

QTL Mapping II:

Hidden Markov model technology and The pseudomarker algorithm

Karl W Broman

Department of Biostatistics
Johns Hopkins University

kbroman@jhsph.edu

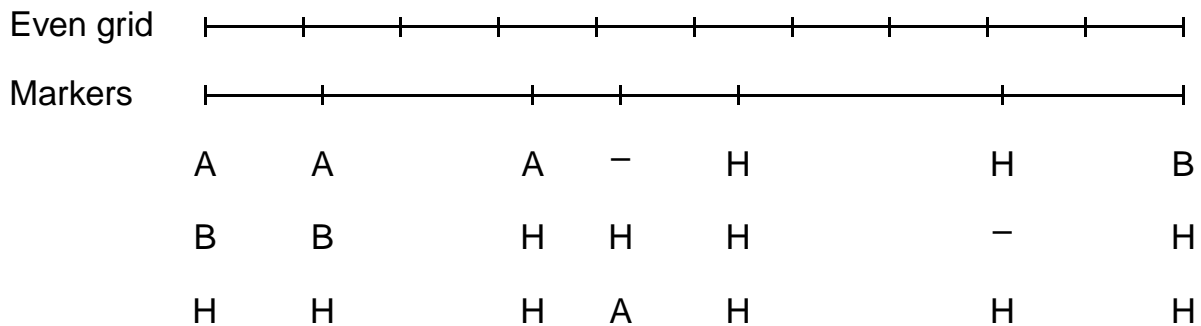
www.biostat.jhsph.edu/~kbroman

HMM technology: Outline

- The problems
- A simple solution
- Why a complex solution?
- The hidden Markov model
- Backcross, intercross
- QTL genotype probabilities
- Simulation of QTL genotypes

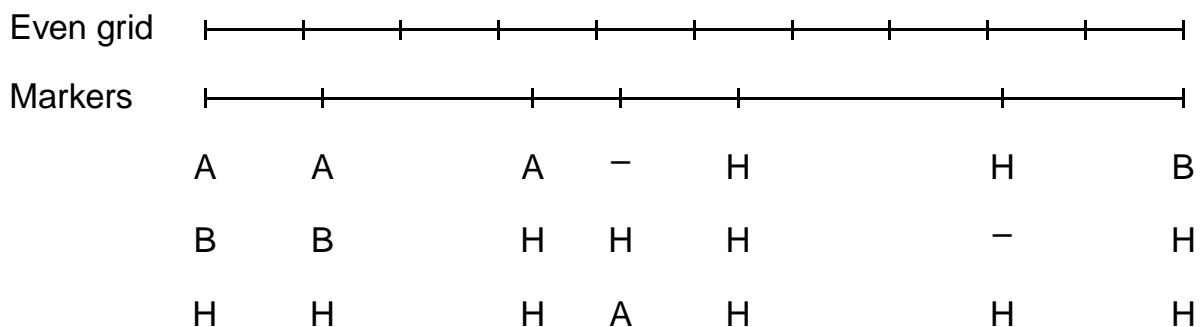
The problems

- Calculate genotype probability at an arbitrary location, conditional on multipoint marker data.
- Simulate from the joint genotype distribution on a grid, given multipoint marker data.



A simple solution

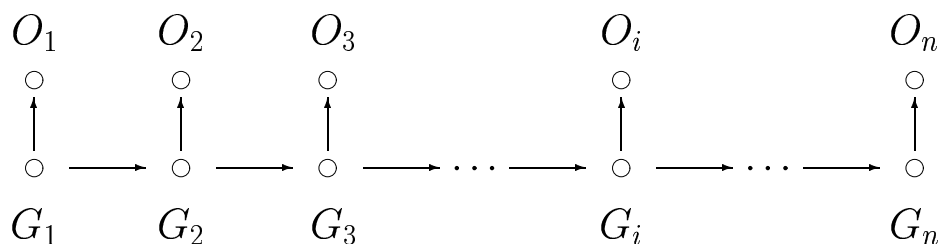
- Under the no interference (NI) model, the genotypes follow a Markov chain.
- Thus, the genotype probability depends only on the nearest flanking typed markers



Why a complex solution?

- Allow for the presence of genotyping errors
- Simply deal with partially informative genotypes (e.g., C = H or B)
- Simplify bookkeeping tasks in the implementation
- Easily extend algorithms to more complex experimental crosses (such as the four-way cross)

The hidden Markov model



- The $\{G_i\}$ (hidden states) form a Markov chain, with values in some finite set, \mathcal{G} .

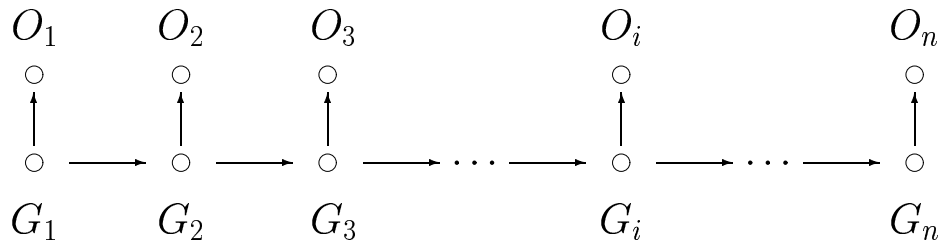
$$\Pr(G_{i+1} \mid G_i, \dots, G_1) = \Pr(G_{i+1} \mid G_i)$$

- The observable random variables, $\{O_i\}$, take values in another finite set, \mathcal{O} .

O_i depends only on G_i

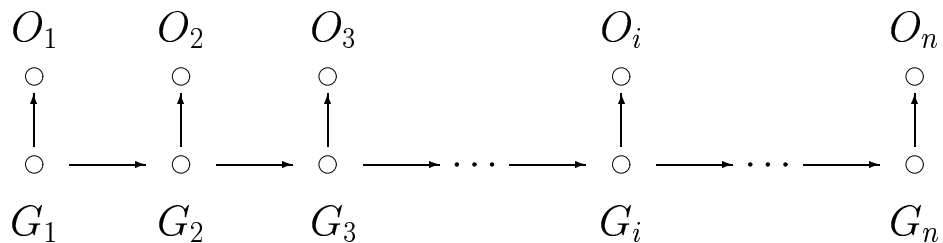
- G_i = “true” genotype at marker i
- O_i = “observed genotype” (marker phenotype) at i

Model parameters



- **Initiation probabilities:** $\pi(g) = \Pr(G_1 = g)$
for $g \in \mathcal{G}$
- **Transition probabilities:** $t_i(g, g') = \Pr(G_{i+1} = g' \mid G_i = g)$
for $i = 1, \dots, n - 1$ and $g, g' \in \mathcal{G}$
- **Emission probabilities:** $e_i(g, o) = \Pr(O_i = o \mid G_i = g)$
for $i = 1, \dots, n$, $g \in \mathcal{G}$, and $o \in \mathcal{O}$
(We assume $e_i(g, o) \equiv e(g, o)$ for all i .)

Joint probability



$$\begin{aligned}
 \Pr(\mathbf{G} = \mathbf{g}, \mathbf{O} = \mathbf{o}) &= \Pr(G_1 = g_1, \dots, G_n = g_n, O_1 = o_1, \dots, O_n = o_n) \\
 &= \Pr(G_1 = g_1) \Pr(G_2 = g_2 \mid G_1 = g_1) \cdots \\
 &\quad \cdots \Pr(G_n = g_n \mid G_{n-1} = g_{n-1}) \cdot \Pr(O_1 = o_1 \mid G_1 = g_1) \cdots \\
 &\quad \cdots \Pr(O_n = o_n \mid G_n = g_n) \\
 &= \pi(g_1) \prod_{i=1}^{n-1} t_i(g_i, g_{i+1}) \prod_{i=1}^n e(g_i, o_i)
 \end{aligned}$$

The backcross

$$\mathcal{G} = \{AA, AB\} \quad \mathcal{O} = \{A, H, -\} \quad (- = \text{missing})$$

Initiation probabilities:

$$\pi(AA) = \pi(AB) = 1/2$$

Transition probabilities:

r_i = recombination fraction for interval i .

$$t_i(AA, AB) = t_i(AB, AA) = r_i$$

$$t_i(AA, AA) = t_i(AB, AB) = 1 - r_i$$

Emission probabilities:

ϵ = genotyping error rate

$$e(AA, A) = e(AB, H) = 1 - \epsilon, \quad e(AA, -) = e(AB, -) = 1$$

$$e(AA, H) = e(AB, A) = \epsilon$$

The intercross

We'll consider phase-unknown genotypes.

$$\mathcal{G} = \{AA, AB, BB\}$$

Initiation probabilities:

$$\pi(AA) = \pi(BB) = 1/4, \quad \pi(AB) = 1/2$$

Transition probabilities, $t_i(g, g') = \Pr(G_{i+1} = g' \mid G_i = g)$:

g	g'		
	AA	AB	BB
AA	$(1 - r_i)^2$	$2r_i(1 - r_i)$	r_i^2
AB	$r_i(1 - r_i)$	$(1 - r_i)^2 + r_i^2$	$r_i(1 - r_i)$
BB	r_i^2	$2r_i(1 - r_i)$	$(1 - r_i)^2$

The intercross (cont.)

$$\mathcal{O} = \{A, H, B, C, D, -\}$$

$$- = \text{missing} = \{A \text{ or } H \text{ or } B\}$$

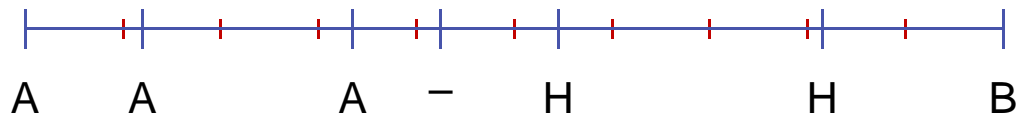
$$C = \text{not } A = \{H \text{ or } B\}$$

$$D = \text{not } B = \{A \text{ or } H\}$$

Emission probabilities, $e(g, o) = \Pr(O_i = o \mid G_i = g)$:

g	o					
	A	H	B	C	D	$-$
AA	$1 - \epsilon$	$\epsilon/2$	$\epsilon/2$	ϵ	$1 - \epsilon/2$	1
AB	$\epsilon/2$	$1 - \epsilon$	$\epsilon/2$	$1 - \epsilon/2$	$1 - \epsilon/2$	1
BB	$\epsilon/2$	$\epsilon/2$	$1 - \epsilon$	$1 - \epsilon/2$	ϵ	1

QTL genotype probabilities



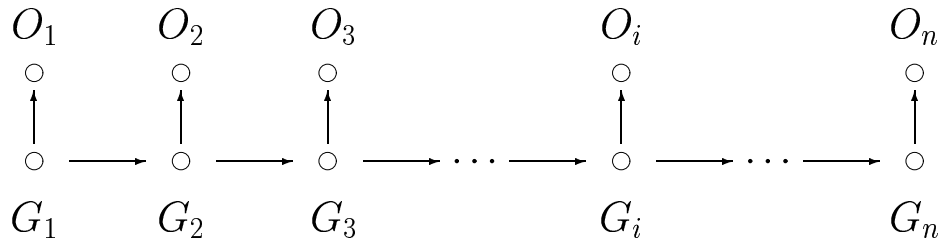
We seek to calculate $\Pr(G_i = g \mid \mathbf{O})$, where $\mathbf{O} = (O_1, O_2, \dots, O_n)$ is the observed multipoint marker data.

Brute force:

$$\begin{aligned} \Pr(G_i = g_i \mid \mathbf{O}) &= \sum_{g_1} \dots \sum_{g_{i-1}} \sum_{g_{i+1}} \dots \sum_{g_n} \Pr(G_1 = g_1, \dots, G_n = g_n \mid \mathbf{O}) \\ &\propto \sum_{g_1} \dots \sum_{g_{i-1}} \sum_{g_{i+1}} \dots \sum_{g_n} \pi(g_1) \prod_{j=1}^{n-1} t_j(g_j, g_{j+1}) \prod_{j=1}^n e(g_j, O_j) \end{aligned}$$

For the phase-unknown intercross, this is a sum with 3^{n-1} terms; clearly this is unwieldy and unnecessary. **But, of course, there is a simpler way!**

The forward and backward equations



Our approach makes use of the following two sets of probabilities:

$$\begin{aligned}\alpha_i(g) &= \Pr(O_1, \dots, O_i, G_i = g) \\ \beta_i(g) &= \Pr(O_{i+1}, \dots, O_n | G_i = g)\end{aligned}$$

Note that once the α 's and β 's have been calculated, the probability that is our focus follows directly:

$$\begin{aligned}\Pr(G_i = g | \mathbf{O}) &= \Pr(G_i = g, \mathbf{O}) / \Pr(\mathbf{O}) \\ &= \alpha_i(g)\beta_i(g) / \sum_{g'} \alpha_i(g')\beta_i(g')\end{aligned}$$

The forward equations

The α 's are calculated inductively.

First, note that

$$\alpha_1(g) = \Pr(O_1, G_1 = g) = \pi(g) e(g, O_1)$$

Now, assume that we've calculated $\alpha_i(g)$ for each $g \in \mathcal{G}$. Then

$$\begin{aligned}\alpha_{i+1}(g) &= \Pr(O_1, \dots, O_i, O_{i+1}, G_{i+1} = g) \\ &= \sum_{g'} \Pr(O_1, \dots, O_i, O_{i+1}, G_i = g', G_{i+1} = g) \\ &= \sum_{g'} \Pr(O_1, \dots, O_i, G_i = g') \Pr(G_{i+1} = g | G_i = g') \Pr(O_{i+1} | G_{i+1} = g) \\ &= e(g, O_{i+1}) \sum_{g'} \alpha_i(g') t_i(g', g)\end{aligned}$$

The backward equations

The β 's are calculated similarly, but moving backward.

First, we define $\beta_n(g) \equiv 1$ for all $g \in \mathcal{G}$.

Now, assume that we've calculated $\beta_i(g)$ for each $g \in \mathcal{G}$. Then

$$\begin{aligned}\beta_{i-1}(g) &= \Pr(O_i, \dots, O_n | G_{i-1} = g) \\ &= \sum_{g'} \Pr(O_i, \dots, O_n, G_i = g' | G_{i-1} = g) \\ &= \sum_{g'} \Pr(O_{i+1}, \dots, O_n | G_i = g') \Pr(G_i = g' | G_{i-1} = g) \Pr(O_i | G_i = g') \\ &= \sum_{g'} \beta_i(g') t_{i-1}(g, g') e(g', O_i)\end{aligned}$$

QTL genotype probabilities

1. Calculate the α 's and β 's, simultaneously, via the forward and backward equations.
2. Calculate, for each i and g ,

$$\begin{aligned}\Pr(G_i = g | \mathbf{O}) &= \Pr(G_i = g, \mathbf{O}) / \Pr(\mathbf{O}) \\ &= \alpha_i(g)\beta_i(g) / \sum_{g'} \alpha_i(g')\beta_i(g')\end{aligned}$$

Simulation of QTL genotypes

We seek to simulate from the joint distribution, $\Pr(G_1, \dots, G_n \mid \mathbf{O})$

[Why? We'll explain shortly.]

First draw g_1^* from the distribution

$$\Pr(G_1 = g \mid \mathbf{O}) = \frac{\alpha_1(g)\beta_1(g)}{\sum_{g'} \alpha_1(g')\beta_1(g')}$$

Genotypes for further loci are drawn iteratively:

having drawn g_1^*, \dots, g_i^* , draw g_{i+1}^* from

$$\begin{aligned} \Pr(G_{i+1} = g \mid \mathbf{O}, G_i = g_i^*) &= \frac{\Pr(G_{i+1} = g, G_i = g_i^* \mid \mathbf{O})}{\Pr(G_i = g_i^* \mid \mathbf{O})} \\ &= \frac{\alpha_i(g_i^*) t_i(g_i^*, g) e(g, O_{i+1}) \beta_{i+1}(g)}{\alpha_i(g_i^*) \beta_i(g_i^*)} \\ &= t_i(g_i^*, g) e(g, O_{i+1}) \beta_{i+1}(g) / \beta_i(g_i^*) \end{aligned}$$

Note that we need to first calculate the β 's (via the backward equations).

A practical issue

In the case of **many** genetic markers (or pseudomarkers), the direct calculation of α and β , as described above, will result in **underflow**.

$\alpha_n(g) = \Pr(O_1, O_2, \dots, O_n, G_n = g)$ can be **extremely** small!

One method to deal with this is to work with $\alpha' = \log \alpha$ and $\beta' = \log \beta$.

But in the forward equations, we need

$$\alpha'_{i+1}(g) = \log e(g, O_{i+1}) + \log\{\sum_{g'} \alpha_i(g') t_i(g', g)\}$$

This leads to the problem of calculating $\log(f_1 + f_2)$ on the basis of $g_i = \log f_i$, which may be facilitated with the following trick:

$$\begin{aligned} \log(f_1 + f_2) &= \log(e^{g_1} + e^{g_2}) \\ &= \log\{e^{g_1}(1 + e^{g_2 - g_1})\} \\ &= g_1 + \log(1 + e^{g_2 - g_1}) \end{aligned}$$

A problem occurs when $g_2 \gg g_1$: the above formula will result in an overflow. In such a case one simply notes that $\log(f_1 + f_2) \approx g_2$.

The pseudomarker algorithm: 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

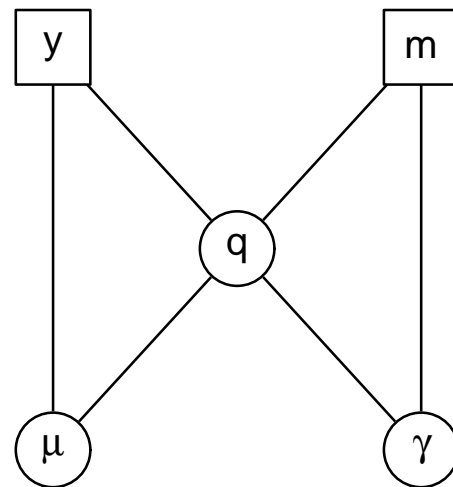
m = observed marker genotypes

q = unobserved QTL genotypes

μ = model parameters

γ = QTL locations

H = QTL model



The factorization

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

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

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

The **unobserved QTL genotypes** play a **central role**.

If the QTL genotypes were **known**, the problem reduces to

model selection in 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.

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 slow, for one- or two-dimensional genome scans

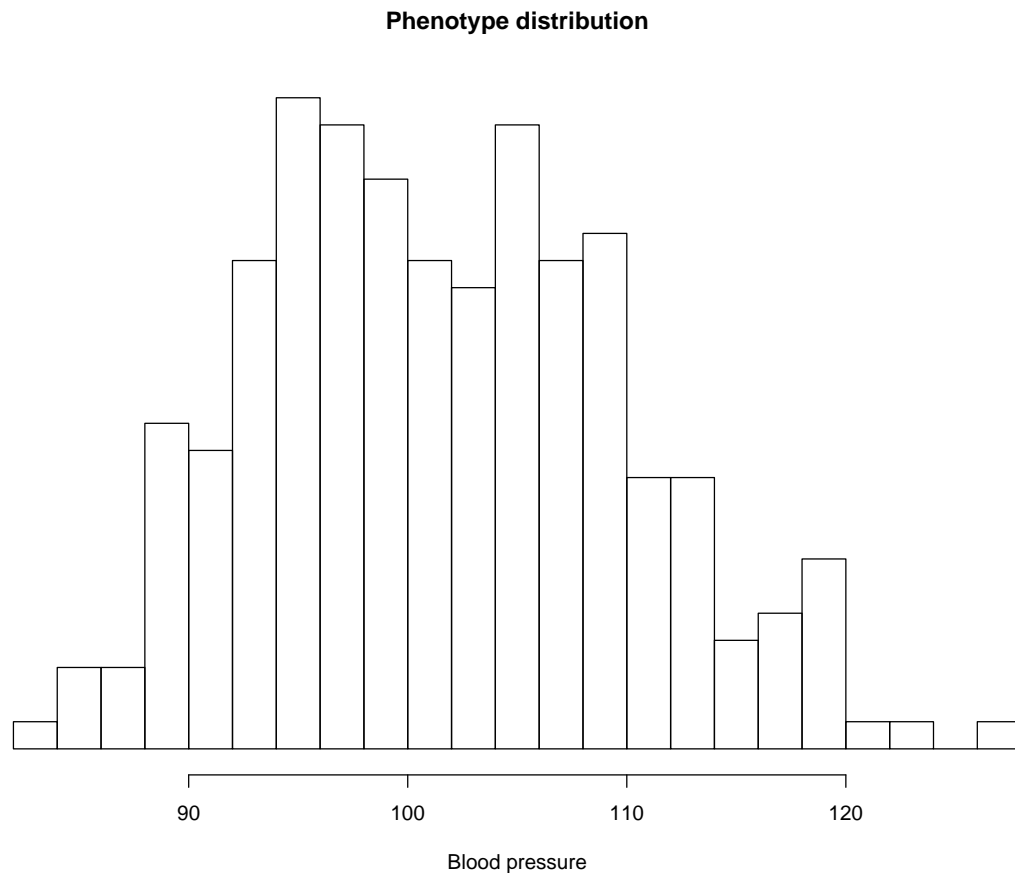
An example

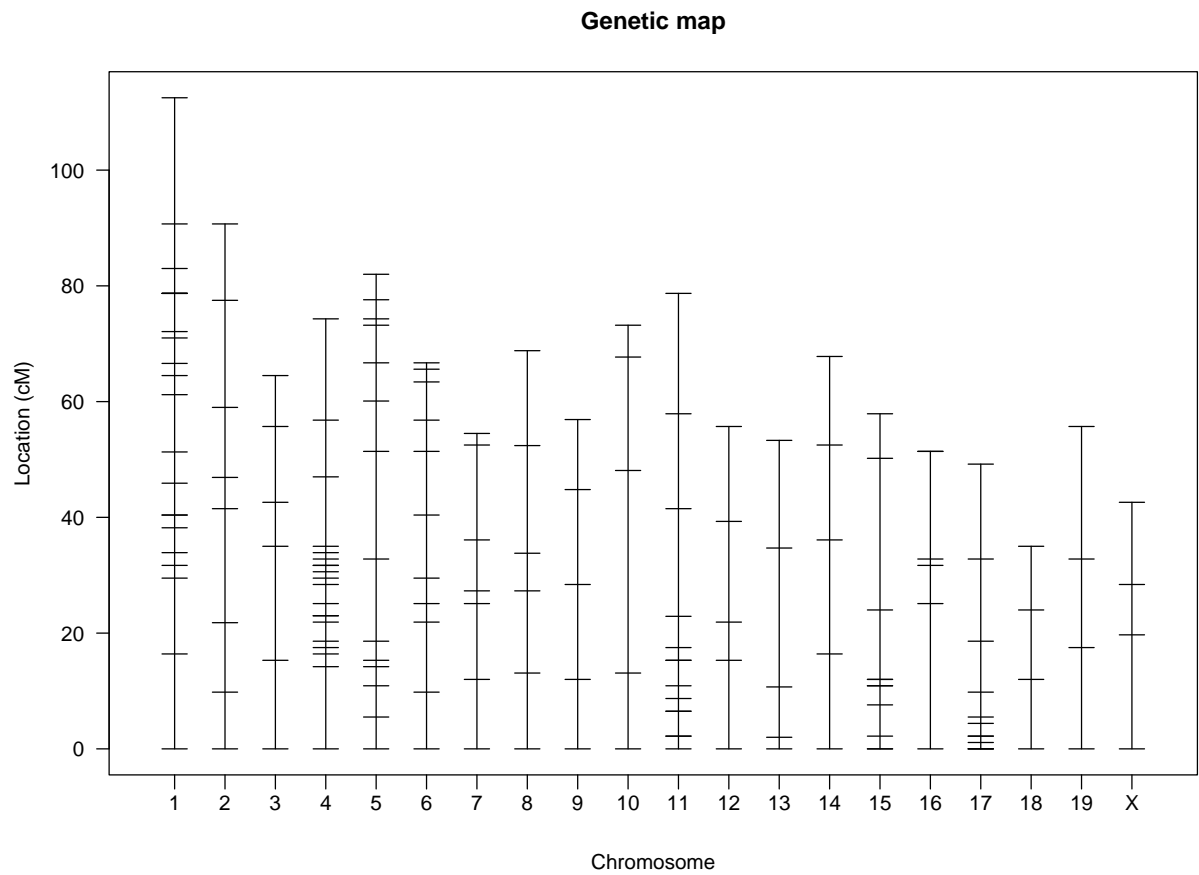
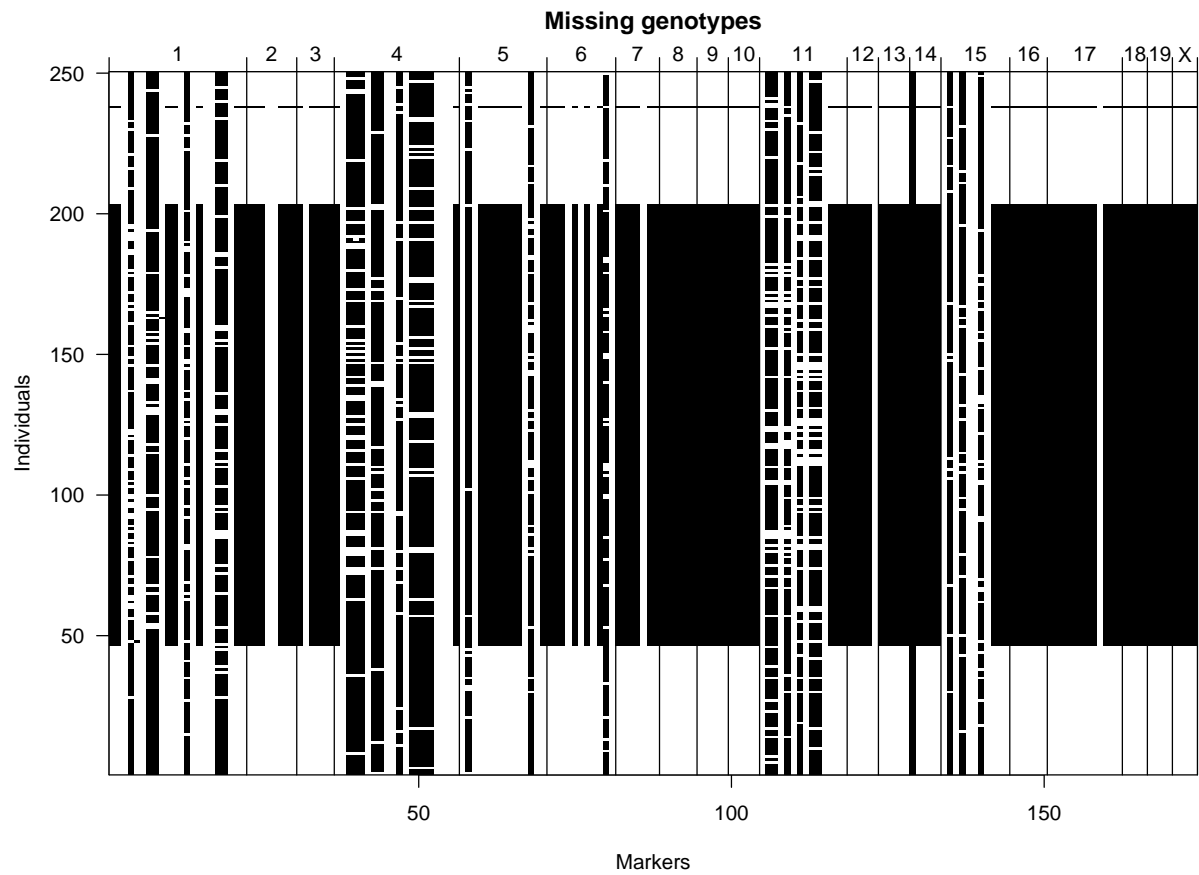
Sugiyama et al. (2001) Genomics 71:70–77

Salt-induced hypertension in the mouse.

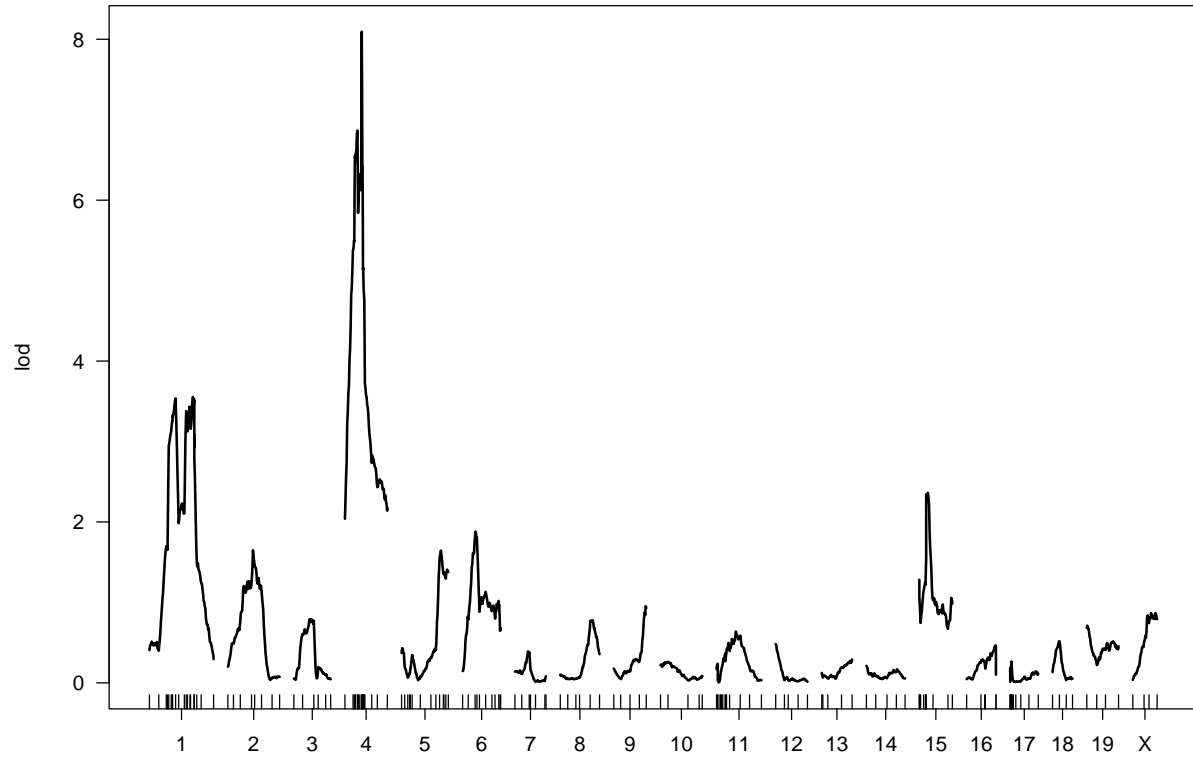
Backcross with 250 individuals.

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

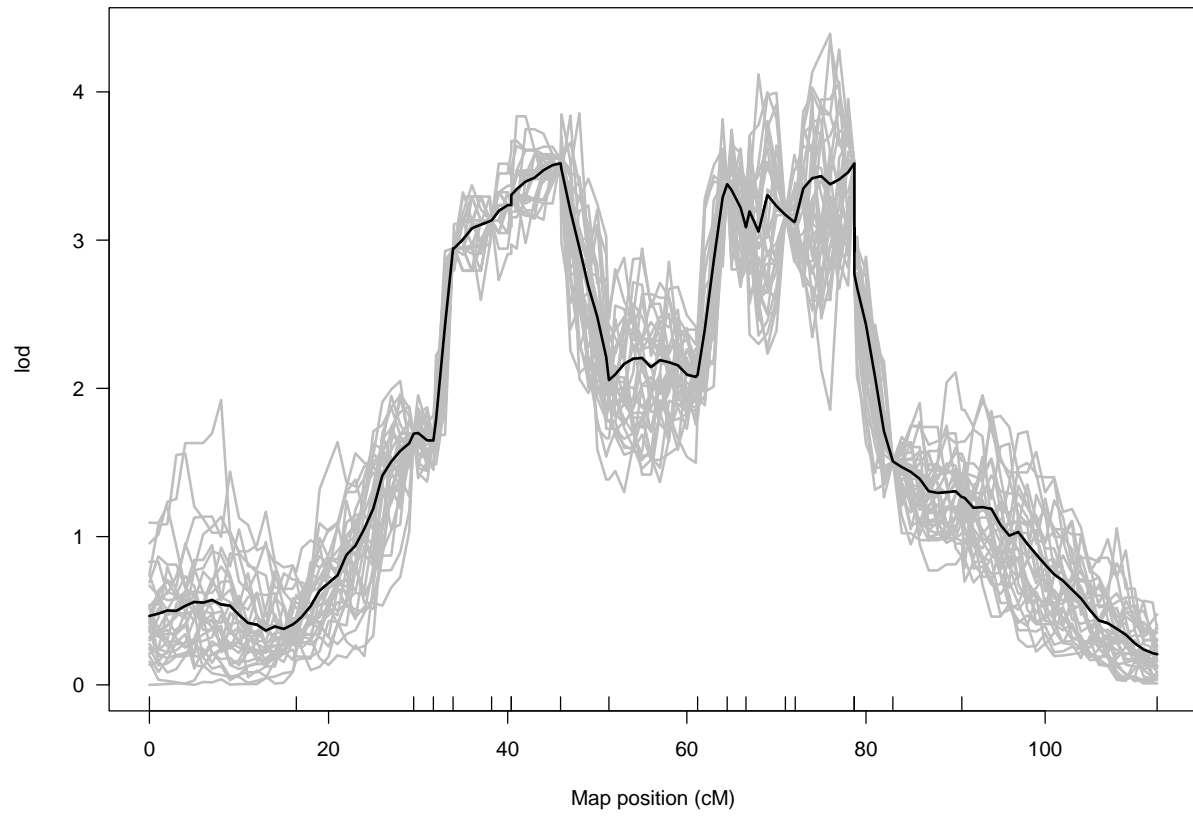




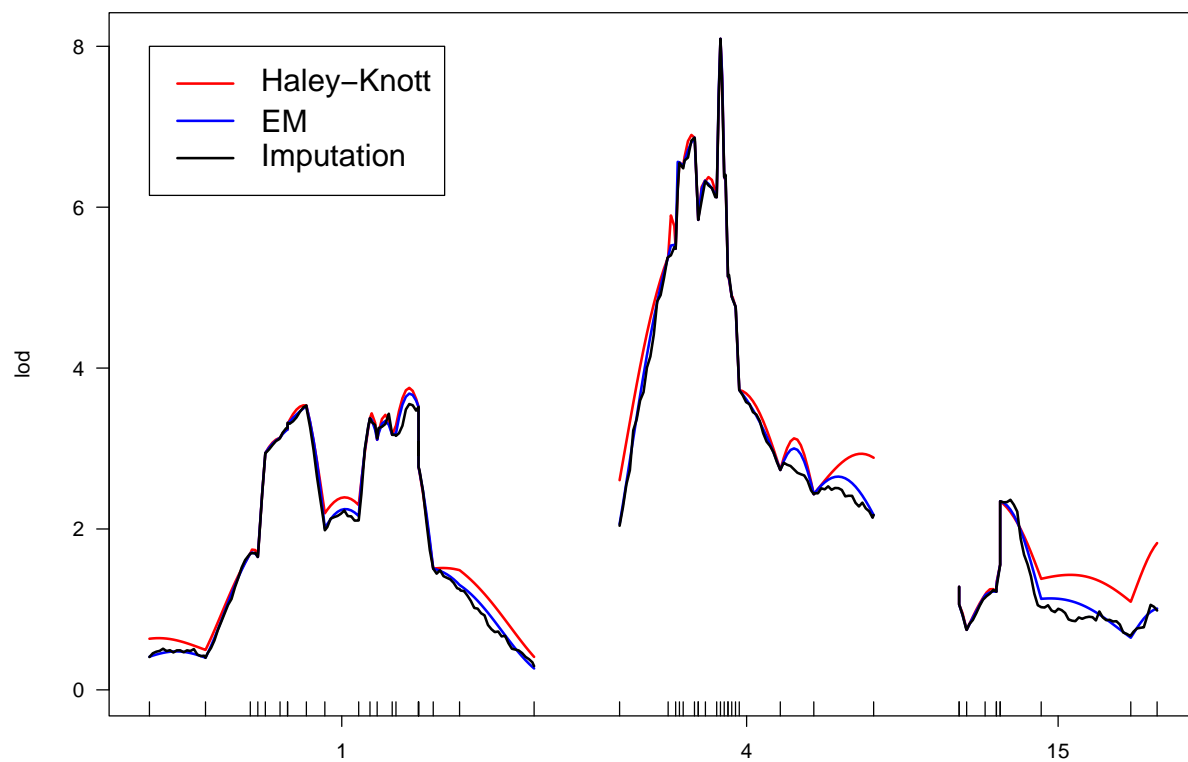
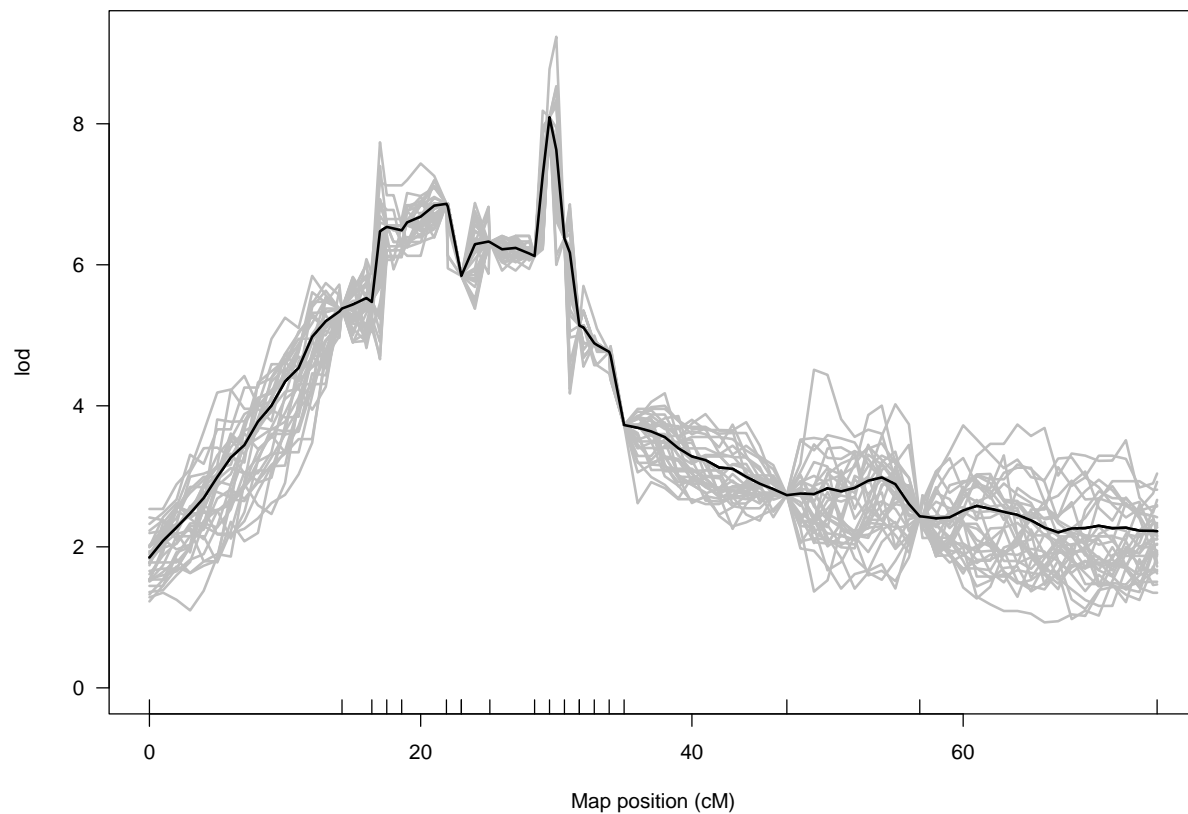
All chromosomes



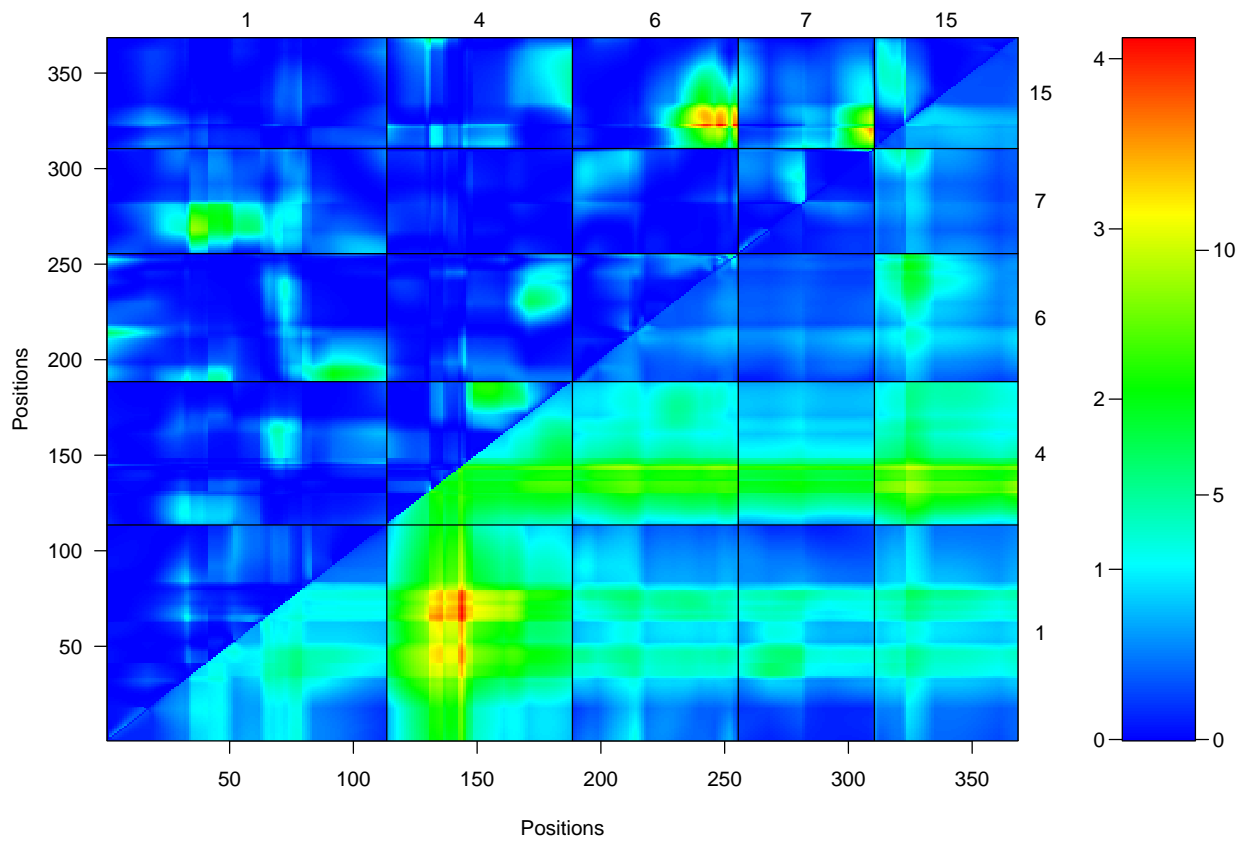
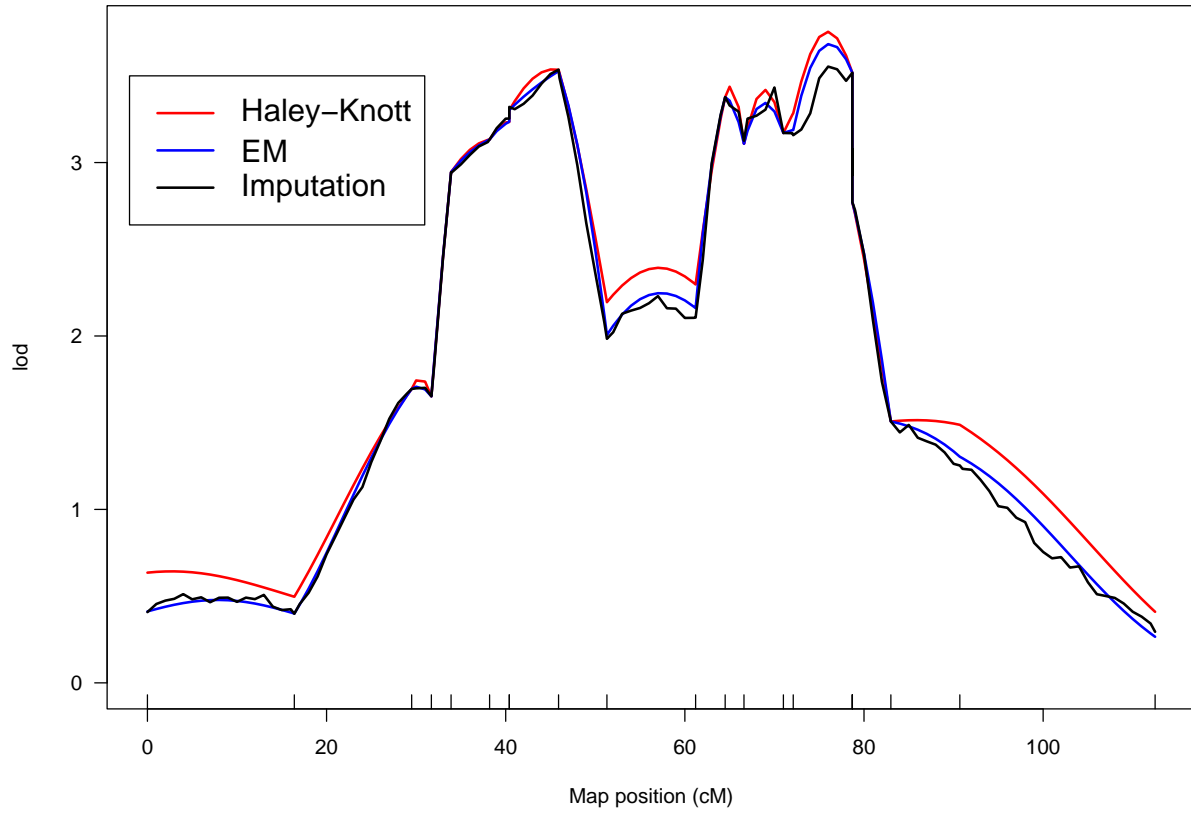
Chromosome 1

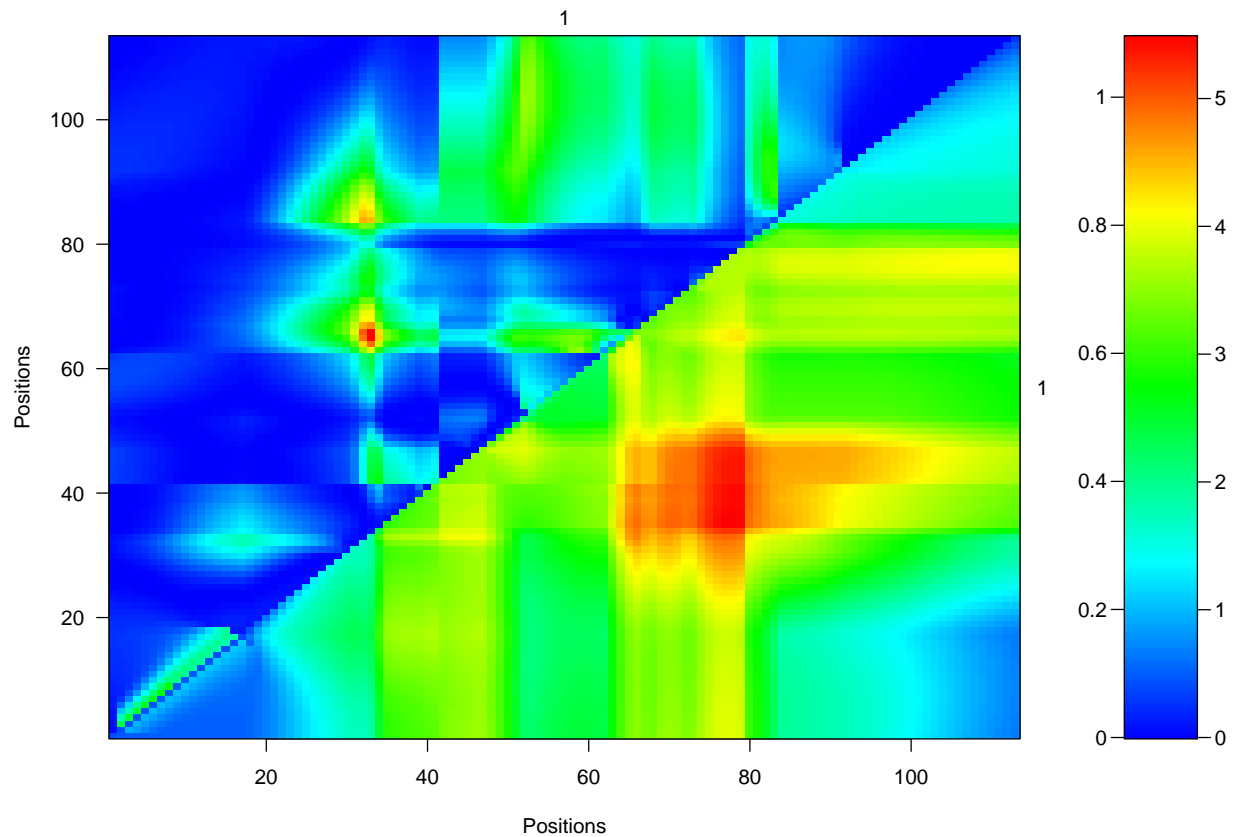


Chromosome 4



Chromosome 1





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

- Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat* 41:164–171
[The first paper on hidden Markov models.](#)
- Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77:257–286
[A quite readable review of HMMs.](#)
- Lange K (1999) *Numerical analysis for statisticians*. Springer, New York, section 23.3.
[Review of HMMs.](#)
- Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* 51:79–94
[The first application of HMMs in biology.](#)
- Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 84:2363–2367
[First use of HMMs for genetic mapping.](#)

- Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. *Genomics* 14: 604–610.
[Paper describing how to deal with genotyping errors in experimental crosses.](#)
- Jiang C, Zeng ZB (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* 101:47–58
[An alternative approach for dealing with missing and partially missing genotype data.](#)
- 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, Higgins 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.](#)