Nonlinear Switching State-Space Models for aCGH Analysis

Jeffrey Dunn

CMU-CS-09-124

April 2009

School of Computer Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

Thesis Committee:
Eric P. Xing, Chair
Russell Schwartz

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Keywords: Array comparative genome hybridization, microarray, switching state-space model, colorectal cancer, non-small cell lung cancer, whole-genome analysis, expectation-maximization

Abstract

DNA copy number aberrations serve as key biological markers for cancer and many other diseases and determining their location has important applications to cancer diagnosis, drug development, and molecular therapy. Analysis of a variety of cancers have revealed that gains present in proto-oncogenes and losses of tumor suppressor genes have serious impacts on growth-limiting functions, cell-death programs, and self-repair processes of cancerous regions. Methods to efficiently and accurately detect these aberrations serve as an important step in understanding the behavior of cancer and have significant consequences in cancer diagnosis and development of treatments.

Array comparative genome hybridization (aCGH) provides an efficient method for full genome analysis of DNA copy numbers, but is corrupted by serious systematic errors such as impurity of the DNA sample, heterogeneity of copy numbers among defective cells, and measurement noise. Previous methods have shortcomings in inference of dosage states, failing to properly annotate genomes displaying spatially correlated samples and large spikes in log-ratio. A new model, the Nonlinear Genome Imbalance Scanner (NL-GIMscan), is proposed which captures both nonlinear spatial drift of aCGH intensities and measurement noise through fitting state-specific non-linear hidden trajectories with an overlying first-order Markov switching process. NL-GIMscan is demonstrated on two different datasets of malignant tumors and resulted in improved performance over existing models such as hidden Markov models (HMM). A software implementation of the model is available from the author.

Acknowledgments

I would like to thank my advisor, Eric Xing, for his advice and feedback throughout my research. His guidance was crucial to my development of this model and his insights were invaluable to my progress. I would also like to thank Russel Schwartz for his helpful comments and suggestions and for bringing a new perspective as a member of my thesis committee.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

DNA copy number aberrations serve as key biological markers for cancer and many other diseases and determining their location has important applications to cancer diagnosis, drug development, and molecular therapy. Analysis of a variety of cancers have revealed that gains present in proto-oncogenes and losses of tumor suppressor genes have serious impacts on growth-limiting functions, cell-death programs, and self-repair processes of cancerous regions. Methods to efficiently and accurately detect these aberrations serve as an important step in understanding the behavior of cancer and have significant consequences in cancer diagnosis and development of treatments.

Array comparative genome hybridization (aCGH) provides an efficient method for full genome analysis of DNA copy numbers of individuals [12]. In an aCGH experiment, samples of DNA from two different populations, the "test" or tumor population and the "control" or normal population, are labeled with differing fluorescence and cohybridized to normal metaphase chromosomes. Then, an array of probes uniformly spanning the entire genome measure the ratio of the fluoresence intensities, forming the log-ratio output. These log-ratio values measure the relative difference in copy number between the two populations at each measured location. In theory, the copy numbers are completely determined by this log-ratio output from the aCGH assay, but real

measurements exhibit severe deviations due to problems such as impurity of the DNA sample, heterogeneity of copy numbers among defective cells, and noise due to the method of measurement. Thus techniques to determine the true dosage states are required to infer biologically plausible dosage states from the data.

Early methods of dosage state annotation focused on thresholding, selection of value-windows based on the relationship between log-ratio value and copy number, to assign dosage states to clones. These methods tend to largely ignore the actual range of log-ratio values which may not directly correspond to the theoretical values and fail to capture changes in signal intensity due to factors beyond copy number aberrations. More modern approaches apply statistical techniques to perform automatic annotation and fall into four main categories:

Mixture models: Mixture models assume the log-ratio values of the clones are independent samples generated by a mixture of distributions, each corresponding to a dosage state. This type of model can be learned using the EM algorithm and clones are assigned to the dosage state with the highest probability of generating them. For example, Hodgson *et al.* [5] applied a mixture of Gaussians and assigned clones to the normal state if they were within three standard deviations of the mean of the normal state.

Regression models: Regression models aim to fit curves to noisy log-ratio values to capture overall trend. Such curves facilitate manual annotation via visual inspection or thresholding through smoothing. However, these types of methods are not suitable for automatic annotation and mainly facilitate visualization and denoising.

Segmentation models: Due to biological processes, consecutive clones tend to remain in the same dosage state and changes are typically marked by abrupt changes in log-ratio value. Thus,

correct annotation often results in dividing the complete genome into segments which are internally uniform in copy number. Segmentation methods focus on identification of segments and minimizing within segment variance of log-ratio values. There are many algorithms in this class including popular methods such as GLAD [6] and DNACopy [11]. Following segmentation, each segment must be assigned to a dosage state. Several heuristic algorithms have been developed for this purpose, however, many suffer from identification of numerous states without clear biological meanings. Algorithms to merge these spurious states have been developed to attempt to remedy this deficiency. More recently, Lai *et al.* have taken a more statistical approach to segmentation models with the stochastic change-point model which provides a posterior distribution for segmentation [8].

Spatial models: Spatial models jointly address clone sequence segmentation and dosage-state annotation under unified models of aCGH data. Intuitively, joint estimation will likely produce superior annotations as the quality of the solution for each subproblem directly impacts that of the other. Fridlyand *et al.* [2] applied hidden Markov models (HMM) to aCGH profiles, modeling log-ratio values as outputs generated by state-dependent Gaussian distributions. Through standard HMM algorithms, the methods of Fridlyand *et al.* learn these emission distributions and use the Viterbi algorithm to decode the most probable underlying dosage state sequence. Later extensions to this model include BioHMM developed by Marioni *et al.* [9] which includes clone length and position details into the inference algorithm. More recently, the GIMscan algorithm developed by Shi *et al.* [15] employs a switching-state space model allowing for linear spatial drift of state-specific distributions.

Previous computational methods for aCGH analysis remain limited in terms of accuracy and robustness in annotation of aCGH data. Most previous methods fail to capture spatially correlated samples through independence assumptions. Also, many types of random fluctuations are

not captured through methods such as mixture and thresholding methods by ignoring the spatial relationship of clones. Many of these methods provide annotations which exhibit frequent switching between dosage states or fail to capture large spikes in log-ratio, both leading to biologically implausible annotations.

The proposed model, Nonlinear Genome Imbalance Scanner (NL-GIMscan), extends the switching state-space model of Shi *et al.* [15] to allow for non-linear spatial drift of the state-specific distributions and leverage chemical similarity of chromosomes across individuals. The linear spatial drift incorporated into the GIMscan model is highly dependent on the exact log-ratio values, allowing for varying degrees of drift based on the observed intensities of each state. NL-GIMscan corrects these shortcomings in a robust manner and provides high-quality dosage state annotations.

Chapter 2

NL-GIMscan: Nonlinear Switching State-Space Model for aCGH Analysis

NL-GIMscan uses an adaptation of a switching state-space model to describe both spatial drift of the hybridization signals as well as transitions between dosage states. Within the state-space model, a two-dimensional state is used to allow for non-linear drift and a parameter sharing scheme is applied to model similarities across individuals and chromosomes. Section 2.1 describes the NL-GIMscan switching state-space model for non-linear spatial drift as well as the aforementioned parameter sharing scheme and Section 2.2 describes the applied learning algorithm.

2.1 Model Formulation

The spatial drift of the hybridization signals within each dosage state is modeled by the hidden trajectory of a state-space model and the uncertainty in log-ratio values is modeled by zero-mean Gaussian noise about the hidden trajectory. Each state in the state-space model corresponds to a two-dimensional vector containing a "position" term corresponding to the spatial drift and a "velocity" term corresponding to the per-measurement change in spatial drift. Similar to models

used in many physical processes, an "acceleration" input term is introduced which models non-linear spatial drift and is learned via EM as described in Section 2.2.

A hidden trajectory is assumed for each unique combination of person, dosage state, and chromosome. Acceleration terms are shared across trajectories for different individuals since hybridization signals follow a generally comparable pattern across individuals for a given chromosome as the chemical properties of a particular chromosome are similar among individuals of the same species. However, the exact magnitude of a given dosage state varies across individuals due to normal cell contamination. Thus, the mean initial value as well as starting variance for each distinct dosage state trajectory is shared across chromosomes, but not across individuals, to account for this variability. Each individual's aCGH profile corresponds to one specific laboratory experiment so the output variance, corresponding to the variance of the log-ratio values about the hidden trajectory, is assumed to be specific to an individual. A similar parameter sharing scheme is applied by Shi *et al.* [15], however sharing of acceleration terms does not necessarily impose linear spatial drift.

For an individual j, chromosome k, and dosage state m, the state-space model is defined by

$$X_{t+1,k}^{(m,j)} = AX_{t,k}^{(m,j)} + Bu_{t,k}^{(m)} + w_{t,k}^{(m,j)} \qquad A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix} \tag{2.1}$$

where $X_{t,k}^{(m,j)}$ is the state of the hidden trajectory at measurement t, $u_{t,k}^{(m)}$ is the acceleration input term, and $w_{t,k}^{(m,j)}$ is the position noise. [1] Gaussian noise is assumed thus $w_k \sim N(0, \Sigma_x)$. The initial state $X_{1,k}^{(m,j)}$ is assumed to be distributed $N(\mu^{(m,j)}, \sigma^{(m,j)})$.

A first-order Markov process over the state-space model is applied to model the dosage state transitions. At each clone t, the multinomial variable $(S_{t,k}^{(1,j)}, ..., S_{t,k}^{(M,j)})$ predicts the dosage state

[1] A table of variables and common notation is provided in Appendix A.

for the clone where the binary variable $S_{t,k}^{(m,j)}$ is one if the predicted state at position t for an individual j on chromosome k is m and zero otherwise. The transition matrix Φ contains the first-order transition probabilities, i.e. $\Phi_{m_1 m_2} = \mathbb{P}\left(S_{t+1,k}^{(m_2,j)} = 1 | S_{t+1,k}^{(m_1,j)} = 1\right)$. The initial state is parameterized such that $(S_{1,k}^{(1,j)}, ..., S_{1,k}^{(M,j)}) \sim Multinomial(1, \pi)$. The model assumes the observed log-ratio values are generated by the dosage state and state-space model positions by

$$y_{t,k}^{(j)} = \sum_{m=1}^{M} C X_{t,k}^{(m,j)} S_{t,k}^{(m,j)} + z_{t,k}^{(j)} \qquad\qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{2.2}$$

where $z_{t,k}^{(j)} \sim N(0, \Sigma_y)$ is the Gaussian measurement noise. This corresponds to the log-ratio value being the position term of the state-space model selected by the dosage state at clone t plus zero-mean Gaussian measurement noise.

The log-likelihood for each state-space model is

$$\begin{aligned}
\mathcal{L}_k^{(m,j)}(\theta) = &- \frac{1}{2} \sum_{m=1}^{M} \left(X_{1,k}^{(m,j)} - \mu_1^{(m,j)}\right)' (\sigma^{(m,j)})^{-1} \left(X_{1,k}^{(m,j)} - \mu_1^{(m,j)}\right) \\
&- \frac{1}{2} \sum_{m=1}^{M} \sum_{t=2}^{T} \left(X_{t,k}^{(m,j)} - A X_{t-1,k}^{(m,j)} - B u_{t-1,k}^{(m)}\right)' \left(\Sigma_{x_k}^{(m)}\right)^{-1} \left(X_{t,k}^{(m,j)} - A X_{t-1,k}^{(m,j)} - B u_{t-1,k}^{(m)}\right) \\
&- \frac{1}{2} \sum_{m=1}^{M} \log |\Sigma_{x_k}^{(m)}| - \frac{1}{2} \sum_{m=1}^{M} \log |\sigma^{(m,j)}| + C_1
\end{aligned} \tag{2.3}$$

where C_1 is a constant. Combining the likelihood of the state-space models with the Markov switching process yields a complete log-likelihood of

$$\begin{aligned}
\mathcal{L}(\theta) = &\sum_{k,j,m} \mathcal{L}_k^{(m,j)}(\theta) - \frac{1}{2} \sum_{k,j,m} \sum_{t=1}^{T} S_{t,k}^{(m,j)} \left(y_{t,k}^{(j)} - C X_{t,k}^{(m,j)}\right)' (\Sigma_y^{(j)})^{-1} \left(y_{t,k}^{(j)} - C X_{t,k}^{(m,j)}\right) \\
&+ \sum_{k,j,m} S_{1,k}^{(m,j)} \log \pi^{(m)} - \frac{TK}{2} \sum_{j} \log |\Sigma_y^{(j)}| + \sum_{k,j} \sum_{t=2}^{T} \sum_{m=1}^{M} \sum_{n=1}^{M} S_{t,k}^{(m,j)} S_{t-1,k}^{(n,j)} \log \phi_{m,n} + C_2
\end{aligned} \tag{2.4}$$

where C_2 is a constant.

2.2 EM Algorithm

The model can be learned efficiently using a generalization of the EM algorithm. The exact E-step for switching state-space models in general is intractable as the hidden state variables become conditionally dependent given the observation sequence as shown in Figure 2.1. Instead, the posterior distribution $\mathcal{P}$ is approximated by a distribution $\mathcal{Q}(\mathbf{h}, \mathbf{q})$ from a tractable subfamily of distributions [4]. The modified EM algorithm updates the variational parameters $\{\mathbf{h}, \mathbf{q}\}$ to minimize the KL-divergence between the approximate and true posterior distributions. The tractable subfamily used is a set of uncoupled state-space models with a discrete Markov chain. The updates of the variational parameters are carried out using fixed-point equations [4] which iteratively increase a lower-bound on the log-likelihood.

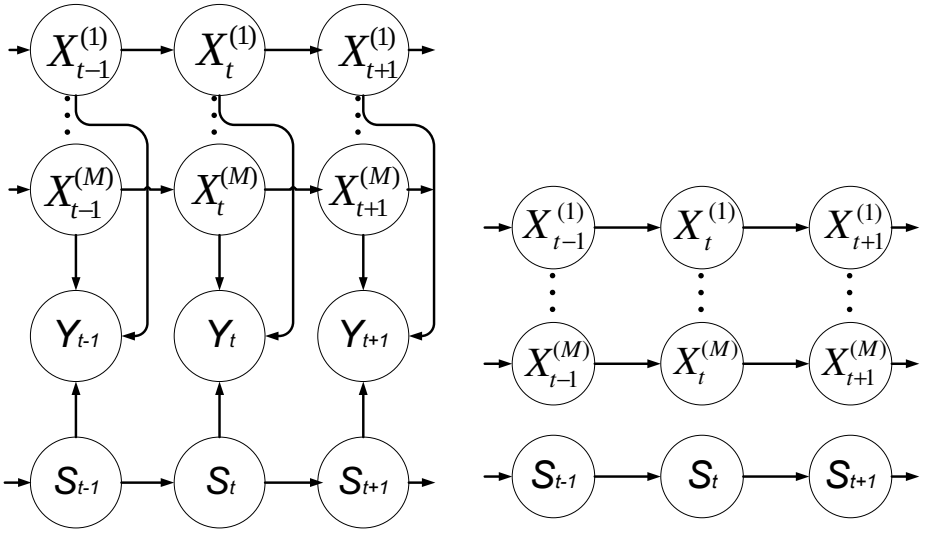


Figure 2.1: (a) Graphical structure of NL-GIMscan depicting M chains of state-space models, a Markov chain of switching processes, and the series of observed log-ratio values for a specific individual's chromosome. (b) Graphical structure of the uncoupled switching state-space model representing the tractable subfamily of distributions employed in the learning algorithm.

In the generalized EM algorithm, the E-step iteratively updates the variational parameters to find the best approximate posterior distribution. The M-step then maximizes the aforementioned lower-bound on the log-likelihood. In general, the KL-divergence minimized may have multiple minima thus deterministic annealing is applied to avoid local minima [4]. Appendix C contains the equations necessary to compute the E-step and M-step of the algorithm.

Chapter 3

Experiments and Results

3.1 Simulated Data

The learning algorithm and model were first validated on simulated "high quality" aCGH datasets containing thirty individuals with 200 clones per chromosome. The datasets were generated using five dosage states corresponding to deletion, loss, normal, gain, and amplification and the means of the latent trajectories were assigned to the theoretical values with a standard deviation of 0.05. The "acceleration" terms were randomly generated from zero-mean Gaussian distributions with a standard deviation of 5×10^{-5}, simulating a low degree of spatial drift. Twelve datasets were generated according to this scheme and the model was learned on each. On average, NL-GIMscan predicted the states with 99.41% accuracy. When the same datasets were learned using HMMs, accuracy decreased slightly to 98.06%. Although the relative difference is small, the HMM suffers in light of a very small degree of spatial drift present in a high quality sample. Real-world datasets display much higher drift and increased correlation between successive clones which the HMM cannot properly annotate as shown in Sections 3.2 and 3.3.

Figure 3.1 shows the true and NL-GIMscan annotated simulated datasets for one individual. As can be seen, there is some visual ambiguity as to the correct states on chromosomes 7, 18, and

19 due to drift of the log-ratio values with dosage states appearing to converge. However, the model does not exactly infer the acceleration terms across the datasets with an average squared error of 2.6206×10^{-9} per acceleration term. Instead, NL-GIMscan learns a smoother trajectory which, although not exactly reproducing the true hidden trajectory, still provides highly accurate state predictions accomplishing the intended goal.

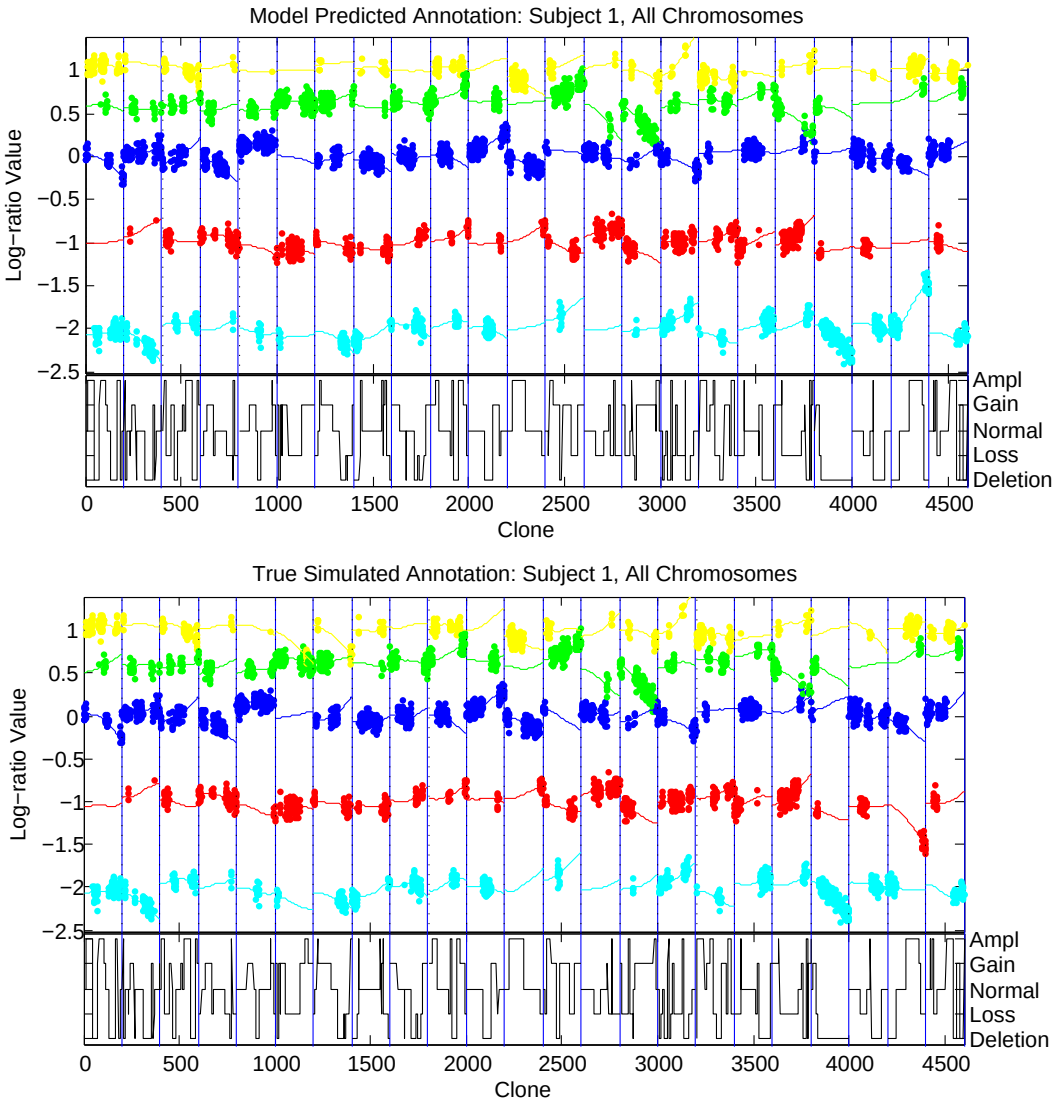


Figure 3.1: True state annotations for a individual from a simulated "high quality" aCGH dataset with five dosage states (top) and NL-GIMscan annotations for the individual (bottom). For each plot, the upper portion depicts the log-ratio values color-coded by dosage state with curves depicting the hidden trajectories of the dosage state-space models and vertical lines denoting divisions between chromosomes. The bottom portion depicts the dosage state of each clone.

3.2 Colorectal Cancer Dataset

NL-GIMscan was learned for a dataset of aCGH profiles for 125 primary colorectal tumors hybridized onto an array consisting of 2,463 clones collected by Nakao *et al* [10]. Nakao *et al.* identified losses often occurring in 8p, 17p, 18p, and 18q and gains often occurring in 8q and 20q via thresholding methods. On average, Nakao *et al.* observed that 17.3% of the entire genome was altered in the samples.

NL-GIMscan identified 29.6% of the entire genome on average as either gained or lost (13.04% and 16.56% respectively). Figure 3.2 shows the fraction of cases gained or lost as identified by the model as well as by Nakao *el al*. As can be seen, similar patterns are observed and the aforementioned losses and gains are identified by the model although the exact percentage varies slightly. Across the population, the model identified high occurances of gains or amplifications in chromosome 7, 8q, 13, 20, and 23 as well as high occurances of losses or deletions in 1p, 5q, 8p, 11p, 14, 15, 17p, 18, and 21. These results concur with those of Nakao *et al.* [10], in addition to identifying additional abberrations which likely were missed due to use of methods ignoring spatial drift. The HMM identified similar patterns of aberrations as NL-GIMscan, but predicted significantly greater occurances of aberrations.

To demonstrate the robustness of the NL-GIMscan, a small-scale study of representative chromosomes containing typical spatial drift patterns demonstrating shortcomings in conventional methods is provided. For comparison purposes, HMM methods were re-implemented according to Fridlyand *et al.* [2] with parameter sharing extensions similar to our model to allow for full genome analysis.

Pattern I: Flat-Arch. Figure 3.3 displays the flat-arch pattern present in the log-ratio measurements of chromosome 4 of individual X77. This pattern is defined by elevated log-ratio values in

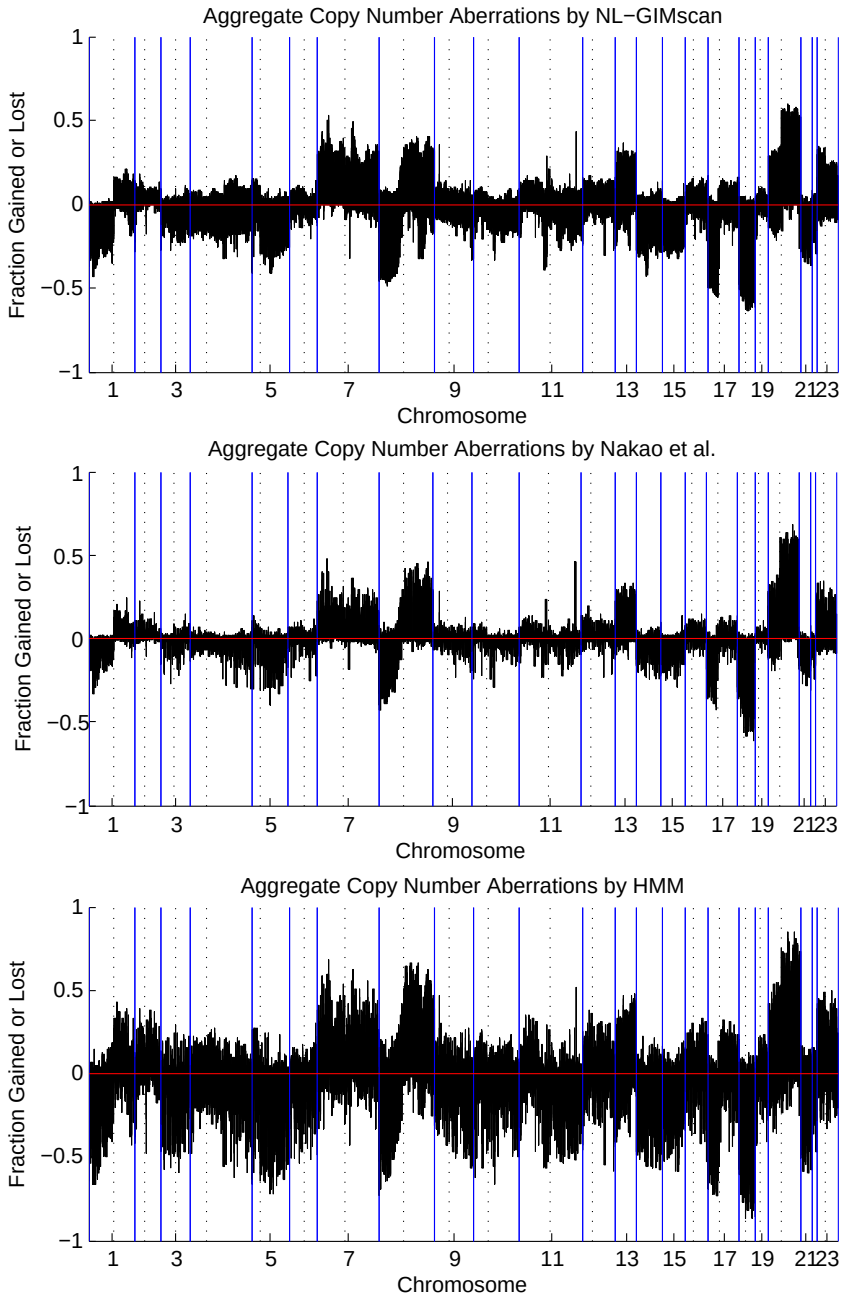
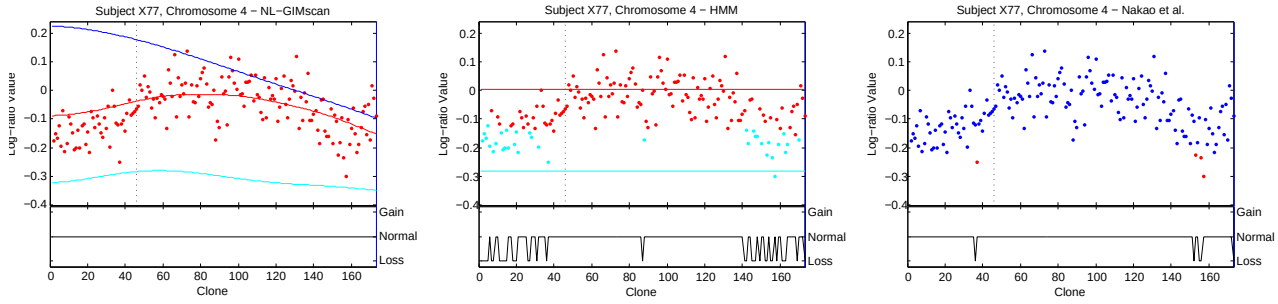


Figure 3.2: Fraction of cases gained or lost as identified by NL-GIMscan (top), by Nakao *et al* [10] (middle), and by HMM (bottom) for colorectal cancer dataset. Data are ordered by chromosomal position of the clones. Solid vertical lines denote the beginning and end of chromosomes and dotted lines denote the location of the centromeres. Lower bars represent losses or deletions while upper bars represent gains or amplifications.

the central region of the chromosome surrounded by lower magnitudes at the telomere regions. Along the chromosome, the hybridization intensities continuously evolve spatially with no clear abrupt changes that signal dosage-state aberrations. NL-GIMscan fits this pattern well, inferring a single non-linear trajectory to the dosage-state. However, methods with invariant state-specific hybridization intensities fail to capture the spatial drift due to the high dispersion of log-ratio values. The HMM may either fit the pattern as a single high-variance Gaussian distribution or divide the pattern into two overlapping Gaussians. As shown in Figure 3.3(b), this may result in inferior state estimation and biologically implausible results due to frequent switching between dosage states along the chromosome.

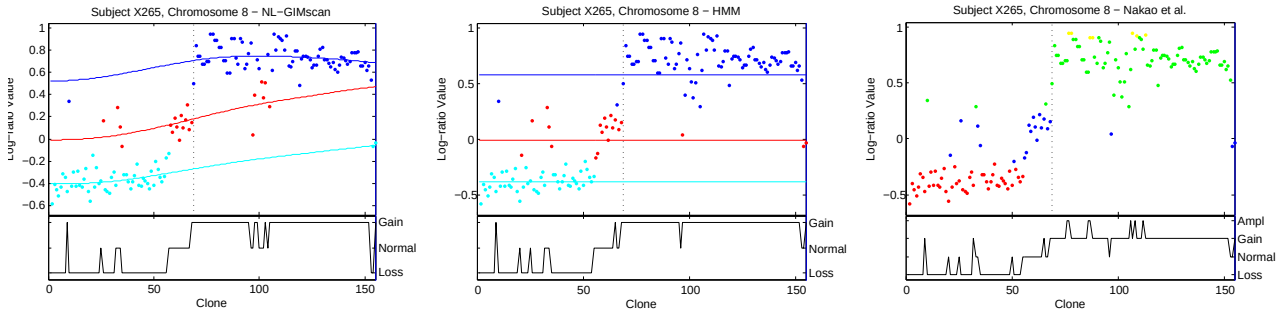


(a) (b) (c)

Figure 3.3: Log-ratio and state estimates for chromosome four of individual X77 displaying the flat-arch pattern. Figure 3.3(a) shows state estimates under NL-GIMscan with curves showing the trajectories of the state-space models. Figure 3.3(b) shows estimates under the HMM with horizontal lines showing the means of the state-dependent Gaussian output distributions. Figure 3.3(c) shows estimates under the thresholding method of Nakao *el al.* For each plot, the upper portion shows the log-ratio values color-coded by state prediction with a vertical dashed line at the centromere while the lower portion shows the state prediction for each clone in the series.

Pattern II: Step. The step pattern occurs in moderately noisy log-ratio measurements in locations where a quantum change of log-ratio magnitudes occurs between ends of a chromosome but the

exact boundary lacks sharpness. Figure 3.4 displays the step pattern present in the log-ratio measurements of chromosome 8 of individual X265. This specific sample also exhibits several local spikes, which are potential dosage-state changes. Both NL-GIMscan and the HMM provide reasonable annotations with minor errors. The HMM does not fully capture the spikes, especially in the region surrounding clone 100. NL-GIMscan better captures the spikes present in the sample and the state change between ends of the chromosome, but predicts too many switches between the normal and gain states around the spike.

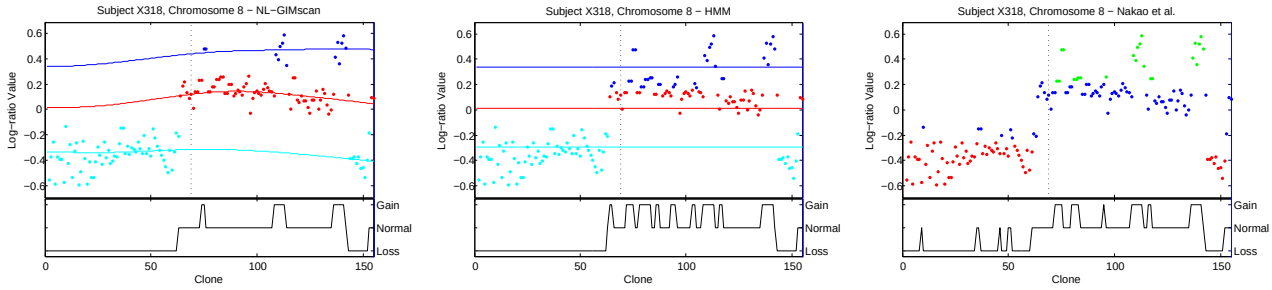


(a) (b) (c)

Figure 3.4: Log-ratio and state estimates for chromosome eight of individual X265 displaying the step pattern. Figure 3.4(a) shows state estimates under NL-GIMscan with curves showing the trajectories of the state-space models. Figure 3.4(b) shows estimates under the HMM with horizontal lines showing the means of the state-dependent Gaussian output distributions. Figure 3.4(c) shows estimates under the thresholding method of Nakao *el al.* For each plot, the upper portion shows the log-ratio values color-coded by state prediction with a vertical dashed line at the centromere while the lower portion shows the state prediction for each clone in the series.

Pattern III: Spikes. The spike pattern is characterized by short sequences of sudden increases in log-ratio value as shown in Figure 3.5, depicting chromosome 8 of individual X318. In this sample, loss occurs in the majority of the 8p arm with spikes near clones 75, 110, and 140 on the 8q arm. The HMM fails to properly annotate the states, assigning two overlapping Gaussians to the

normal and gain states causing biologically implausible state-switching. However, NL-GIMscan properly captures all the aforementioned spikes as well as the state change at the end of the 8q arm.

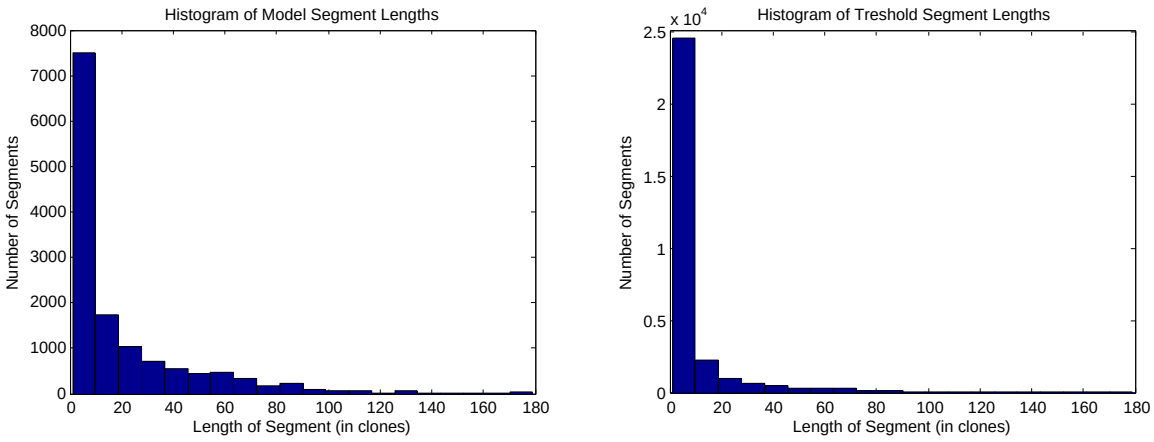


(a) (b) (c)

Figure 3.5: Log-ratio and state estimates for chromosome eight of individual X318 displaying the spike pattern. Figure 3.5(a) shows state estimates under NL-GIMscan with curves showing the trajectories of the state-space models. Figure 3.5(b) shows estimates under the HMM with horizontal lines showing the means of the state-dependent Gaussian output distributions. Figure 3.5(c) shows estimates under the thresholding method of Nakao *el al.* For each plot, the upper portion shows the log-ratio values color-coded by state prediction with a vertical dashed line at the centromere while the lower portion shows the state prediction for each clone in the series.

When compared with both the HMM method and the thresholding approach of Nakao *et al.* [10], NL-GIMscan infers smoother, more biologically plausible dosage states across the whole genome by capturing the spatial drift present in the aforementioned patterns. Figure 3.7 shows the whole genome results for individuals X31, X40, and X77 predicted by the model and the thresholding methods reported in [10]. A visual inspection reveals that the thresholding method predicts a greater degree of state switching than NL-GIMscan in segments which do not appear to correspond to actual dosage state changes. Since the thresholding method inferred five dosage states opposed to the model which inferred three, for comparison purposes, the gain and amplification states as well as the loss and deletion states of the thresholding method will each

be considered single states. The thresholding method predicted 27,516 copy-number breakage points across the 2,463 clones and 125 individuals while NL-GIMscan predicted 10,693, a reduction of 61%. The length of predicted segments by NL-GIMscan tend to be longer as compared to the thresholding method which are more dominated by single clone segments. Figure 3.6 shows the distribution of segment lengths for both NL-GIMscan and the thresholding method.



(a) (b)

Figure 3.6: Histogram of lengths of segments (in clones) as inferred by NL-GIMscan (a) and by thresholding performed by Nakao *et al* [10] (b).

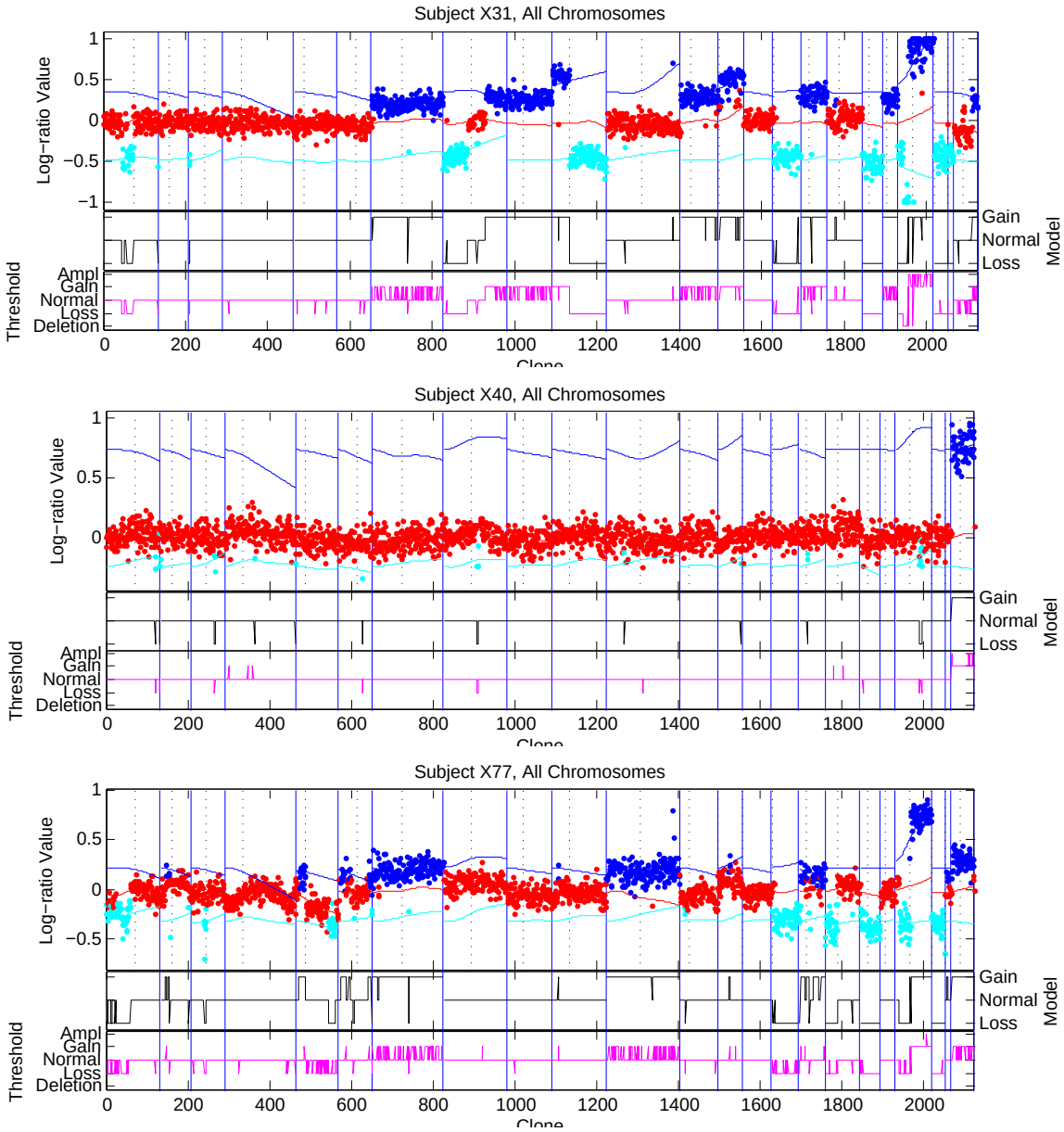


Figure 3.7: Dosage state annotations by NL-GIMscan and threshold for individuals X31 (top), X40 (middle), and X77 (bottom). For each plot, the upper portion shows the log-ratio values color-coded by state prediction with a vertical dashed line at the centromere and solid vertical lines at the end of each chromosome while the lower portions show the state prediction for each clone in the series by NL-GIMscan and thresholding methods.

3.3 Non-Small Cell Lung Cancer Dataset

NL-GIMscan was learned for a dataset of aCGH profiles for 23 non-small cell lung cancer (NSCLC) tumors hybridized onto an array consisting of 12,625 clones collected by Dehan *et al* [1]. Dehan *et al.* identified frequent amplifications of 3q and 8q and deletions of 3p21.31 as well as less common amplifications of 7q22.3-31.31 and 12p11.23-13.2 and deletion of 11q12.3-13.3. Previous studies have revealed gains of 1q21-31, 3q26-qter, 5p13-14, and 8q23-qter as well as loss of 3p14-21, 8p21-23, and 17p12-13 are common in NSCLC samples.

NL-GIMscan identified 15.82% of the entire genome as altered. Figure 3.8 shows the fraction of cases gained or lost as identified by the model as well as by the HMM and Dehan *et al* [1] . As can be seen, NL-GIMscan predicts many short segments of gains or losses in common across the individuals. The model does not predict nearly as many losses or deletions among the tumors as gains or amplifications. Visual inspection of the log-ratio data does not reveal many clear breakage points signaling a change in dosage state beyond short sequences of spikes, reflected in the aggregate results.

To evaluate the dosage state annotations of the model on this dataset, consider two NSCLC tumors, 2002T and 2075T, with log-ratio values typical of this dataset. The full genome annotations for these tumors are below in Figure 3.9. Through AIC model selection, a model with four dosage states was selected corresponding to loss/deletion, normal, gain, and amplification. A visual inspection reveals that the vast majority of values appear to be clustered around zero, the theoretical value for the normal dosage state, with a moderately high variance. A few values appear to be slightly lower than the main cluster which are assigned to the loss state while a few values appear slightly higher which are assigned to the gain state. There are several spikes with values well above points assigned to the gain state which the model assigns to the amplification state with log-ratios ranging from 1.5 to 2.5. Chromosome 5 of tumor 2002T and chromosome

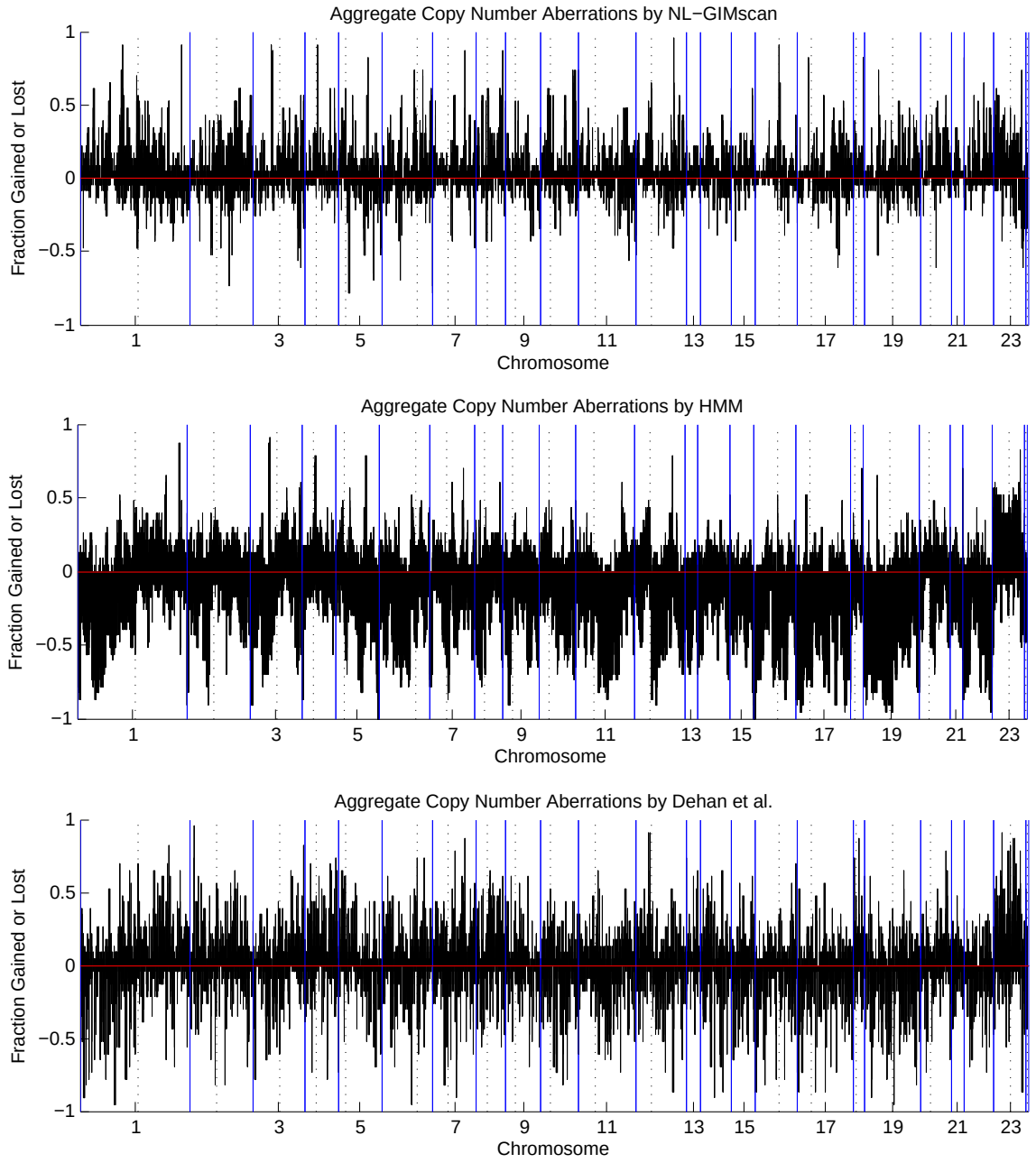


Figure 3.8: Fraction of cases gained or lost as identified by NL-GIMscan (top), by HMM (middle), and by Dehan *et al.* (bottom) for non-small cell lung cancer dataset [1]. Data are ordered by chromosomal position of the clones. Solid vertical lines denote the beginning and end of chromosomes and dotted lines denote the location of the centromeres. Lower bars represent losses or deletions while upper bars represent gains or amplifications.

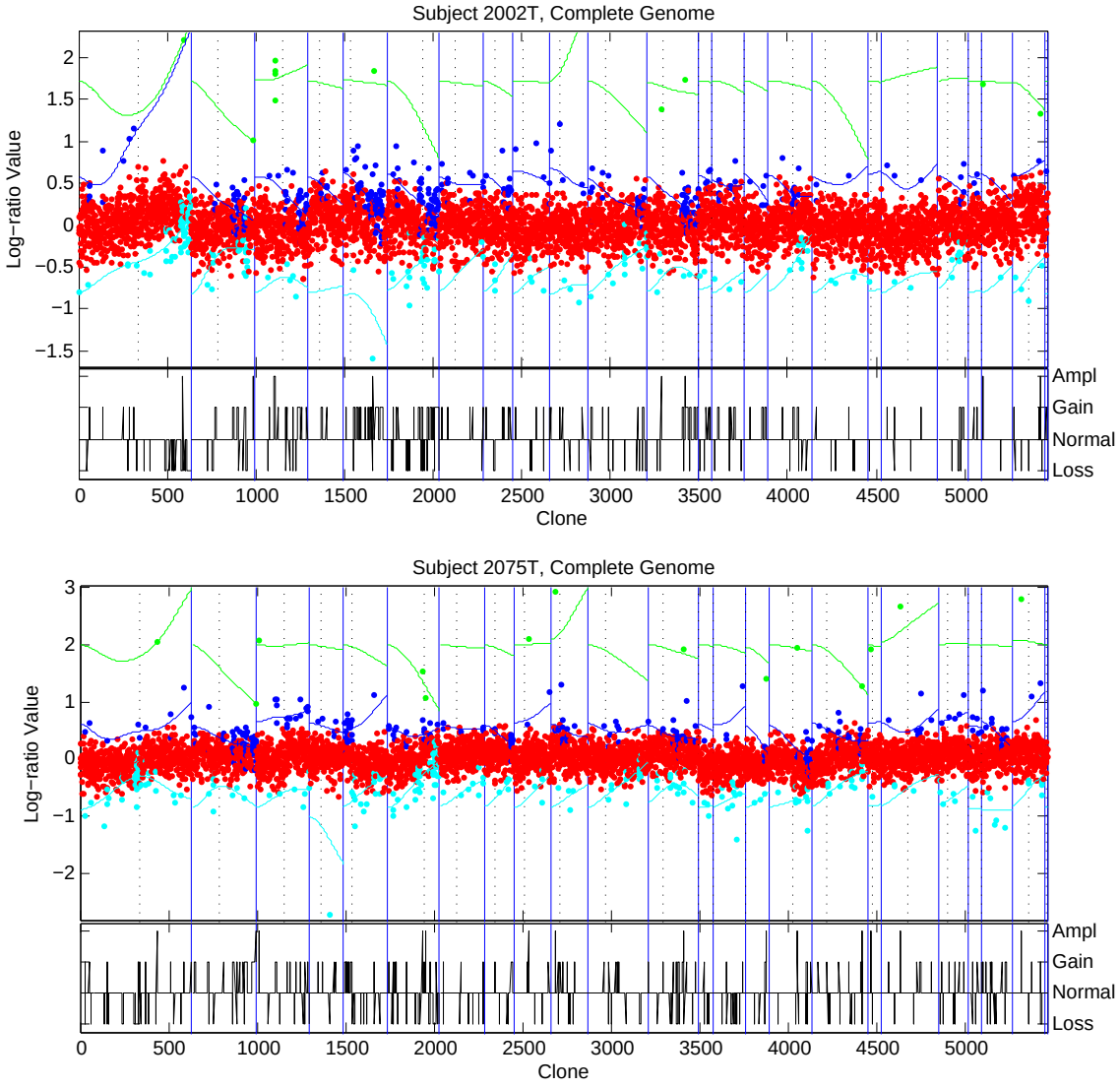


Figure 3.9: Dosage state annotations for tumors 2002T (top) and 2075T (bottom) as predicted by NL-GIMscan with curves showing the state-specific trajectories. For each plot, the upper portion shows the log-ratio values color-coded by state prediction with a vertical dashed line at the centromere and solid vertical lines at the end of each chromosome while the lower portion shows the state prediction for each clone.

4 of 2075T each contain a single point with a significantly smaller log-ratio value than the others. If the model had adopted five states, these points likely would be assigned to the deletion state. Unlike the colorectal tumors examined in Section 3.2, the log-ratio values exhibit few clear segmentation points corresponding to dosage state changes beyond the aforementioned spike pattern.

For comparison purposes, Figure 3.10 contains dosage state annotation for tumors 2002T and 2075T via HMM. AIC model selection chose a model with three dosage states corresponding to loss, normal, and gain. Since the samples lack clearly defined segmentation points, the HMM fits two highly overlapping Gaussians to the central cluster causing frequent switching between dosage states. The spikes with intensities well above the majority of other points are correctly assigned to the gain state. The HMM's overlapping Gaussians do capture the trend of the main cluster to alternate between slightly higher and slightly lower bands of log-ratio values, however, the overall continuous flow of points does not appear to indicate a dosage state change and instead likely suggests detection of some type of systematic error.

There are several genes which NL-GIMscan identifies as either gained or amplified in at least 80% of tumors analyzed and as lost in at least 55% of tumors analyzed. Table 3.1 contains a list of these genes as well as brief descriptions of each. Five genes were identified as having high occurances of either gains/amplifications or losses/deletions occur in areas with known copy number polymorphisms (CNPs) which may account for the model's detection of aberrations due to CNPs in the control sample.

NL-GIMscan infers a somewhat different dosage state annotation from the z-score method employed by Dehan *et al*. The z-score method compares the measured signals of the test samples to the measured signals of four different normal male placentas [1].

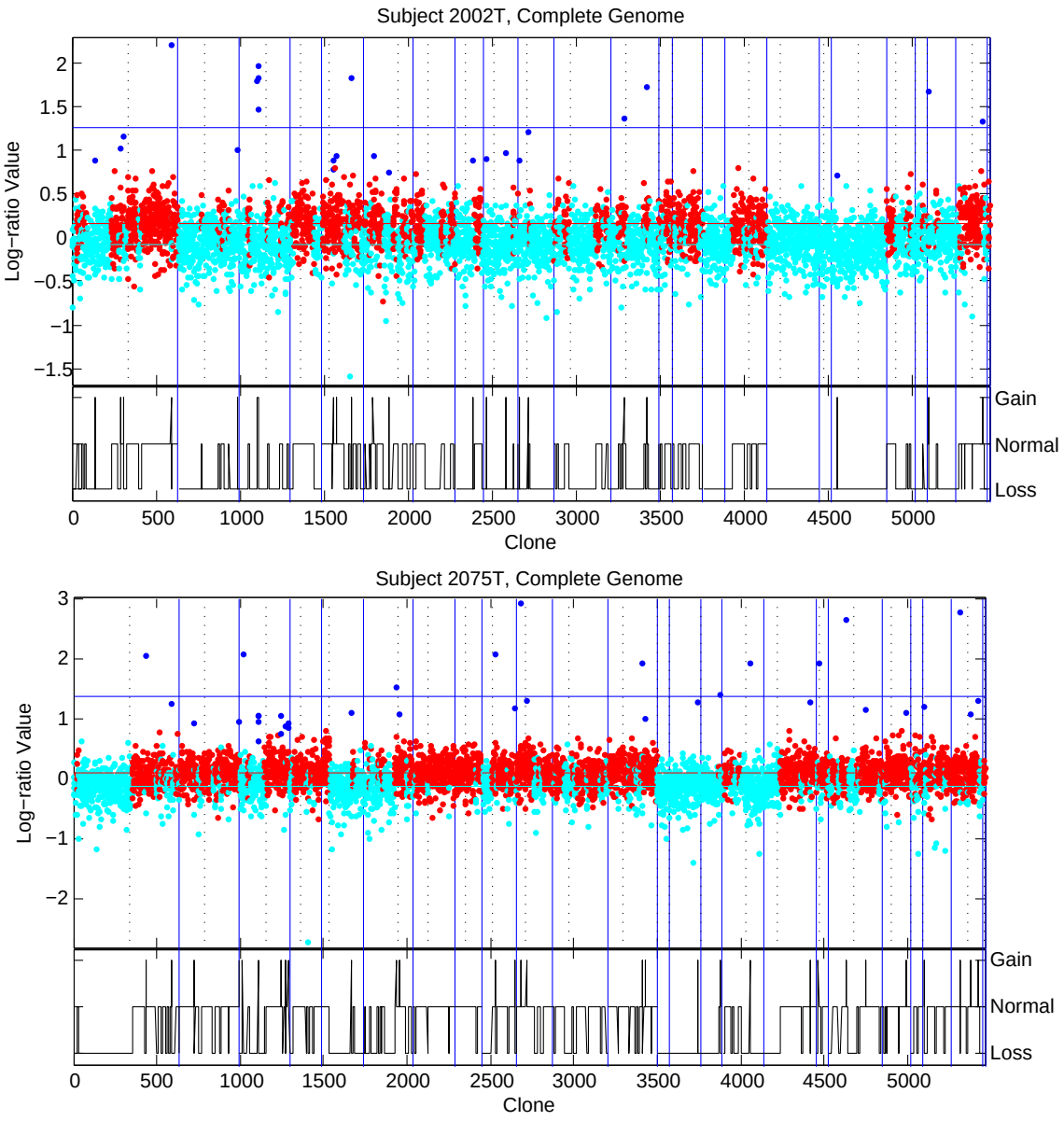


Figure 3.10: Dosage state annotations for tumors 2002T (top) and 2075T (bottom) as predicted by HMM with curves showing the state-specific Gaussian means. For each plot, the upper portion shows the log-ratio values color-coded by state prediction with a vertical dashed line at the centromere and solid vertical lines at the end of each chromosome while the lower portion shows the state prediction for each clone.

At each clone, a z-score is computed by

$$Z_{t,k}^{(j)} = \frac{y_{t,k}^{(j)} - \bar{y}_{t,k}^{*}}{\sigma_{t,k}^{*}} \tag{3.1}$$

where $\bar{y}_{t,k}^{*}$ is the mean signal of the normal male samples at the clone and $\sigma_{t,k}^{*}$ is the standard deviation of the signals of the normal male samples at the clone. A threshold for the Z values is selected for determination of potential gains and losses. This method has the advantage of using multiple normal samples, decreasing the impact of copy number polymorphisms and other noise in the control samples.

Figure 3.11 contains dosage state predictions using the single-clone z-score method when compared to four normal males with a threshold of $|Z| > 3$. A visual inspection reveals that this method roughly corresponds to a thresholding method (with some exceptions) on the modified log-ratio values consisting of the log-ratio of the tumor signal to the mean normal male signal. The z-score method predicts a large degree of oscillation between dosage states while the model predicts a smoother annotation, however, this may be due to averaging and variance across the normal male samples. The overall pattern of each method is similar, though the model incorporates correlation between samples which the z-score method lacks, likely accounting for some of the additional smoothness. Unlike the model, the z-score method can be used for annotation of sequences of clones rather of single clones. This provides an advantage at determining population aberrations of entire arms of chromosomes. The model and z-score method agree on the dosage state of 79.25% of clones across the 23 individuals on average.

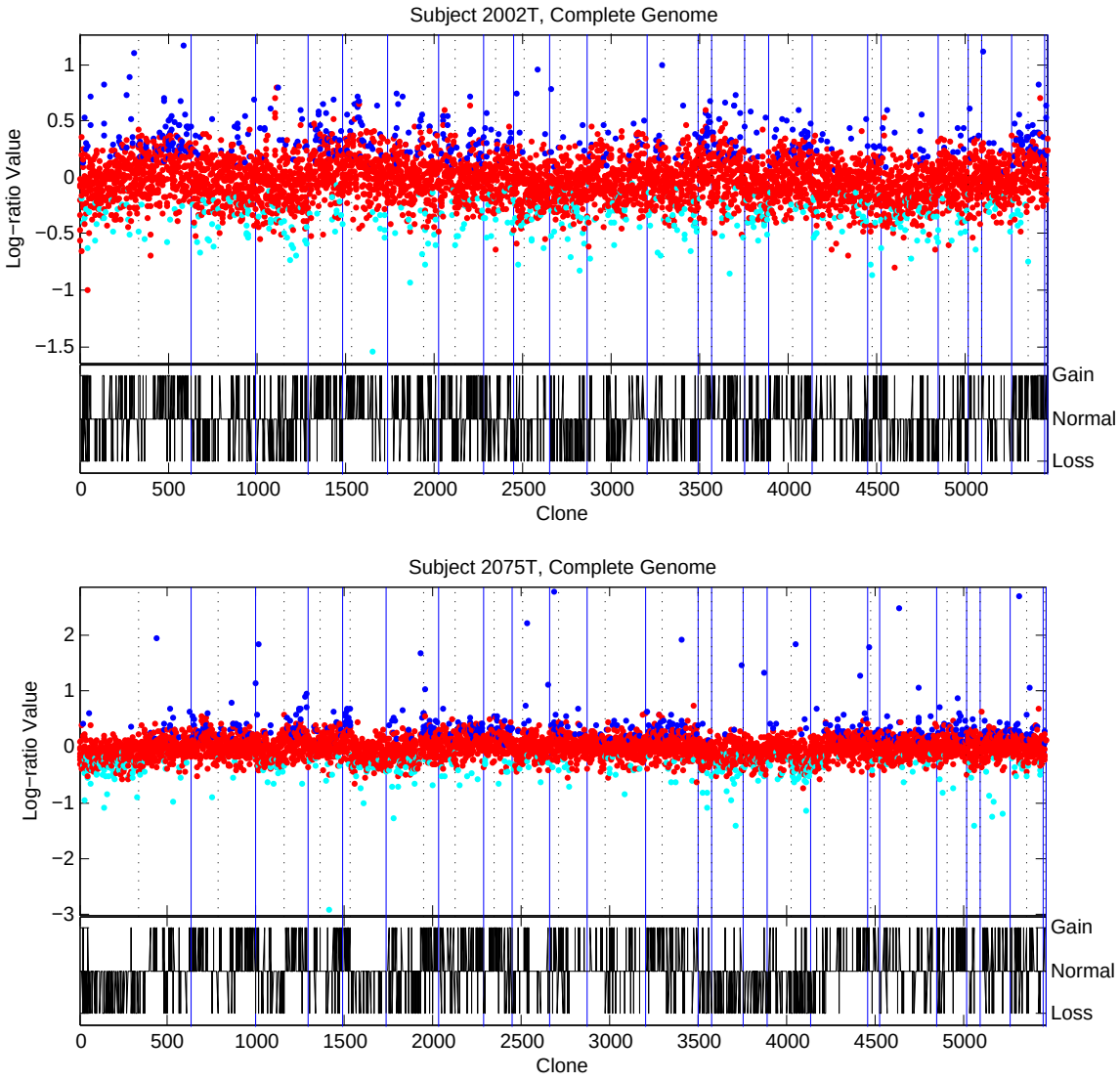


Figure 3.11: Dosage state annotations for tumors 2002T (top) and 2075T (bottom) as predicted by the Z-score method with $|Z| > 3$. For each plot, the upper portion shows the LR values computed as the log ratio of the tumor sample to the mean of the four normal male samples color-coded by state prediction with a vertical dashed line at the centromere and solid vertical lines at the end of each chromosome while the lower portion shows the state prediction for each clone.

Table 3.1: Genes identified as gained/amplified in 80% of NSCLC tumors and genes identified as lost/deleted in at least 55% of NSCLC tumors. Genes identified whose dosage state may be influenced by copy number polymorphisms in the control sample according to the UCSC Genome Browser (`http://genome.ucsc.edu/index.html`) [7] are denoted below.

Clone	Chr	Type	Gene	Cytoband	CNP	Description
250	1	Gain	NFIA	1p31.3-p31.2	-	nuclear factor I/A
590	1	Gain	LEFTB	1q42.1	-	left-right determination, factor B
1106	3	Gain	ACY1	3p21.1	-	aminoacylase 1
1107	3	Gain	ACY1	3p21.1	-	aminoacylase 1
1108	3	Gain	ACY1	3p21.1	-	aminoacylase 1
1109	3	Gain	ACY1	3p21.1	-	aminoacylase 1
1373	4	Gain	PROL3	4q13.3	-	proline rich 3
2221	7	Gain	AASS	7q31.3	-	aminoadipate-semialdehyde synthase
2434	8	Gain	ADCY8	8q24	-	adenylate cyclase 8 (brain)
3425	12	Gain	SLC25A3	12q23	-	solute carrier family 25, member 3
4197	17	Gain	PIGL	17p12-p11.2	Redon [13]	phosphatidylinositol glycan, class L
4517	18	Gain	SCOP	18q21.32	-	SCN Circadian Oscillatory Protein (SCOP)
5100	22	Gain	GGT2	22q11.23	-	gamma-glutamyltransferase 2
861	2	Loss	MGC33864	2q23.3	Redon [13]	hypothetical protein MGC33864
862	2	Loss	REPRIMO	2q23.3	-	candidate mediator of the p53-dependent G2 arrest
863	2	Loss	GALNT5	2q24.1	Sebat [14]	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 5 (GalNAc-T5)
1258	3	Loss	KIAA0861	3q27.3	-	KIAA0861 protein
1272	3	Loss	EPHB3	3q21-qter	-	EphB3
1552	5	Loss	BIRC1	5q13.1	Sebat [14]	baculoviral IAP repeat-containing 1
1691	5	Loss	TTC1	5q32-q33.2	-	tetratricopeptide repeat domain 1
1692	5	Loss	SLU7	5q34	-	step II splicing factor SLU7
1693	5	Loss	ATP10B	5q34	-	ATPase, Class V, type 10B
1694	5	Loss	ATP10B	5q34	-	ATPase, Class V, type 10B
1736	5	Loss	GFPT2	5q34-q35	-	glutamine-fructose-6-phosphate transaminase 2
1737	5	Loss	MGC1127	5q35.3	-	hypothetical gene MGC1127
1848	6	Loss	C6orf10	6p21.3	Redon [13]	chromosome 6 open reading frame 10
2031	6	Loss	XAP135	6q27	-	PHD zinc finger protein XAP135
2032	6	Loss	PSMB1	6q27	-	proteasome subunit, beta type, 1
3162	11	Loss	FLJ10726	11q23.2	-	hypothetical protein FLJ10726
4368	17	Loss	RISC	17q23.2	-	likely homolog of rat and mouse retinoid-inducible serine carboxypeptidase
4937	20	Loss	GHRH	20q11.2	-	growth hormone releasing hormone

Chapter 4

Discussion

Visual inspections of the results reveal that NL-GIMscan performs well at identifying discrete changes in copy number. However, the EM algorithm employed is highly sensitive to parameter initialization. Choice of the initial position of the trajectory μ and output variance Σ_y greatly change the quality of the annotations. However, simple human intervention can significantly improve results. A k-means algorithm was employed to segment the observations into multiple states to initialize the individual-specific μ values, but results were significantly improved after manual adjustment of these means based on quick visual inspections. Given this finding, future work will include software packages to allow users to effectively improve annotations through adjustment of initialization parameters in an efficient manner.

The parameter-sharing scheme is crucial for the success of the model in annotating dosage states of populations, however, it also offers some potential weaknesses which were not addressed by the results. As presented, the model cannot be fit for a single individual as the input terms u can be adapted to fit a single trajectory to all log-ratio values as this will represent the global maximum of the log-likelihood in this situation. Similar issues will likely result in annotating dosage states for small populations, though the minimum acceptable size is unclear. In addition, the parameter sharing scheme of the input terms assumes similar patterns of aberrations throughout

the population. Although such a scheme is appropriate when examining tumors of similar types, it remains to be seen if this scheme will produce high-quality annotations when applied to aCGH profiles of tumors of differing types.

Future extensions of NL-GIMscan may consider the distance between adjacent clones and the length of each clone in estimation, similar to that of BioHMM [9]. In addition, incorporating joint estimation of copy numbers and clone-sequence segmentation within the context of the aCGH model may improve accuracy.

Appendix A

Notation and Variables

Symbol	Size	Description
Parameters/Variables		
A	2×2	State dynamics matrix
B	2×1	State input matrix
C	1×2	Output matrix
$u_{t,k}^{(m)}$	1×1	Input at position t on chromosome k for dosage state m
$\sigma^{(m,j)}$	1×1	Standard deviation of initial state for dosage state m, individual j
$X_{t,k}^{(m,j)}$	2×1	State vector at position t on chromosome k for dosage state m, individual j
$y_{t,k}^{(j)}$	2×1	Observed LR at position t on chromosome k for individual j
$\Sigma_{x_k}^{(j)}$	2×2	State noise covariance matrix on chromosome k for individual j
Dimensions		
J		Number of individuals
K		Number of chromosomes
T		Length of observation sequence
M		Number of dosage states

Appendix B

Software

The provided analyses were produced using a package developed for Java with an interface to MATLAB. The source code and binary packages for both Java and MATLAB are available online at `http://www.andrew.cmu.edu/~jsd1/Thesis`. Both datasets have been made available on this page as well as instructions for using the software packages. The Java software package implements the EM algorithm described in Appendix C in a multi-threaded approach to leverage multiple cores.

Appendix C

EM Algorithm

The equations below are required to compute the expectations in the E-step of the learning algorithm as well as re-estimate the parameters in the M-step. For additional details to implement the learning algorithm including computation of variational parameters through fixed-point equations and order of computations, refer to [4].

C.1 E-Step

Consider the states and predicted log-ratio values for a given hidden trajectory. Let $\mathbf{x}_t^\tau$ denote $\mathbb{E}\left[X_{t,k}^{(m,j)}|\{\mathbf{y}_{1,k}^{(j)},...,\mathbf{y}_{\tau,k}^{(j)}\}\right]$ and let $V_t^\tau = \text{var}\left(X_{t,k}^{(m,j)}|\{\mathbf{y}_{1,k}^{(j)},...,\mathbf{y}_{\tau,k}^{(j)}\}\right)$. In order to calculate the expected log-likelihood, the following expectations are necessary and computed by the Kalman filter forward recursions [3] with the data weighted by the variational parameter $\mathbf{h}$:

$$\mathbf{x}_t^{t-1} = A\mathbf{x}_{t-1}^{t-1} + Bu_{t-1,k}^{(m)} \tag{C.1}$$

$$V_t^{t-1} = AV_{t-1}^{t-1}A' + {\Sigma_x}^{(j)} \tag{C.2}$$

$$K_t = V_t^{t-1}C'\left(CV_t^{t-1}C' + \frac{{\Sigma_y}^{(j)}}{h_t}\right)^{-1} \tag{C.3}$$

$$\mathbf{x}_t^t = \mathbf{x}_t^{t-1} + K_t(y_{t,k}^{(m)} - C\mathbf{x}_t^{t-1}) \tag{C.4}$$

$$V_t^t = V_t^{t-1} - K_t C V_t^{t-1} \tag{C.5}$$

where $\mathbf{x}_1^0 = \mu^{(m,j)}$ and $V_1^0 = \sigma^{(m,j)}$. In order to compute $\mathbf{x}_t^T$, $\mathbb{E}\left[X_{t,k}^{(m,j)} X_{t,k}^{(m,j)\prime}\right] = V_t^T + \mathbf{x}_t^T \mathbf{x}_t^T$, and $\mathbb{E}\left[X_{t-1,k}^{(m,j)} X_{t,k}^{(m,j)\prime}\right] = V_{t,t-1}^T + \mathbf{x}_t^T \mathbf{x}_{t-1}^T{}'$ where T is the total number of observations the following Kalman filter backward recursions are required [3]:

$$J_t = V_t^t A' (V_{t+1}^t)^{-1} \tag{C.6}$$

$$\mathbf{x}_t^T = \mathbf{x}_t^t + J_t(\mathbf{x}_{t+1}^T - A\mathbf{x}_t^t) \tag{C.7}$$

$$V_t^T = V_t^t + J_t(V_{t+1}^T - V_{t+1}^t)J_{t-1}' \tag{C.8}$$

$$V_{t,t-1}^T = V_t^t J_{t-1}' + J_t(V_{t+1,t}^T - AV_t^t)J_{t-1}', V_{T,T-1}^T = (I - K_T C)AV_{T-1}^{T-1}. \tag{C.9}$$

Expectations for the hidden Markov model component of the switching state-space model are computed using the forward-backward algorithm with observations weighted by $\mathbf{q}$. Let $\alpha_{t,k}^{(m,j)} = \mathbb{P}\left(y_{1\ldots t,k}^{(j)}, S_{t,k}^{(m,j)} = 1\right)$ and let $\beta_{t,k}^{(m,j)} = \mathbb{P}\left(y_{t+1\ldots T,k}^{(j)} | S_{t,k}^{(m,j)} = 1\right)$. The following formulas implement the forward-backward algorithm:

$$\alpha_{1,k}^{(m,j)} = \pi_m q_{1,k}^{(m,j)}, \qquad \alpha_{t,k}^{(m,j)} = q_{t,k}^{m,j} \sum_i \alpha_{t-1,k}^{i,j} \Phi_{i,m} \tag{C.10}$$

$$\beta_{T,k}^{(m,j)} = 1, \qquad \beta_{t,k}^{(m,j)} = \sum_i \Phi_{m,i} q_{t+1,k}^{i,j} \beta_{t+1,k}^{i,j}. \tag{C.11}$$

The expectations $\mathbb{E}\left[S_{t,k}^{(m,j)}\right]$ and $\mathbb{E}\left[S_{t,k}^{(m_1,j)} S_{t+1,k}^{(m_2,j)}\right]$ can be computed using the probabilities in Equations C.10 and C.11 as follows:

$$\mathbb{E}\left[S_{t,k}^{(m,j)}\right] = \frac{\alpha_{t,k}^{(m,j)} \beta_{t,k}^{(m,j)}}{\sum_i \alpha_{t,k}^{(i,j)} \beta_{t,k}^{(i,j)}} \tag{C.12}$$

$$\mathbb{E}\left[S_{t,k}^{(m_1,j)} S_{t+1,k}^{(m_2,j)}\right] = \frac{\mathbb{E}\left[S_{t,k}^{(m_1,j)}\right] \Phi_{m_1,m_2} q_{t+1,k}^{(m_2,j)} \beta_{t+1,k}^{(m_2,j)}}{\beta_{t,k}^{(m_1,j)}}. \tag{C.13}$$

C.2 M-Step

Using the expectations computed in Section C.1 for each hidden trajectory, the parameters $\pi, \Phi, \mu, \mathbf{u}, \Sigma_y$ are updated to maximize the expected log-likelihood. The resulting update equations are given below.

$$u_{t,k}^{(m),new} = \frac{B' \sum_{j=1}^{J} {\Sigma_{x_k}^{(j)}}^{-1} \mathbb{E}\left[AX_{t,k}^{(m,j)} - X_{t+1,k}^{(m,j)} | \{y_{1,k}^{(j)}, ..., y_{T,k}^{(j)}\}\right]}{B' \left(\sum_{j=1}^{J} {\Sigma_{x_k}^{(j)}}^{-1}\right) B} \tag{C.14}$$

$$\mu^{(m,j),new} = \frac{1}{K} \sum_{k=1}^{K} X_{1,k}^{(m,j)} \tag{C.15}$$

$$\pi_m^{new} = \frac{\sum_{k,j} \mathbb{E}\left[S_{1,k}^{(m,j)} | \{y_{1,k}^{(j)}, ..., y_{T,k}^{(j)}\}\right]}{\sum_{m,k,j} \mathbb{E}\left[S_{1,k}^{(m,j)} | \{y_{1,k}^{(j)}, ..., y_{T,k}^{(j)}\}\right]} \tag{C.16}$$

$$\Phi_{m_1,m_2}^{new} = \frac{\sum_{t=1}^{T-1} \sum_{j,k} \mathbb{E}\left[S_{t,k}^{(m_1,j)} S_{t+1,k}^{(m_2,j)} | \{y_{1,k}^{(j)}, ..., y_{T,k}^{(j)}\}\right]}{\sum_{t=1}^{T-1} \sum_{j,k,m_3,m_4} \mathbb{E}\left[S_{t,k}^{(m_3,j)} S_{t+1,k}^{(m_4,j)} | \{y_{1,k}^{(j)}, ..., y_{T,k}^{(j)}\}\right]} \tag{C.17}$$

$$\Sigma_y^{(j)new} = \frac{1}{TK} \sum_{m,k,t} \mathbb{E}\left[S_{t,k}^{(m,j)} \left({y_{t,k}^{(j)}}^2 - 2CX_{t,k}^{(m,j)} y_{t,k}^{(j)} + CX_{t,k}^{(m,j)} {X_{t,k}^{(m,j)}}' C'\right) | \{y_{1,k}^{(j)}, ..., y_{T,k}^{(j)}\}\right] \tag{C.18}$$

Bibliography

[1] E. Dehan, A. Ben-Dor, W. Liao, D. Lipson, H. Frimer, S. Rienstein, D. Simansky, M. Krupsky, P. Yaron, E. Friedman, G. Rechavi, M. Perlman, A. Aviram-Goldring, S. Izraeli, M. Bittner, Z. Yakhini, and N. Kaminski. Chromosomal aberrations and gene expression profiles in non-small cell lung cancer. *Lung cancer (Amsterdam, Netherlands)*, 56(2):175–184, May 2007. ISSN 0169-5002. doi: http://dx.doi.org/10.1016/j.lungcan.2006.12.010. URL `http://dx.doi.org/10.1016/j.lungcan.2006.12.010`. 3.3, 3.8, 3.3

[2] Jane Fridlyand, Antoine M. Snijders, Dan Pinkel, Donna G. Albertson, and Ajay N. Jain. Hidden markov models approach to the analysis of array cgh data. *J. Multivar. Anal.*, 90 (1):132–153, 2004. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2004.02.008. 1, 3.2

[3] Zoubin Ghahramani and Geoffrey E. Hinton. Parameter estimation for linear dynamical systems. Technical report, University of Toronto, 1996. C.1, C.1

[4] Zoubin Ghahramani and Georey E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12:963–996, 1998. 2.2, 2.2, C

[5] G. Hodgson, J. H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, N. Nowak, D. G. Albertson, D. Pinkel, C. Collins, D. Hanahan, and J. W. Gray. Genome scanning with array cgh delineates regional alterations in mouse islet carcinomas. *Nat Genet*, 29(4):459–464, December 2001. ISSN 1061-4036. doi: http://dx.doi.org/10.1038/ng771. URL `http://dx.doi.org/10.1038/ng771`. 1

[6] P. Hupé, N. Stransky, J. P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–3422, December 2004. ISSN 1367-4803. URL http://view.ncbi.nlm.nih.gov/pubmed/15381628. 1

[7] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Res*, 12(6):996–1006, Jun 2002. 3.1

[8] Lai, Tze Leung, Xing, Haipeng, Zhang, and Nancy. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*, 9(2):290–307, April 2008. ISSN 1465-4644. doi: 10.1093/biostatistics/kxm031. URL http://dx.doi.org/10.1093/biostatistics/kxm031. 1

[9] J. C. Marioni, N. P. Thorne, and S. Tavaré. Biohmm: a heterogeneous hidden markov model for segmenting array cgh data. *Bioinformatics*, 22(9):1144–1146, May 2006. ISSN 1367-4803. URL http://view.ncbi.nlm.nih.gov/pubmed/16533818. 1, 4

[10] Kentaro Nakao, Kshama R. Mehta, Jane Fridlyand, Dan H. Moore, Ajay N. Jain, Amalia Lafuente, John W. Wiencke, Jonathan P. Terdiman, and Frederic M. Waldman. High-resolution analysis of dna copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, 25(8):1345–1357, August 2004. doi: http://dx.doi.org/10.1093/carcin/bgh134. URL http://dx.doi.org/10.1093/carcin/bgh134. 3.2, 3.2, 3.6

[11] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, October 2004. ISSN 1465-4644. URL http://view.ncbi.nlm.nih.gov/pubmed/15475419. 1

[12] Daniel Pinkel, Richard Segraves, Damir Sudar, Steven Clark, Ian Poole, David Kowbel,

Colin Collins, Wen-Lin Kuo, Chira Chen, Ye Zhai, Shanaz H. Dairkee, Britt-Marie Ljung, Joe W. Gray, and Donna G. Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20(2):207–211, October 1998. doi: http://dx.doi.org/10.1038/2524. URL `http://dx.doi.org/10.1038/2524`. 1

[13] Richard Redon, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, Daniel T. Andrews, Heike Fiegler, Michael H. Shapero, Andrew R. Carson, Wenwei Chen, Eun K. Cho, Stephanie Dallaire, Jennifer L. Freeman, Juan R. Gonzalez, Monica Gratacos, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R. Macdonald, Christian R. Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J. Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F. Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P. Carter, Hiroyuki Aburatani, Charles Lee, Keith W. Jones, Stephen W. Scherer, and Matthew E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, November 2006. doi: 10.1038/nature05329. URL `http://dx.doi.org/10.1038/nature05329`. 3.1

[14] Jonathan Sebat, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Par Lundin, Susanne Maner, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy, James Hicks, Kenny Ye, Andrew Reiner, Conrad C. Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, July 2004. doi: 10.1126/science.1098918. URL `http://dx.doi.org/10.1126/science.1098918`. 3.1

[15] Yanxin Shi, Fan Guo, Wei Wu, and Eric P. Xing. Gimscan: A new statistical method for analyzing whole-genome array cgh data. In *Proceedings of RECOMB 2007*, 2007. 1, 2.1