Multivariate Statistics in Ecology and Quantitative Genetics
Quantitative Traits Loci (QTL) Mapping

Dirk Metzler & Noémie Becker

9. Juli 2014

Contents

📕 K.W. Broman, S. Sen (2009) *A guide to QTL Mapping with R/qtl.*
Springer, New York.

Contents

Contents

A   **×**   **B**   **inbred lines**

A   **×**   **F1**

back cross

A   B   inbred lines

F1   F1

F2
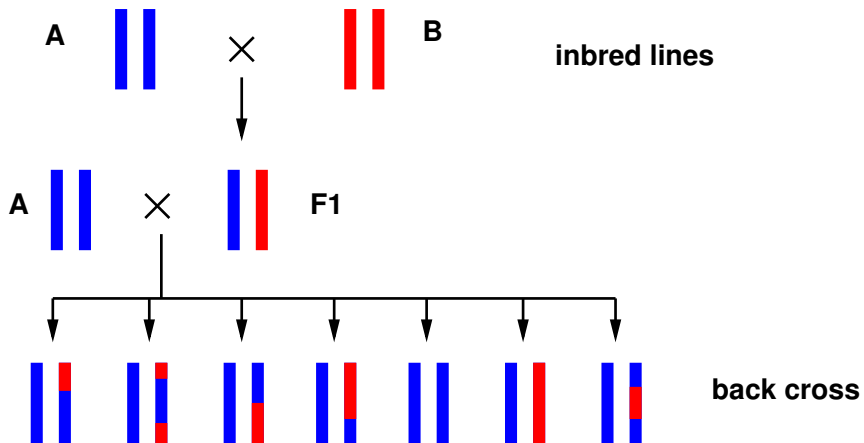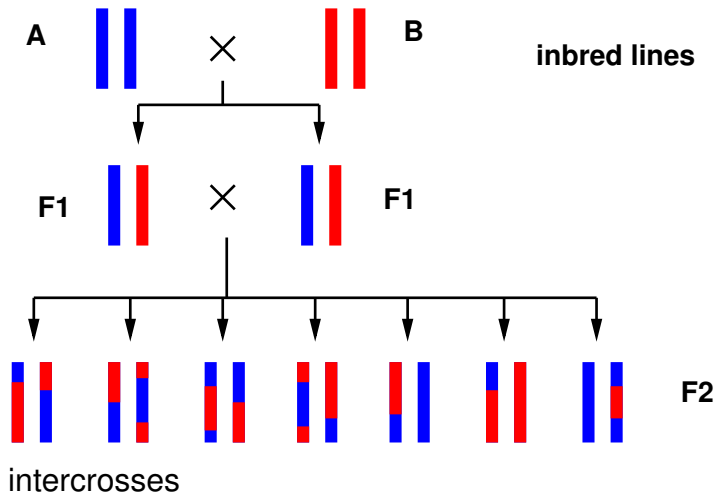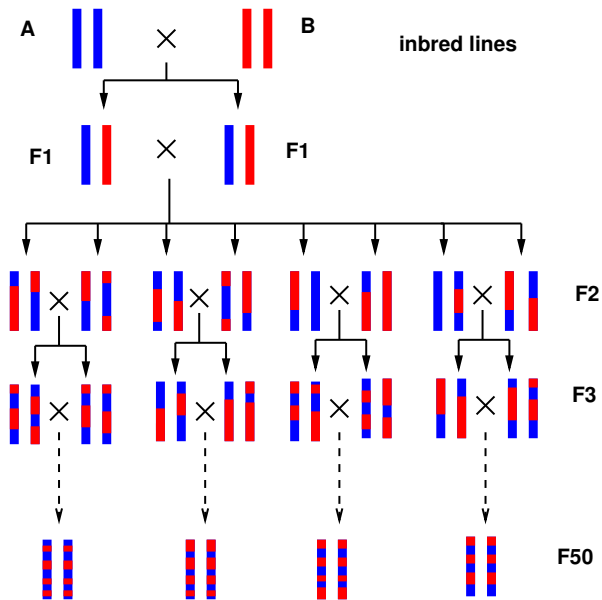
intercrosses

Recombinant Inbred Lines (RILs)

Contents

Example dataset with backcrosses

```
> library(qtl)
> data(hyper)
> summary(hyper)
    Backcross

    No. individuals:      250

    No. phenotypes:       2
    Percent phenotyped:   100 100

    No. chromosomes:      20
        Autosomes:        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
        X chr:            X

    Total markers:        174
    No. markers:          22 8 6 20 14 11 7 6 5 5 14 5 5 5 11 6 12 4 4 4
    Percent genotyped:    47.7
    Genotypes (%):        BB:50.2  BA:49.8
```
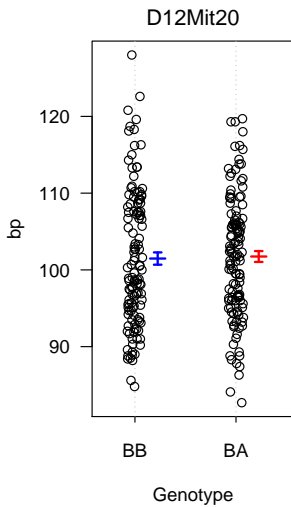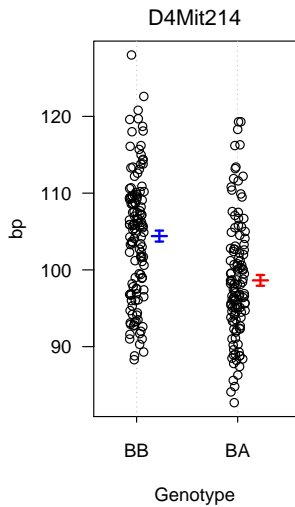
D4Mit214

D12Mit20

```
par(mfrow=c(1,2))
plot.pxg(hyper,"D4Mit214")
plot.pxg(hyper,"D12Mit20")
par(mfrow=c(1,1))
```

Assume that p sites have an influence on the quantitative trait y of interest and denote an individual's genotype at these sites by $g = (g_1, g_2, \ldots, g_p)$

$$
\begin{aligned}
\mu_g &:= \mathbb{E}(y|g) \\
\sigma_g^2 &:= \mathrm{var}(y|g) \\
\text{we assume: } y|g &\sim \mathcal{N}(\mu_g, \sigma_g^2) \\
\text{additive model: } \mu_g &= \mu + \sum_{j=1}^{p} z_j \cdot \Delta_j,
\end{aligned}
$$

whereas z_j is 0 or 1 according to the genotype of g_j, and Δ_j is the effect of the QTL at position j.

In a strict sense, *epistasis* means that the effect of a mutation can be masked by a mutation at a different loci.

However, in the context of QTL mapping, the word epistasis if often used to express that there is a non-additive interaction between two loci. (Problem: whether effects are additive or not depends on how the trait is scaled.)

In a strict sense, *epistasis* means that the effect of a mutation can be masked by a mutation at a different loci.

However, in the context of QTL mapping, the word epistasis if often used to express that there is a non-additive interaction between two loci. (Problem: whether effects are additive or not depends on how the trait is scaled.)

Main problem: We do not know where the QTLs are. We only have genetic markers to determine for several sites whether they stem from A or B.

Contents

Contents

Assume a backcross experiment with n F2 individuals
Let $y = (y_1, \ldots, y_n)$ be their phenotypes for the trait of interest.

Null hypothesis H_0: no QTL
Residual sum of squares under H_0:

$$\mathrm{RSS}_0 = \sum_{k=1}^{n}(y_k - \bar{y})^2$$

Very simple alternative H_1: single QTL at marker position i

$$y|g_i \sim \mathcal{N}(\mu_{g_i}, \sigma^2)$$

Likelihood function:

$$
\begin{aligned}
L_1(\mu_{AA}, \mu_{AB}, \sigma^2) &= \Pr(y|\text{QTL marker}, \mu_{AA}, \mu_{AB}, \sigma^2) \\
&= \Pi_{k=1}^{n}\phi(y_k; \mu_{g_{ik}}, \sigma^2),
\end{aligned}
$$

whereas ϕ is the density of the normal distribution and g_{ik} is the genotype of individual k at marker position i.

The maximal likelihood under H_1 is $\text{RSS}_1^{-n/2}$, with

$$\text{RSS}_1 = \sum_{k=1}^{n} (y_k - \widehat{\mu_{g_{ik}}})^2,$$

where $\mu_{g_{ik}}$ is the mean trait value over all individuals that have type g_{ik} at marker position i.

The LOD score is the $\log_{10}$ of the likelihood ratio of H_1 and H_0:

$$\text{LOD} = \frac{n}{2} \log_{10} \left(\frac{\text{RSS}_0}{\text{RSS}_1} \right)$$

The LOD score is traditionally used in QTL mapping. However, it is equivalent to the classical anova *F*-statistic:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{df}}{\mathrm{RSS}_1/(n - \mathrm{df} - 1)} = \left(10^{2 \cdot \mathrm{LOD}/n} - 1\right) \cdot \frac{n - \mathrm{df} - 1}{\mathrm{df}}$$

$$\mathrm{LOD} = \frac{n}{2} \log_{10}\left(\frac{F \cdot \mathrm{df}}{n - \mathrm{df} + 1} + 1\right)$$

So, if the marker positions are our candidates for the QTLs we just perform anovas.

Contents

- ▶ The QTLs may be between the marker positions, and their genotypes can only be estimated from the markers.

- ▶ The QTLs may be between the marker positions, and their genotypes can only be estimated from the markers.
- ▶ Let M_k be the multipoint marker genotype of individual k and $g_{\ell k}$ its QTL genotype at candidate position ℓ, and

$$p_{kj} := \Pr(g_{\ell k} = j | M_k).$$

(Computation uses recombination rates.)

- The QTLs may be between the marker positions, and their genotypes can only be estimated from the markers.
- Let M_k be the multipoint marker genotype of individual k and $g_{\ell k}$ its QTL genotype at candidate position ℓ, and

$$p_{kj} := \Pr(g_{\ell k} = j | M_k).$$

(Computation uses recombination rates.)

- Probability density of an individual's phenotype (at candidate locus ℓ) is a mixture of normal distribution densities:

$$\sum_j p_{kj} \cdot \phi(y_k; \mu_j, \sigma^2)$$

EM algorithm for ML-estimation of μ_j and σ

Start with initial estimates $\mu_j^{(0)}$ and $\sigma^{(0)}$ and iterate the following steps for $s = 1, \ldots, N$:

E-step

$$
\begin{aligned}
w_{kj}^{(s)} &:= \Pr(g_{\ell k} = j | M_k, y_k, \mu_j^{(s-1)}, \sigma^{(s-1)}) \\
&= \frac{p_{kj}\phi(y_k; \mu_j^{(s-1)}, \sigma^{(s-1)})}{\sum_h p_{kh}\phi(y_k; \mu_h^{(s-1)}, \sigma^{(s-1)})}
\end{aligned}
$$

M-step

$$
\begin{aligned}
\mu_j^{(s)} &:= \sum_k w_{kj}^{(s)} y_i / \sum_h w_{hj}^{(s)} \\
\sigma^{(s)} &:= \sqrt{\sum_{kj} w_{kj}^{(s)}(y_k - \mu_{g_{kj}}^{(s)})^2 / n}
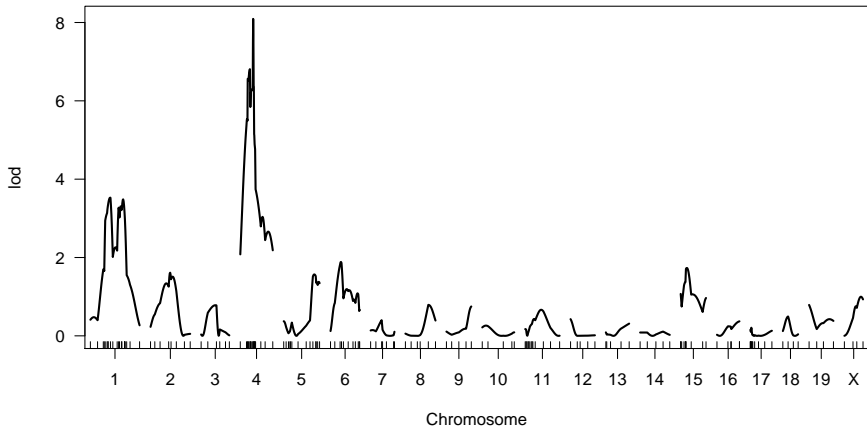\end{aligned}
$$

The aim of the EM algorithm is that $\mu_j^{(s)}$ and $\sigma^{(s)}$ converge against the ML estimators $\widehat{\mu}$ and $\widehat{\sigma}$.

The aim of the EM algorithm is that $\mu_j^{(s)}$ and $\sigma^{(s)}$ converge against the ML estimators $\widehat{\mu}$ and $\widehat{\sigma}$.

Then, the LOD score can be computed:

$$\text{LOD} = \log_{10}\left(\frac{\Pi_i \sum_j p_{ij}\phi(y_i; \widehat{\mu}_j, \widehat{\sigma}^2)}{\Pi_i\phi(y_i; \widehat{\mu}_0, \widehat{\sigma}_0^2)}\right)$$

```
## calculate p_{kj}
hyper <- calc.genoprob(hyper,step=1,error.prob=0.001)
out.em <- scanone(hyper,method="em")
plot(out.em)
```

Sometimes EM can be very slow.
Haley-Knott (HK) regression is a fast approximation:
For each point on the grid calculate $p_{kj} = \Pr(g_i = j|M)$ and
estimate μ_j and σ by fitting a linear model

$$y_k|M_k \sim \mathcal{N}\left(\sum_j p_{kj}\mu_j, \sigma^2\right)$$

Sometimes EM can be very slow.
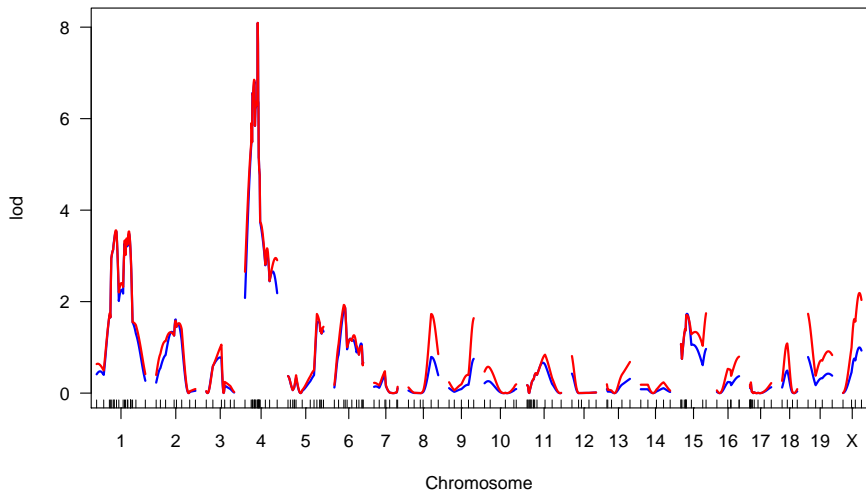*Haley-Knott (HK) regression* is a fast approximation:
For each point on the grid calculate $p_{kj} = \Pr(g_i = j|M)$ and
estimate μ_j and σ by fitting a linear model

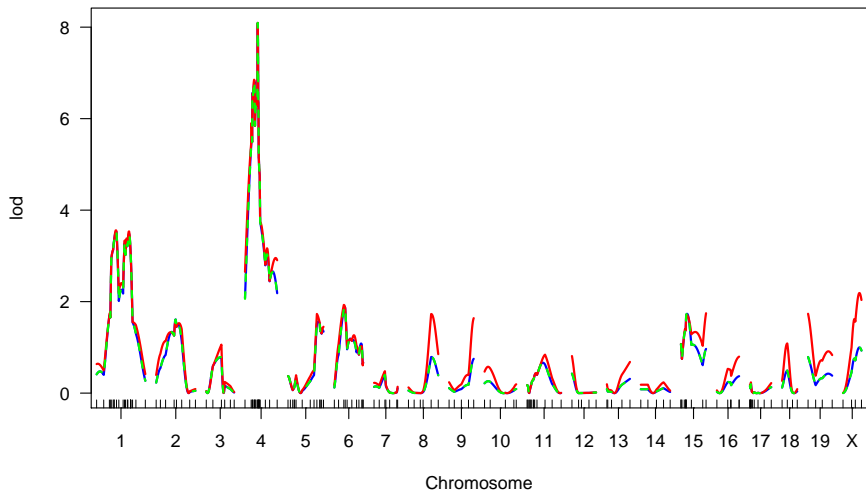$$y_k|M_k \sim \mathcal{N}\left(\sum_j p_{kj}\mu_j, \sigma^2\right)$$

Extended Haley-Knott (EHK) regression: Takes into account that
p_{kj} and μ_j have an influence on the variance:

$$y_k|M_k \sim \mathcal{N}\left(\sum_j p_{kj}\mu_j\,,\; \sum_j p_{kj}\left(\mu_j - \sum_h p_{kh}\mu_h\right)^2 + \sigma^2\right)$$

```
out.hk <- scanone(hyper,method="hk")
plot(out.em,out.hk,col=c("blue","red"))
```

```
out.ehk <- scanone(hyper,method="ehk")
plot(out.em,out.hk,out.ehk,col=c("blue","red","green"),lty=c(1,1,2))
```
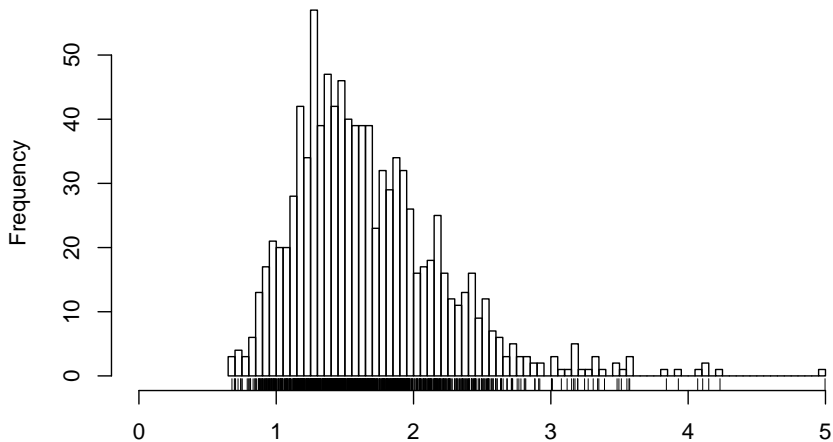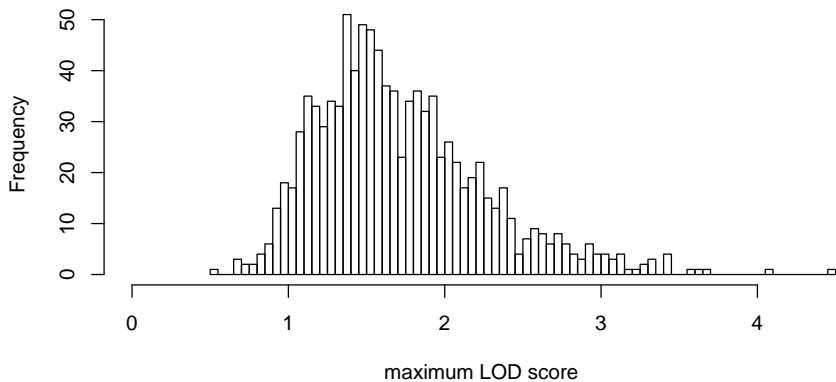
Which LOD scores are significant?

Which LOD scores are significant?
Assess this by a permutation test: shuffle the phenotype column.

```
## next command will take time
out.hk.perm <- scanone(hyper,method="hk",n.perm=1000)
plot(out.hk)
```

```
## this will take even longer:
out.perm <- scanone(hyper,n.perm=1000)
plot(out.perm)
```

Contents

More than one QTL

Composite Interval Mapping While searching for a QTL in one interval use other markers as proxies for nearby QTLs. Thus, markers are used as covariates. Specify maximal number of covariates and how far they should be away from the interval under examination.

two-QTL models search for interacting pairs of QTLs. Same methods like in 1-QTL model: EM, HK, EHK

multiple QTLs When candidate loci are found, fit regression models allowing for interactions and do variable selection.

```
out.cim <- cim(hyper)
plot(out.cim)
```
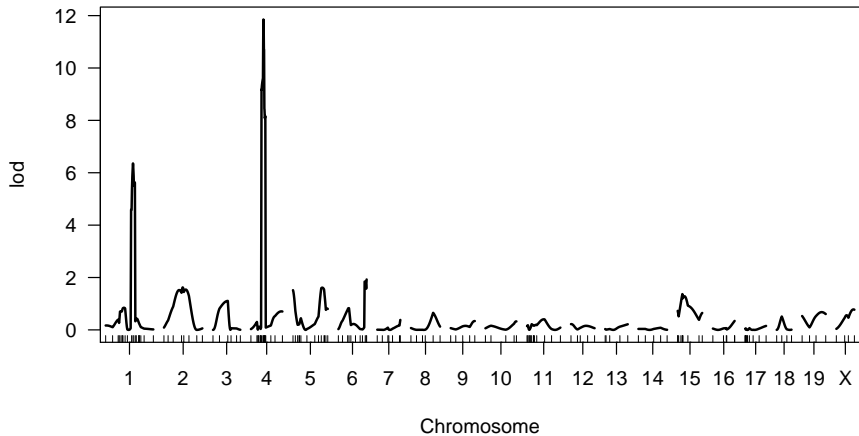
```
out2 <- scantwo(hyper) ## takes quite long
plot(out2)
```
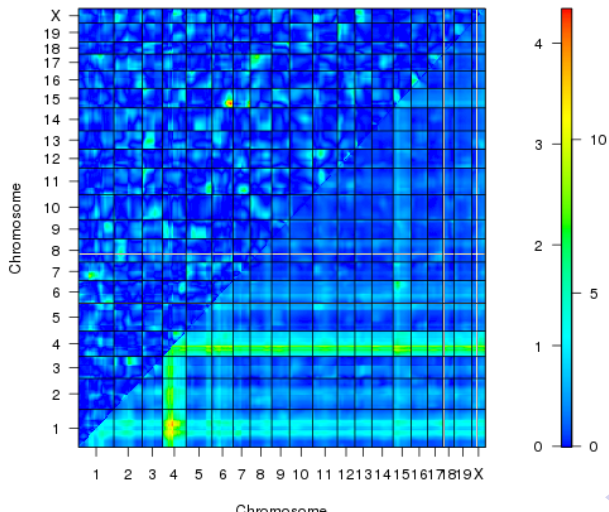
```
out2.hk <- scantwo(hyper,method="hk") ## much faster
plot(out2.hk)
```
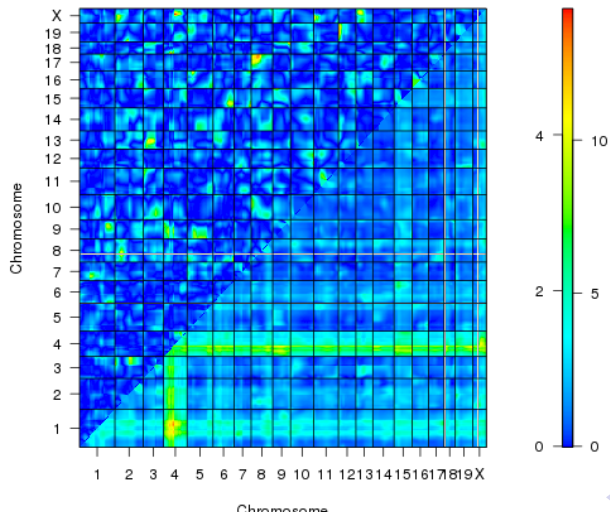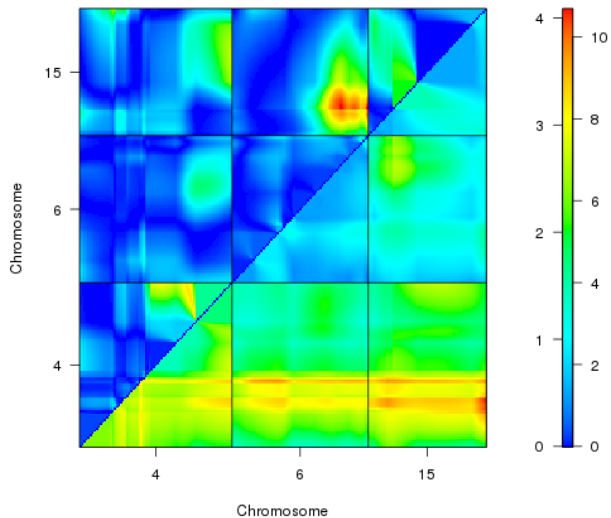
```
plot(out2.hk,chr=c(4,6,15))
```
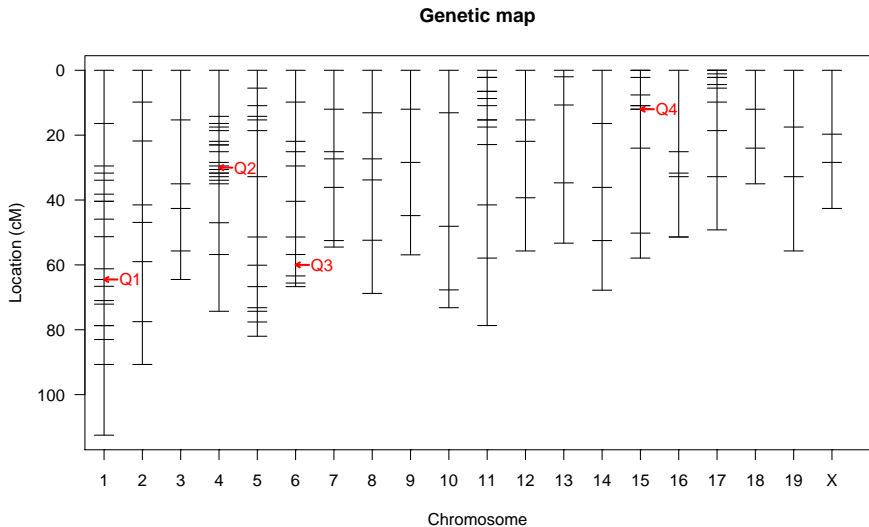
```
> hyper <- sim.geno(hyper,step=2,n.draws=128,err=0.001)
> qtl <- makeqtl(hyper,chr=c(1,4,6,15),pos=c(68.3,30,60,18))
> qtl
  QTL object containing imputed genotypes, with 128 imputation

      name chr  pos n.gen
Q1  1@67.8   1 67.8     2
Q2  4@30.0   4 30.0     2
Q3  6@60.0   6 60.0     2
Q4 15@17.5  15 17.5     2
```

plot(qtl)

Genetic map

```
> out.fq <- fitqtl(hyper,qtl=qtl,formula= y~(Q1+Q2+Q3+Q4)^2)
> summary(out.fq)

fitqtl summary

Method: multiple imputation
Model: normal phenotype
Number of observations : 250

Full model result
----------------------------------
Model formula: y ~ Q1 + Q2 + Q3 + Q4 + Q1:Q2 + Q1:Q3 + Q1:Q4 + Q2:Q3 + Q2:Q4 +
                   Q3:Q4

        df       SS        MS      LOD     %var Pvalue(Chi2) Pvalue(F)
Model   10 6113.512 611.35116 23.05306 34.60034            0         0
Error  239 11555.425  48.34906
Total  249 17668.936


Drop one QTL at a time ANOVA table:
----------------------------------
              df Type III SS     LOD    %var F value Pvalue(Chi2) Pvalue(F)
1@67.8         4     1548.22  6.8258  8.7624  8.0054        0.000  4.51e-06 ***
4@30.0         4     3184.90 13.2152 18.0254 16.4683        0.000  6.23e-12 ***
6@60.0         4     1671.00  7.3321  9.4573  8.6403        0.000  1.58e-06 ***
15@17.5        4     1504.34  6.6437  8.5140  7.7785        0.000  6.57e-06 ***
```

```
> out.fq <- fitqtl(hyper,qtl=qtl,formula= y~(Q1+Q2+Q3+Q4)^2)
> summary(out.fq)

        .
        .
        .

Drop one QTL at a time ANOVA table:
----------------------------------
             df Type III SS    LOD    %var F value Pvalue(Chi2) Pvalue(F)
1@67.8        4     1548.22  6.8258  8.7624  8.0054        0.000  4.51e-06 ***
4@30.0        4     3184.90 13.2152 18.0254 16.4683        0.000  6.23e-12 ***
6@60.0        4     1671.00  7.3321  9.4573  8.6403        0.000  1.58e-06 ***
15@17.5       4     1504.34  6.6437  8.5140  7.7785        0.000  6.57e-06 ***
1@67.8:4@30.0  1      79.45  0.3720  0.4496  1.6432        0.191     0.201
1@67.8:6@60.0  1      50.96  0.2389  0.2884  1.0540        0.294     0.306
1@67.8:15@17.5 1      57.42  0.2691  0.3250  1.1877        0.266     0.277
4@30.0:6@60.0  1      54.02  0.2532  0.3057  1.1172        0.280     0.292
4@30.0:15@17.5 1      29.70  0.1393  0.1681  0.6143        0.423     0.434
6@60.0:15@17.5 1    1071.15  4.8124  6.0623 22.1544        0.000  4.26e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- ▶ Candidate loci and interactions found by scanone and scantwo can then be used in multiple QTL analysis.
- ▶ Then, p-values from multiple QTL analysis are not reliable because not multiple-testing corrected. Massive multiple-testing problem is caused by preselection by scanone and scantwo.
- ▶ If two QTL are close to each other with only few marker loci inbetween, scanone may falsely indicate strong evidence for one QTL between the two.