Bayesian Two-way Clustering for Gene Expression Data

Graeme Ambler

July 12, 2003

MacKay and Miskin's model

- MacKay and Miskin (2001) proposed the following model for gene expression data:

$$y_{gs} = \sum_{h=1}^{H} a_{sh} b_{gh} + \varepsilon_{gs}.$$

  where g = gene, s = sample, and ε is a noise term.

- We think that this model is rather too general to use as-is.

- We propose some simplifications of the model which (we hope) will lead to more interpretable results.

A single sample model

- We first consider a simple model for the single-sample case:

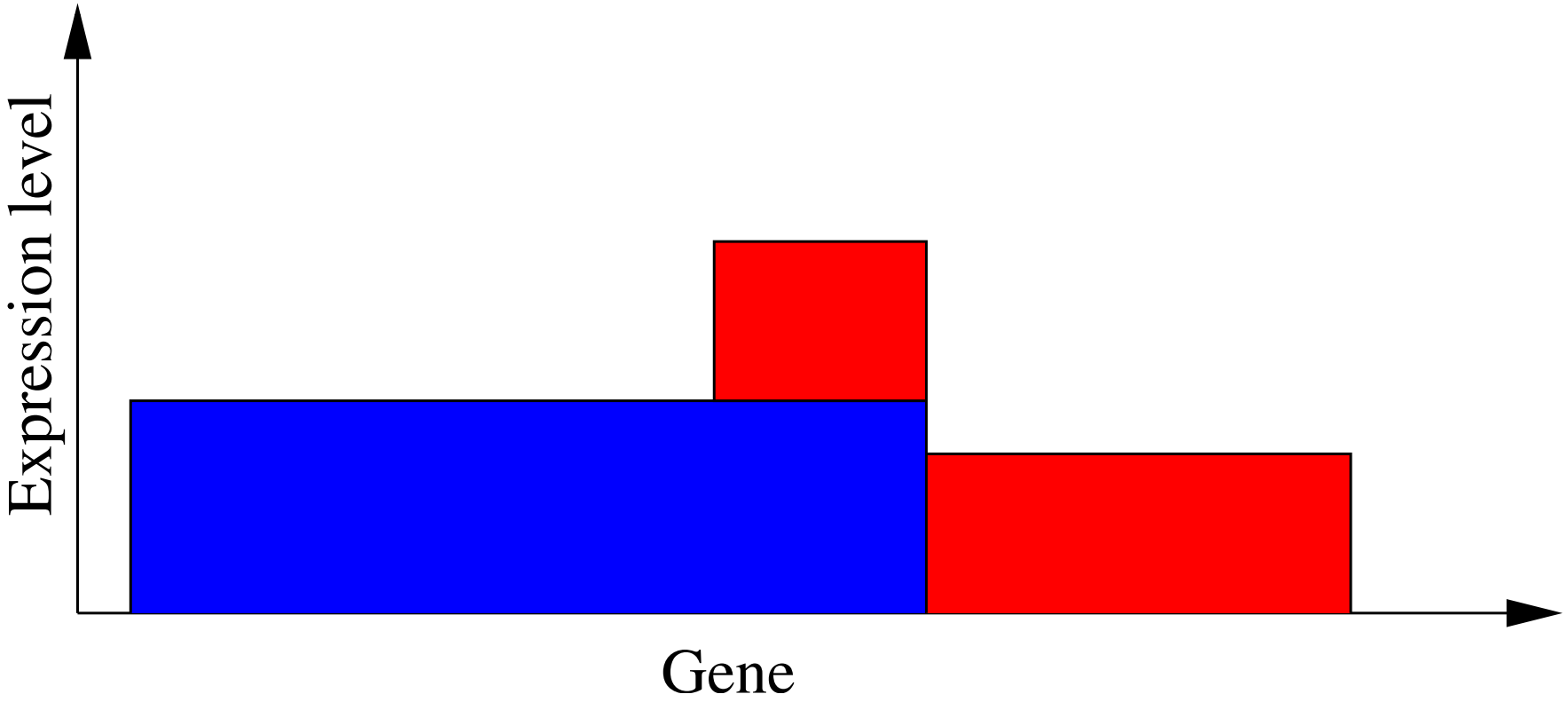
$$y_g = \sum_{h=1}^{H} z_{gh} b_h + \varepsilon_g,$$

where $z_{gh} \in \{0, 1\}$ is an allocation variable and $b_h \in \mathbb{R}$ is a constant level for layer h.

- This is very similar to the standard mixture model with constant variance:

$$y_g = \sum_{h=1}^{H} z_{gh} \mu_h + \varepsilon_g$$

except that the allocations for a mixture model are constrained to give $\sum_h z_{gh} \equiv 1 \ \forall g$.


Expression level
Gene

Prior specification

We propose a fully Bayesian MCMC implementation using the following priors:

$$\begin{aligned} z_{gh} &\sim \text{Bernoulli}\left(\frac{q}{H+1}\right) \\ b_h &\sim N(0, \tau^2) \\ H &\sim \text{Poisson}(\lambda) \\ \varepsilon &\sim N(0, \sigma^2). \end{aligned}$$

We also use a conjugate hyperprior on σ:

$$\sigma^{-2} \sim \Gamma(\alpha, \beta).$$

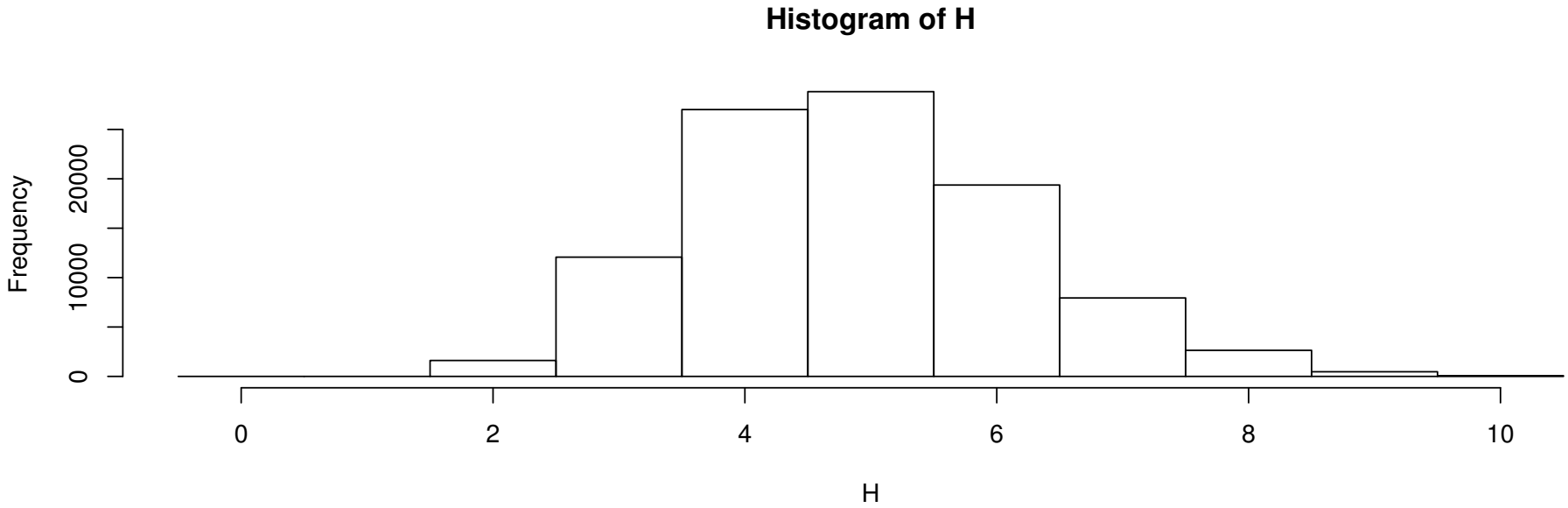
All of the other hyperparameters are held constant.

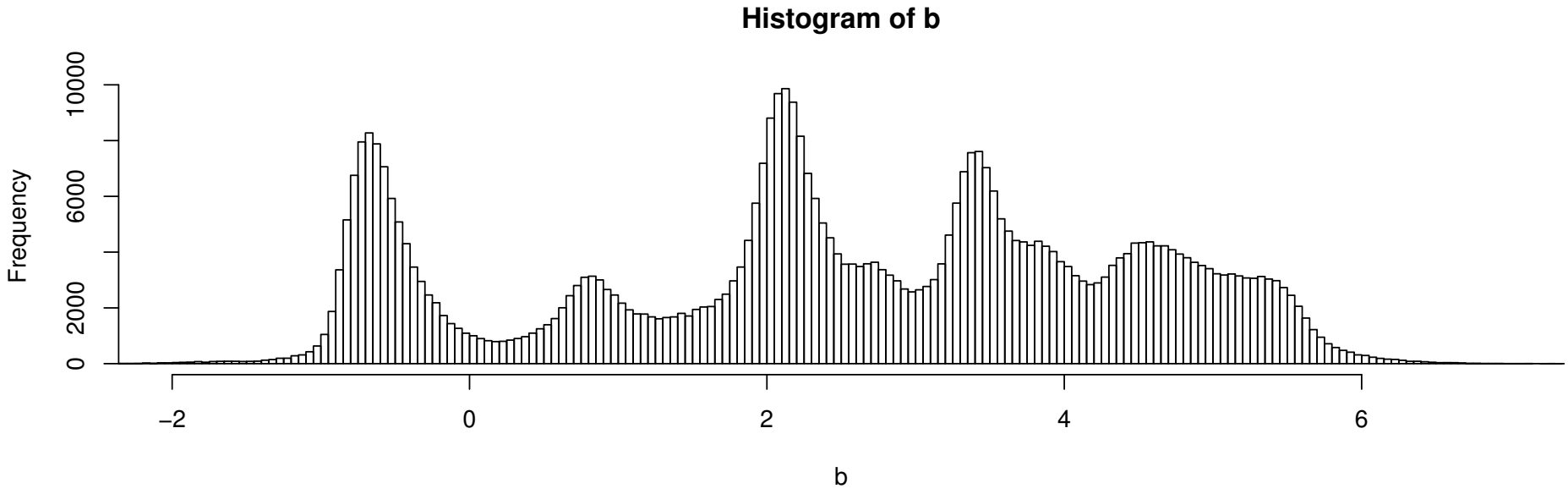
Implementation

- I have implemented an MCMC sampler of this model in C++.
- The sampler performs updates of z, b and σ using Gibbs sampler steps.
- The number (and composition) of layers is updated using both birth-death and split-merge moves.
- There are two different split proposals: one to symmetric values of b_h, and a "bud" move, where one component retains the old value of b_h from the pre-split layer.

Example: synthetic data

- The test data set consisted of 500 genes with layers at 2.2, 3.4 and 4.7.
- There were also composite values at 5.6 ($= 2.2 + 3.4$), 6.9 ($= 2.2 + 4.7$), 8.1 ($= 3.4 + 4.7$) and 10.3 ($= 2.2 + 3.4 + 4.7$), as well as some values at 0.0.
- Under this model, the number of layers was over-estimated.


Histogram of H
Frequency
0
10000
20000
0
2
4
6
8
10
H


Histogram of b
Frequency
0
2000
6000
10000
−2
0
2
4
6
b

Repulsive b's(!)

- We wanted to eliminate the possibility of getting several layers with very similar values of b allocated to distinct groups of genes.

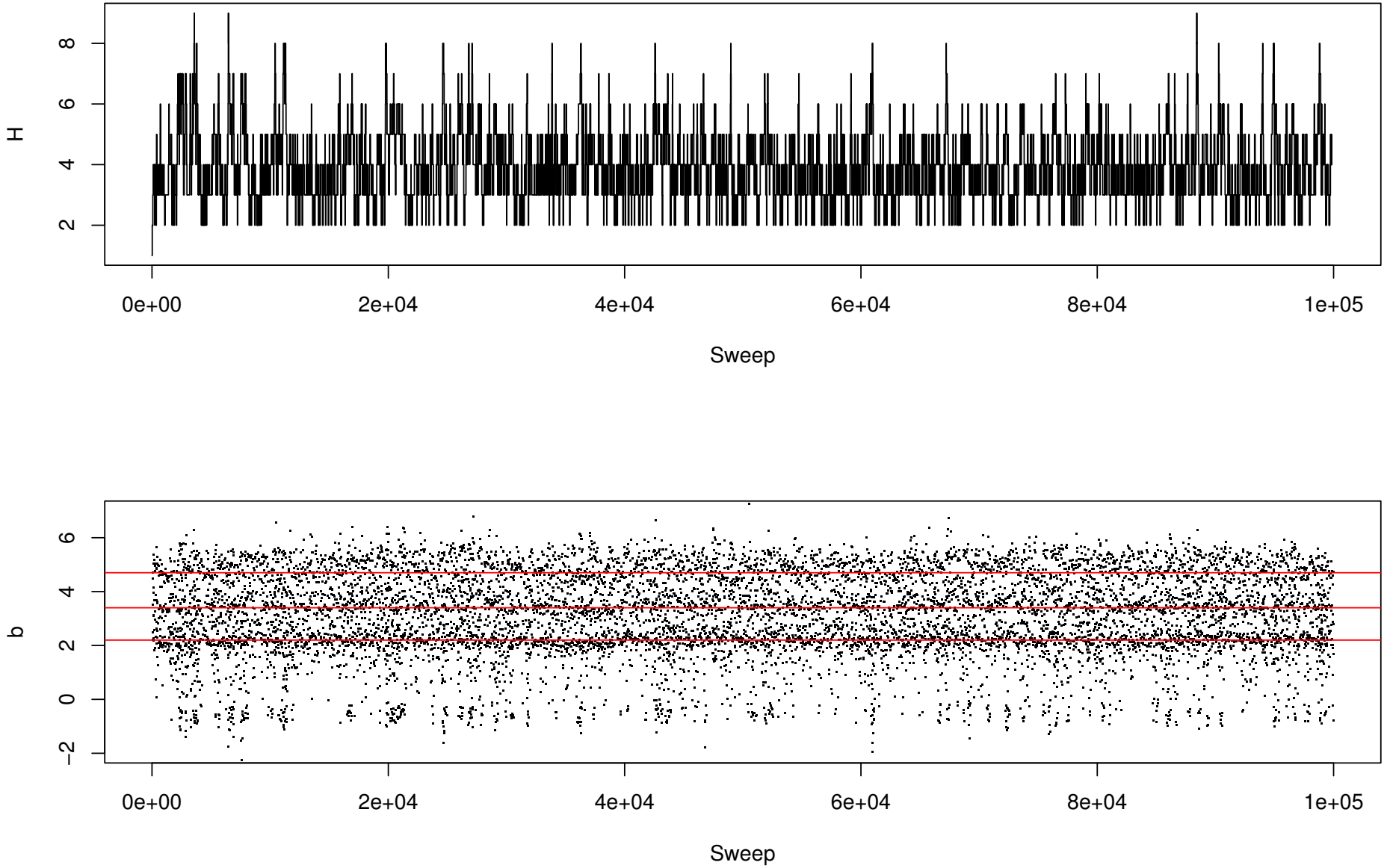
- To solve this problem we propose to use some sort of repulsion between values of b.

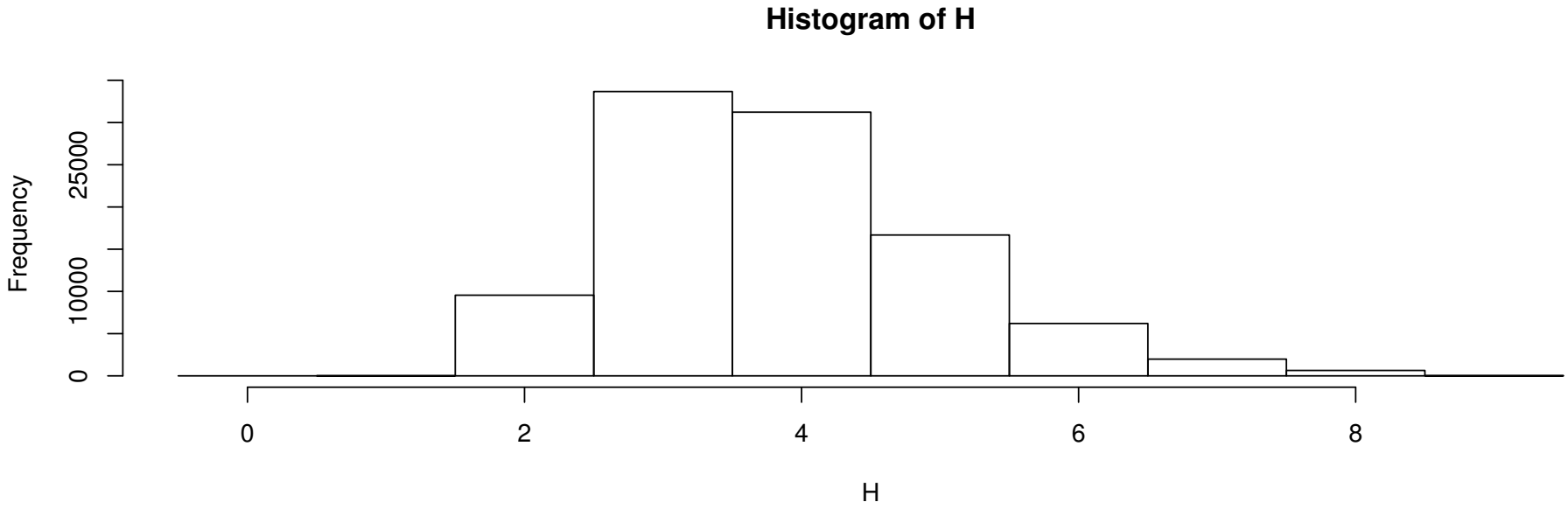
- Since I have worked with area-interaction point processes before (Baddeley and van Lieshout 1995), it seemed natural to me to use a 1-d area-interaction process for this purpose.

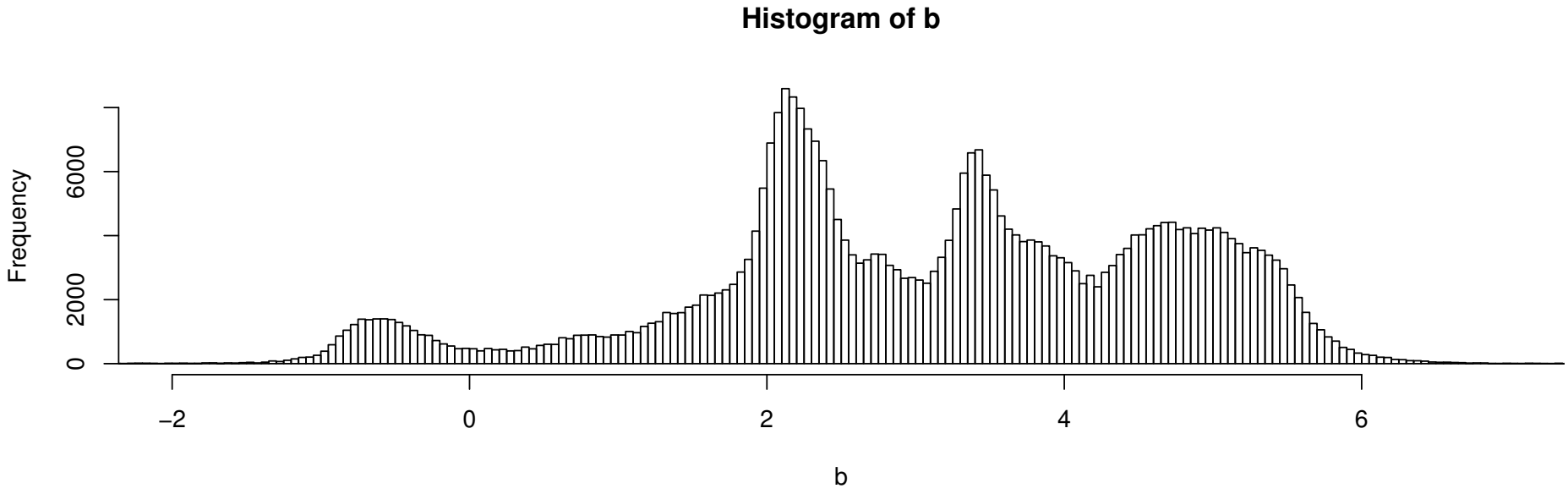

b_1
b_2
b_3
b

More programming...

- I have also implemented this additional feature.
- The b-updates are now 'vanilla' random-walk Metropolis moves.
- The sampler now finds the correct number and composition of layers with high probability.
- Between-model moves are now accepted less frequently.


H
2
4
6
8
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Sweep
b
−2
0
2
4
6
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Sweep


Histogram of H
Frequency
0
10000
25000
0
2
4
6
8
H


Histogram of b
Frequency
0
2000
6000
−2
0
2
4
6
b

The full model

- Having considered the single-sample case, we now extend the model to the multiple-sample case by adding per-sample allocations:

$$y_{gs} = \sum_{h=1}^{H} v_{sh} z_{gh} b_h + \varepsilon_{gs},$$

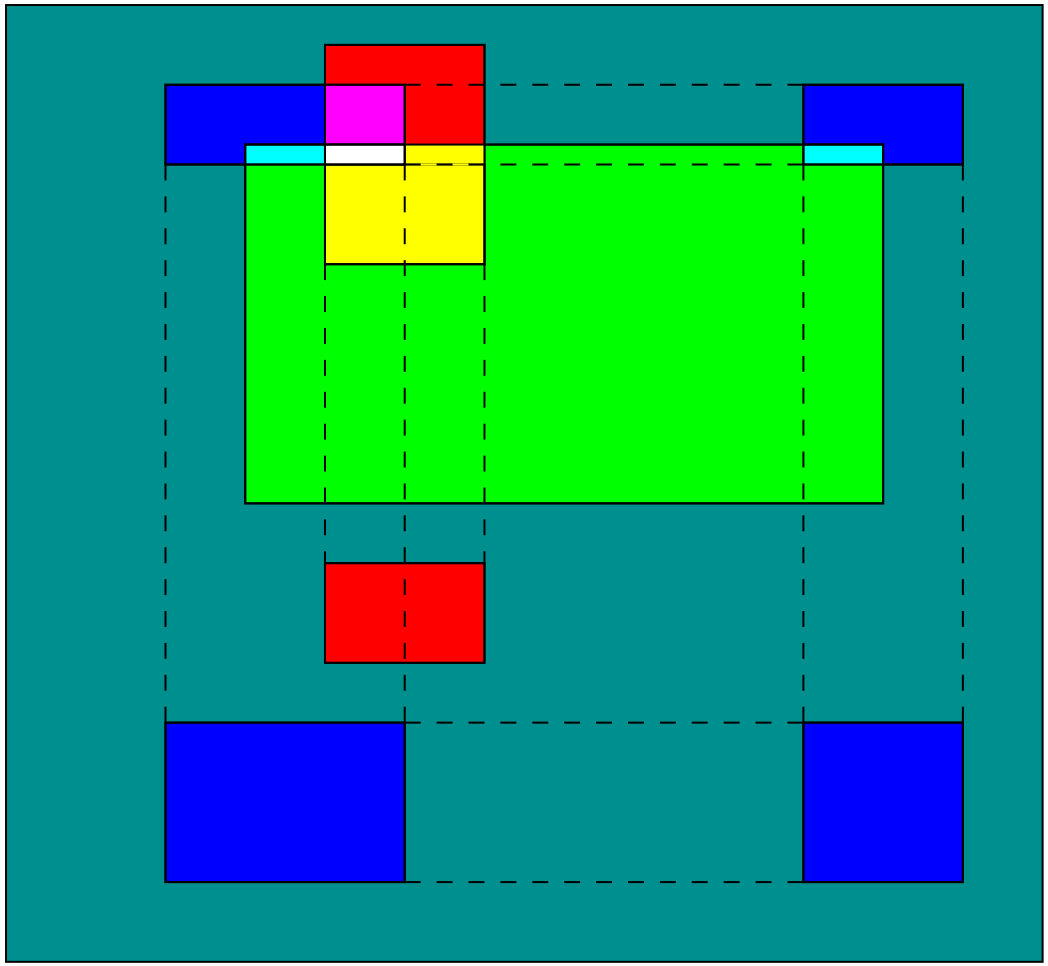
  where s =sample and $v_{sh} \in \{0, 1\}$ is an allocation variable with a simple Bernoulli prior for the v's.

- The layers are constant-height "rectangles".

- This model is capable of describing basic differential expression of groups of genes between samples.

Samples

Genes

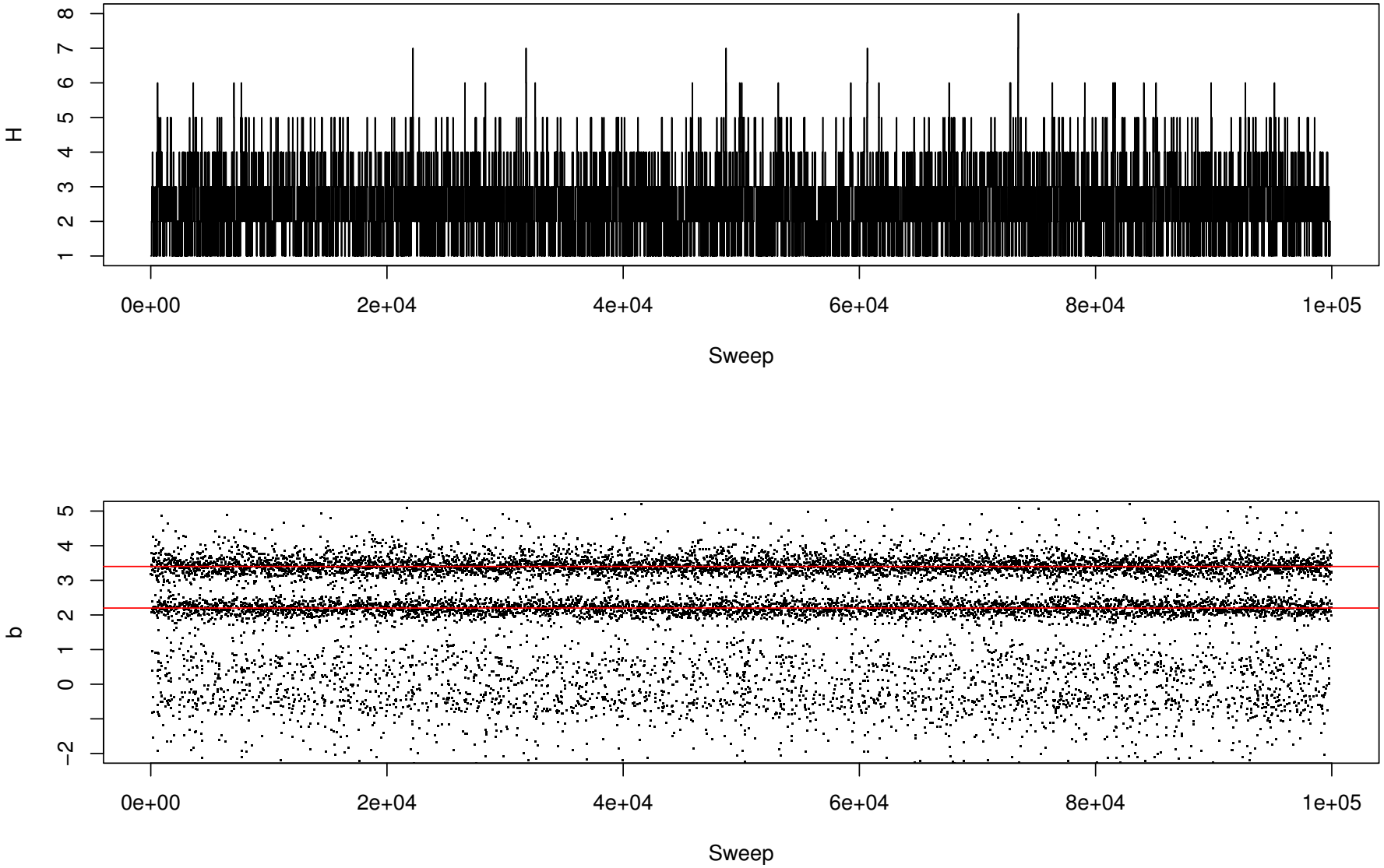


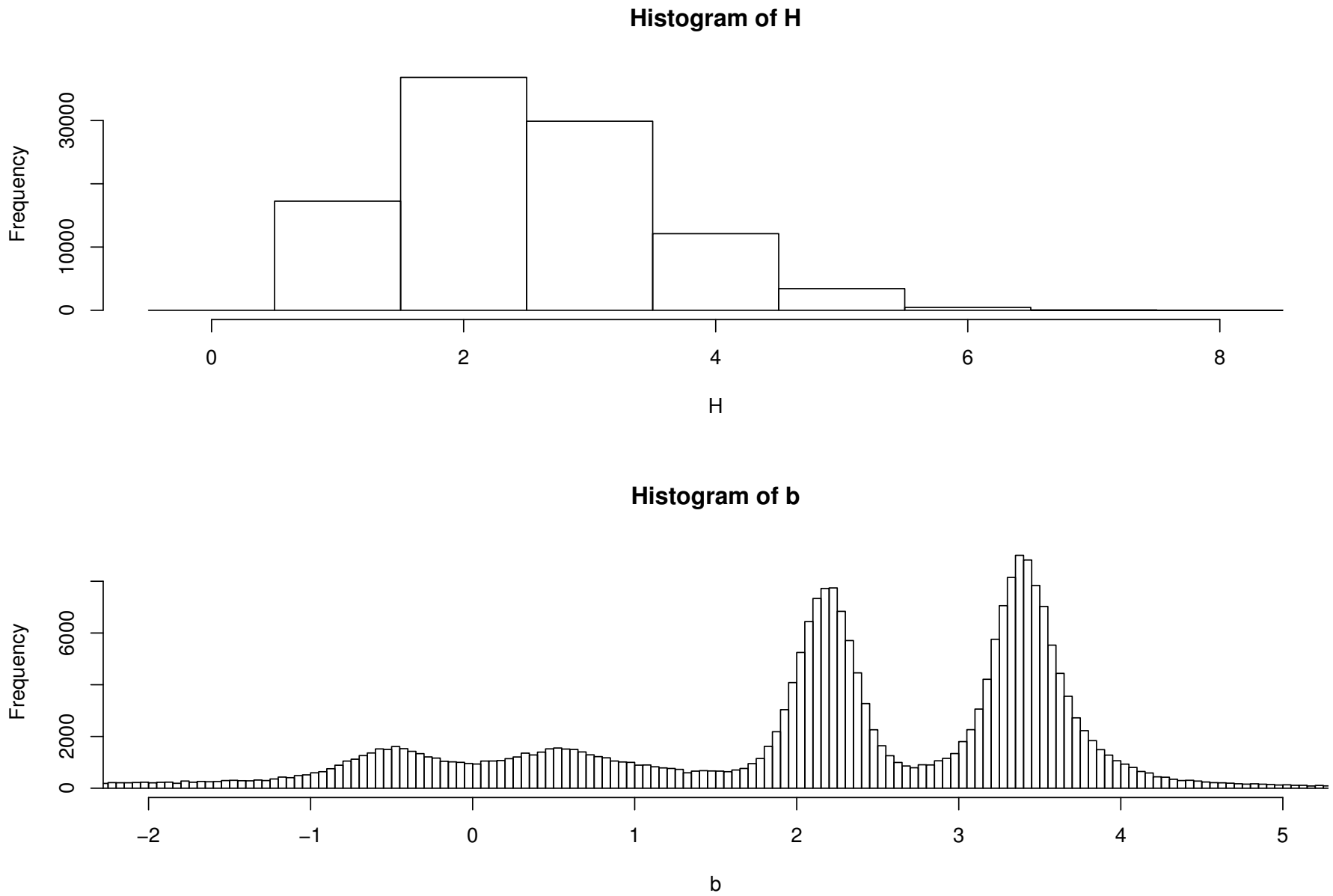
A toy example of the full model

- We next simulated a dataset containing 10 genes for each of 5 samples with 2 layers at 2.2 and 3.4:

2.2	2.2	0	0	2.2
2.2	5.6	3.4	3.4	2.2
0	3.4	3.4	3.4	0
0	3.4	3.4	3.4	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	3.4	3.4	3.4	0
2.2	5.6	3.4	3.4	2.2
2.2	2.2	0	0	2.2

- We added $N(0, (\frac{1}{2})^2)$ noise to this dataset as well.
- The algorithm found the correct number and composition of the layers without difficulty.


H
1
2
3
4
5
6
7
8
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Sweep
b
−2
0
1
2
3
4
5
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Sweep

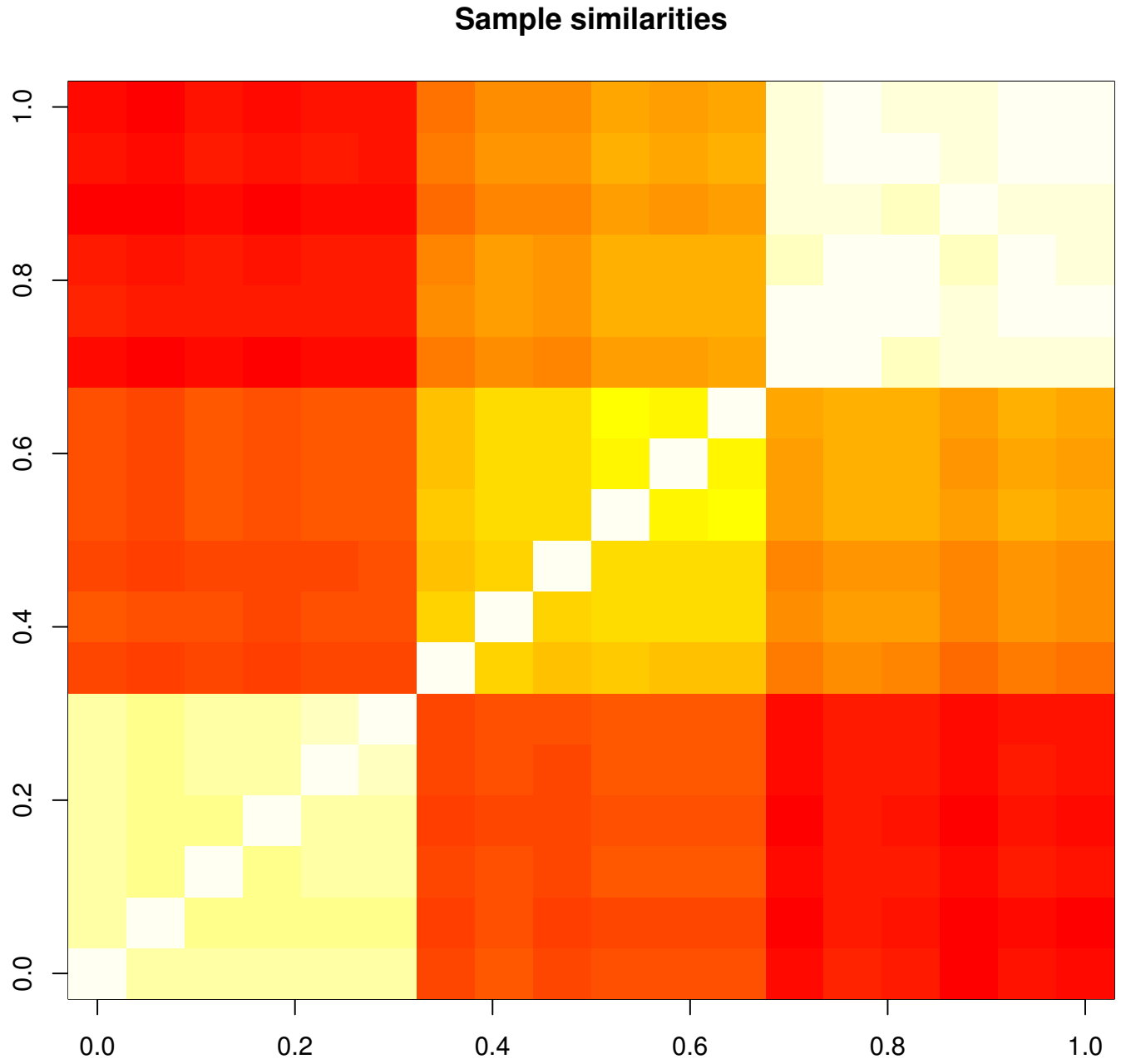

Histogram of H
Frequency
0
10000
30000
0
2
4
6
8
H
Histogram of b
Frequency
0
2000
6000
−2
−1
0
1
2
3
4
5
b

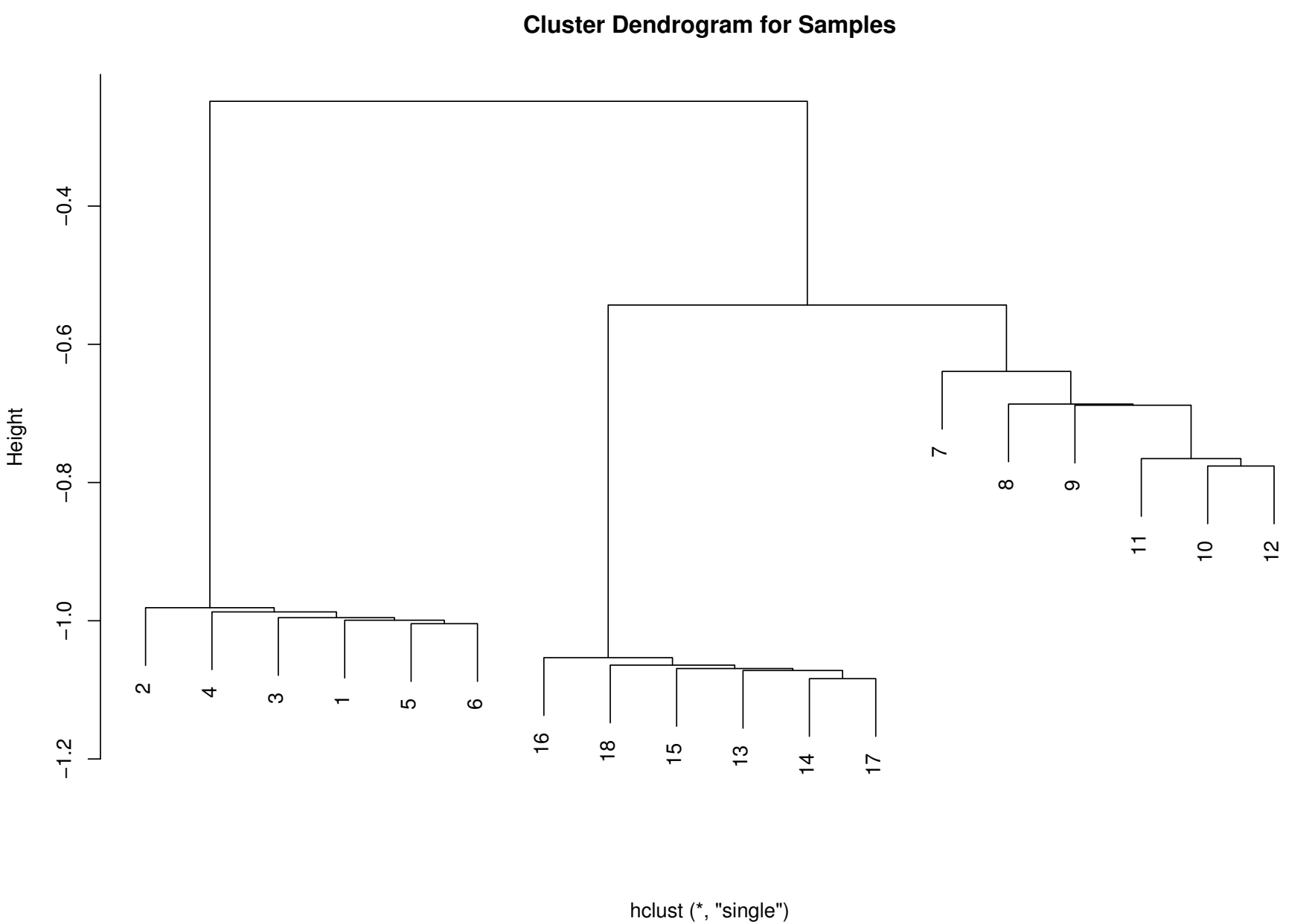
A real dataset

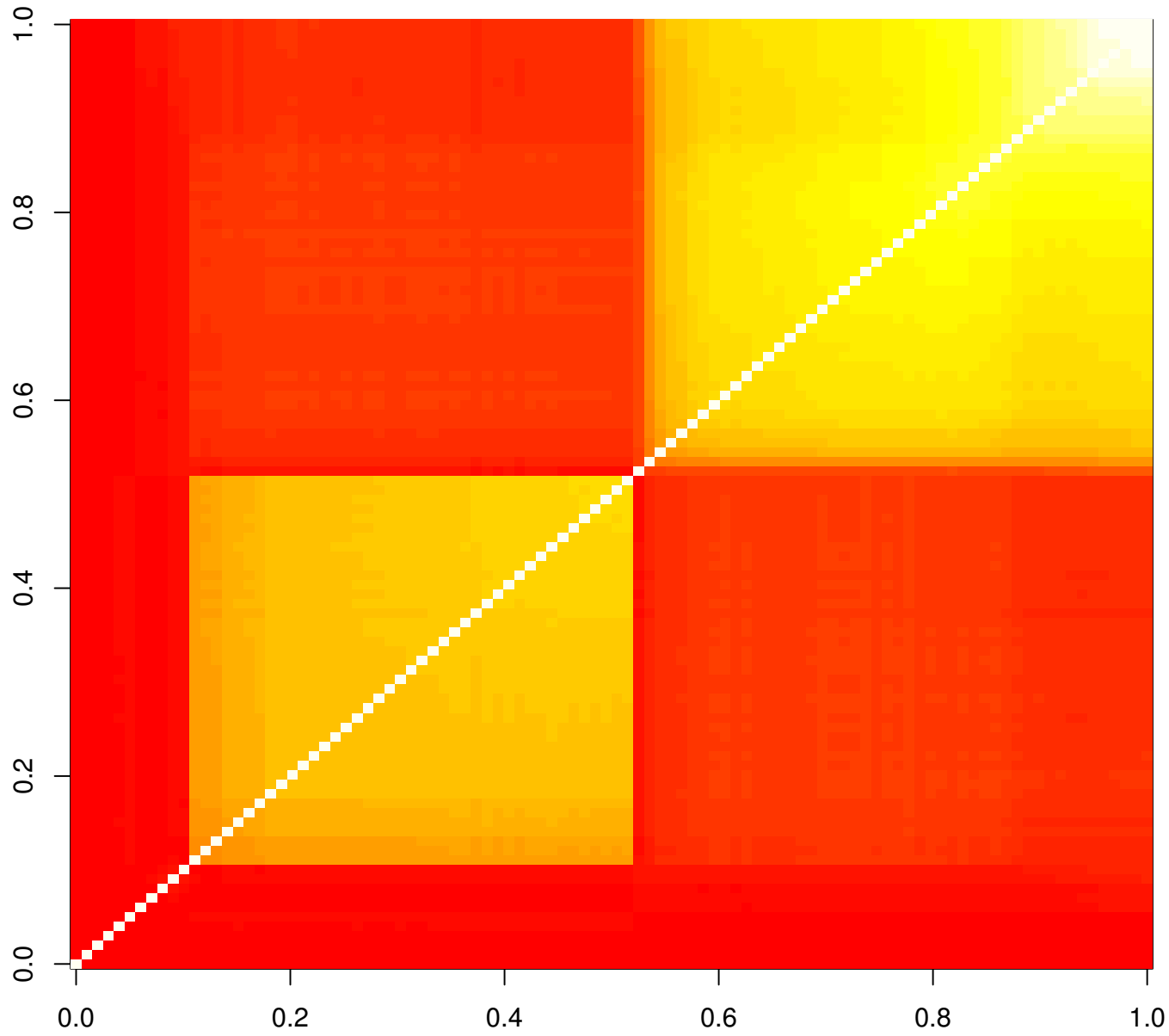
- In order to demonstrate our method with a real dataset, we looked at the human fibroblast data introduced by Lemon et al. (2002).
- This data set consists of 18 samples split into 3 categories: serum starved, serum stimulated and a 50:50 mix of starved/stimulated.

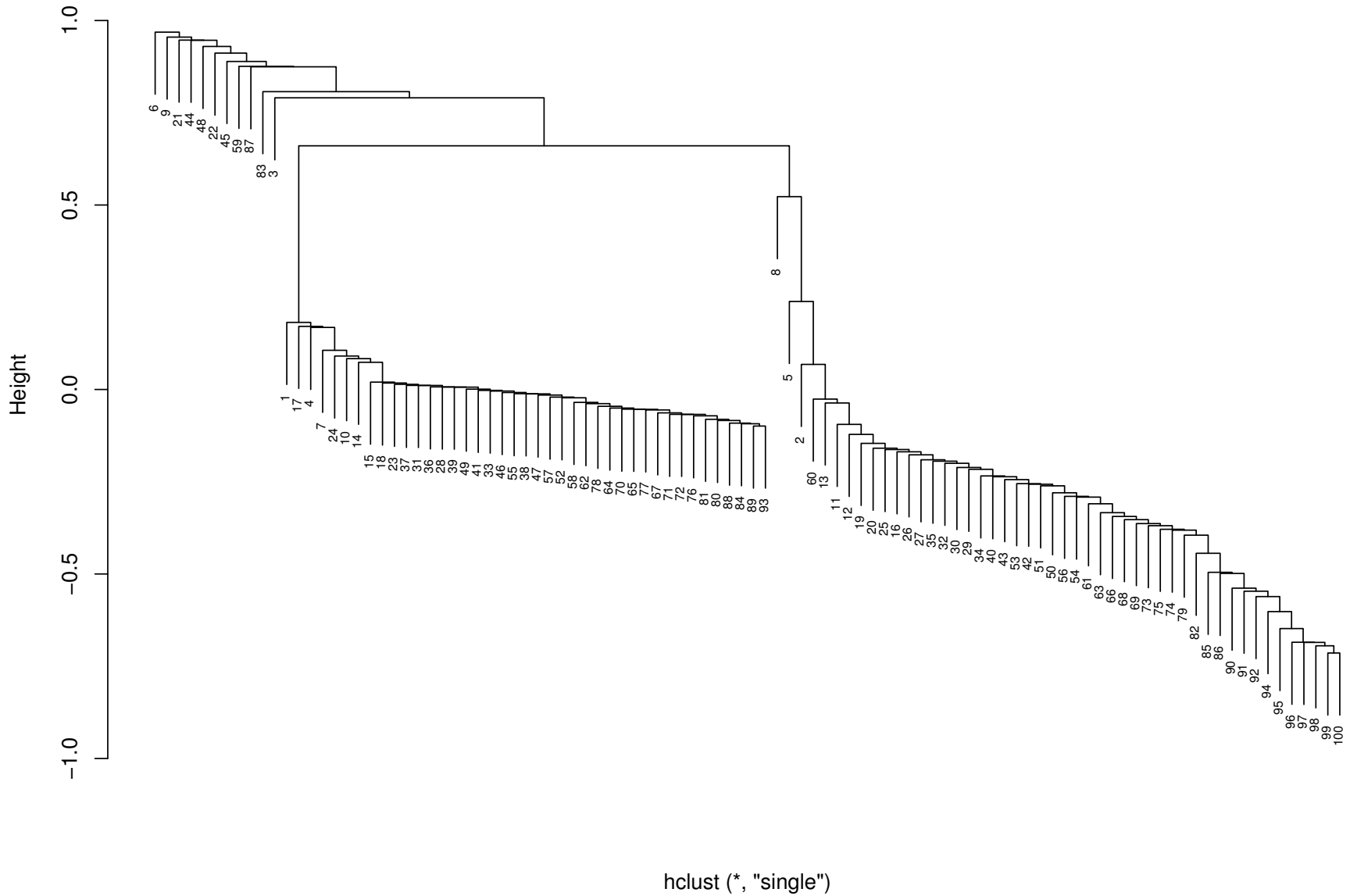
Pre-processing

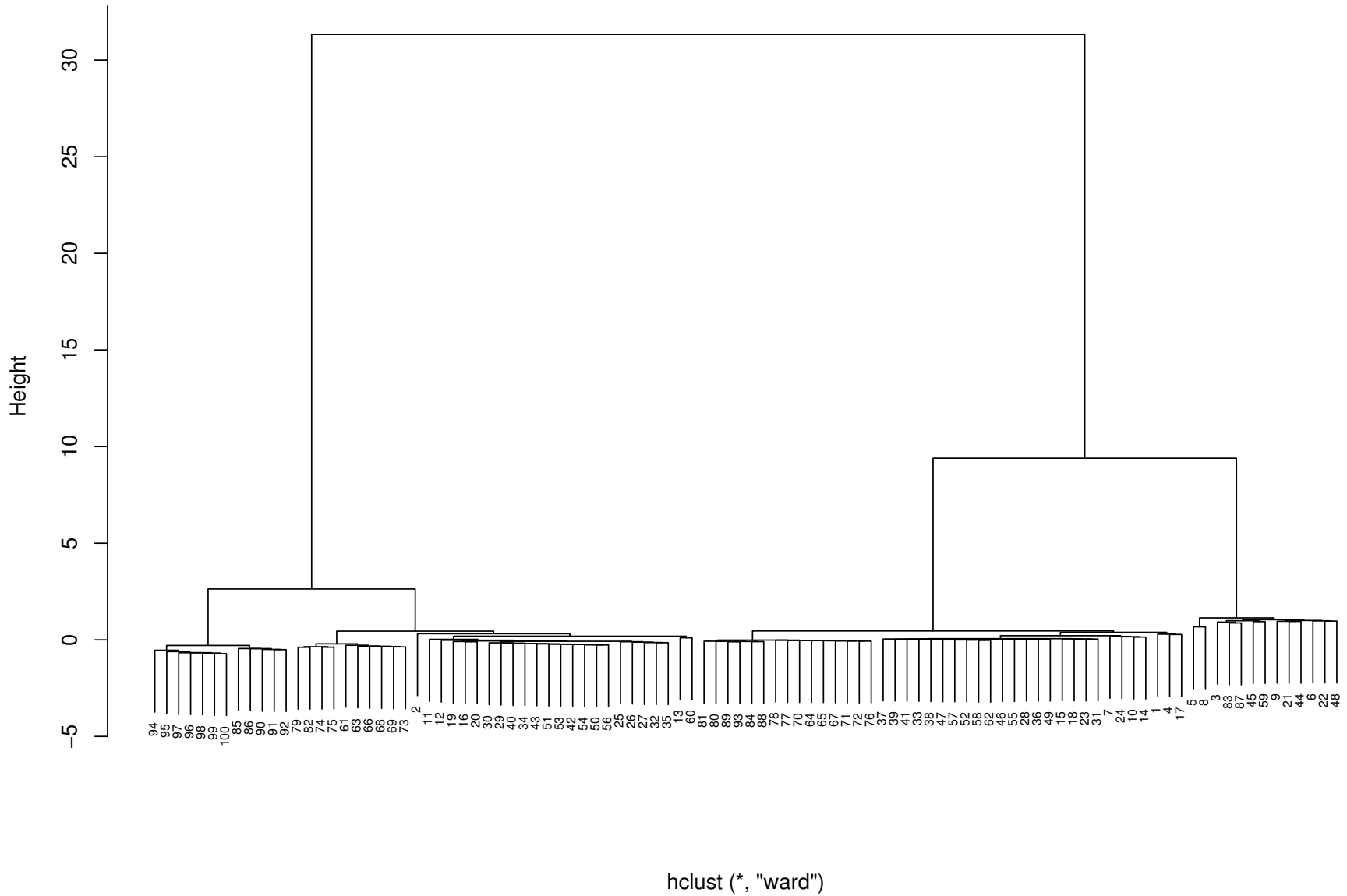
- We used the natural logarithm of Lemon et al.'s calculated LWF values as our measure of expression and subtracted gene and sample mean levels.
- We then selected the 100 most variable genes across all 18 samples and used this 18×100 array as the input to our sampler.
- The following plots use the expected number of common layers that a pair of samples (genes) are in to cluster the samples (genes).

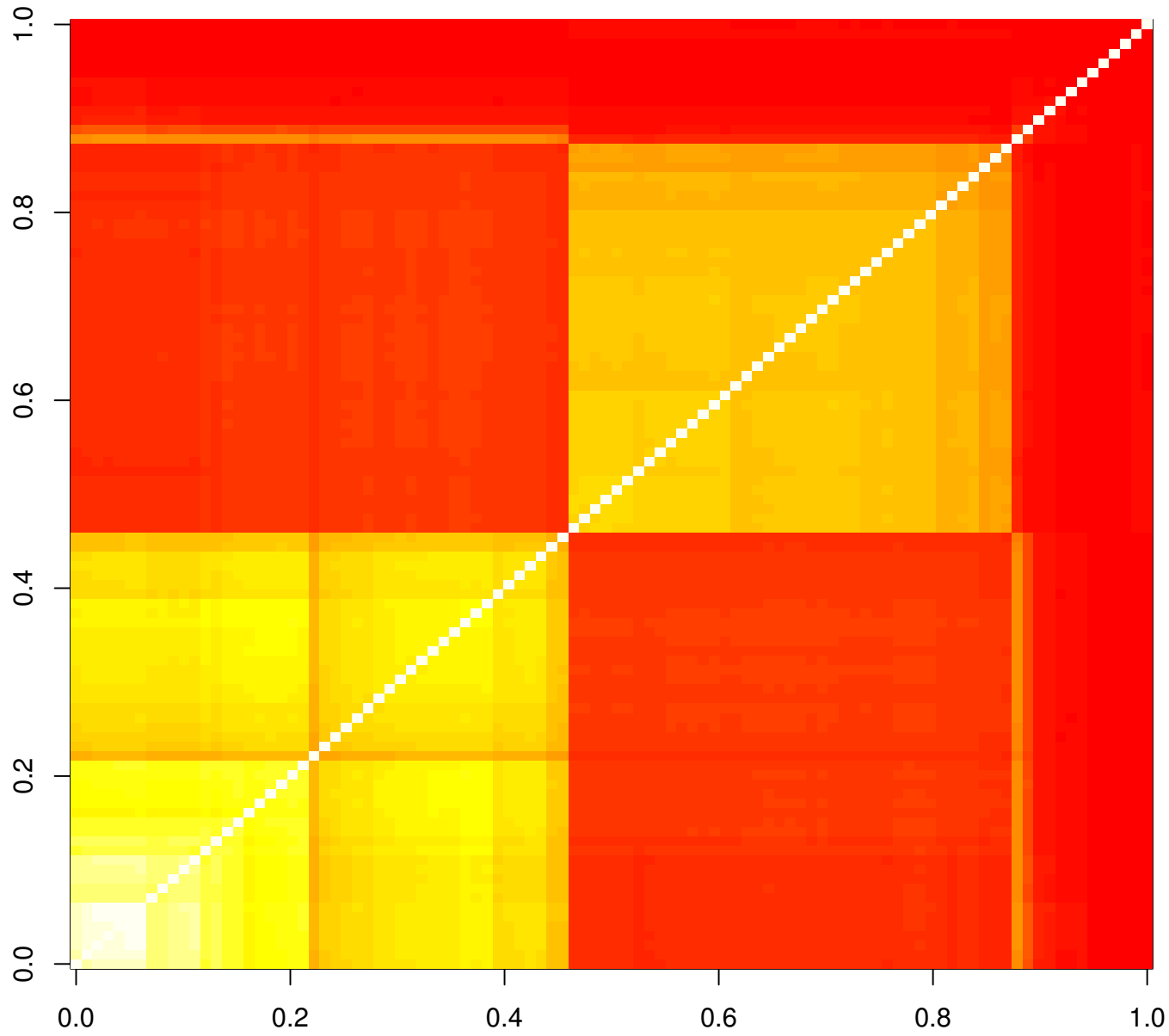

Sample similarities
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Cluster Dendrogram for Samples
Height
−0.4
−0.6
−0.8
−1.0
−1.2
2
4
3
1
5
6
16
18
15
13
14
17
7
8
9
11
10
12
hclust (*, "single")


Sorted (using hclust) gene similarities
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Cluster Dendrogram for Genes
Height
1.0
0.5
0.0
−0.5
−1.0
hclust (*, "single")


Cluster Dendrogram for Genes
Height
30
25
20
15
10
5
0
−5
94 95 97 96 98 99 100 85 86 90 91 92 79 82 74 75 61 63 66 68 69 73 2 11 12 19 16 20 30 29 40 34 43 51 53 42 54 50 56 25 26 27 32 35 13 60 81 80 89 93 84 88 78 77 70 64 65 67 71 72 76 37 39 41 33 38 47 57 52 58 62 46 55 28 36 49 15 18 23 31 7 24 10 14 1 4 17 5 8 3 83 87 45 59 9 21 44 6 22 48
hclust (*, "ward")


Sorted (using hclust) gene similarities
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Future work

- Work on intelligent methods for choosing the parameters.

- Investigate methods for displaying the posterior in a visually-informative way.

References

Baddeley, A. J. and M. N. M. van Lieshout (1995). Area-interaction point processes. *Annals of the Institute for Statistical Mathematics 47*, 601–619.

Lemon, W. J., J. J. T. Palatini, R. Krahe, and F. A. Wright (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics 18*, 1470–1476.

MacKay, D. J. C. and J. Miskin (2001). Latent variable models for gene expression data. Technical report, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE. United Kingdom.