

Detecting QTLs

Workshop on Polygenic Adaption
ICTS, Bangalore
6 – 17 May 2024

Bruce Walsh
University of Arizona
jbwalsh@Arizona.edu

Motivation

- To ask deep questions about polygenic adaptation, we need insight into the **genetic architecture** (joint distribution of allele-frequencies and effects) of traits
- We can do this by **tightly tagging causative sites with scorable genetic markers**
- Linkage mapping ("QTL mapping")
 - Need families/pedigrees, "tagged" regions are megabases in size
 - Usually based on line crosses
 - Captures **between-population** (~ fixed) differences
- Association mapping ("GWAS" – genome-wide association study)
 - Random population sample, "tagged" regions kilobases in size
 - Captures **within-population** (segregating) variation

Background, Additional reading

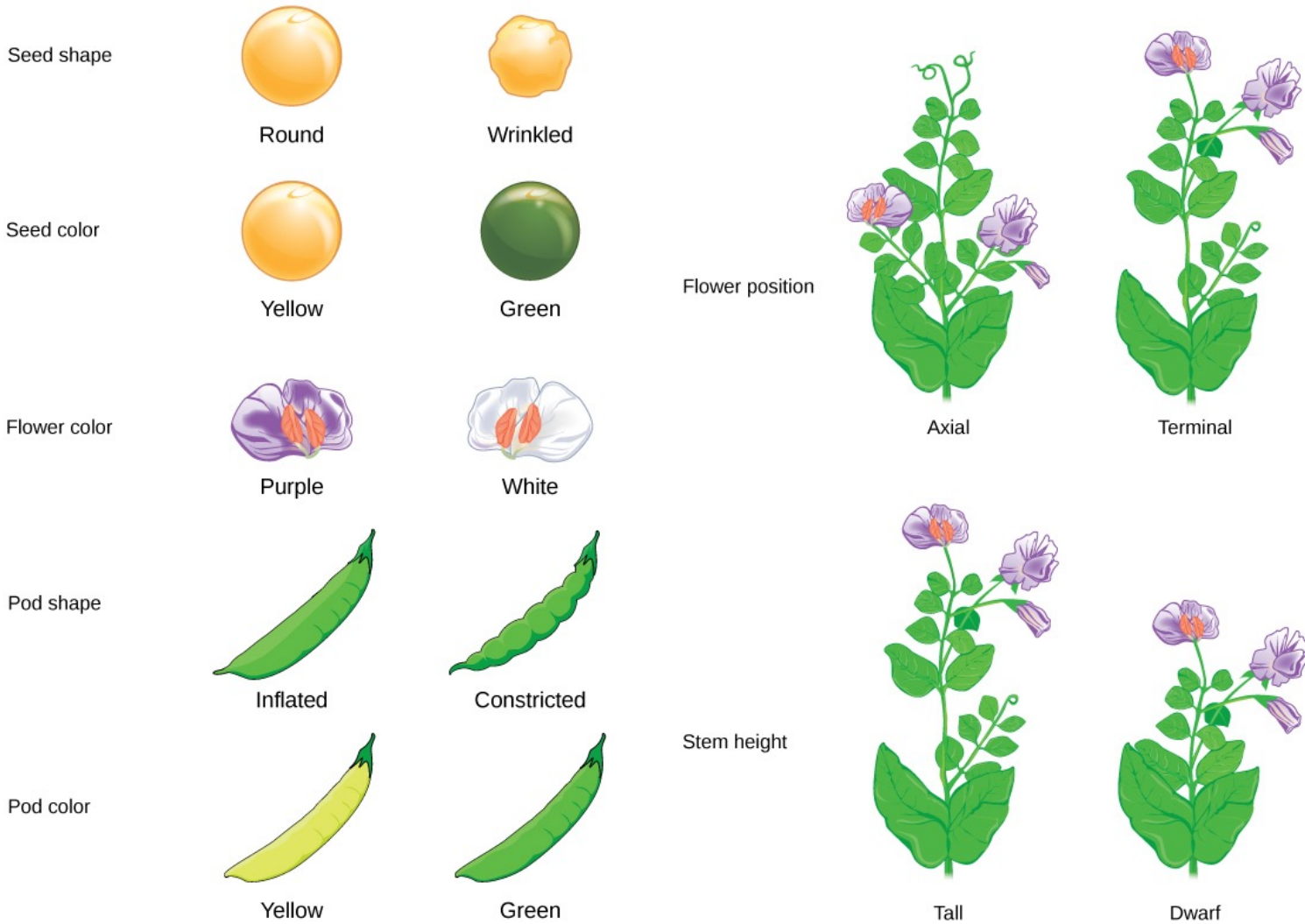
- WVL (Walsh, Visscher, Lynch) 2024.
 - Chapter 5: linkage, LD
 - Chapter 18: QTL mapping
 - Chapter 20: GWAS

Overview

- Genetic Markers (SNPs, STRs, WGS)
- Linkage and linkage disequilibrium (LD)
- Linkage mapping
 - Marker-trait associations
 - Hypothesis testing
 - Examples and Limitations
 - Beavis effects
- Association (LD) mapping (Intro)
 - Marker-trait associations
 - Correcting for population structure

Part I:
Genetic Markers,
Linkage,
Linkage disequilibrium (LD)

Mendel's original seven genes



Molecular Markers

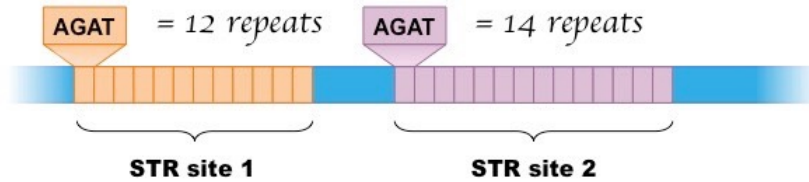
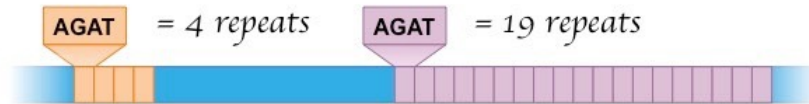
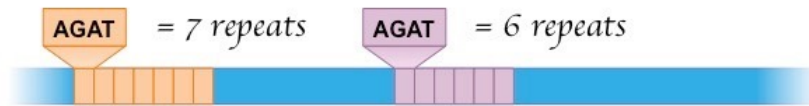
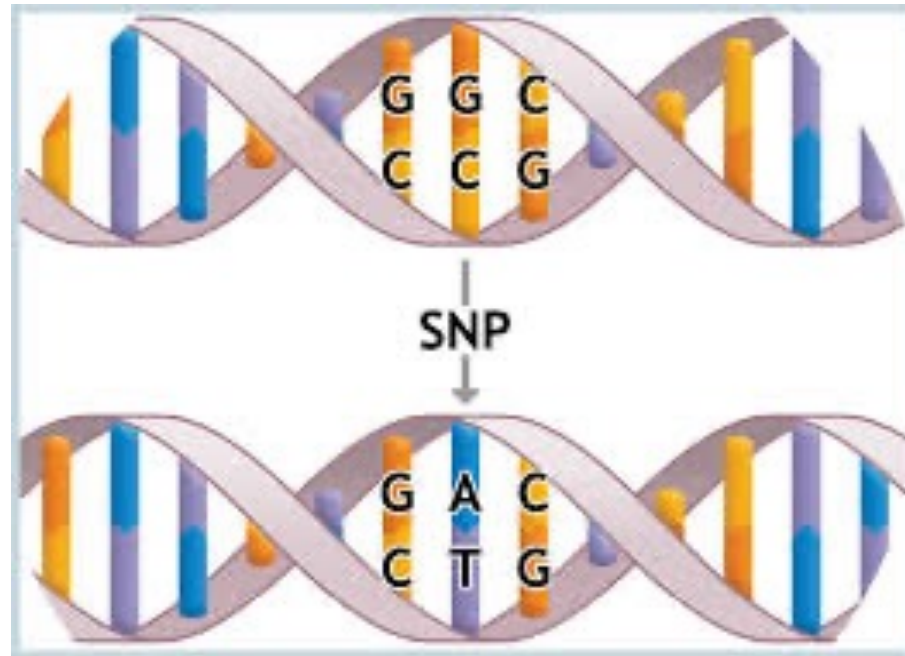
In the molecular era, genetic maps are based not on alleles with large phenotypic effects (i.e., green vs. yellow peas), but rather on molecular markers

SNP -- single nucleotide polymorphism. A particular position on the DNA (say base 123,321 on chromosome 1) that has two different nucleotides (say G or A) segregating

STR -- simple tandem arrays. An STR locus consists of a number of short repeats, with alleles defined by the number of repeats. For example, you might have 6 and 4 copies of the repeat on your two chromosome 7s

Even with whole-genome sequencing (WGS), sites are still classified into these two classes (plus other types)

SNPs



STRs

SNPs

SNPs vs STRs

Cons: Less polymorphic (~ 2 alleles)

Pros: Low mutation rates, alleles very stable

Excellent for looking at historical long-term associations (association mapping)

Cheap to score 100,000s (+) on a single SNP Chip

STRs (= SSR)

Cons: High mutation rate

Pros: Very highly polymorphic (more information/site)

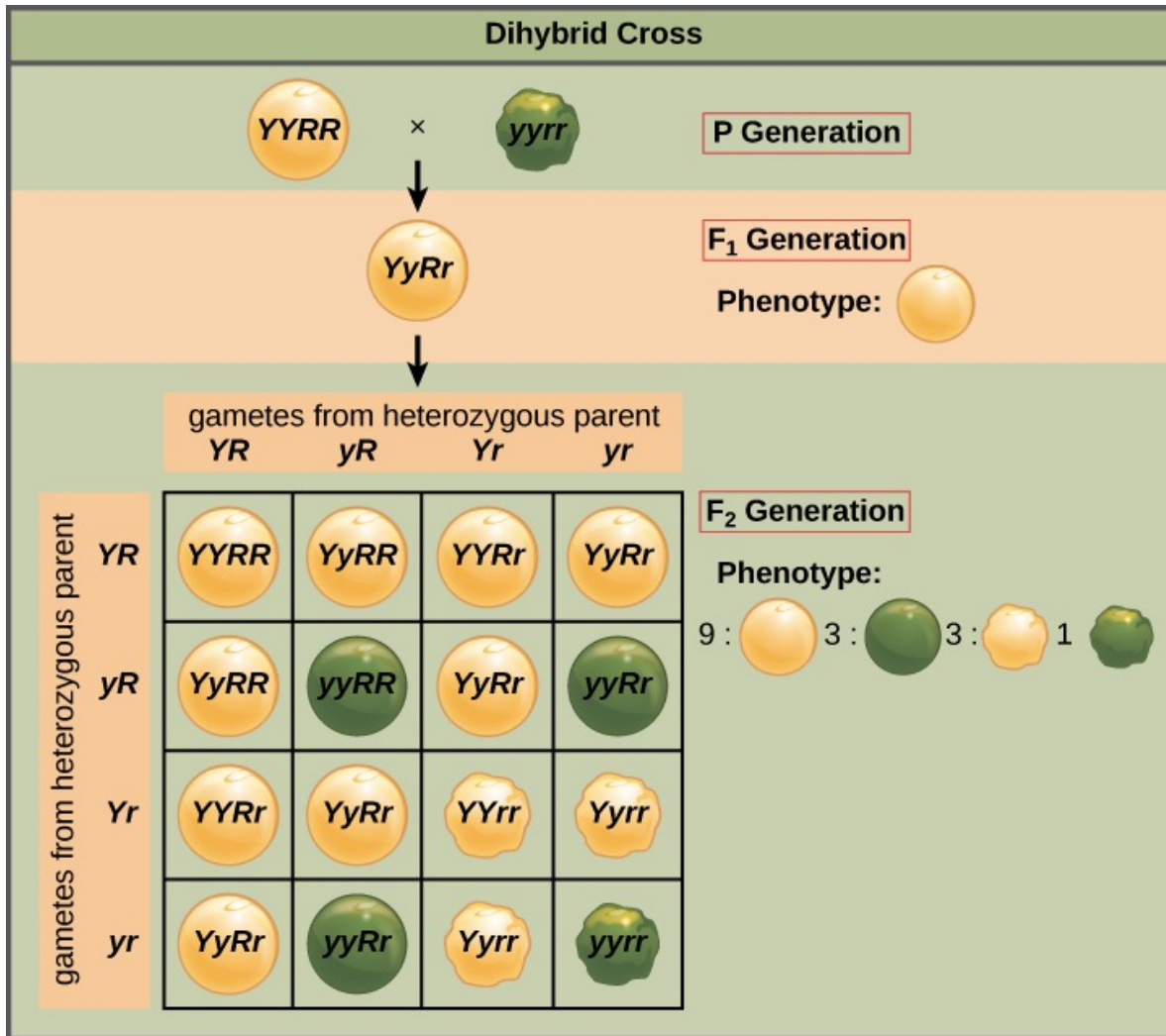
Excellent for linkage studies within an extended pedigree (QTL mapping in families or pedigrees)

Linkage

If genes are located on different chromosomes they (with very few exceptions) show **independent assortment**.

Indeed, peas have only 7 chromosomes, so was Mendel lucky in choosing seven traits at random that happen to all be on different chromosomes?

However, genes on the same chromosome, especially if they are close to each other, tend to be passed onto their offspring in the same configuration as on the parental chromosomes.



Independent
assortment

$$\text{Pr}(YyRr) = \text{Pr}(Yy)\text{Pr}(Rr)$$

Dependent
assortment

$$\text{Pr}(YyRr) = \text{Pr}(Yy|Rr) \text{Pr}(Rr)$$

$$\begin{aligned} \text{Pr}(yyRR) &= \text{Pr}(yR,yR) \quad \text{with linkage, deal with } \underline{\text{gametes}} \\ &= [\text{Pr}(y|R) \text{Pr}(R)] [\text{Pr}(y|R)] \text{Pr}(R) \end{aligned}$$

Mendel was wrong: Linkage

Bateson and Punnett looked at

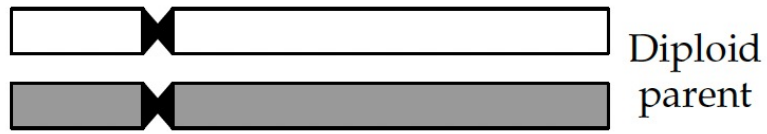
flower color: P (purple) dominant over p (red)

pollen shape: L (long) dominant over l (round)

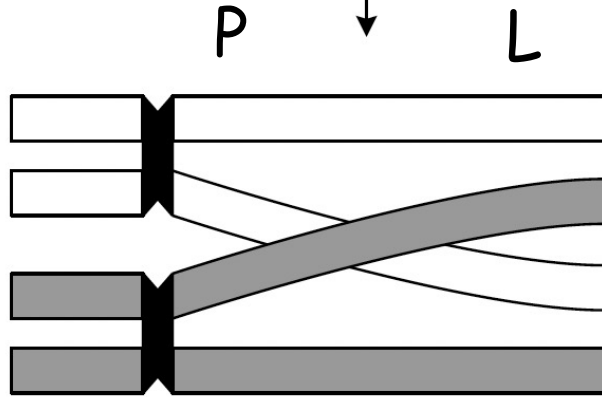
Phenotype	Genotype	Observed	Expected
Purple long	P-L-	284	215
Purple round	P-ll	21	71
Red long	ppL-	21	71
Red round	ppll	55	24

Excess of PL, pl **gametes** over Pl, pL

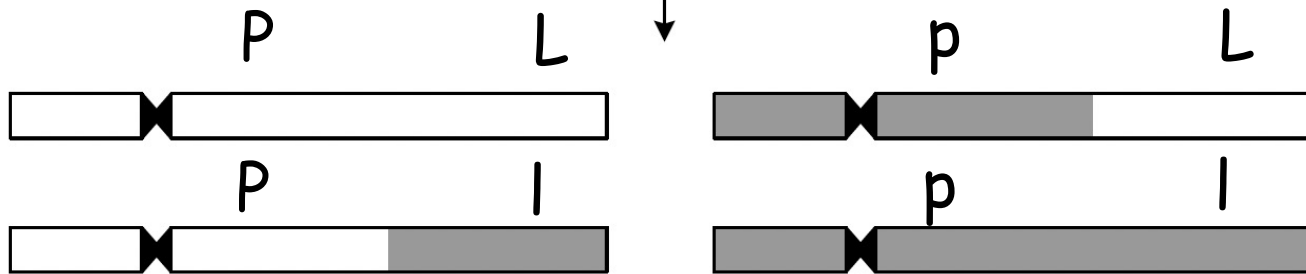
Departure from independent assortment



Chromosome duplication
and
crossing-over



Gamete
production



4 Haploid gametes

Consider the Bateson-Punnett pea data

Let PL / pl denote that in the parent, one chromosome carries the P and L alleles (at the flower color and pollen shape loci, respectively), while the other chromosome carries the p and l alleles.

Unless there is a **recombination** event, one of the two parental chromosome types (PL or pl) are passed onto the offspring. These are called the **parental gametes**.

However, if a recombination event occurs, a PL/pl parent can generate Pl and pL **recombinant chromosomes** to pass onto its offspring.

Let c denote the **recombination frequency** --- the probability that a randomly-chosen gamete from the parent is of the recombinant type (i.e., it is not a parental gamete).

For a PL/pl parent, the gamete frequencies are

Gamete type	Frequency	Expectation under independent assortment
PL	$(1-c)/2$	1/4
pl	$(1-c)/2$	1/4
pL	$c/2$	1/4
Pl	$c/2$	1/4

Parental gametes in excess, as $(1-c)/2 > 1/4$ for $c < 1/2$

Gamete type	Frequency	Expectation under independent assortment
PL	$(1-c)/2$	1/4
pl	$(1-c)/2$	1/4
pL	$c/2$	1/4
Pl	$c/2$	1/4

Recombinant gametes in deficiency, as $c/2 < 1/4$ for $c < 1/2$

Linkage vs. LD

Linkage considers the gametes from a SINGLE Parent

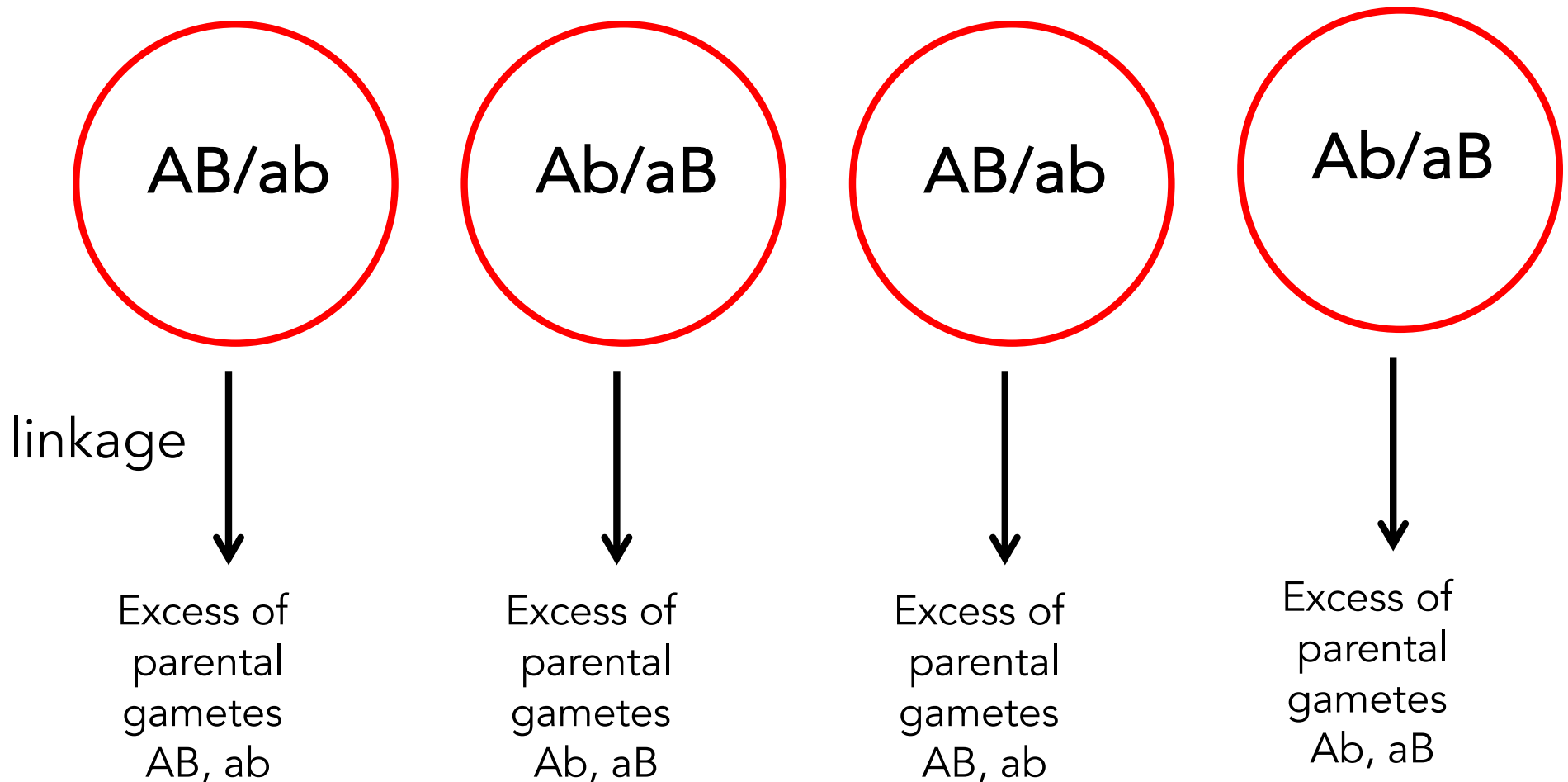
Linkage disequilibrium (LD) concerns a POPULATION SAMPLE of gametes (think chromosomes or haplotypes)

Can have linkage without LD, and LD without linkage

Linkage Disequilibrium

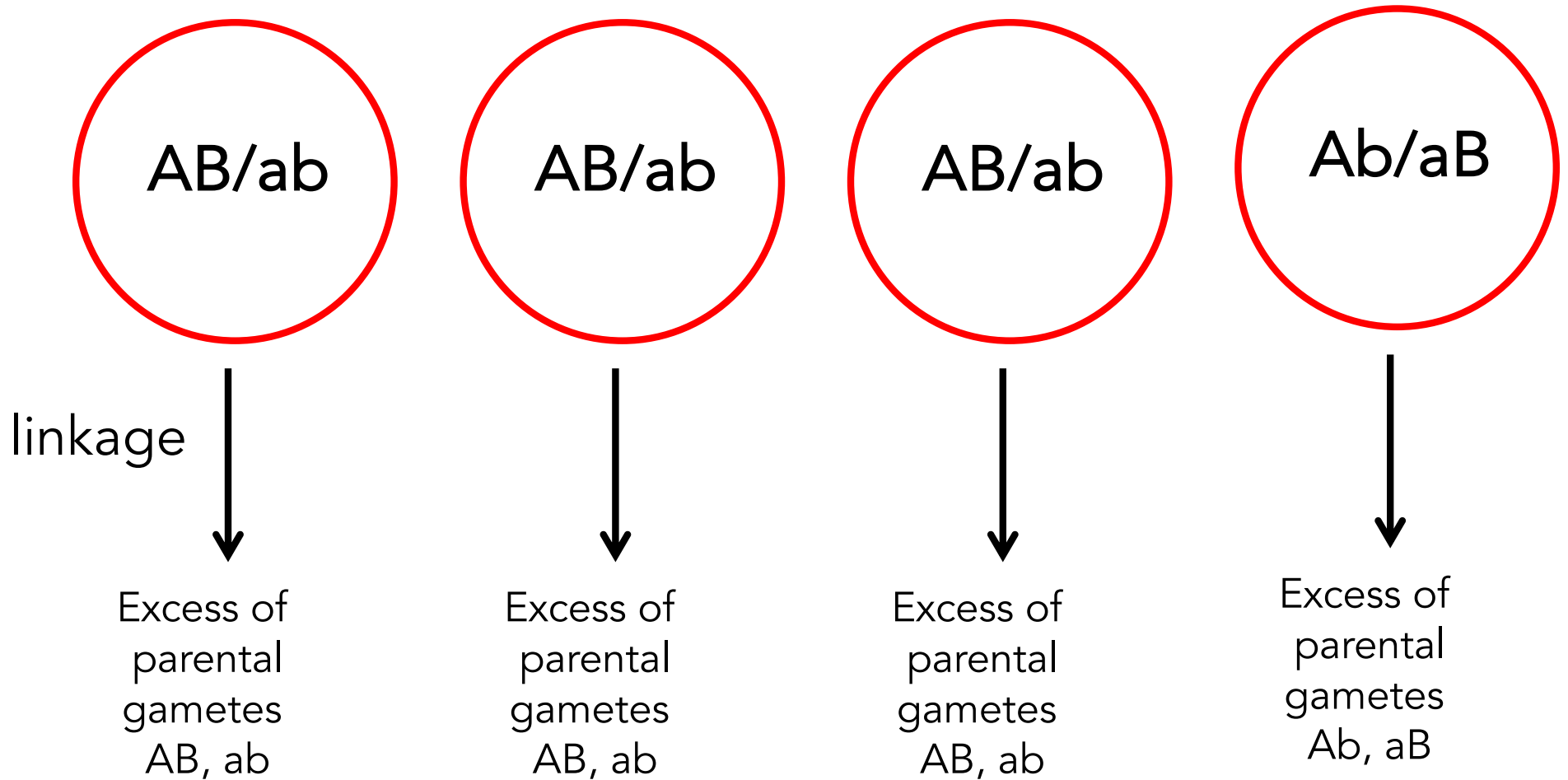
- Under linkage equilibrium, the frequency of gametes is the product of allele frequencies,
 - e.g. $\text{Freq}(AB) = \text{Freq}(A) * \text{Freq}(B)$
 - A and B are **independent** of each other
- If the linkage phase of parents in some set or population departs from random (alleles not independent), linkage disequilibrium (LD) is said to occur
- The amount D_{AB} of disequilibrium for the AB gamete is given by
 - $D_{AB} = \text{Freq}(AB) \text{ gamete} - \text{Freq}(A) * \text{Freq}(B)$
 - $D > 0$ implies AB gamete more frequent than expected
 - $D < 0$ implies AB less frequent than expected

No LD: random distribution of linkage phases



Pool all gametes: AB, ab, Ab, aB equally frequent

With LD, nonrandom distribution of linkage phase



Pool all gametes: Excess of AB, ab due to an excess of AB/ab parents

Dynamics of D

- Under random mating in a large population, allele frequencies do not change. However, gamete frequencies do if there is any LD
- The amount of LD decays by $(1-c)$ each generation
 - $D(t) = (1-c)^t D(0)$
- The expected frequency of a gamete (say AB) is
 - $\text{Freq}(AB) = \text{Freq}(A) * \text{Freq}(B) + D$
 - $\text{Freq}(AB \text{ in gen } t) = \text{Freq}(A) * \text{Freq}(B) + (1-c)^t D(0)$

Part II:
QTL mapping and the use of
inbred line crosses

- QTL mapping tries to detect small (20-40 cM) chromosome segments influencing trait variation
 - Relatively crude level of resolution
- QTL mapping performed either using inbred line crosses or sets of known relatives (pedigrees)
 - Uses the simple fact of an excess of parental gametes

Key idea: Looking for marker-trait associations in collections of relatives

If (say) the mean trait value for marker genotype MM is statistically different from that for genotype mm, then the M/m marker is linked to a QTL

Sax (1923) spotted peas and weight

One can use a random collection of such markers spanning a genome (a genomic scan) to search for QTLs

Inbred lines

$$\begin{array}{c} \frac{M Q}{M Q} \\ \times \\ \frac{m q}{m q} \end{array}$$



$$\begin{array}{c} F_1 \\ \frac{M Q}{m q} \end{array}$$



gametes

freq

M Q

$(1-c)/2$

m q

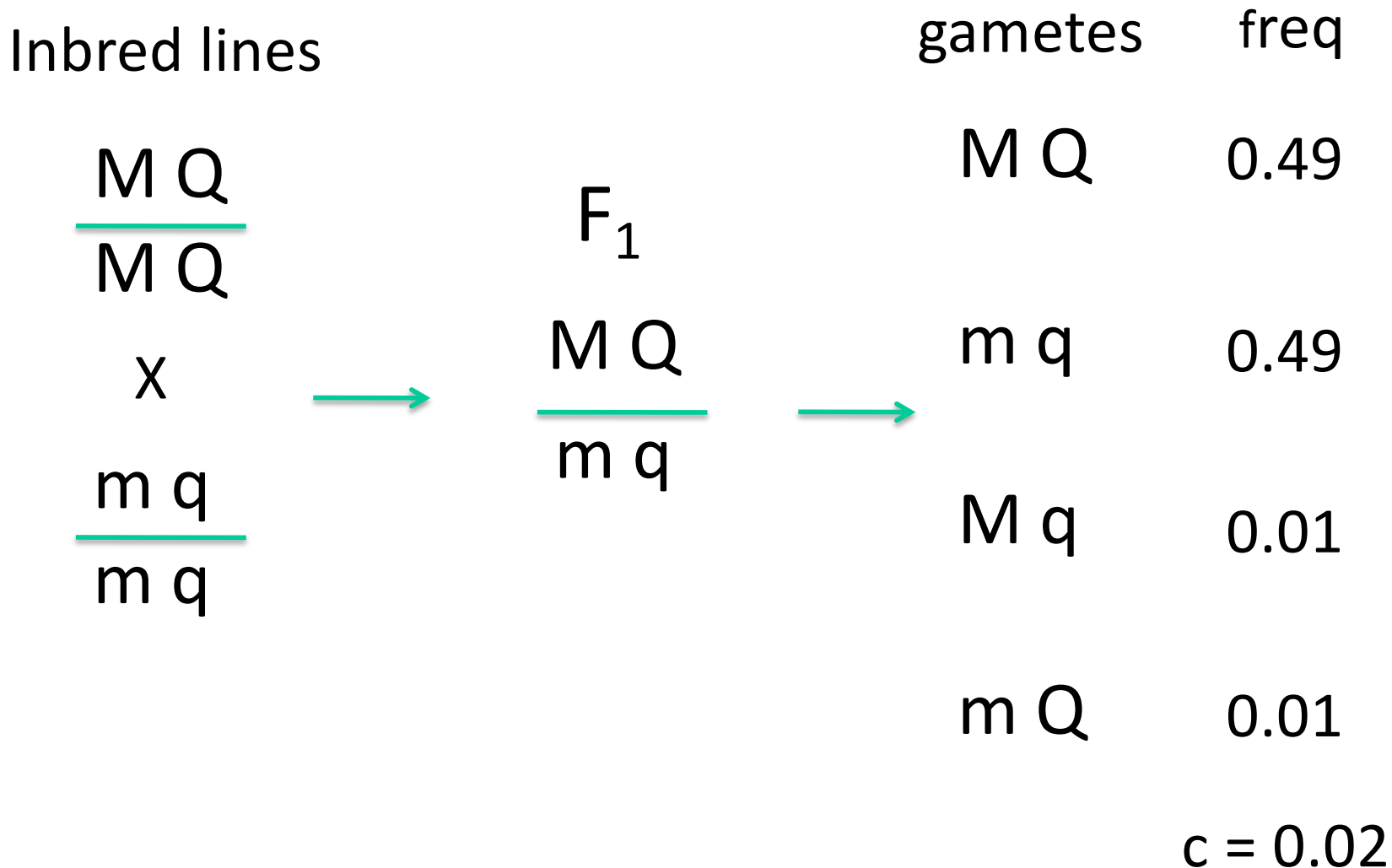
$(1-c)/2$

M q

$c/2$

m Q

$c/2$



Creates a marker-trait association in offspring, with M-bearing chromosomes co-segregating with Q, so that M-bearing gametes will (on average) yield larger trait values (here 98% of M are Q)

Conditional Probabilities of QTL Genotypes

The basic building block for all QTL methods is $\Pr(Q_k | M_j)$ --- the probability of QTL genotype Q_k given the marker genotype is M_j .

$$\Pr(Q_k | M_j) = \frac{\Pr(Q_k M_j)}{\Pr(M_j)}$$

Consider a QTL linked to a marker (recombination Fraction = c). Cross $MMQQ \times mmqq$. In the F_1 , all gametes are MQ and mq

In the F_2 , $\text{freq}(MQ) = \text{freq}(mq) = (1-c)/2$,
 $\text{freq}(mQ) = \text{freq}(Mq) = c/2$

Hence, $\Pr(\text{MMQQ}) = \Pr(\text{MQ})\Pr(\text{MQ}) = (1-c)^2/4$

$$\Pr(\text{MMQq}) = 2\Pr(\text{MQ})\Pr(\text{Mq}) = 2c(1-c)/4$$

$$\Pr(\text{MMqq}) = \Pr(\text{Mq})\Pr(\text{Mq}) = c^2/4$$

Why the 2? MQ from father, Mq from mother, OR MQ from mother, Mq from father

Since $\Pr(\text{MM}) = 1/4$, the conditional probabilities become

$$\Pr(\text{QQ} \mid \text{MM}) = \Pr(\text{MMQQ})/\Pr(\text{MM}) = (1-c)^2$$

$$\Pr(\text{Qq} \mid \text{MM}) = \Pr(\text{MMQq})/\Pr(\text{MM}) = 2c(1-c)$$

$$\Pr(\text{qq} \mid \text{MM}) = \Pr(\text{MMqq})/\Pr(\text{MM}) = c^2$$

How do we use these?

Expected Marker Means

The expected trait mean for marker genotype M_j is just

$$\mu_{M_j} = \sum_{k=1}^N \mu_{Q_k} \Pr(Q_k | M_j)$$

For example, if $QQ = 2a$, $Qq = a(1+k)$, $qq = 0$, then in the F2 of an $MMQQ/mmqq$ cross,

$$(\mu_{MM} - \mu_{mm})/2 = a(1 - 2c)$$

- If the trait mean is significantly different for the genotypes at a marker locus, it is linked to a QTL
- A small MM - mm difference could be (i) a tightly-linked QTL of small effect or (ii) loose linkage to a large QTL

Linear Models for QTL Detection

The use of differences in the mean trait value for different marker genotypes to detect a QTL and estimate its effects is a use of **linear models**.

One-way ANOVA.

Value of trait in kth
individual of marker
genotype type i



$$z_{ik} = \mu + b_i + e_{ik}$$



Effect of marker
genotype i on trait
value

$$z_{ik} = \mu + b_i + e_{ik}$$

Detection: a QTL is linked to the marker if at least one of the b_i is significantly different from zero


Estimation: (QTL effect and position): This requires relating the b_i to the QTL effects and map position

Detecting epistasis


One major advantage of linear models is their flexibility. To test for epistasis between two QTLs, use ANOVA with an interaction term

$$z = \mu + a_i + b_k + d_{ik} + e$$

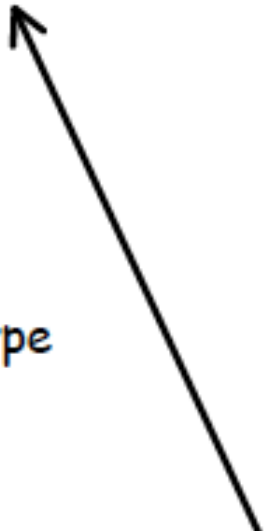
Effect from marker genotype
at first marker set (can be > 1 loci)



Effect from marker genotype
at second marker set



Interaction between marker genotypes i in 1st
marker set and k in 2nd marker set



Detecting epistasis

$$z = \mu + a_i + b_k + d_{ik} + e$$

- At least one of the a_i significantly different from 0
---- QTL linked to first marker set
- At least one of the b_k significantly different from 0
---- QTL linked to second marker set
- At least one of the d_{ik} significantly different from 0
---- interactions between QTL in sets 1 and two

Problem: Huge number of potential interaction terms (order m^2 , where m = number of markers)

Detecting QTLs for Dichotomous Traits: The Cochran-Armitage Trend Test

Many QTL experiments are concerned with dichotomous (binary) traits, such as disease or pest resistance in crop plants or disease susceptibility in humans. In many cases, one can score quantitative physiological traits contributing to the binary trait, such as blood pressure or number of lesions per leaf, and the above methodology for QTL detection with continuous traits applies. However, such underlying variables are often either unknown or unmeasured, and the data are simply scored as presence/absence values (or **cases** versus **controls** in the medical literature). The simplest procedure to detect marker-trait associations in this setting is to test for independence using standard association tables (such as χ^2 or Fisher's exact tests). As shown in the table below, the n total observations are partitioned into counts for each particular class, e.g., n_{P1} is the sample number of Mm individuals showing the trait.

	Marker Genotype			Totals
	mm	Mm	MM	
Present	n_{P0}	n_{P1}	n_{P2}	n_P
Absent	n_{A0}	n_{A1}	n_{A2}	n_A
Totals	n_0	n_1	n_2	n

This same approach easily extended to polychotomous (ordinal) characters. With three marker genotypes and two trait values, the result χ^2 test has two degrees of freedom, with a significant value indicating linkage to one (or more) QTLs.

Maximum Likelihood Methods

ML methods use the entire distribution of the data, not just the marker genotype means.

More powerful than linear models, but not as flexible in extending solutions (new analysis required for each model)

Basic likelihood function:

Trait value given
marker genotype is
type j

$$\ell(z | M_j) = \sum_{k=1}^N \varphi(z, \mu_{Q_k}, \sigma^2) \Pr(Q_k | M_j)$$

This is a **mixture model**

Maximum Likelihood Methods

Sum over the N possible linked QTL genotypes

Probability of QTL genotype k given marker genotype j --- genetic map and linkage phase enter : here

$$\ell(z | M_j) = \sum_{k=1}^N \varphi(z, \mu_{Q_k}, \sigma^2) \Pr(Q_k | M_j)$$

Distribution of trait value given QTL genotype is k is normal with mean μ_{Q_k} . (QTL effects enter here)

ML methods combine both detection and estimation of QTL effects/position.

Test for a linked QTL given from by the **Likelihood Ratio** (or **LR**) **test**

$$LR = -2 \ln \frac{\max l_r(\mathbf{z})}{\max l(\mathbf{z})}$$

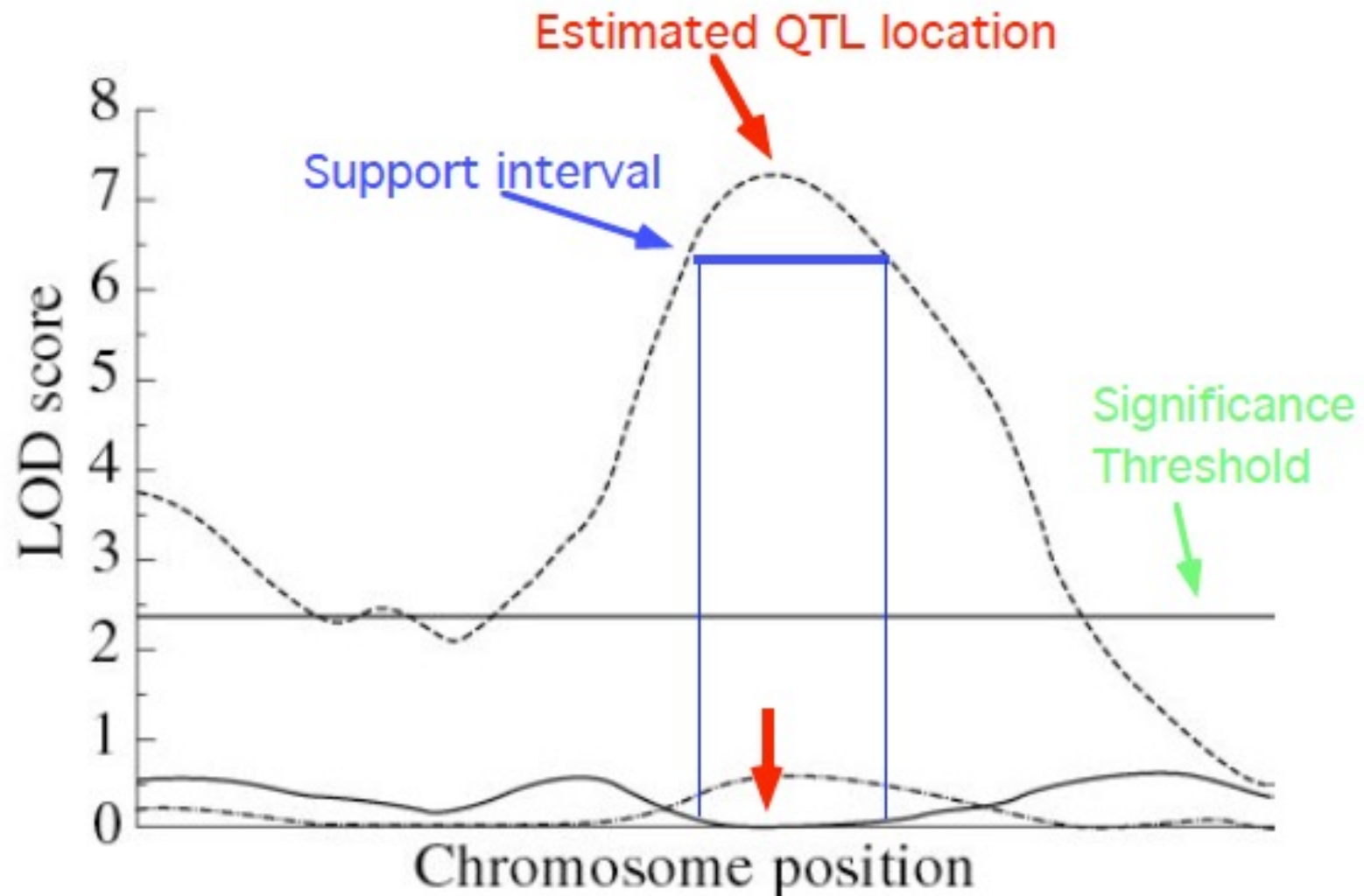
Maximum of the likelihood
under a no-linked QTL
model

Maximum of the
full likelihood

The LR score is often plotted by trying different locations for the QTL (i.e., values of c) and computing a LOD score for each

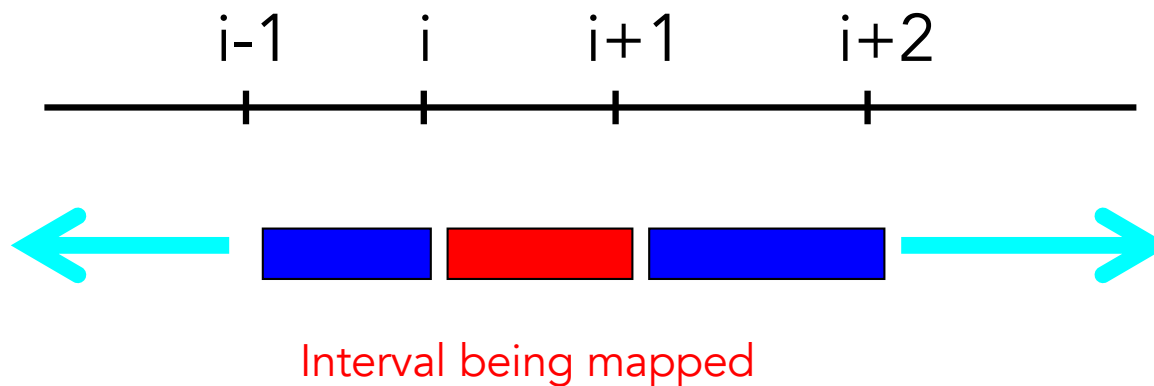
$$LOD(c) = -\log_{10} \left[\frac{\max l_r(\mathbf{z})}{\max l(\mathbf{z}, c)} \right] = \frac{LR(c)}{2 \ln 10} \approx \frac{LR(c)}{4.61}$$

A typical QTL map from a likelihood analysis

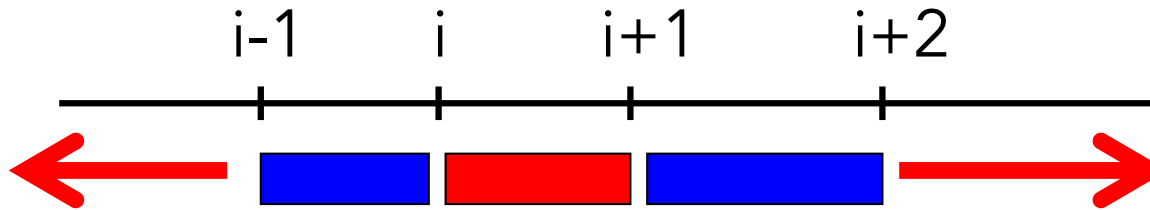


Interval Mapping with Marker Cofactors

Consider interval mapping using the markers i and $i+1$. QTLs linked to these markers, but outside this interval, can contribute (falsely) to estimation of QTL position and effect



Now suppose we also add the two markers flanking the interval ($i-1$ and $i+2$)



Inclusion of markers $i-1$ and $i+2$ fully account for any linked QTLs to the left of $i-1$ and the right of $i+2$

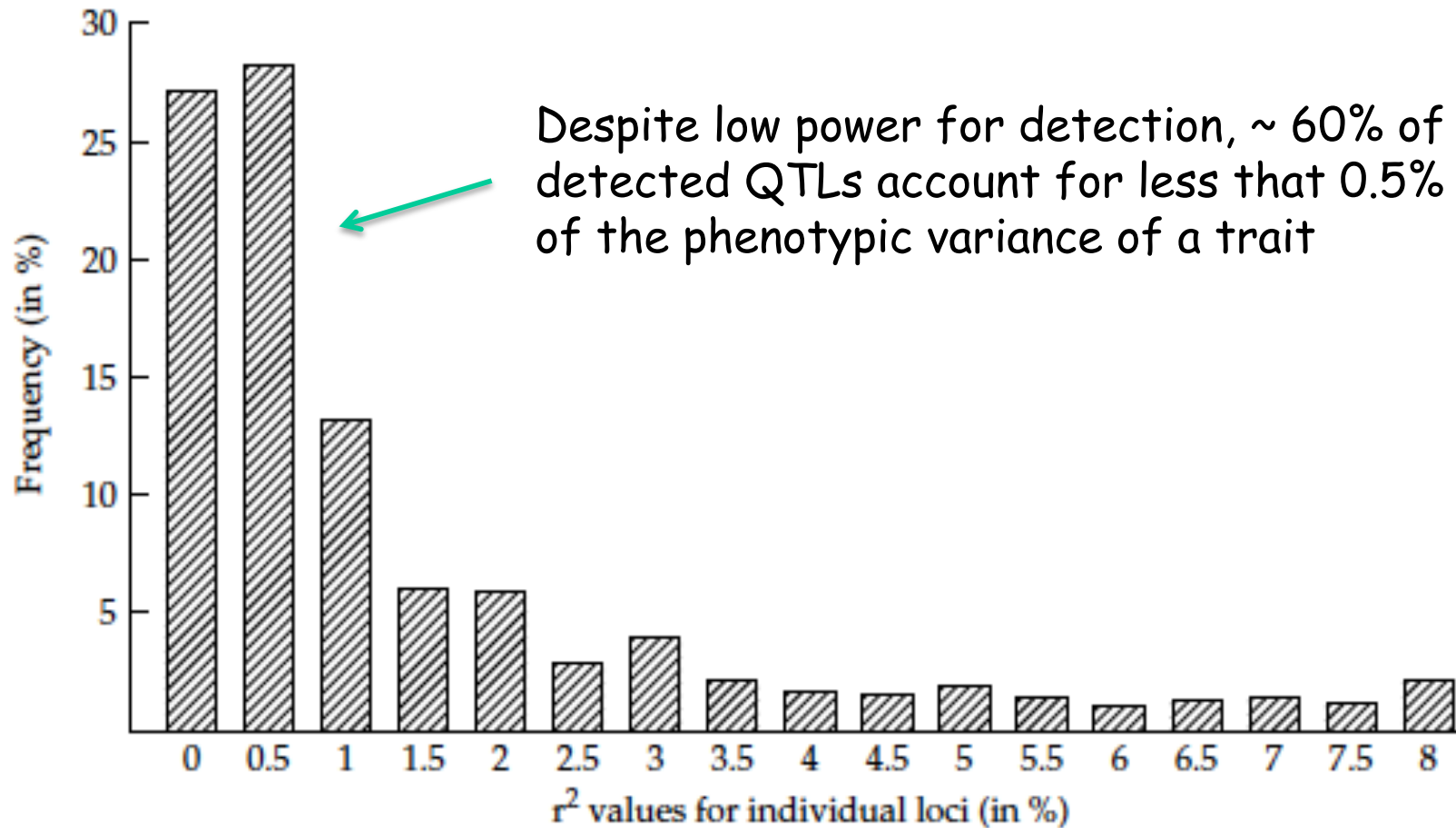
Interval mapping + marker cofactors is called **Composite Interval Mapping (CIM)**

CIM works by adding an additional term to the linear model,

$$\sum_{k \neq i, i+1} b_k x_{kj}$$

CIM also (potentially) includes unlinked markers to account for QTL on other chromosomes.

Some early studies suggested an infinitesimal-like genetic architecture, with the majority of genes having small effects



Edwards et al (1987) vegetative traits in maize

M. lewisii



M. cardinalis



Table 18.4 Number of detected QTLs influencing pollination characters involved in reproductive isolation between *Mimulus cardinalis* and *Mimulus lewisii* and their estimated individual effects (measured by % of variance explained). Due to sampling error, the sum of individual r^2 values exceeds 100% in a few cases. (After Bradshaw et al. 1995.)

	Number of QTLs	% Phenotypic Variance ($r^2 \times 100$)
Pollinator attraction characters		
Petal anthocyanins	2	33.5, 21.5
Petal carotenoids	1	88.3
Corolla width	3	68.7, 33.0, 25.7
Petal width	3	42.4, 41.2, 25.2
Pollinator reward		
Nectar volume	2	53.1, 48.9
Nectar concentration	2	28.5, 23.9
Pollination efficiency		
Stamen length	4	27.7, 27.5, 21.3, 18.7
Pistil length	2	51.9, 43.9

1990's: The age of semi-major genes

- By the early-mid 1990's, extensive QTL studies suggested that genes of major effect are common and underlie many of the fixed differences between crossed lines
- Roughly exponential ("L-shaped") distribution of effects,
 - many genes of small effect
 - a few genes of large effect
 - Usually detected in line-cross populations (hence $MAF = 1/2$)

Power and Precision

While modest sample sizes are sufficient to **detect** a QTL of modest effect (power), large sample sizes are required to **map it** with any precision

With 200-300 F_2 , a QTL accounting for 5% of total variation can be mapped to a 40cM interval

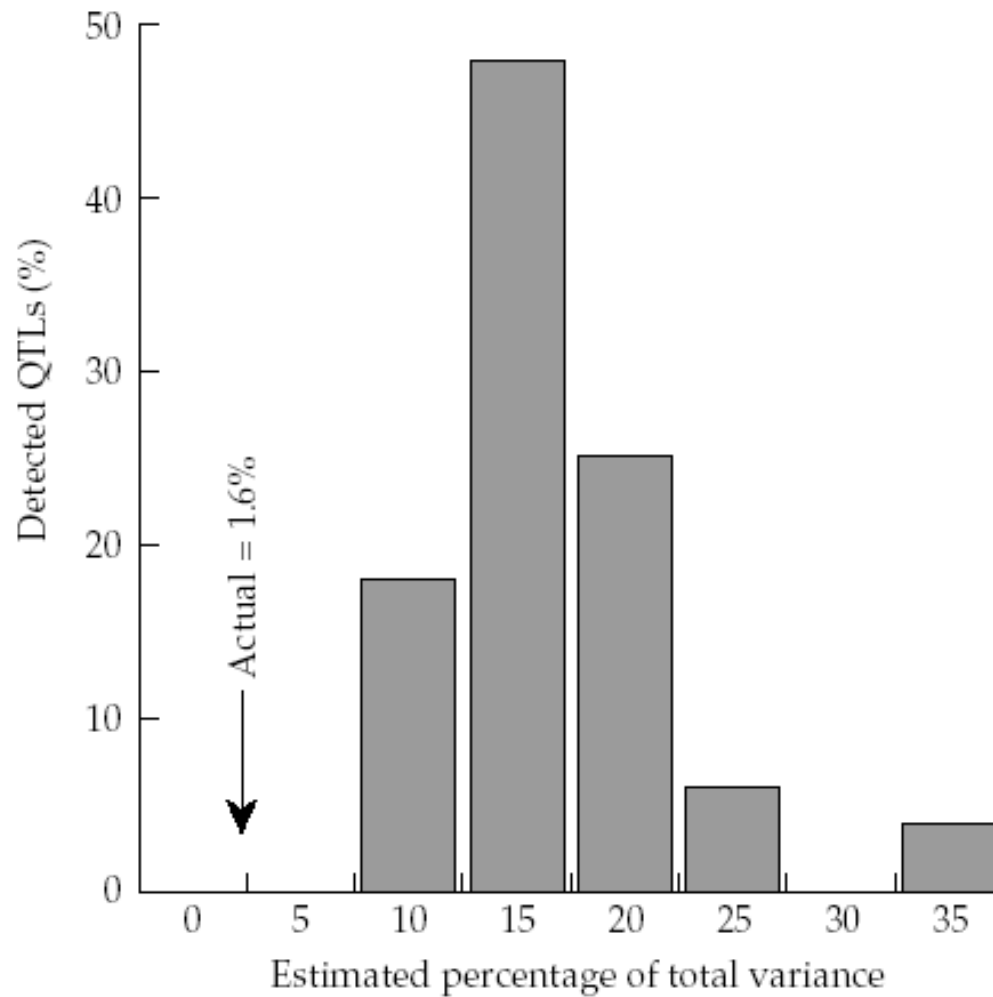
Over 10,000 F_2 individuals are required to map this QTL to a 1cM interval

Power and Repeatability: The Beavis Effect

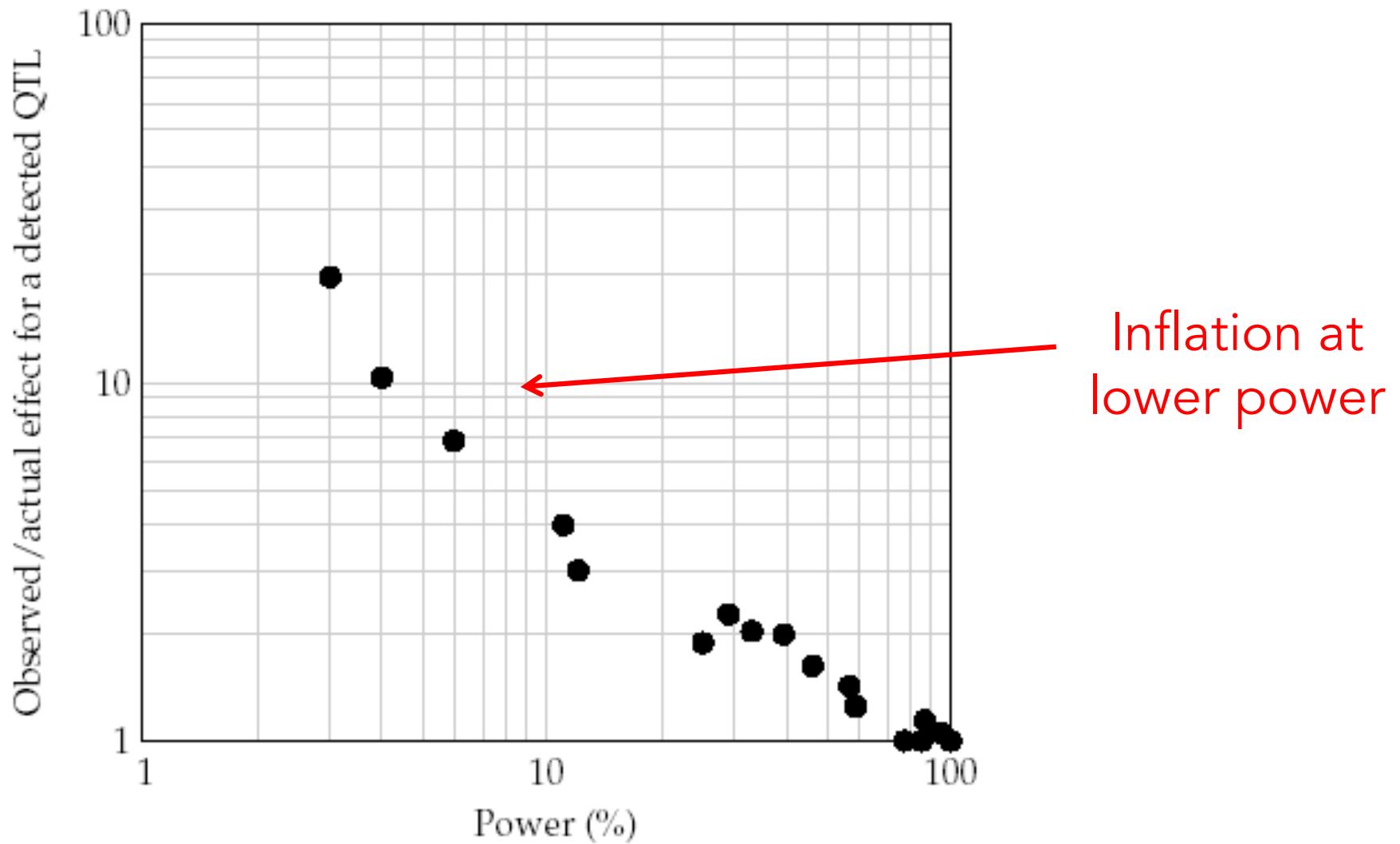
QTLs with low power of detection tend to have their effects overestimated, often very dramatically

As power of detection increases, the overestimation of detected QTLs becomes far less serious

This is often called the **Beavis Effect**, after Bill Beavis who first noticed this in simulation studies. This phenomena is also called the **winner's curse** in statistics (and GWAS)



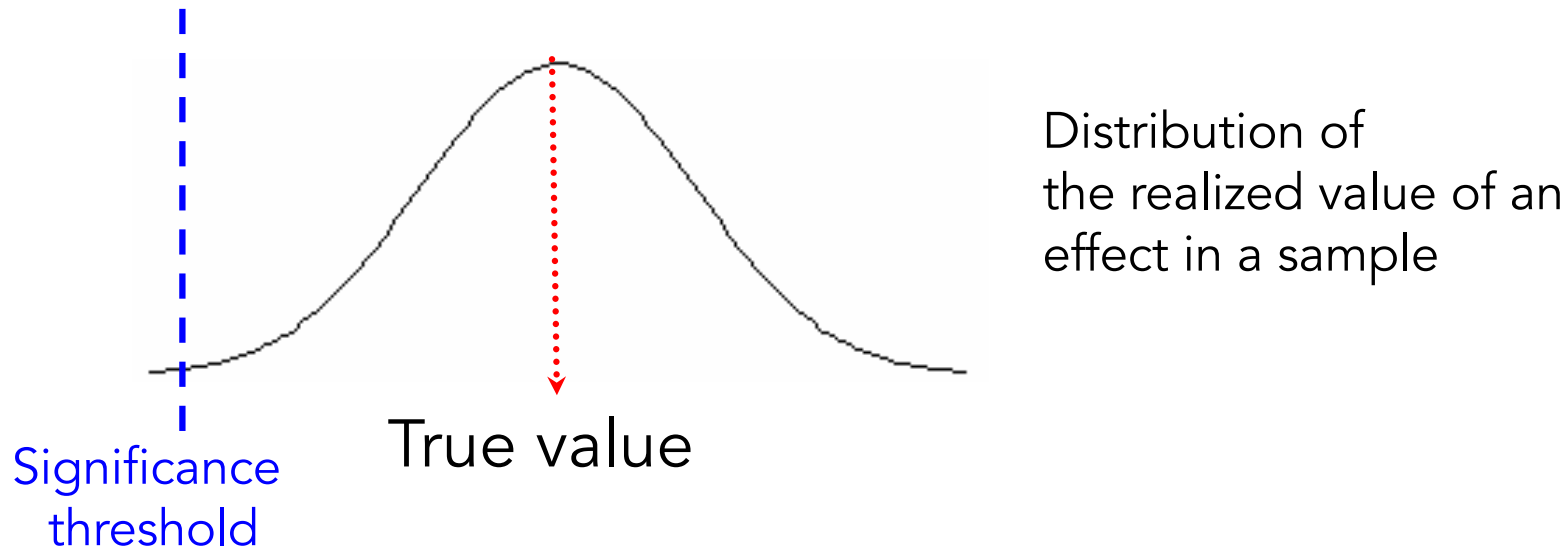
Beavis simulation: actual effect size is 1.6% of variation. Estimated effects (at significant markers) much higher



Inflation can be significant, esp. with low power

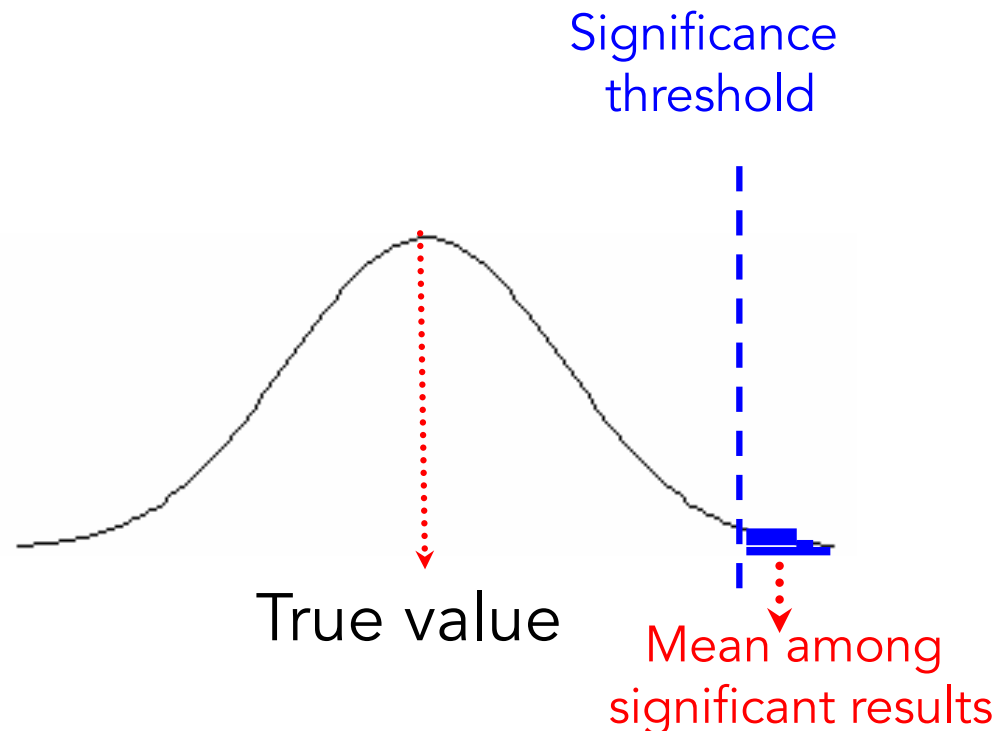
Beavis Effect

Also called the “winner’s curse” in the GWAS literature



High power setting: Most realizations are to the right of the significance threshold. Hence, the average value given the estimate is declared significant (above the threshold) is very close to the true value.

In **low power settings**, most realizations are below the significance threshold, hence most of the time the effect is scored as being nonsignificant



However, the mean of those **declared significant** is much larger than the true mean

M. lewisii



M. cardinalis



Table 18.4 Number of detected QTLs influencing pollination characters involved in reproductive isolation between *Mimulus cardinalis* and *Mimulus lewisii* and their estimated individual effects (measured by % of variance explained). Due to sampling error, the sum of individual r^2 values exceeds 100% in a few cases. (After Bradshaw et al. 1995.)

	Number of QTLs	% Phenotypic Variance ($r^2 \times 100$)
Pollinator attraction characters		
Petal anthocyanins	2	33.5, 21.5
Petal carotenoids	1	88.3
Corolla width	3	68.7, 33.0, 25.7
Petal width	3	42.4, 41.2, 25.2
Pollinator reward		
Nectar volume	2	53.1, 48.9
Nectar concentration	2	28.5, 23.9
Pollination efficiency		
Stamen length	4	27.7, 27.5, 21.3, 18.7
Pistil length	2	51.9, 43.9

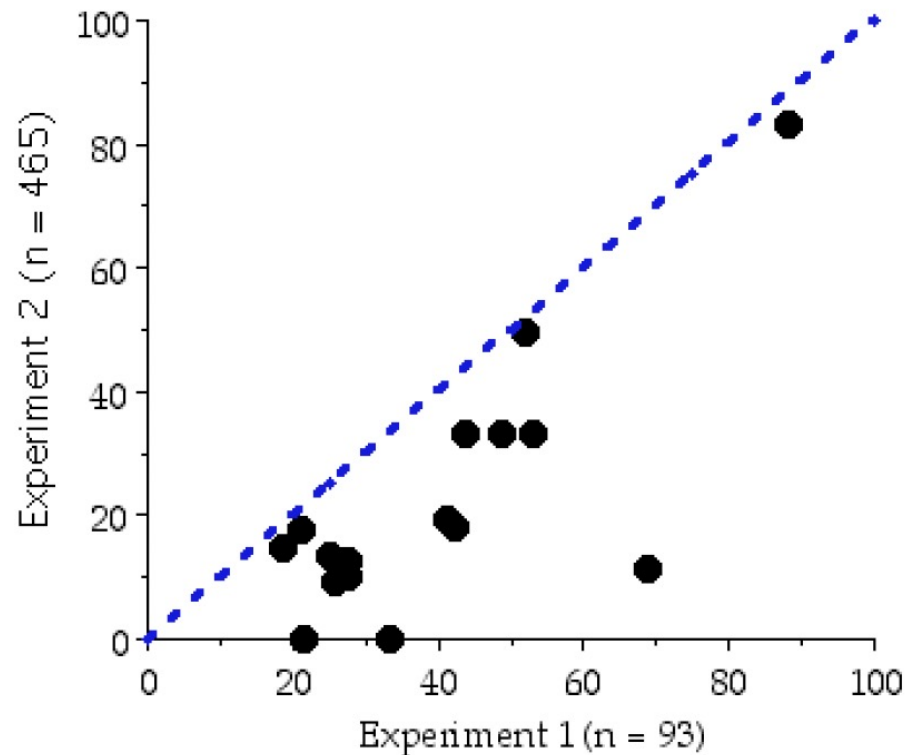


Figure 18.14 Relationship between the effects of detected QTLs for *Mimulus* pollination traits, expressed as percent of F_2 phenotypic variation (r^2), over two crosses with different sample sizes (Bradshaw et al. 1995, 1998). Experiment one measured 96 F_2 plants while experiment two measured 465. Note that all detected QTLs had larger estimated values in the smaller experiment, a clear example of the Beavis effect. The two values of zero from experiment two correspond to the petal anthocyanin QTLs that were not replicated.

What is a "QTL"

- A detected "QTL" in a mapping experiment is a region of a chromosome detected by linkage.
- Usually large (typically 10-40 cM)
- When further examined, most "large" QTLs turn out to be a linked collection of locations with increasingly smaller effects
- The more one localizes, the more subregions that are found, and the smaller the effect in each subregion
- This is called **fractionation**

Limitations of QTL mapping

- **Poor resolution** (~20 cM or greater in most designs with sample sizes in low to mid 100's)
 - Detected “QTLs” are thus large chromosomal regions
- Fine mapping requires either
 - Further crosses (recombinations) involving regions of interest (i.e., RILs, NILs)
 - Enormous sample sizes
 - If marker-QTL distance is 0.5cM, require sample sizes in excess of 3400 to have a 95% chance of 10 (or more) recombination events in sample
 - 10 recombination events allows one to separate effects that differ by ~ 0.6 SD

Limitations of QTL mapping (cont)

- “Major” QTLs typically **fractionate**
 - QTLs of large effect (accounting for $> 10\%$ of the variance) are routinely discovered.
 - However, a large QTL peak in an initial experiment generally becomes a series of smaller and smaller peaks upon subsequent fine-mapping.
- The Beavis effect:
 - When power for detection is low, marker-trait associations declared to be statistically significant **significantly overestimate** their true effects.
 - This effect can be very large (order of magnitude) when power is low.

QTL mapping in outbred populations

- Much lower power than line-cross QTL mapping
- Each parent must be separately analyzed (linkage phase can vary over parents)
- We focus on an approach for general pedigrees, as this leads us into association mapping

General Pedigree Methods

Random effects (hence, variance component) method for detecting QTLs in general pedigrees

Trait value for individual i → $z_i = \mu + A_i + A'_i + e_i$

Genetic effect of chromosomal region of interest

Genetic value of other (background) QTLs

The diagram shows the equation $z_i = \mu + A_i + A'_i + e_i$ with arrows pointing to each term. An arrow from the text 'Trait value for individual i' points to z_i . An arrow from the red text 'Genetic effect of chromosomal region of interest' points to A_i . An arrow from the purple text 'Genetic value of other (background) QTLs' points to A'_i .

The model is rerun for each marker

$$z_i = \mu + A_i + A'_i + e_i$$

The covariance between individuals i and j is thus

Variance explained by the **region** of interest

Resemblance between relatives correction

$$\sigma(z_i, z_j) = R_{ij} \sigma_A^2 + 2\Theta_{ij} \sigma_{A'}^2$$

Fraction of chromosomal region shared IBD between individuals i and j .

Variance explained by the **background** polygenes

Assume \mathbf{z} is MVN, giving the covariance matrix as

$$\mathbf{V} = \mathbf{R} \sigma_A^2 + \mathbf{A} \sigma_{A'}^2 + \mathbf{I} \sigma_e^2$$

Here

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \hat{R}_{ij} & \text{for } i \neq j \end{cases}, \quad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases}$$

Estimated from marker
data

Estimated from
the pedigree

The resulting likelihood function is

$$\ell(\mathbf{z} | \mu, \sigma_A^2, \sigma_{A'}^2, \sigma_e^2) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp \left[-\frac{1}{2} (\mathbf{z} - \mu)^T \mathbf{V}^{-1} (\mathbf{z} - \mu) \right]$$

A significant σ_A^2 indicates a linked QTL.