Implementation of robust methods for locating quantitative trait loci in R

Andreas Baierl and Andreas Futschik

Institute of Statistics and Decision Support Systems
University of Vienna

Overview

- Introduction to QTL mapping
- Analysis of QTL data
  - modified BIC
  - Robust methods
- Implementation and Simulations in R

Locating quantitative trait loci (QTL)

Quantitative trait:

evolution occurred in small steps

characters, that are influenced by many genes

Many relevant traits are quantitative: height, yield, ...

Quantitative trait locus (QTL):

gene (functional sequence of bases) that influences a certain quantitative trait

Relevant questions:

- How many genes influence a trait (How many QTL)
- Find exact positions of QTL
- (- estimate size of genetic effects)

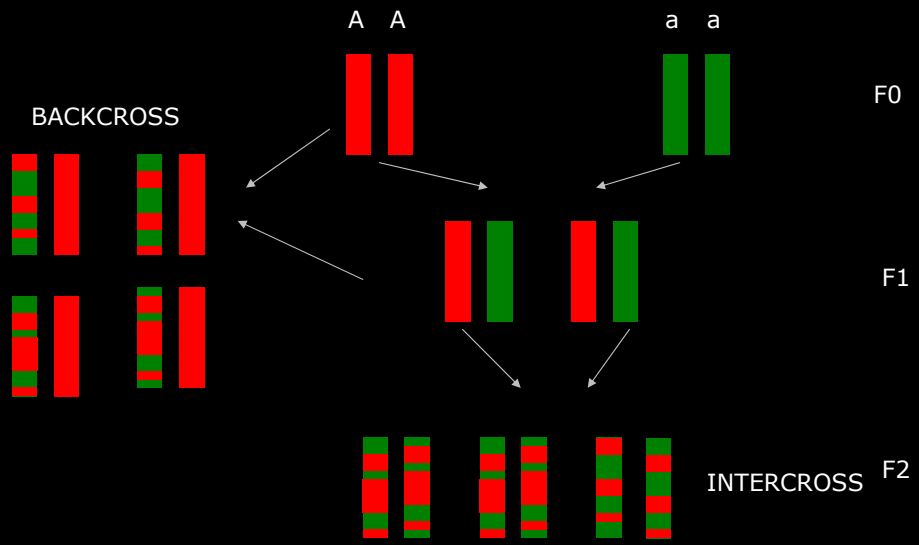
Background

- A gene can obtains different forms (alleles)
- contribution of genetic effects to total (phenotypic) variation of a trait (heritability) determines rate at which characters respond to selection. (environmental variance reduces efficiency of response)

  trait value = genetic influence + environmental influence
- partitioning genotypic variance into components with different impact on selection: additive, non-additive gene effects (epistasis)
  -> dependency on background population
  evolutionary reason: stabilization of phenotype

phenotype: the form taken by some character in a specific individual.
genotype: genetic makeup of individual

Data from experimental crosses



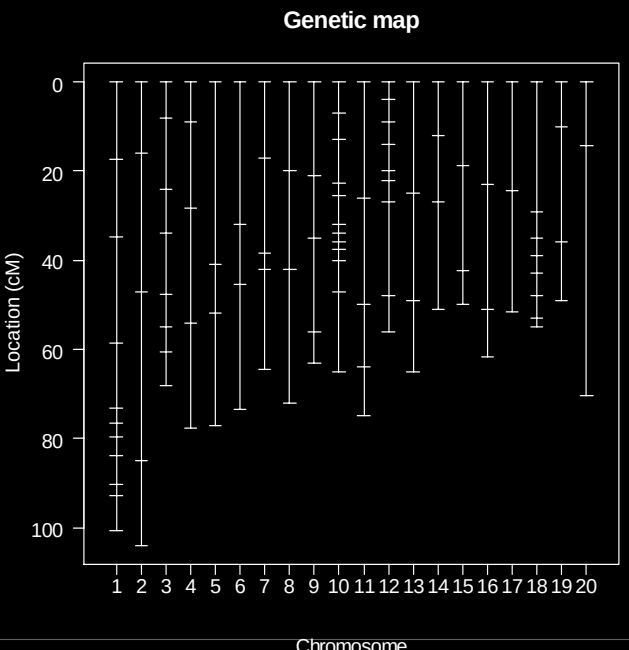
Data matrix for backcross design

Indiv.	QT	marker.1	marker.2	...	marker.m
1	34.3	AA	Aa	...	AA
2	65.4	Aa	AA	...	*
3	23.2	Aa	*	...	Aa
4	45.4	AA	AA	...	Aa
...	....	...	...	...	...

~ 50-500 markers

~ 200 – 1000 individuals

Genetic map



Distance between markers is usually estimated from recombination frequency

If marker is close to QTL, then marker genotype will be associated with QTL genotype (There would be a *1-1* correspondence, if there were no recombinations)

No linkage between chromosomes

Analysis of QTL data

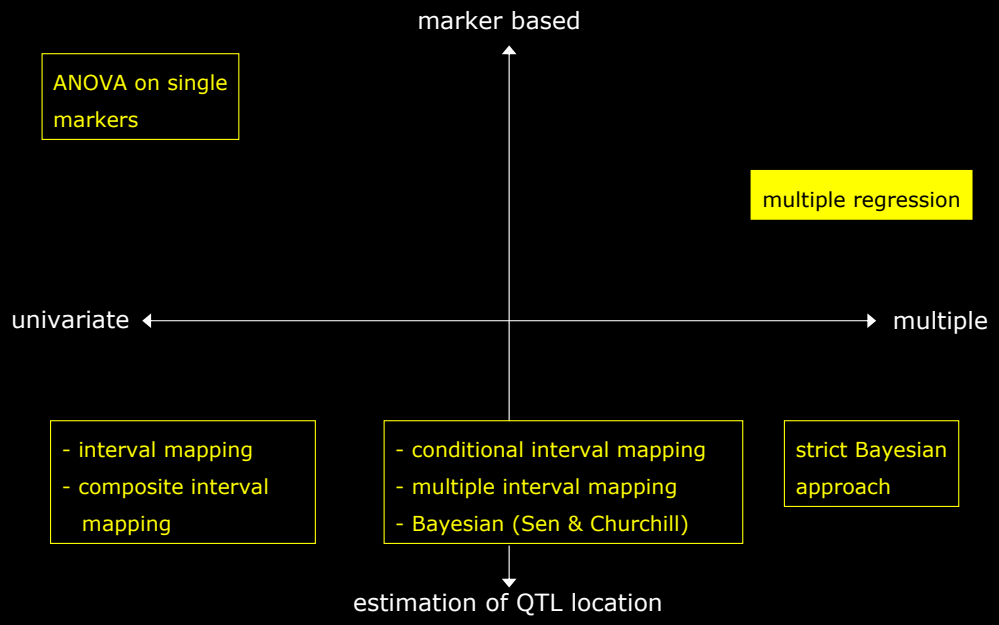
Find NUMBER, POSITIONS, EFFECT TYPES and SIZES of QTL

Challenges:

- large number of possible models
  (main effects + interactions = m + m(m-1)/2 ~ 100 + 5.000)
  -> efficient search strategy
  -> correct for test multiplicity
- deviation from normality of conditional distribution of trait given marker genotypes (especially when heavy tails or outliers)

- recover unobserved / wrong / missing genotype information
- confounding of effect types
- selection bias for effect sizes, especially for small effects

Methods for QTL mapping



Multiple regression approach

$$Y_i = \mu + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv} + \varepsilon_i$$

X_{ij}: genotype of the i^{th} individual (out of n) at the j^{th} marker (out of m).

X_{ij} = ½ if individual has genotype AA (homozygous)

X_{ij} = -½ if individual has genotype Aa (heterozygous)

I: subset of the set $N = \{1,...,m\}$ marker

U: subset of $N \times N$

ε_i : random error term with distribution *f*

Model selection

aim: identify correct model, not minimise prediction error

-> criterion for inclusion and exclusion of variables

- cross validation / bootstrap
- AIC: *n log (RSS) + 2k/n* minimises prediction error
- BIC: *n log (RSS) + k log(n)* more conservative than AIC, especially for small *n*

n: sample size

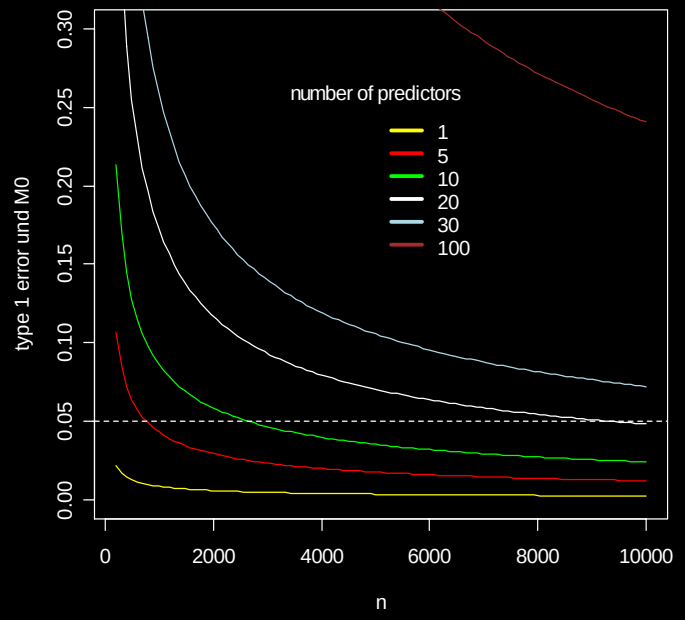
k = *p* + *q* = number of main effects (*p*) and interaction effects (*q*) under consideration

RSS: residual sum of squares (assuming normal error distribution !)

-> efficient search strategy

forward selection + backward elimination step

Behaviour of BIC depending on n & # of predictors



$$P(BIC_{M_i} > BIC_0) \leq$$

$$\leq N \cdot 2P(Z > \sqrt{\log(n)})$$

BIC_{Mi}: BIC of 1-dimensional Model M_i

N: Number of 1-dim models

n: sample size

BIC chooses too many QTL

every model has the same probability to be selected

-> more likely to select large model.

modified BIC

Additional penalty term dependent on number of predictors under consideration (Bogdan et al 2004)

modified BIC

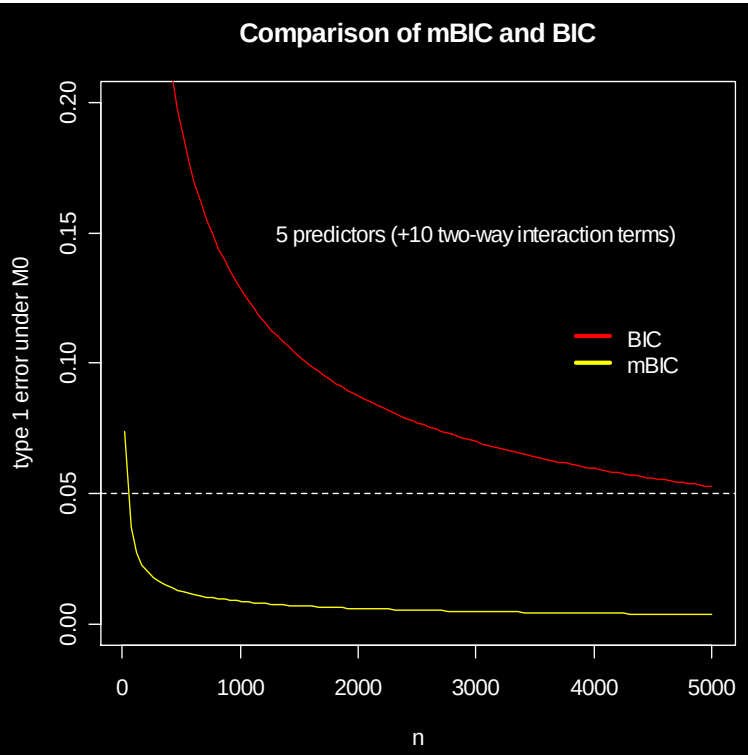
$$mBIC = n \log RSS + (p+q)\log n + 2p\log(m/E(p)-1) + 2q\log(m(m-1)/2/E(q)-1)$$

with

E(p): expected number of main effects

E(q): expected number of epistasis (=interaction) effects

E(p) = *E(q)* = 2.2 controls the Type I error at a level of of 5% (for n = 200)



Deviations from Normality

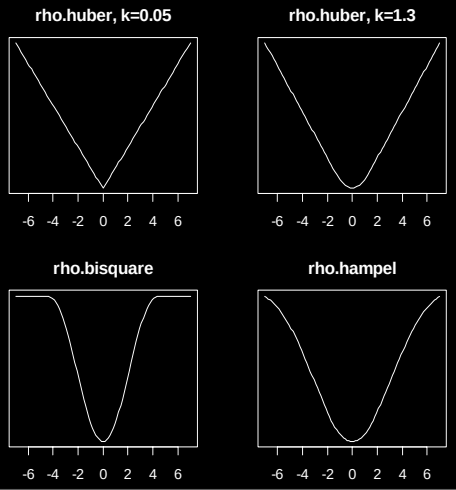
- Typically, non-parametric methods based on ranks are used
- Here we use robust regression techniques, in particular M-Estimators: minimise other measure of distance instead of residual sum of squares.

popular alternatives are:

$$\rho_{Huber}(x) := \begin{cases} k|x| - k^2/2 & \text{for } |x| > k \\ x^2/2 & \text{for } |x| \le k \end{cases}$$

$$\rho_{Bisquare}(x) := \begin{cases} k^2/6 & \text{for } |x| > k \\ \frac{k^2}{6}[1-(1-(\frac{x}{k})^2)^3] & \text{for } |x| \le k \end{cases}$$

$$\rho_{Hampel}(x) := \begin{cases} a(b-a+c)/2 & \text{for } |x| > c \\ a(b-a+c)/2 - \frac{a(|x|-c)^2}{2(c-b)} & \text{for } b < |x| \le c \\ a|x| - a^2/2 & \text{for } a < |x| \le b \\ x^2/2 & \text{for } |x| \le a \end{cases}$$



Robust model selection criterion

$$BIC^*_\rho := n \log \sum_{i=1}^{n} \rho(Y_i - x_i'\hat{\theta}) + k\log(n)$$

still consistent under quite general conditions on the error distribution (Martin, 1980)

but performance of BIC^*_ρ depends on ρ and error distribution:

Jurečkova and Sen (1996) derived limiting distribution for

$$\sum_{i=1}^{n}\left(\rho(Y_i - x_i'\hat{\theta}_1) - \rho(Y_i - x_i'\hat{\theta}_2)\right)$$

Limiting Distribution

We showed that

$$D_n = n(\log \sum \rho(Y_i - x_i'\hat{\theta}_1) - \log \sum \rho(Y_i - x_i'\hat{\theta}_2))$$

has the following property:

$$c_e D_n \xrightarrow{d} \chi^2_{(p_2+q_2)-(p_1+q_1)}$$

with $c_e = \frac{2\gamma\delta}{\sigma_\psi^2}$

$$\psi(x) = \rho'(x)$$
$$\gamma = \int \psi'(x) f(x) dx$$
$$\sigma_\psi^2 = \int \psi(x)^2 f(x) dx$$
$$\delta = \int \rho(x) f(x) dx$$

and error distribution *f(x)*

Values for normalisation constant c_e

error distr.	Huber$_{k=0.05}$	Huber$_{k=1.345}$	Bisquare	Hampel
Normal	1.267	1.096	1.105	1.037
Laplace	1.970	1.436	1.410	1.291
Cauchy	*	*	2.199	2.407
Tukey	1.768	1.925	1.359	1.564
χ^2	1.197	1.148	1.161	1.145
χ^2_{med}	1.291	1.259	1.254	1.192

for L_2 c_e = 1

Robust mBIC

In practice, c_e and therefore the error distribution *f(x)* have to be estimated.

This leads to a robust version of the mBIC:

$$mBIC = \hat{c}_e n \log \sum \rho(Y_i - x_i'\theta) + (p+q)\log n + 2p\log(m/E(p) - 1) + 2q\log(m(m-1)/2/E(q) - 1)$$

with $\hat{c}_e = \frac{2\hat{\gamma}\hat{\delta}}{\hat{\sigma}_\psi^2}$

Simulation Setup

2 chromosomes with 11 marker each (m=22)

200 individuals (n=200)

1 additive effect

1 epistasis effect

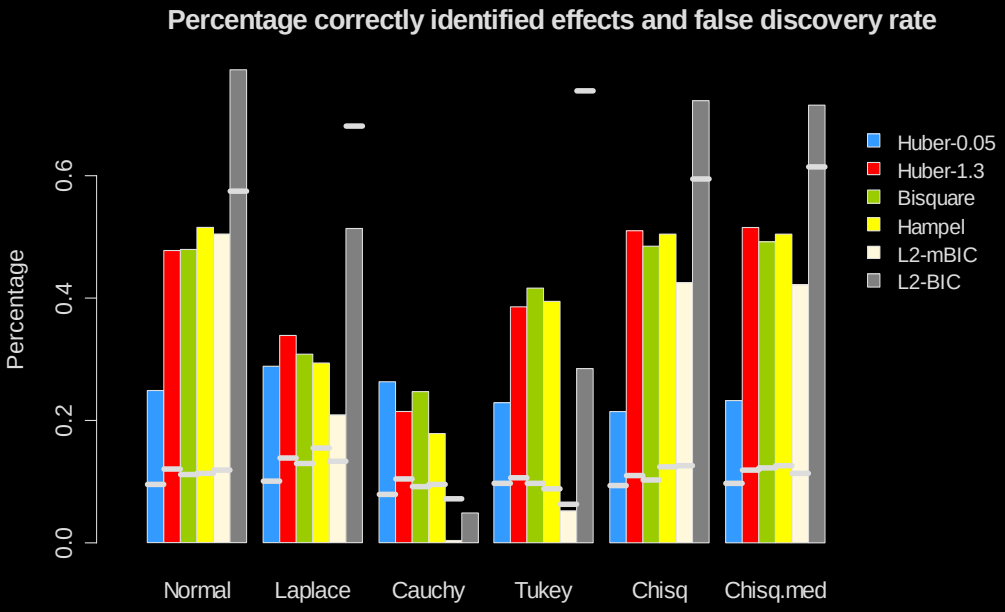
error distributions:

Normal, Laplace, Cauchy, Tukey, χ^2

estimators:

L_2, Huber (k=0.05) ~ L_1, Huber (k=1.3), Bisquare, Hampel

Simulation Results



Implementation in R

- Robust regression using procedure *rlm* of package *MASS*
- program structure:
  - parameter specification
  - generate realisation of genetic setup
  - estimation of error distribution and c_e
  - in each forward step: estimate likelihood for m + m(m-1)/2 models
  - generate output
- simulations:
  - 1000 replications
  - n=200-500, m=20-120

References

- Baierl, A., Bogdan, M., Frommlet, F., Futschik, A., 2006. On Locating multiple interacting quantitative trait loci in intercross designs. To appear in Genetics.
- Bogdan, M., J. K. Ghosh and R. W. Doerge, 2004. Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. Genetics, **167**: 989-999.
- Broman, K. W. and T. P. Speed, 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. J Roy Stat Soc B, **64**: 641-656.
- Jureckova, J., Sen, P.K., 1996. Robust statistical procedures: asymptoticsand interrelations. Wiley, New York.
- Sen and Churchill (2001), A Statistical framework for quantitative trait mapping, Genetics, **159**:371-387.