)}{\beta_2 + \beta_1(1-2\theta)}}\right]$$

$$\alpha = \sqrt{\frac{[\beta_1 + (1-2\theta)\beta_2][[\beta_2 + (1-2\theta)\beta_1}{1-2\theta}}$$

Maximum likelihood estimation

Prob. Dens. Function: $F(y_i) = P(y|\theta)$

$$f(y_i|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\frac{1}{2}(y-\mu)^2}{\sigma^2}} = L(\mu,\sigma|y_i)$$

The total likelihood of data set **y** is calculated as the product of all likelihoods for each observation.

$$L(\mu,\sigma|\mathbf{y}) = \Pi_i L(\mu,\sigma|y_i)$$

Likelihood for QTL model

$$L(\mu_1, \mu_2, \sigma | y_i) = P(\mu_1).\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{\frac{1}{2}(y-\mu_1)^2}{\sigma^2}} + P(\mu_2).\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{\frac{1}{2}(y-\mu_2)^2}{\sigma^2}}$$

Sum over the possibilities for the different QTL genotypes

Solve for μ, μ, σ by EM algorithm

Example of Likelihood calculation

the QTL position M1-Q = 0.1

Phenotype	Marker haplotye	Prob(Q|markers)	Expected phenotype (H1-model)	LogL0	LogL
50.98	M1M2	0.9718	50.43	-1.18884	-0.81727
49.98	M1M2	0.9718	50.43	-0.4575	-0.65658
50.75	M1m2	0.7451	50.34	-0.73859	-0.59655
49.75	M1m2	0.7451	50.34	-0.73859	-0.91164
50.75	m1M2	0.2549	50.16	-0.73859	-0.91152
49.75	m1M2	0.2549	50.16	-0.73859	-0.59663
50.52	m1m2	0.0282	50.07	-0.4575	-0.65648
49.52	m1m2	0.0282	50.07	-1.18884	-0.81739
			sum	-6.24705	-5.96407

No QTL-model: mean = 50.25,

SST = 2.21 > var = 0.316

QTL-model: mu = 50.06 α = 0.386 > means: 50.06 and 50.44

SST = 2.05 > var = 0.292

- LogL value of H0 model: -6.247
  under H1: -5.964

LR = -2*(L0 - L) = -2 (-6.247 + 5.964) = 0.57

The approximate LR value from regression was

appr.LR = $n \ln(\frac{SSE_{reduced}}{SSE_{full}})$ = 8.ln(2.21/2.05) = 0.63.

Multiple family testing

- Sum over families:
  - Prob (hetereozygote sire) X
  - Prob(phase) X
  - - x

Comparison regression-ML

- ML takes into account that within a marker type there are really two normal distributions
- Most of the variation comes from between marker type differences as ML ~ Regress.
- Difference is largest with big QTL and with QTL further from markers
- Xu (1995) suggested a correction to avoid upward bias in estimate of variance

$$\sigma^2_{e\_corrected} = \sigma^2_e - a^2 \sum_{i=1}^{4} p_i(1-p_i)$$

Comparison regression-ML

- ML is computationally and practically a bigger task
- ML needed for across family analysis!
- Regression more robust against non-normality
- ML uses more information (I.e. segregation analysis!) but regression model with genotype probability routine is similar

Other methods

- Regression on Q probability, last obtained with segregation analysis (genotype probability)
- MCMC
- QTL as random effect
  - GRM = Gametic Relationship Matrix

QTL as random "GRM method"

Model:

Individual animal phenotype =

Fixed Environmental effects

\+ Sum of average effects of 'polygenic' alleles

\+ ***Average effect of paternal QTL allele***

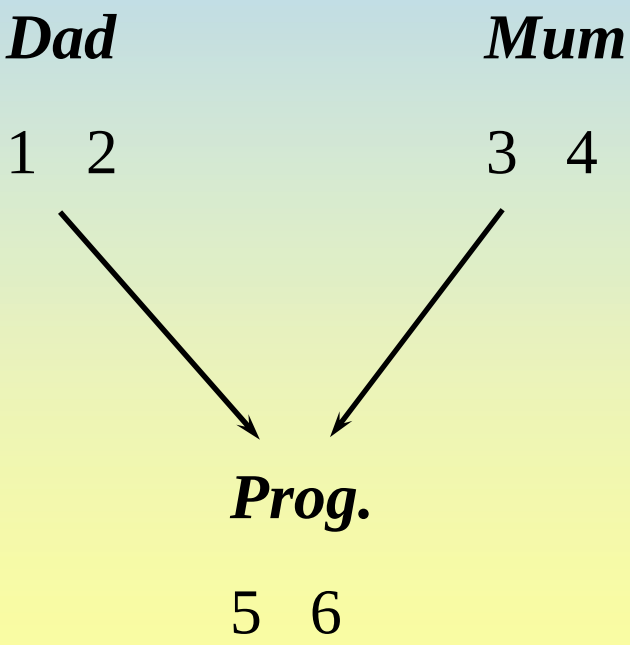
\+ ***Average effect of maternal QTL allele***

\+ Random Error

} Estimate VC's

Gametic relationship matrix

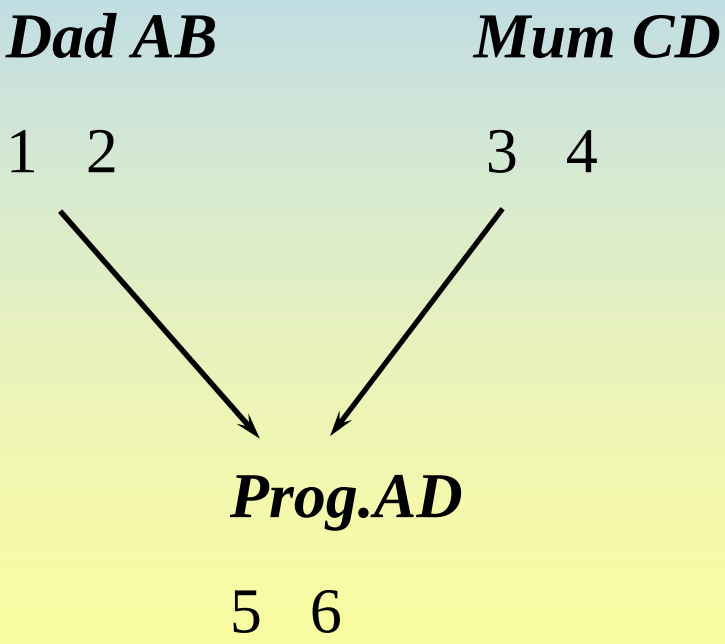
With no markers:



		Dad		Mum		Prog.	
	Site	1	2	3	4	5	6
Dad	1	1	0	0	0	.5	0
	2	0	1	0	0	.5	0
Mum	3	0	0	1	0	0	.5
	4	0	0	0	1	0	.5
Prog	5	.5	.5	0	0	1	0
	6	0	0	.5	.5	0	1

Gametic relationship matrix

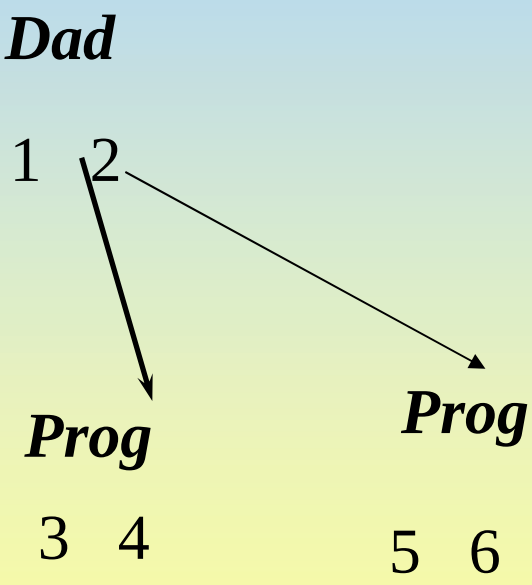
With markers: r = 0.1



		Dad		Mum		Prog.	
	Site	1	2	3	4	5	6
Dad	1	1	0	0	0	.9	0
	2	0	1	0	0	.1	0
Mum	3	0	0	1	0	0	.1
	4	0	0	0	1	0	.9
Prog	5	.9	.1	0	0	1	0
	6	0	0	.1	.9	0	1

Gametic relationship matrix for MAS

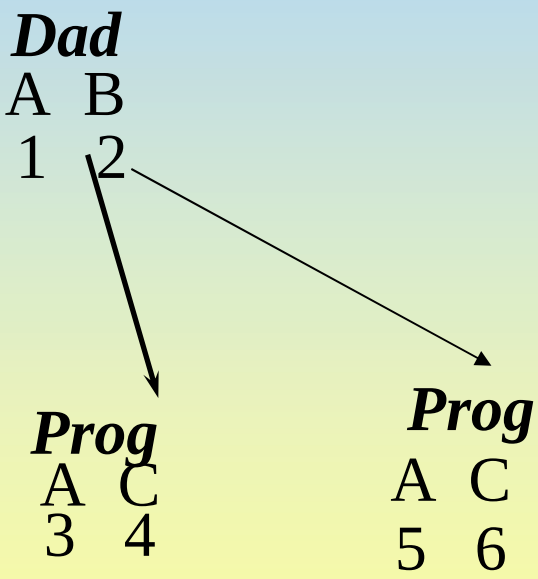
With no markers



		Dad		Proq1		Proq2	
	Site	1	2	3	4	5	6
Dad	1	1	0	.5	0	.5	0
	2	0	1	.5	0	.5	0
Proq1	3	.5	.5	1	0	.5	0
	4	0	0	0	1	0	0
Proq2	5	.5	.5	.5	0	1	0
	6	0	0	0	0	0	1

Gametic relationship matrix for MAS

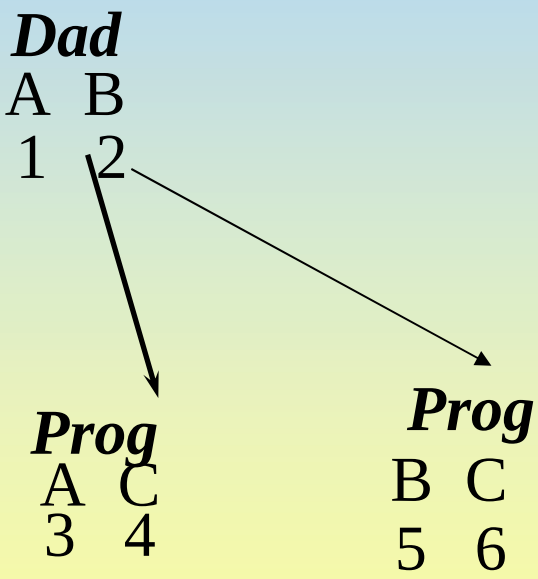
With markers: r = 0.1



		Dad		Proq1		Proq2	
	Site	1	2	3	4	5	6
Dad	1	1	0	.9	0	.9	0
	2	0	1	.1	0	.1	0
Proq1	3	.9	.1	1	0	.81	0
	4	0	0	0	1	0	0
Proq2	5	.9	.1	.81	0	1	0
	6	0	0	0	0	0	1

Gametic relationship matrix for MAS

With markers: r = 0.1



		Dad		Prog1		Prog2	
	Site	1	2	3	4	5	6
Dad	1	1	0	.9	0	.1	0
	2	0	1	.1	0	.9	0
Prog1	3	.9	.1	1	0	.81	0
	4	0	0	0	1	0	0
Prog2	5	.1	.9	.18	0	1	0
	6	0	0	0	0	0	1

The GRM gives more accurate relationships at the QTL!

- True covariance (at the QTL) rather than the one based on average effects

- Relationships matrix = $\Sigma\ A_{QTL}|$markers

Hypothesis testing

- LR tests have chi-squared distribution
- 1-LOD-rule gives 95% CI

- These tests are not exact a we compare normal with a non-normal distribution or normally distributed errors
  - alternative: empirical testing

Hypothesis Testing

Significance thresholds based on Permutation test (Churchill and Doerge, 199?)

Original data

Animal ID	Marker Genotype	Pheno-type
1	Mmnn	9.8
2	mmnn	10.4
3	mmnn	9.3
4	Mmnn	8.5
5	MmNn	11.3
6	MmNn	9.6
7	MmNn	9.9
8	mmnn	7.6
9	MmNn	8.0
10	mmNn	10.7

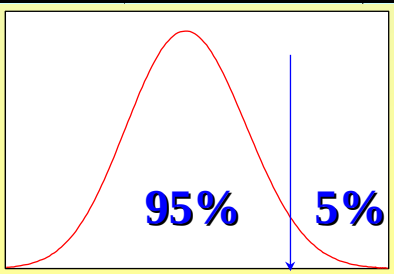
Randomly permuted data

Animal ID	Marker Genotype	Pheno-type
1	MmNn	9.8
2	mmNn	10.4
3	Mmnn	9.3
4	MmNn	8.5
5	mmnn	11.3
6	MmNn	9.6
7	Mmnn	9.9
8	mmnn	7.6
9	MmNn	8.0
10	mmnn	10.7

Test statistic under Null Hypothesis

Replicate

Distribution of test statistic



Threshold

bootstrapping

- Analyze a set of data, with obs'ns taken from the original data *with replacement*

Account for multiple testing!!

$\alpha = 1 - (1 - \gamma)^{1/n} \approx \gamma/n$

n = number of tests

$\gamma = 1 - (1 - \alpha)^{n}$ = prob of at least one test positive

e.g. 200 tests: use significance level of 0.05/200 = 0.000025

Informative F_2 Families for QTL Mapping in Pig at Hohenheim University

♂ ♀ ♂ ♀ Sources

X X X Mating

F_1 F_1 F_1

F_2 (> 300) F_2 (> 300) F_2 (> 300) Offspring Generation (Number of Pigs)

Performance Traits (Fattening, Carcass Composition, Stress Reaction) — Station Test

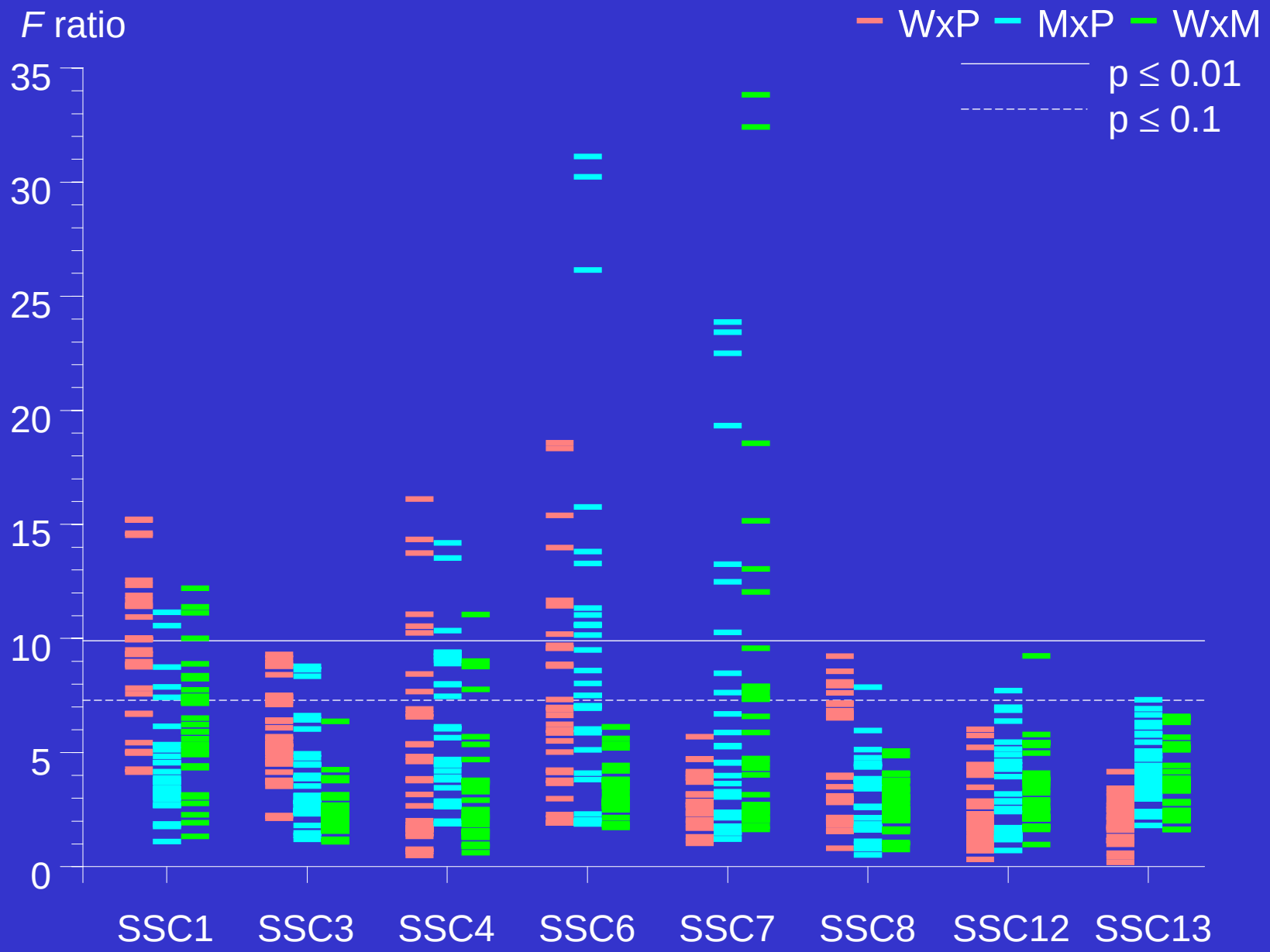
⇩

Recording of Marker Loci and Detailed Trait Criteria — Laboratory Analysis

⇩

Gene and QTL Mapping

Summary of F Ratio Maxima for Carcass Traits



Summary methods

- Can do a quick scan with single marker regression
- In promising regions can use interval mapping, either ML or multiple marker regression

- Need to account for additional QTL (see next)

- Use empirical test statistics (permutation tests)