QTL Mapping II:

The pseudomarker algorithm

Karl W Broman

Department of Biostatistics
Johns Hopkins University

kbroman@jhsph.edu

www.biostat.jhsph.edu/~kbroman

Outline

Sen & Churchill (2001) Genetics 159:371–387

- Data structure and notation
- Basic idea
- Advantages and cautions
- An example

Data structure and notation

y = phenotypes

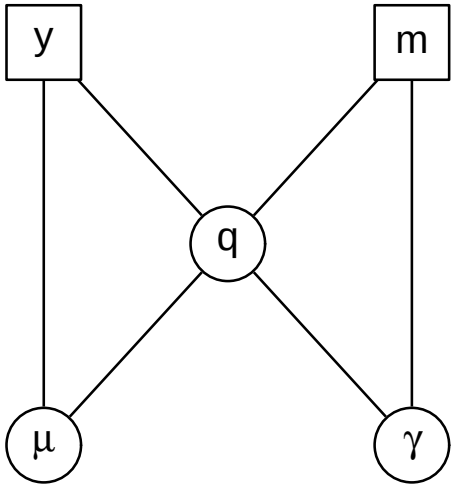
m = observed marker genotypes

q = unobserved QTL genotypes

μ = model parameters

γ = QTL locations

H = QTL model



The factorization

$$\Pr(y, m, q, \mu, \gamma) = \{\Pr(y \mid q, \mu)\ \Pr(\mu)\}\ \{\Pr(q \mid m, \gamma)\ \Pr(m)\ \Pr(\gamma)\}$$

$\Pr(y \mid q, \mu)\ \Pr(\mu)$ = genetic model part

$\Pr(q \mid m, \gamma)\ \Pr(m)\ \Pr(\gamma)$ = linkage part

The unobserved QTL genotypes play a central role.

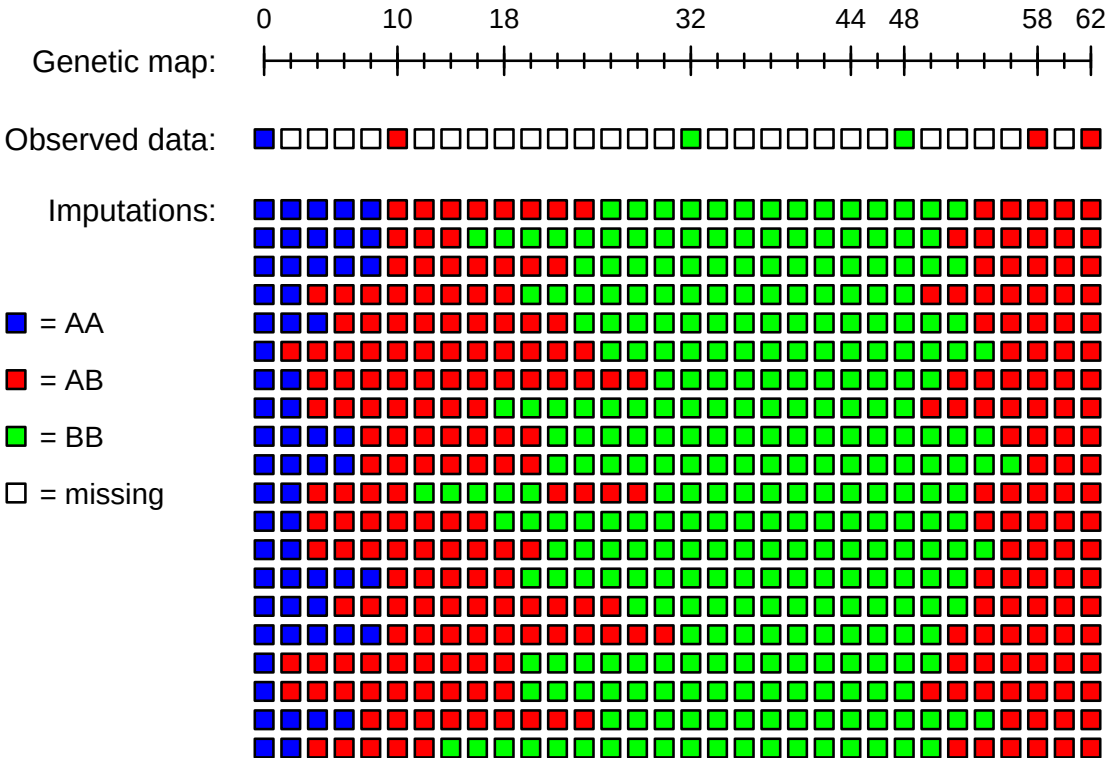
If the QTL genotypes were known, the problem reduces to

linear regression

The basic idea

- Simulate multiple realizations of the joint genotypes on a uniform grid, conditional on the observed multipoint marker data.
- Fit a QTL model with each realization, one at a time.
- Combine the realizations to get an estimate of the posterior probability of the QTL model.

Imputation illustration



Advantages

- Simple computation (just regression)
- Handle missing genotype data
- Covariates
- Any phenotype distribution
- Multi-dimensional genome scans
- Linked QTL; interacting QTL
- Modular algorithm
- No MCMC worries

Cautions

- Monte carlo error (number of imputations)
- Numerical integration error (density of pseudomarker grid)
- Model selection (as usual)
- Relatively large up-front cost for the imputations
  (biggest advantage in case of many phenotypes or many alternative models)

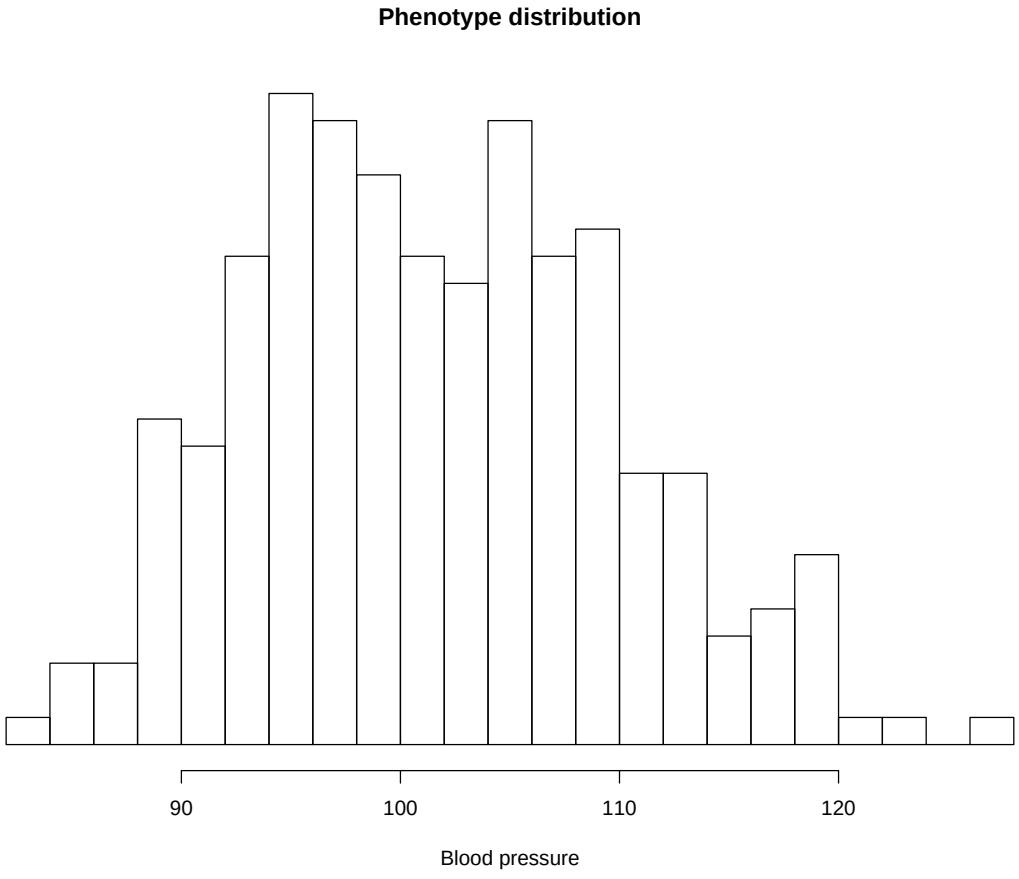
An example

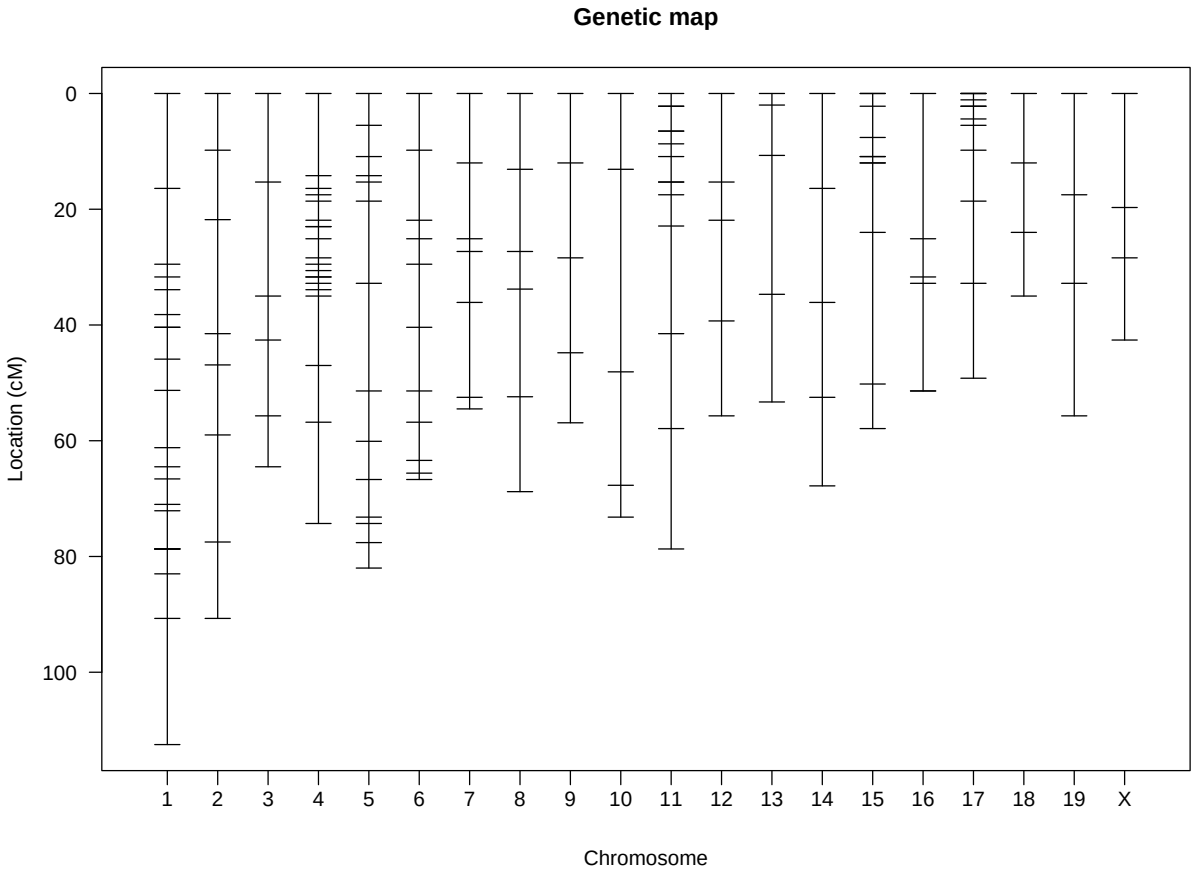
Sugiyama et al. (2001) Genomics 71:70–77

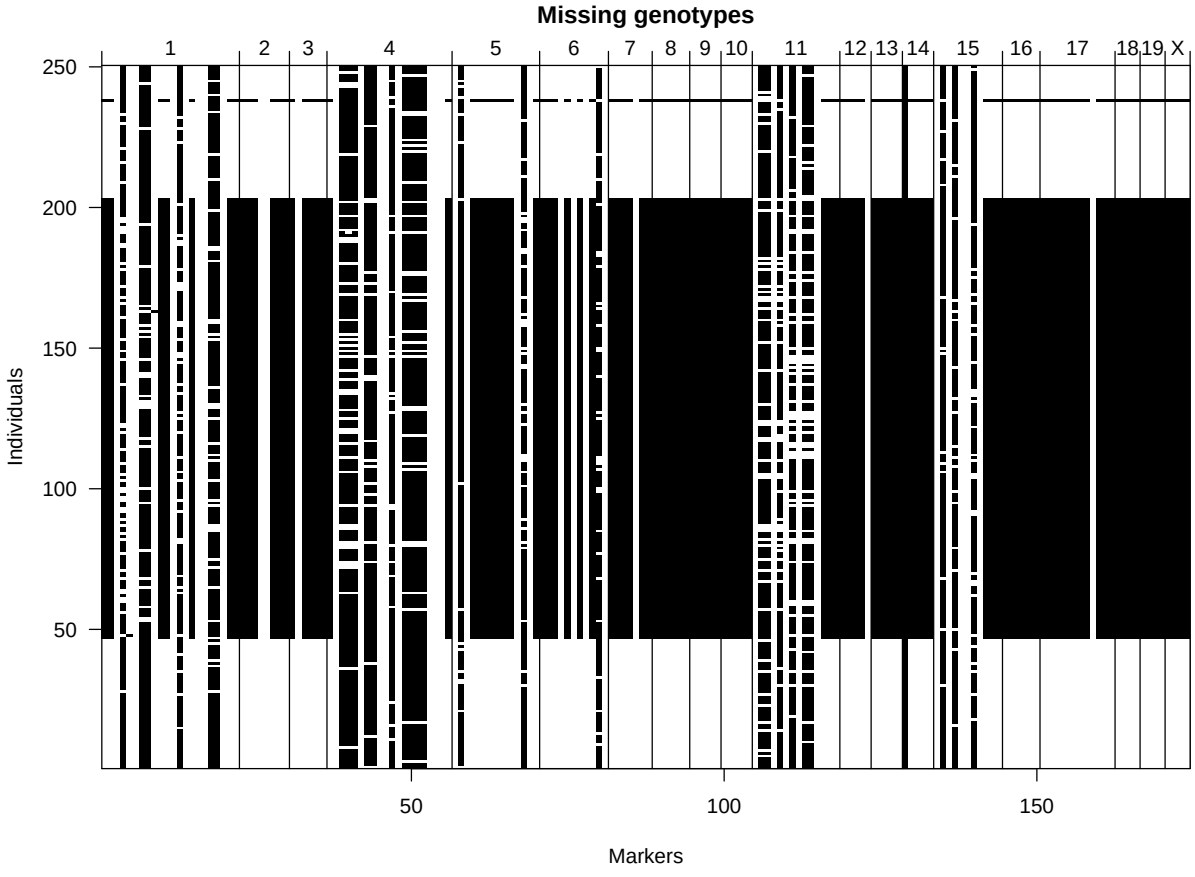
Salt-induced hypertension in the mouse.

Backcross with 250 individuals.

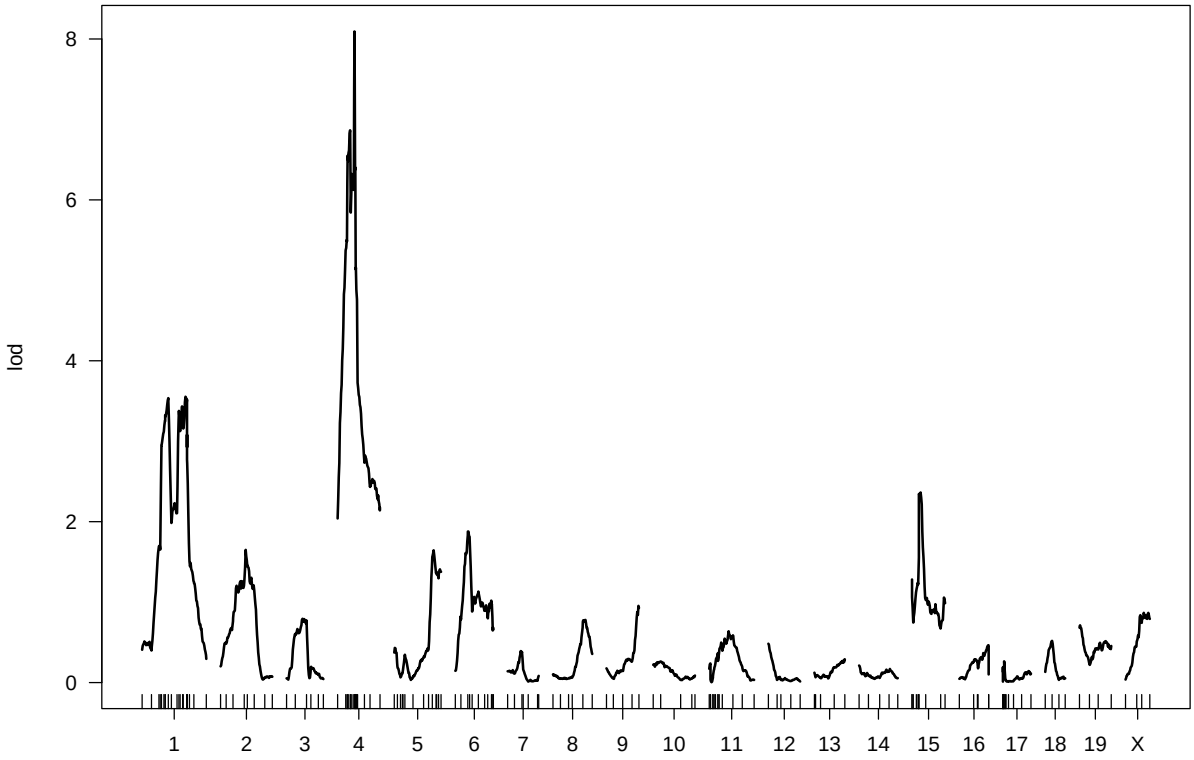
174 markers (for most, only genotyped the extremes).



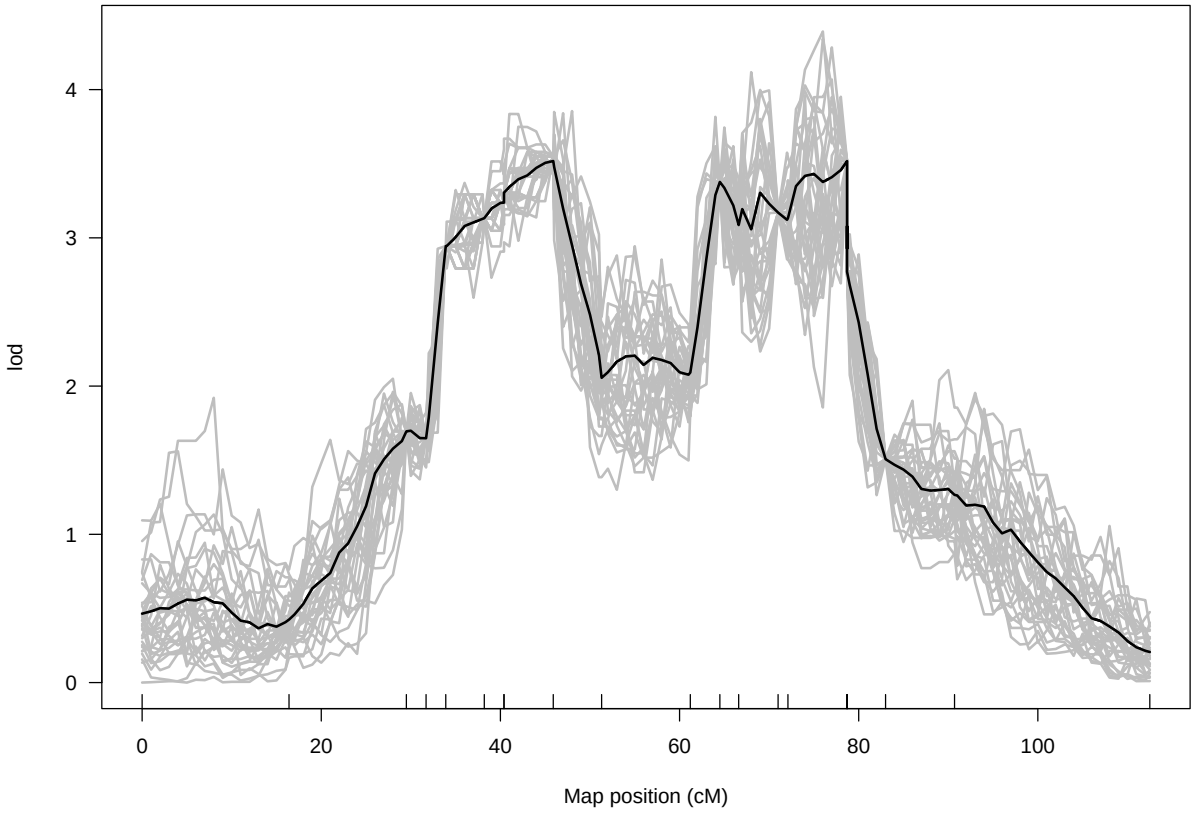

Genetic map
Location (cM)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

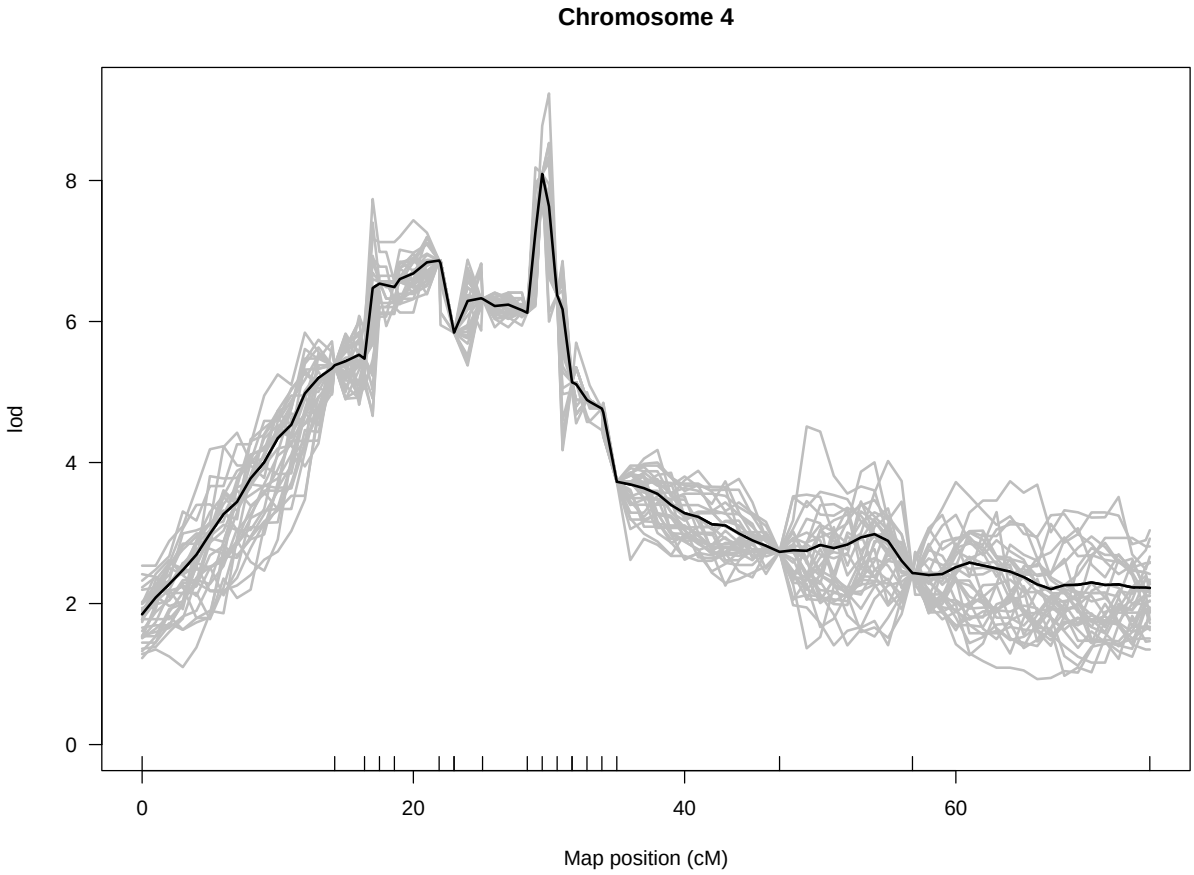

Missing genotypes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Individuals
250
200
150
100
50
50
100
150
Markers

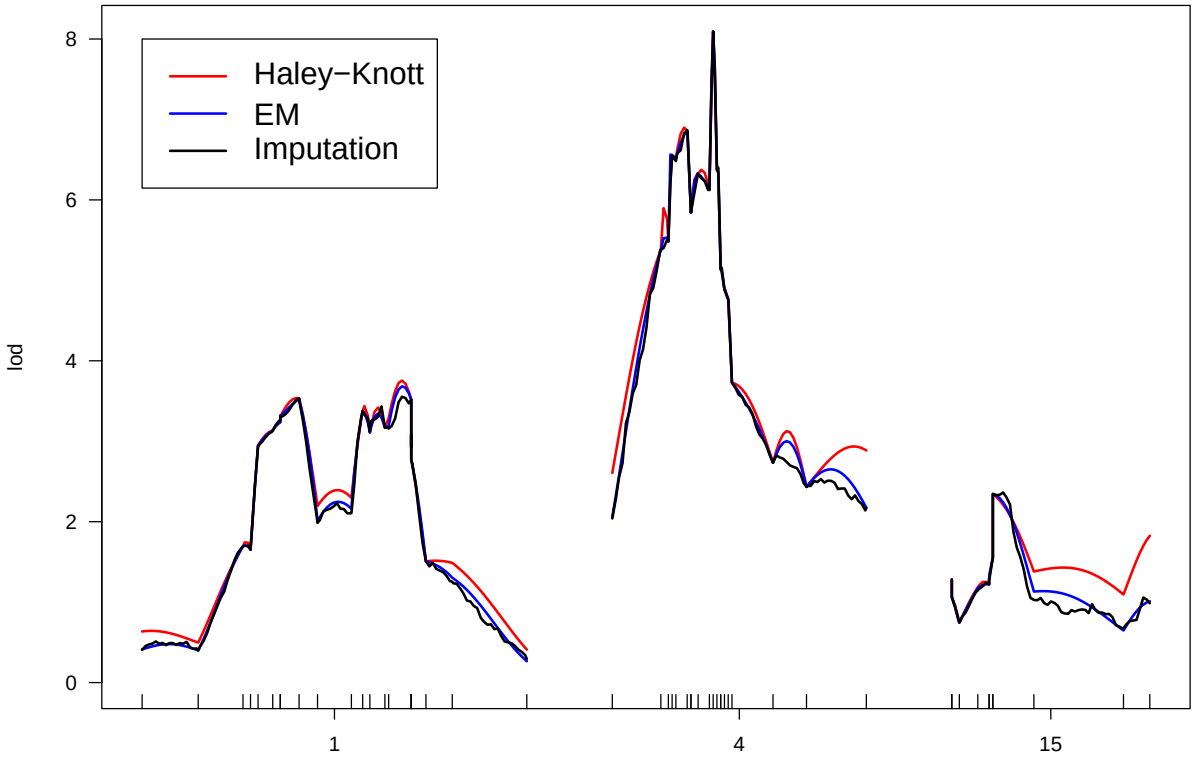
All chromosomes

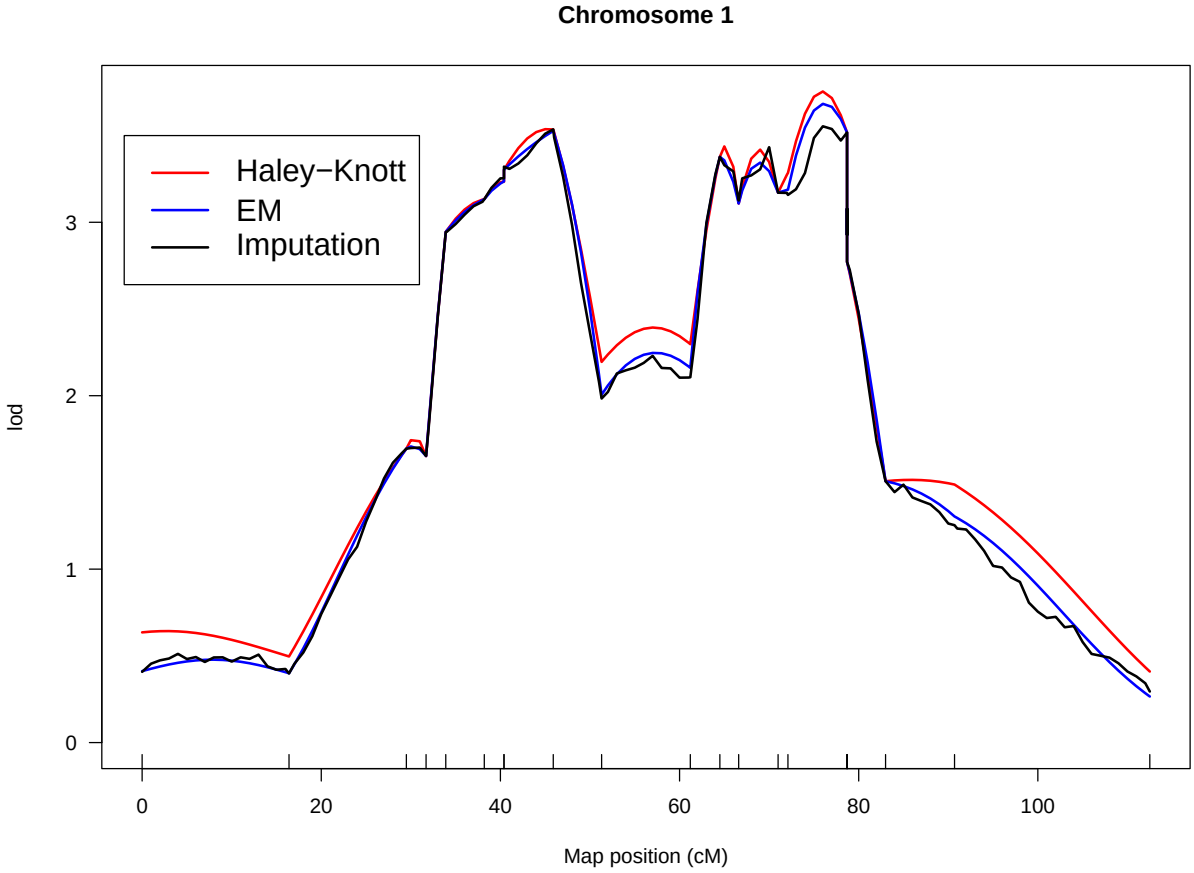


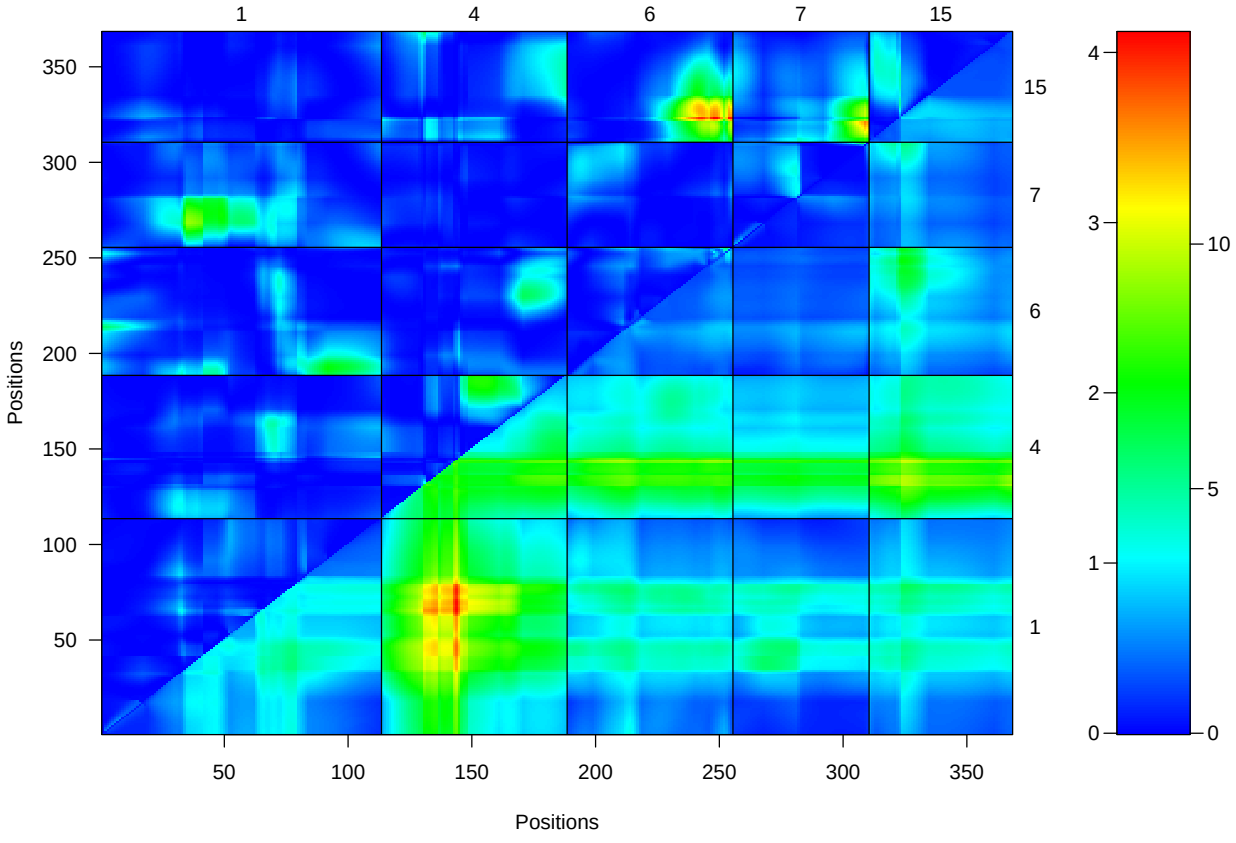
Chromosome 1

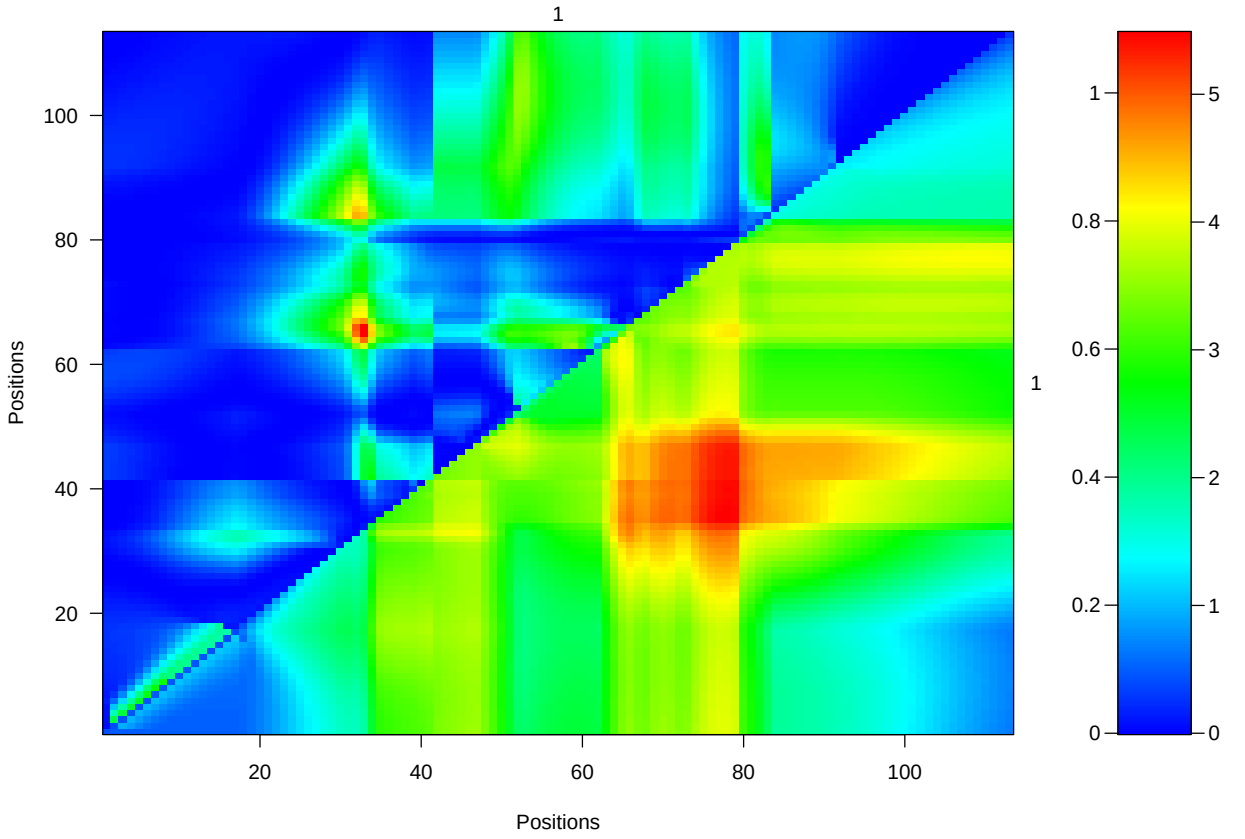



Chromosome 4
lod
0
2
4
6
8
0
20
40
60
Map position (cM)


Haley−Knott
EM
Imputation
lod
0
2
4
6
8
1
4
15


Chromosome 1
Haley−Knott
EM
Imputation
lod
0
1
2
3
0
20
40
60
80
100
Map position (cM)


1
4
6
7
15
350
300
250
200
150
100
50
Positions
15
7
6
4
1
50
100
150
200
250
300
350
Positions
4
3
2
1
0
10
5
0



Drop-one-term table

Term	df	LOD	% variance explained
c1@37	1	1.9	2.3
c1@80	1	3.1	3.8
c4@30	1	9.5	12.3
c6@60	2	5.7	7.1
c7@54	2	2.0	2.4
c15@18	3	7.6	9.6
c6@60 : c15@18	1	3.8	4.6
c7@54 : c15@18	1	1.7	2.1

References

- Sen S, Churchill G (2001) A statistical framework for quantitative trait mapping. Genetics 159:371–387

  The paper on the imputation method (the “pseudomarker algorithm”).

- Sugiyama F, Churchill GA, Higgens DC, Johns C, Makaritsis KP, Gavras H, Paigen B (2001) Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics 71:70–77

  The salt-induced hypertension example.