

Comparison Between MSCBS and the Hierarchical Hidden Markov Method of Shah et al. (2007)

We did a detailed comparison between our method and the method of Shah et al. (2007) on a representative region taken from the Stanford quality control data set used in our paper. The region spans the first 2000 SNPs covering Chromosome 22q, and contains three visually identifiable CNVs, including a complex deletion that has several variants among the 62 samples. Note that this type of complexity, although not commonly found in normal samples, is very common in cancer samples, which is the focus of Shah et al. (2007). Figure 1, which is also given in our paper, shows the data for this region and the segmentation obtained by our second algorithm.

We applied the hierarchical hidden Markov model method from Shah et al. (2007) to this data with their recommended settings: The initialization is done using the function provided in their software package. The parameter ϵ , which controls the coupling of the individual samples to the master underlying state, is set to 0.7. We set a burn-in time of 500 iterations for the MCMC chain (their recommended burn-in time is shorter, at 200 iterations). Their hidden Markov model contains four states: Normal, Gain, Loss, and Uncoupled. In the uncoupled state, each individual sample varies freely. According to Shah et al. (2007), the informative quantities estimated from their model are the posterior probabilities of gain and loss given the data, which should be the focus of interpretation.

Figure 2 shows the estimated posterior probabilities of gain and loss, aligned with the heatmap of this 2000 SNP region. The figure also shows the posterior probabilities of the normal and uncoupled state. As we can clearly see, *none* of the three visually identifiable CNVs are detected by the hierarchical hidden Markov model. Furthermore, there are many spurious detections, as evident in the many spikes in the posterior probabilities of gain and loss.

The hierarchical hidden Markov model of Shah et al. (2007) failed to detect the CNVs because its model is too restrictive. It assumes the following:

1. For any given CNV region, all carriers must be gains, or all carriers must be losses. There can not be a mixture of gains or losses. Clearly, all of the visually identifiable CNVs in this region break this rule. CNVs whose carriers are a mixture of gains and losses is quite common in the genome.
2. For any given sample, there is only one loss state and one gain state. Thus, all losses (and gains) must have the same mean. We showed in our paper that this is not true for real data. There are two reasons for this: (1) The true copy number levels may not be the same, i.e. when there is a loss the copy number may be 1 or 0. (2) Even when the true copy number levels are the same, the SNPs in different regions have different response rates, leading to different mean levels.
3. All copy number variant regions have the same proportion carrier probability.

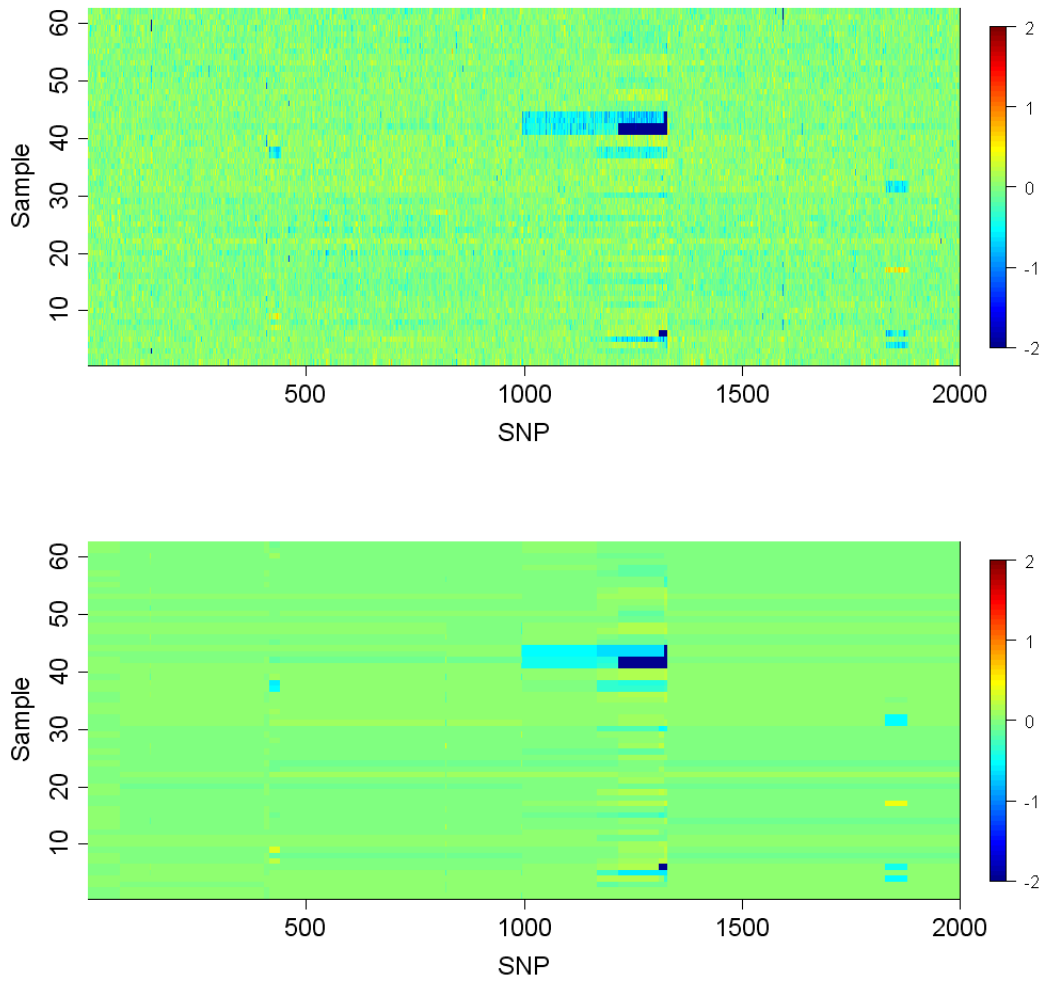


Figure 1. Example 2000 SNP region in cytoband 22q11 of Stanford Quality control data. This region contains a complex CNV with nested deletions. Bottom panel shows segmentation given by Algorithm 2.

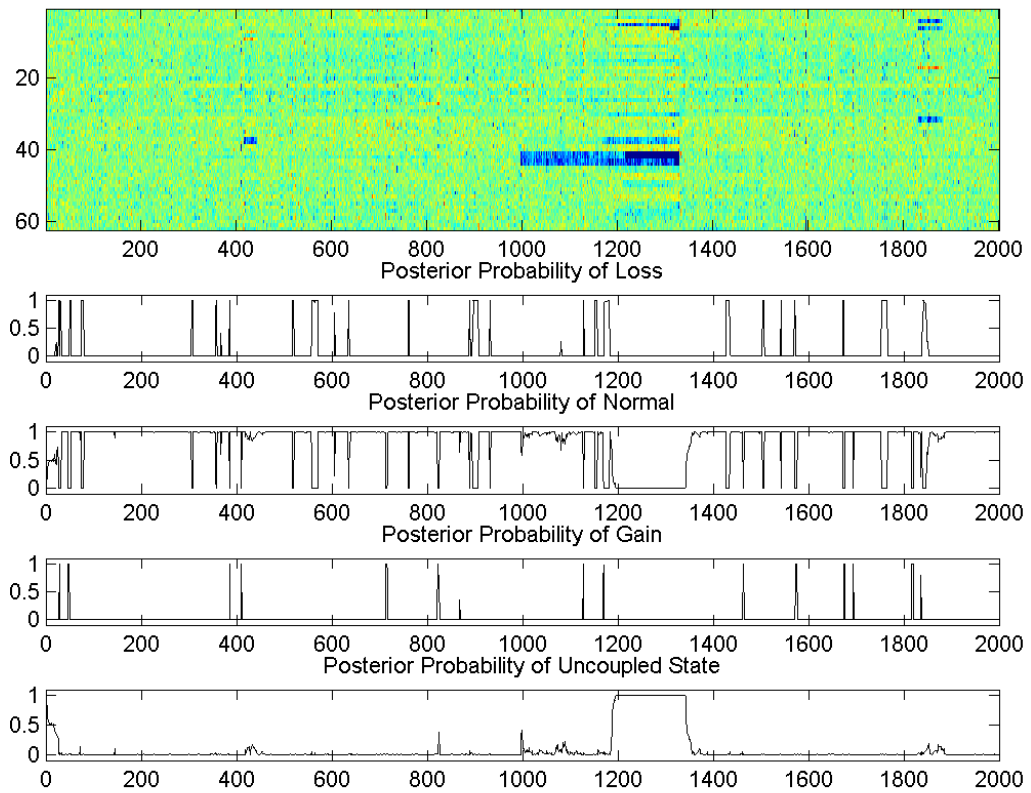


Figure 2. Top panel: 2000 SNP region in cytoband 22q11 for Stanford quality control data. Panels 2-5 are the posterior state probabilities obtained from the hierarchical hidden Markov model of Shah et al. (2007).

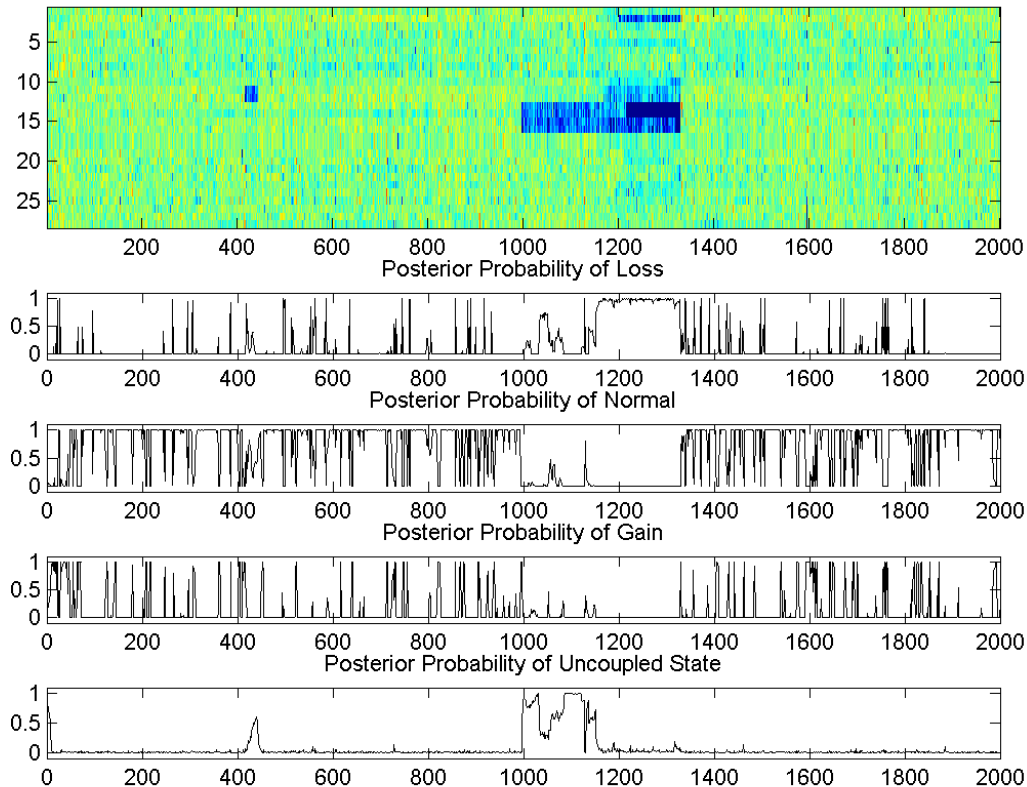


Figure 3. Top panel: 28 samples selected from the data set in Figure 2 that contain a deletion in the 1000-1400 SNP region. Panels 2-5 are the posterior state probabilities obtained from the hierarchical hidden Markov model of Shah et al. (2007).

Our model does not assume (1) and (2). While (3) is also assumed in our model, it is used merely in the weighting of the chi-square statistic. Thus, when the true proportion of carriers is different from the prior proportion, if the chi-square values are moderately large the prior probability has little effect on the overall scan statistic. For the hierarchical hidden Markov model, the prior carrier probability seems to play a bigger role in the final results.

Since assumption (1) above for Shah et al. (2007) seems to be the main limitation in the analysis of this data set, we decided to give it an easier task. We took the subset of the 62 samples that, by visual inspection, contain deletions in the 1000-1400 SNP region. In other words, we manually curated a smaller sample set so that all carriers should agree on the hidden state in the 1000-1400 region. This smaller sample set is shown in the top panel of Figure 3. The bottom panels show the results from the method of Shah et al. (2007). Note that for this easier data set, the region containing a homozygote deletion between SNPs 1217 and 1309 is detected. However, the regions

between 1000-1217 and between 416-442, which are visually obvious and have sum-of-chisquare values in the thousands, are missed.

This comparison also shows that the output between Shah et al. (2007) and our method is very different. While Shah et al. (2007) gives a continuous profile of posterior state probabilities, our method gives a hard segmentation. The output of Shah et al. (2007) must be further processed to identify the change-point boundaries, and in the process contiguous CNVs are often broken into many segments. Also, while the method in Shah et al. (2007) can certainly be extended to identify the CNV carriers, their current software implementation does not have this functionality. This functionality is needed for a consistency assessment for CNV detections for the method of Shah et al. (2007), as presented in Section 3 of our paper. Thus, which such a consistency comparison would be informative, it is outside the scope of the current study.

References

SHAH, S. P., LAM, W. L., NG, R. T. & MURPHY, K. P. (2007). Modeling recurrent dna copy number alterations in array cgh data. *Bioinformatics* **23**, 450–458.