Multivariate Statistics in Ecology and Quantitative Genetics

Quantitative Traits Loci (QTL) Mapping

Dirk Metzler & Noémie Becker

http://evol.bio.lmu.de/_statgen

July 2013
Contents

Introduction
  Crossing Schemes
  QTL model assumptions

Single-QTL analysis
  LOD score
  Interval mapping

More than one QTL
Contents

Introduction

Crossing Schemes

QTL model assumptions

Single-QTL analysis

LOD score

Interval mapping

More than one QTL
Contents

Introduction

Crossing Schemes

QTL model assumptions

Single-QTL analysis

LOD score

Interval mapping

More than one QTL
A

\[ \times \]

B

inbred lines

F1

\[ \times \]

F1

intercrosses

F2
Recombinant Inbred Lines (RILs)
Contents

Introduction
  Crossing Schemes
  QTL model assumptions

Single-QTL analysis
  LOD score
  Interval mapping

More than one QTL
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
Introduction

QTL model assumptions

```
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 \( g = (g_1, g_2, \ldots, g_p) \)

\[
\begin{align*}
\mu_g & := \mathbb{E}(y|g) \\
\sigma^2_g & := \text{var}(y|g)
\end{align*}
\]

we assume: \( y|g \sim \mathcal{N}(\mu_g, \sigma^2_g) \)

additive model: \( \mu_g = \mu + \sum_{j=1}^{p} z_j \cdot \Delta_j \),

whereas \( z_j \) is 0 or 1 according to the genotype of \( g_j \), and \( \Delta_j \) 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 is 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

Introduction
Crossing Schemes
QTL model assumptions

Single-QTL analysis
LOD score
Interval mapping

More than one QTL
Contents

Introduction
Crossing Schemes
QTL model assumptions

Single-QTL analysis
LOD score
Interval mapping

More than one QTL
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 $H_0$:

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

Very simple alternative $H_1$: single QTL at marker position $i$

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

Likelihood function:

$$L_1(\mu_{AA}, \mu_{AB}, \sigma^2) = \Pr(y|\text{QTL marker}, \mu_{AA}, \mu_{AB}, \sigma^2)$$

$$= \prod_{k=1}^{n} \phi(y_k; \mu_{g_{ik}}, \sigma^2)$$

whereas $\phi$ 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

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

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:

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

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

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

Introduction
Crossing Schemes
QTL model assumptions

Single-QTL analysis
LOD score
Interval mapping

More than one QTL
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 $\ell$, 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 $\ell$, and

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

(Computation uses recombination rates.)

Probability density of an individual’s phenotype (at candidate locus $\ell$) 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 $\mu_j$ and $\sigma$

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

**E-step**

$$w_{kj}^{(s)} := \Pr(g_{\ell k} = j | M_k, y_k, \mu_j^{(s-1)}, \sigma^{(s-1)})$$

$$= \frac{p_{kj}\phi(y_k; \mu^{(s-1)}_j, \sigma^{(s-1)})}{\sum_h p_{kh}\phi(y_k; \mu^{(s-1)}_h, \sigma^{(s-1)})}$$

**M-step**

$$\mu_j^{(s)} := \frac{\sum_k w_{kj}^{(s)} y_i}{\sum_h w_{hj}^{(s)}}$$

$$\sigma^{(s)} := \sqrt{\frac{\sum_{kj} w_{kj}^{(s)} (y_k - \mu_{g_{kj}}^{(s)})^2}{n}}$$
The aim of the EM algorithm is that $\mu_j^{(s)}$ and $\sigma^{(s)}$ converge against the ML estimators $\hat{\mu}$ and $\hat{\sigma}$. 
The aim of the EM algorithm is that $\mu_j^{(s)}$ and $\sigma^{(s)}$ converge against the ML estimators $\hat{\mu}$ and $\hat{\sigma}$.

Then, the LOD score can be computed:

$$\text{LOD} = \log_{10} \left( \frac{\prod_i \sum_j p_{ij} \phi(y_i; \hat{\mu}_j, \hat{\sigma}^2)}{\prod_i \phi(y_i; \hat{\mu}_0, \hat{\sigma}_0^2)} \right)$$
## calculate $p_{kj}$

```r
hyper <- calc.genoprob(hyper, step=1, error.prob=0.001)
out.em <- scanone(hyper, method="em")
plot(out.em)
```

### Genome scan

![Genome scan](image.png)
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 $\mu_j$ and $\sigma$ by fitting a linear model

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

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

Introduction
  Crossing Schemes
  QTL model assumptions

Single-QTL analysis
  LOD score
  Interval mapping

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.
More than one QTL

```r
out.cim <- cim(hyper)
plot(out.cim)
```
More than one QTL

```r
out2 <- scantwo(hyper)  ## takes quite long
plot(out2)
```
More than one QTL

```r
out2.hk <- scantwo(hyper, method="hk") ## much faster
plot(out2.hk)
```
More than one QTL

plot(out2.hk, chr=c(4, 6, 15))
```r
> 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 imputations.

<table>
<thead>
<tr>
<th>name</th>
<th>chr</th>
<th>pos</th>
<th>n.gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1</td>
<td>67.8</td>
<td>2</td>
</tr>
<tr>
<td>Q2</td>
<td>4</td>
<td>30.0</td>
<td>2</td>
</tr>
<tr>
<td>Q3</td>
<td>6</td>
<td>60.0</td>
<td>2</td>
</tr>
<tr>
<td>Q4</td>
<td>15</td>
<td>17.5</td>
<td>2</td>
</tr>
</tbody>
</table>
```
More than one QTL

plot(qtl)
More than one QTL

> 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:

<table>
<thead>
<tr>
<th>df</th>
<th>Type III SS</th>
<th>LOD</th>
<th>%var</th>
<th>F value</th>
<th>Pvalue(Chi2)</th>
<th>Pvalue(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67.8</td>
<td>4</td>
<td>1548.22</td>
<td>6.8258</td>
<td>8.7624</td>
<td>8.0054</td>
</tr>
<tr>
<td>4</td>
<td>30.0</td>
<td>4</td>
<td>3184.90</td>
<td>13.2152</td>
<td>18.0254</td>
<td>16.4683</td>
</tr>
<tr>
<td>6</td>
<td>60.0</td>
<td>4</td>
<td>1671.00</td>
<td>7.3321</td>
<td>9.4573</td>
<td>8.6403</td>
</tr>
<tr>
<td>15</td>
<td>17.5</td>
<td>4</td>
<td>1504.34</td>
<td>6.6437</td>
<td>8.5140</td>
<td>7.7785</td>
</tr>
<tr>
<td>1@67.8:4@30.0</td>
<td>1</td>
<td>79.45</td>
<td>0.3720</td>
<td>0.4496</td>
<td>1.6432</td>
<td>0.191</td>
</tr>
<tr>
<td>1@67.8:6@60.0</td>
<td>1</td>
<td>50.96</td>
<td>0.2389</td>
<td>0.2884</td>
<td>1.0540</td>
<td>0.294</td>
</tr>
<tr>
<td>1@67.8:15@17.5</td>
<td>1</td>
<td>57.42</td>
<td>0.2691</td>
<td>0.3250</td>
<td>1.1877</td>
<td>0.266</td>
</tr>
<tr>
<td>4@30.0:6@60.0</td>
<td>1</td>
<td>54.02</td>
<td>0.2532</td>
<td>0.3057</td>
<td>1.1172</td>
<td>0.280</td>
</tr>
<tr>
<td>4@30.0:15@17.5</td>
<td>1</td>
<td>29.70</td>
<td>0.1393</td>
<td>0.1681</td>
<td>0.6143</td>
<td>0.423</td>
</tr>
<tr>
<td>6@60.0:15@17.5</td>
<td>1</td>
<td>1071.15</td>
<td>4.8124</td>
<td>6.0623</td>
<td>22.1544</td>
<td>0.000</td>
</tr>
</tbody>
</table>

---

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.