# QTL: Interval Mapping

Jinliang Yang
Nov. 30th, 2018

# Interval Mapping

Interval mapping is able to more **precisely locate the putative QTL within a marker interval**.

Interval mapping was introduced to remedy the deficiencies of single marker analysis (Lander and Botstein, 1989)

- By taking advantage of information from **flanking markers**,
- It can achieve the **same power with fewer progenies** compared to single marker analysis (Haley and Knott, 1992)

# LOD Score

Interval mapping calculates **the likelihood of a QTL** being at each cM position between two markers

# LOD Score

Interval mapping calculates **the likelihood of a QTL** being at each cM position between two markers

- The position with the highest likelihood is **the maximum likelihood estimate** of the QTL position
- A **likelihood ratio statistic** is calculated and is a function of the likelihood of the data arose from linked QTL divided by the likelihood of the data arose under the absence of a QTL.

$$\frac{L(QTL)}{L(noQTL)}$$

# LOD Score

Interval mapping calculates **the likelihood of a QTL** being at each cM position between two markers

- The position with the highest likelihood is **the maximum likelihood estimate** of the QTL position
- A **likelihood ratio statistic** is calculated and is a function of the likelihood of the data arose from linked QTL divided by the likelihood of the data arose under the absence of a QTL.

$$\frac{L(QTL)}{L(noQTL)}$$

- This is converted to an **LOD (logarithm of the odds)** score.

$$log_{10}(\frac{L(QTL)}{L(noQTL)})$$

# Basics for interval mapping

## Conditional probabilities of QTL genotypes

The foundation of QTL mapping is the conditional probability that the QTL allele is $Q_k$, given the observed marker allele is $M_j$.

From the definition of conditional probability, this is expressed as

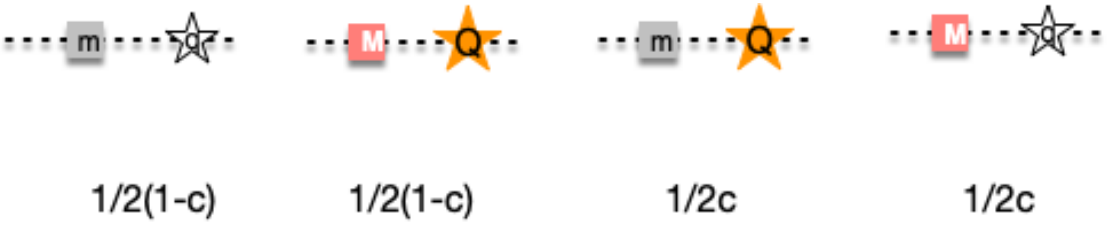
$$P(Q_k|M_j) = \frac{P(Q_k M_j)}{P(M_j)}$$

- where $P(Q_k M_j)$ is the probability of obtaining **the $Q_k M_j$ haplotype**.
  - This is calculated just as before using recombination frequencies between markers and QTL.
  - This is also a function of the experimental design, i.e., BC1 or F2.

# Pearl millet example (F2 population)


1/2(1-c) 1/2(1-c) 1/2c 1/2c

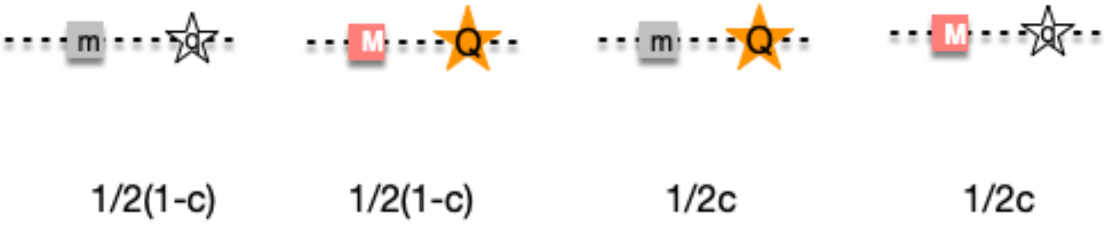
# Pearl millet example (F2 population)




What is the probability of a **Qq QTL genotype** given a **MM marker genotype**?

$$P(Qq|MM) = \frac{P(QqMM)}{P(MM)}$$

# Pearl millet example (F2 population)

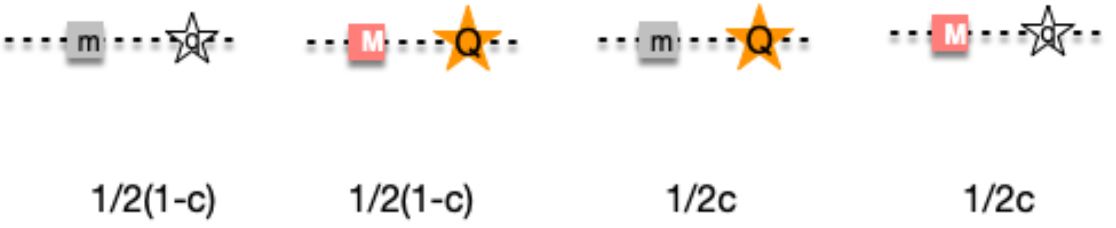



What is the probability of a **Qq QTL genotype** given a **MM marker genotype**?

$$P(Qq|MM) = \frac{P(QqMM)}{P(MM)}$$

| Genotype | Value | Frequency |
|---|---|---|
| $MMQQ$ | a | $\frac{1}{4}(1-c)^2$ |
| $MMQq$ | **d** | $\frac{1}{2}c(1-c)$ |
| $MMqq$ | -a | $\frac{1}{4}c^2$ |
| $MmQQ$ | a | $\frac{1}{2}c(1-c)$ |
| ... | | |

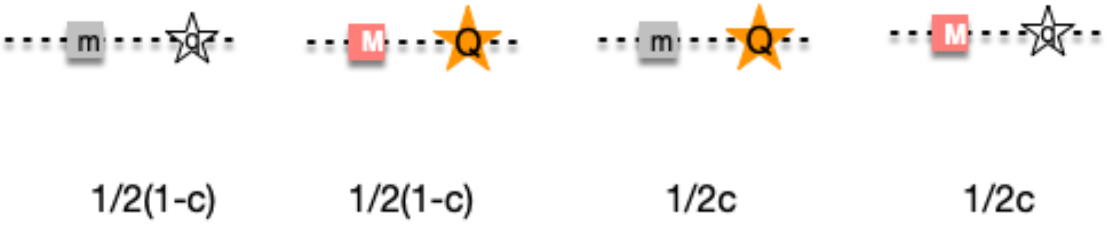
# Conditional probabilities of QTL genotypes




## An example:

Consider a QTL and marker separated by $c = 0.20$ **in an F2 population**.

# Conditional probabilities of QTL genotypes

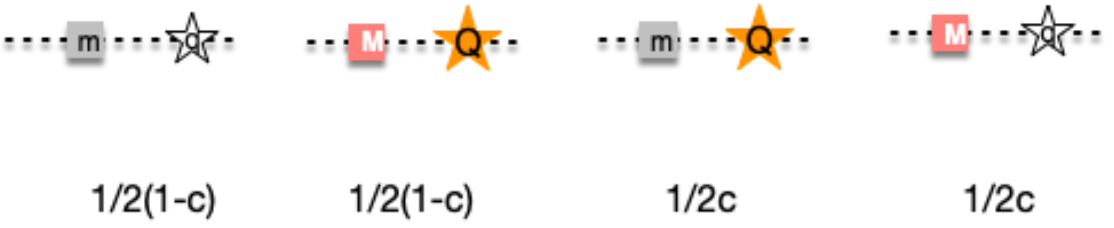



## An example:

Consider a QTL and marker separated by $c = 0.20$ **in an F2 population**.

$$P(MM) = 1/4$$
$$P(QqMM) = 1/2c(1-c)$$
$$P(Qq|MM) = \frac{P(QqMM)}{P(MM)} = \frac{1/2c(1-c)}{1/4} = 2c(1-c)$$

# Conditional probabilities of QTL genotypes




## An example:

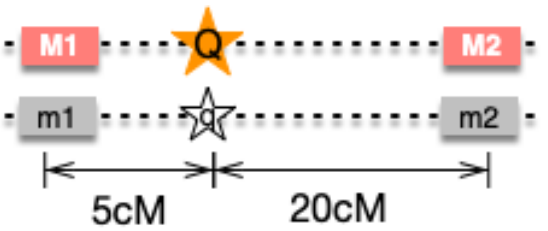
Consider a QTL and marker separated by $c = 0.20$ **in an F2 population**.

$$P(MM) = 1/4$$
$$P(QqMM) = 1/2c(1-c)$$
$$P(Qq|MM) = \frac{P(QqMM)}{P(MM)} = \frac{1/2c(1-c)}{1/4} = 2c(1-c)$$

Substitute in $c = 0.2$ and obtain 0.32.

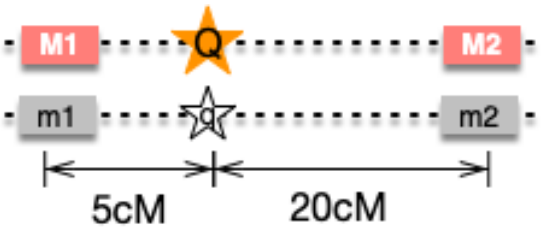
# Interval Mapping

Consider a QTL flanked by two markers segregating in an F2 population.



# Interval Mapping

Consider a QTL flanked by two markers segregating in an F2 population.



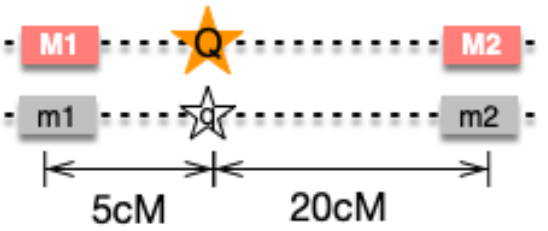

The recombination frequency between **M1 and Q is 0.05** or 5cM and between **Q and M2 is 0.20** or 20cM. Assume no interference.

- What are the probabilities of the three QTL genotypes given a marker genotype of $M_1M_1M_2M_2$?

$$P(QQ|M_1M_1M_2M_2)$$
$$P(Qq|M_1M_1M_2M_2)$$
$$P(qq|M_1M_1M_2M_2)$$

# Interval Mapping

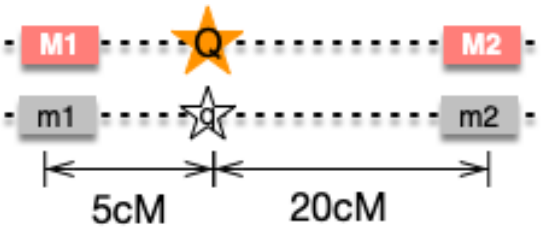
Consider a QTL flanked by two markers segregating in an F2 population.




$$P(QQ|M_1M_1M_2M_2) = \frac{P(M_1M_1QQM_2M_2)}{P(M_1M_1M_2M_2)}$$

# Interval Mapping

Consider a QTL flanked by two markers segregating in an F2 population.

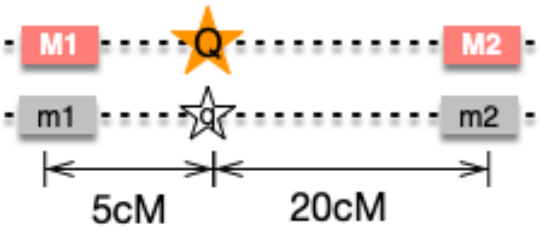



$$P(QQ|M_1M_1M_2M_2) = \frac{P(M_1M_1QQM_2M_2)}{P(M_1M_1M_2M_2)}$$

$$P(M_1M_1M_2M_2) = P(M_1M_2)^2 = (\frac{1-c_{12}}{2})^2$$

# Interval Mapping

Consider a QTL flanked by two markers segregating in an F2 population.



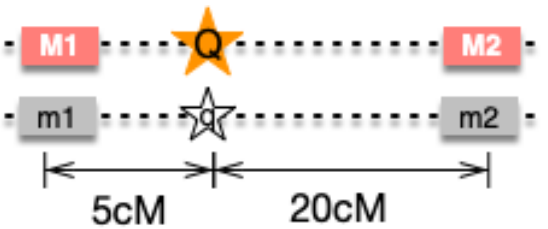

$$P(QQ|M_1M_1M_2M_2) = \frac{P(M_1M_1QQM_2M_2)}{P(M_1M_1M_2M_2)}$$

$$P(M_1M_1M_2M_2) = P(M_1M_2)^2 = (\frac{1-c_{12}}{2})^2$$

$$P(M_1M_1QQM_2M_2) = P(M_1QM_2)^2 = (\frac{(1-c_{1Q})(1-c_{2Q})}{2})^2$$

# Interval Mapping

Consider a QTL flanked by two markers segregating in an F2 population.




$$P(QQ|M_1M_1M_2M_2) = \frac{P(M_1M_1QQM_2M_2)}{P(M_1M_1M_2M_2)}$$

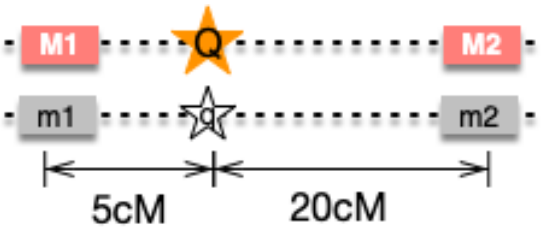
$$P(M_1M_1M_2M_2) = P(M_1M_2)^2 = (\frac{1-c_{12}}{2})^2$$

$$P(M_1M_1QQM_2M_2) = P(M_1QM_2)^2 = (\frac{(1-c_{1Q})(1-c_{2Q})}{2})^2$$

$$P(QQ|M_1M_1M_2M_2) = \frac{(\frac{(1-c_{1Q})(1-c_{2Q})}{2})^2}{(\frac{1-c_{12}}{2})^2} = 0.973$$

# Interval Mapping

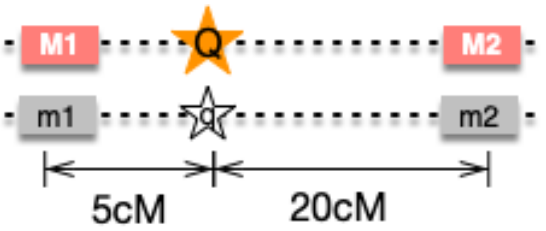
Consider a QTL flanked by two markers segregating in an F2 population.




$$P(QQ|M_1M_1M_2M_2) = \frac{(\frac{(1-c_{1Q})(1-c_{2Q})}{2})^2}{(\frac{1-c_{12}}{2})^2} = 0.973$$

# Interval Mapping

Consider a QTL flanked by two markers segregating in an F2 population.




$$P(QQ|M_1M_1M_2M_2) = \frac{(\frac{(1-c_{1Q})(1-c_{2Q})}{2})^2}{(\frac{1-c_{12}}{2})^2} = 0.973$$

Similarly,

$$P(Qq|M_1M_1M_2M_2) = \frac{P(M_1M_1QqM_2M_2)}{P(M_1M_1M_2M_2)} = 0.026$$

$$P(qq|M_1M_1M_2M_2) = \frac{P(M_1M_1qqM_2M_2)}{P(M_1M_1M_2M_2)} = 1.7 \times 10^{-4}$$

# Interval Mapping

If the genotypic values for each of the QTL genotypes were given as below:

| Genotype | Value | Probability |
|---|---|---|
| $QQ$ | 7 | $P(QQ/M_1M_1M_2M_2) = 0.973$ |
| $Qq$ | 5 | $P(Qq/M_1M_1M_2M_2) = 0.026$ |
| $qq$ | 0 | $P(qq/M_1M_1M_2M_2) = 1.7 \times 10^{-4}$ |

- What is the exected value of individuals with the $M_1M_1M_2M_2$?

# Interval Mapping

If the genotypic values for each of the QTL genotypes were given as below:

| Genotype | Value | Probability |
|---|---|---|
| $QQ$ | 7 | $P(QQ/M_1M_1M_2M_2) = 0.973$ |
| $Qq$ | 5 | $P(Qq/M_1M_1M_2M_2) = 0.026$ |
| $qq$ | 0 | $P(qq/M_1M_1M_2M_2) = 1.7 \times 10^{-4}$ |

- What is the exected value of individuals with the $M_1M_1M_2M_2$?

$$E(M_1M_1M_2M_2) = 0.973 \times 7 + 0.026 \times 5 + 1.7 \times 10^{-4} \times 0 = 6.94$$

# What is the maximum likelihood estimate?

The likelihood funcion is represented as $L(\theta|s) = f_\theta(s)$ This function represents the likelihood of a certain parameter value ( $\theta$ ) given a data vector ( $s$ ).

- The $f_\theta(s)$ represents the **probability density function** with $\theta$ set as the parameter and $s$ set as the observations.
- The value of $L(\theta|s)$ is called the likelihood of $\theta$.
- To find the value of $\theta$ with the maximum likelihood, a range of theta values is tested aginst the observed data, and the $\theta$ giving the **highest** likelihood is determined to be the **maximum likelihood estimator of** $\theta$.

Note: we are fixing the data and varying the parameter.

# Construction of QTL likelihood functions?

When a major bi-allelic locus is segregating in a population. The distrition of the entire population can be broken into three underlying distributions:

- The distribution of the QQ individuals,
- Qq individuals,
- qq individuals.

# Construction of QTL likelihood functions?

When a major bi-allelic locus is segregating in a population. The distrition of the entire population can be broken into three underlying distributions:

- The distribution of the QQ individuals,
- Qq individuals,
- qq individuals.

The likelihood of the genotypic parameters given phenotypic value z is:

$$\begin{aligned} L(z) &= L(\mu_{QQ}, \mu_{Qq}, \mu_{qq}, \sigma^2 | z) \\ &= P(QQ) f(z, \mu_{QQ}, \sigma^2) + P(Qq) f(z, \mu_{Qq}, \sigma^2) + P(qq) f(z, \mu_{qq}, \sigma^2) \end{aligned}$$

# Construction of QTL likelihood functions?

When a major bi-allelic locus is segregating in a population. The distrition of the entire population can be broken into three underlying distributions:

- The distribution of the QQ individuals,
- Qq individuals,
- qq individuals.

The likelihood of the genotypic parameters given phenotypic value z is:

$$\begin{aligned} L(z) &= L(\mu_{QQ}, \mu_{Qq}, \mu_{qq}, \sigma^2 | z) \\ &= P(QQ) f(z, \mu_{QQ}, \sigma^2) + P(Qq) f(z, \mu_{Qq}, \sigma^2) + P(qq) f(z, \mu_{qq}, \sigma^2) \end{aligned}$$

- Where $P(Q_k)$ equals the **probability of a particular genotype**
  - e.g. 1/4 in an F2 population for QQ
- $f(z, \mu_k, \sigma^2)$ is the **probability density function** for a normally distributed random variable with mean $\mu_k$ and variance $\sigma^2$.
  - The mean value of QQ = a, Qq=d and qq=-a.

# Construction of QTL likelihood functions?

Now, let's return to our conditional probabilities, specifically the probability of a QTL genotype given a marker genotype.

The likelihood of an individual with phenotypic value $z$ given a marker genotype $M_i$ is represented as:

$$L(z|M_i) = P(QQ|M_i)f(z, \mu_{QQ}, \sigma^2) + P(Qq|M_i)f(z, \mu_{Qq}, \sigma^2) + P(qq|M_i)f(z, \mu_{qq}, \sigma^2)$$

# Construction of QTL likelihood functions?

Now, let's return to our conditional probabilities, specifically the probability of a QTL genotype given a marker genotype.

The likelihood of an individual with phenotypic value $z$ given a marker genotype $M_i$ is represented as:

$$L(z|M_i) = P(QQ|M_i)f(z, \mu_{QQ}, \sigma^2) + P(Qq|M_i)f(z, \mu_{Qq}, \sigma^2) + P(qq|M_i)f(z, \mu_{qq}, \sigma^2)$$

For example, the likelihood for **genotype MM** is:

$$L(z|MM) = P(QQ|MM)f(z, \mu_{QQ}, \sigma^2) + P(Qq|MM)f(z, \mu_{Qq}, \sigma^2) + P(qq|MM)f(z, \mu_{qq}, \sigma^2)$$

# Construction of QTL likelihood functions?

Now, let's return to our conditional probabilities, specifically the probability of a QTL genotype given a marker genotype.

The likelihood of an individual with phenotypic value $z$ given a marker genotype $M_i$ is represented as:

$$L(z|M_i) = P(QQ|M_i)f(z, \mu_{QQ}, \sigma^2) + P(Qq|M_i)f(z, \mu_{Qq}, \sigma^2) + P(qq|M_i)f(z, \mu_{qq}, \sigma^2)$$

For example, the likelihood for **genotype MM** is:

$$L(z|MM) = P(QQ|MM)f(z, \mu_{QQ}, \sigma^2) + P(Qq|MM)f(z, \mu_{Qq}, \sigma^2) + P(qq|MM)f(z, \mu_{qq}, \sigma^2)$$

The $P(Q_k|M_j)$ parts are a function of the **map positions and experimental design**, so that

$$L(z|MM) = 4(1-c)^2 f(z, \mu_{QQ}, \sigma^2) + 8c(1-c)f(z, \mu_{Qq}, \sigma^2) + 4c^2 f(z, \mu_{qq}, \sigma^2)$$

And the QTL effects enter through the means and variances of the underlying normal distrutions $f_\theta(z)$ or $f(z, \mu_{Qk}, \sigma^2)$.

# Back to interval mapping

To calculate the likelihoods for an interval, we simply insert the probabilities of a QTL genotype given a marker interval genotype.

# Back to interval mapping

To calculate the likelihoods for an interval, we simply insert the probabilities of a QTL genotype given a marker interval genotype.

For example,

$$\begin{aligned} L(z|M_1M_1M_2M_2) = & \frac{(1-c_1)^2(1-c_2)^2}{(1-c_{12})^2} f(z, \mu_{QQ}, \sigma^2) \\ & + \frac{2c_1c_2(1-c_1)(1-c_2)}{(1-c_{12})^2} f(z, \mu_{Qq}, \sigma^2) \\ & + \frac{c_1^2c_2^2}{(1-c_{12})^2} f(z, \mu_{qq}, \sigma^2) \end{aligned}$$

# Back to interval mapping

To calculate the likelihoods for an interval, we simply insert the probabilities of a QTL genotype given a marker interval genotype.

For example,

$$
\begin{aligned}
L(z|M_1M_1M_2M_2) = & \frac{(1-c_1)^2(1-c_2)^2}{(1-c_{12})^2} f(z, \mu_{QQ}, \sigma^2) \\
& + \frac{2c_1c_2(1-c_1)(1-c_2)}{(1-c_{12})^2} f(z, \mu_{Qq}, \sigma^2) \\
& + \frac{c_1^2c_2^2}{(1-c_{12})^2} f(z, \mu_{qq}, \sigma^2)
\end{aligned}
$$

- This likelihood value is calculated for each genetic position in between the two flanking markers by varing value of **recombination rate (c)**.
- The span of the **entire interval ( $c_{12}$ )** is calculated using a mapping function.
- The values of $\mu_{QQ}, \mu_{Qq}, \mu_{qq}, \sigma^2$ are estimated at each genetic position.

# Back to interval mapping

Fundamentally, the likelihood of these values **given $c_1, c_2, c_{12}$ and the phenotypic data** is being calculated.

- This likelihood is contrasted with the likelihood that no underlying QTL exists (the data arose in the absence of a QTL).
- **That is, there are no underlying QTL genotypes and the distribution of individuals with the $M_1M_1M_2M_2$ genotype consists of a single distribution.**

# Back to interval mapping

Fundamentally, the likelihood of these values **given $c_1, c_2, c_{12}$ and the phenotypic data** is being calculated.

- This likelihood is contrasted with the likelihood that no underlying QTL exists (the data arose in the absence of a QTL).
- **That is, there are no underlying QTL genotypes and the distribution of individuals with the $M_1M_1M_2M_2$ genotype consists of a single distribution.**

# LOD score

- This ratio is what provides the likelihood ratio.
- It is converted to the famed LOD score, which indicates:
  - the probability of obtaining the **alternative hypothesis** (QTL present and therefore an underlying mixture distribution)
  - relative to the **null hypothesis** (no QTL, single distribution)

# LOD score

$$LOD = log_{10}(\frac{L_{full}}{L_{reduced}})$$

- Where $L_{full}$ is the likelihood of a QTL at assumed genetic position given the data.
- $L_{reduced}$ is the likelihood of no QTL present given the data.

# LOD score

$$LOD = log_{10}(\frac{L_{full}}{L_{reduced}})$$

- Where $L_{full}$ is the likelihood of a QTL at assumed genetic position given the data.
- $L_{reduced}$ is the likelihood of no QTL present given the data.

If the LOD score is 3, for example, this means that the likelihood for a model including a QTL at the given genetic position is 1,000 times higher than no QTL at that position!

# Statistical significance

QTL mapping involves a large number of tests, which requires adjustments for multiple testing to keep **the experiment-wise error rate** low.

## Bonferroni correction

- Assume **all tests are independent**, which is not the case in QTL mapping because markers are linked.
- Overly conservative for QTL mapping.

# Statistical significance

QTL mapping involves a large number of tests, which requires adjustments for multiple testing to keep **the experiment-wise error rate** low.

## Permutation test

A commonly used technique for QTL mapping.

# Statistical significance

QTL mapping involves a large number of tests, which requires adjustments for multiple testing to keep **the experiment-wise error rate** low.

## Permutation test

A commonly used technique for QTL mapping.

- Basically, the phenotypic data is **randomized** relative to the marker data so that the null hypothesis is established.
- Then, the test statistic for each QTL is calculated and the **largest test statistic** across the genome is tabulated.
- This is **repeated 1,000 or more times** in order to estabilish an empirical distribution of the test statistic under the null hypothesis.
- The test statistics calculated for the real data are compared to this distribution to determine the significance level.