Model based approaches to array CGH data analysis

by

Sohrab P. Shah

B.Sc. (Hons), Queens University, 1996

B.Sc., The University of British Columbia, 2001

M.Sc., The University of British Columbia, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

(Computer Science)

The University Of British Columbia

(Vancouver)

November, 2008

Abstract

DNA copy number alterations (CNAs) are genetic changes that can produce adverse effects in numerous human diseases, including cancer. CNAs are segments of DNA that have been deleted or amplified and can range in size from one kilobases to whole chromosome arms. Development of array comparative genomic hybridization (aCGH) technology enables CNAs to be measured at sub-megabase resolution using tens of thousands of probes. However, aCGH data are noisy and result in continuous valued measurements of the discrete CNAs. Consequently, the data must be processed through algorithmic and statistical techniques in order to derive meaningful biological insights. We introduce model-based approaches to analysis of aCGH data and develop state-of-the-art solutions to three distinct analytical problems.

In the simplest scenario, the task is to infer CNAs from a single aCGH experiment. We apply a hidden Markov model (HMM) to accurately identify CNAs from aCGH data. We show that borrowing statistical strength across chromosomes and explicitly modeling outliers in the data, improves on baseline models.

In the second scenario, we wish to identify recurrent CNAs in a set of aCGH data derived from a patient cohort. These are locations in the genome altered in many patients, providing evidence for CNAs that may be playing important molecular roles in the disease. We develop a novel hierarchical HMM profiling method that explicitly models both statistical and biological noise in the data and is capable of producing a representative profile for a set of aCGH experiments. We demonstrate that our method is more accurate than simpler baselines on synthetic data, and show our model produces output that is more interpretable than other methods.

Finally, we develop a model based clustering framework to stratify a patient

cohort, expected to be composed of a fixed set of molecular subtypes. We introduce a model that jointly infers CNAs, assigns patients to subgroups and infers the profiles that represent each subgroup. We show our model to be more accurate on synthetic data, and show in two patient cohorts of distinct types of lymphomas how the model discovers putative novel subtypes and clinically relevant subgroups.

Table of Contents

List of Tables

List of Figures

Glossary

ACGH array comparative genomic hybridization, molecular technique for measuring DNA copy number alterations

AF alteration frequency

AUC area under curve

BL blastic-type lymphoma

CNA copy number alteration, genomic amplifications or deletions in the DNA of a sample vs a reference

CNV copy number variation

DLBCL diffuse large B-cell lymphoma

EM expectation maximization,, inference framework for model fitting

ETL enteropathy T-cell lymphoma

FFBS Forwards-filtering backwards-sampling

FPR false positive rate

GMM Gaussian mixture model

HL Hodgkin lymphoma

HMM-MIX HMM mixture model, novel model based clustering method

MCL mantle cell lymphoma

MCMC Markov chain Monte Carlo, sampling based inference framework

NSCLC non-small cell lung cancer

ROC receiver operator characteristic

SCLC small cell lung cancer

TPR true positive rate

Acknowledgements

To my supervisors Drs. Kevin Murphy and Raymond Ng, I could not have asked for a better supervisory team. The complementary levels of insight, involvement and mentoring produced productive years of research and invaluable dissemination of knowledge and constructive feedback - thank you.

I wish to thank the following collaborators at the BC Cancer Research Centre who provided critical scientific and clinical insight into this work. To Dr. Wan Lam and Ron DeLeeuw for providing me with a solid introduction to array CGH and for sharing the MCL data. To Drs. Douglas Horsman and K-John Cheung Jr. for a fruitful collaboration on the analysis of the follicular lymphoma cohort and finally to Drs. Randy Gascoyne and Nathalie Johnson for providing me with the DLBCL data. The funding for my work came from a research grant to Dr. Ng from Genome Canada and a scholarship to me from the Michael Smith Foundation for Health Research.

Most importantly, thank you to my family. To my mother for encourging me to reach higher and to my father for always believing that I would get there. To my wife Nikiah and children Zubin and Zahra for unwavering support, tolerance, understanding and patience that was so much more than I could have asked for. I simply could not have succeeded without you. My free time is yours once again!

Statement of Co-authorship

The material included in all chapters except Chapter 4 is work carried out and written by the author, Sohrab P. Shah. Chapter 4 is a co-authored chapter in which the author designed and carried out the analysis of the data as well as co-authored the manuscript with K-John Cheung. All other co-authors are listed at the start of Chapters 2, 3, 4 and 5 in the publication citation. These authors were collaborators who provided access to data and clinical insight into this research.

Chapter 1

Introduction

1.1 DNA copy number alterations and human disease

DNA copy number alterations (CNA) contribute to the pathogenesis of numerous human diseases including cancer. Also called segmental aneuploidies and chromosomal aberrations, CNAs are discrete genomic intervals in a particular sample, ranging in size from 1 kilobase (Kb) to whole chromosomes, where the number of copies of DNA is higher (amplification, or gain) or lower (deletion, or loss) than in a reference sample (usually with two copies) [1]. A schematic diagram showing a CNA deletion and a CNA amplification on one chromosome is shown in Figure 1.1. A genomic region of loss is shown on the left, and a genomic region of gain is shown on the right. CNAs related to human disease can occur as somatic mutations (as in the case of cancer) where tissue-specific cells are affected, or as congenital abnormalities where germline cells are affected [2]. CNAs can also occur in normal, healthy individuals. Such aberrations are known as copy number variations (CNV) and they represent naturally occurring copy number states in the human population. In contrast to disease-related CNAs which are indicative or causative agents of disease, CNVs characterize individual genomic variation. As such, they can be responsible for phenotypic differences between individuals [3, 4] and are usually benign.

Studying CNAs has a broad scope of applicability in the understanding of human genetics and disease. To illustrate this, we consider the example of the trans-

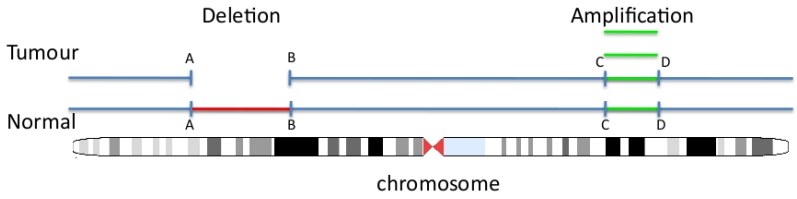


Figure 1.1: A schematic diagram of copy number alterations showing a single chromosome with 2 CNAs. The chromosome is shown on the bottom, the normal reference DNA in the middle and the tumour DNA on the top. Segment A-B in the normal (red) is not present in the tumour representing a CNA loss (deletion) and segment C-D (green) is present in the tumour with 2 additional copies representing a CNA gain (amplification).

formation of a normal human cell to a neoplastic cancerous cell. During this process, a normal cell successively acquires epigenetic changes (that do not modify the genome) and genetic changes (many of which are CNAs), which give rise to growth and proliferation advantages [5]. This is true of nearly all cancer cells. Identifying the CNAs can therefore provide baseline genetic evidence of how a cancerous cell, in terms of its genomic structure, is different from a normal cell. This latter point has implications in our understanding of the clonal evolution of a tumour [2], its molecular characterization, the potential development of diagnostic and prognostic tests, and the development of targeted therapies.

CNAs are often studied concurrently in samples taken from several individuals (from a patient cohort) with the same disease (see for example [6–9]). There are several key pieces of knowledge that can be gained from studying CNAs in a patient cohort. First, we can identify the CNAs for each patient individually. This allows us to relate the presence or absence of a given CNA to clinical outcome/response to therapy data, in order to make inferences about CNAs that may be clinically relevant. A significant complication that arises from determining the CNAs in each patient is that some may be a benign *byproduct* of the clonal evolution of the tumour, and may be irrelevant to the disease. We refer to these alterations as "passengers" [10]. Detection of *recurrent* CNAs (see Figure 3.1 for example) in the patient cohort addresses this issue. These are CNAs that appear more often than

expected in the cohort and provide evidence of disease-specific "driver" CNAs that may have been *selected for* in the clonal evolution of a tumour [2]. Such patterns form a CNA profile for the cohort and can suggest CNA-induced disruption of biochemical mechanisms [11] and/or expression of genes [12] and thus contribute to our understanding of the relevant CNAs for a particular disease.

A confounding issue in the determination of recurrent CNAs is that many cohorts will be composed of molecularly heterogeneous subgroups of patients [13]. If these cohorts are assumed to be homogeneous, CNA patterns may not be visible. Moreover, the neoplastic state of tumours can arise from alternative mechanisms, some of which are mutually exclusive. We can assume that tumours from different patients can have arisen through alternate evolutionary paths [2]. Identifying these alternate paths by defining different patterns of alterations within a cohort can reveal important recurrent CNAs that may have been undetectable under the assumption of molecular homogeneity. Importantly, if clinical outcome data is available, molecular subtypes can be correlated in order to test how molecular alterations confer differential prognoses or response to therapy (see [14, 15] for landmark examples from gene expression data).

1.1.1 Key biological questions related to CNAs

We focus on three key biological questions related to the analysis of CNAs in a patient cohort: a) Where are the CNAs in each patient (Chapter 2)? b) Where are the recurrent CNAs in the cohort (Chapter 3)? c) What are the molecular sub-groups in the cohort and what are their CNA patterns (Chapter 5)? These questions form the clinical and biological motivation behind the body of work presented in this dissertation. Before providing the specifics of our proposed solutions, we first discuss how CNAs can be measured using array comparative genomic hybridization technology given DNA samples.

1.2 Measuring CNAs with array comparative genomic hybridization

CNAs can be measured by a number of different technologies, collectively called genomic hybridization arrays. The general experimental protocol is to select a set

of DNA probes that cover small intervals of the genome. The number (30,000-500,000) and size (10-150Kb) of the probes vary, depending on the specific platform. We will consider whole genome array comparative genomic hybridization (ACGH) [16, 17] which has 30,000 probes of $\sim$100Kb in size which cover the whole human genome in an overlapping tiling arrangement. Other platforms such as SNP genotyping arrays are also used for detecting CNAs and we will discuss these technologies in Section 2.4. In aCGH (depicted schematically in Figure 1.2), the probes are spotted on a glass slide to form the array. Sample DNA and reference DNA are then differentially fluorescently labeled, mixed together and hybridized to the slide. The relative fluorescence intensity of the sample vs. reference (measured by image analysis) is taken as a log ratio for each probe in the array. This results in an ordered set of measurements (by physical genomic location) covering the genome. The measurements have a noisy correspondence to the relative number of copies of DNA of the sample compared to the reference.

An example of the data for one chromosome is plotted in Figure 1.3 (a). This data (see DeLeeuw *et al* [19]) is generated from a mantle cell lymphoma cell line HBL2 and will serve as a running example throughout the text. The horizontal axis represents physical location on the chromosome from p-arm to q-arm and the vertical axis represents the log ratio. The large gap in the center of the plot corresponds to the location of the centromere, where for technical reasons it is difficult to map probes. In Figure 1.3 (b), the data are shown as labeled by an expert [19]: CNA gain probes are indicated by green circles, and CNA loss probes are indicated by red squares. CNA neutral probes are blue crosses.

1.2.1 Statistical characteristics of aCGH data

There are three main characteristics of the data that we can observe from Figure 1.3. First, log ratio levels correspond to CNAs: losses result in negative ratios, gains result in positive ratios, and neutral regions in zero ratios, but the correspondence is noisy. Second, CNAs tend to occur in runs, spanning several contiguous probes. Therefore the CNA state (loss, neutral, gain) is generally dependent on its neighbours. Third, some probes do not follow this trend and produce outlier log ratios (indicated by arrows in Figure 1.3 (b)). These outliers are important to con-

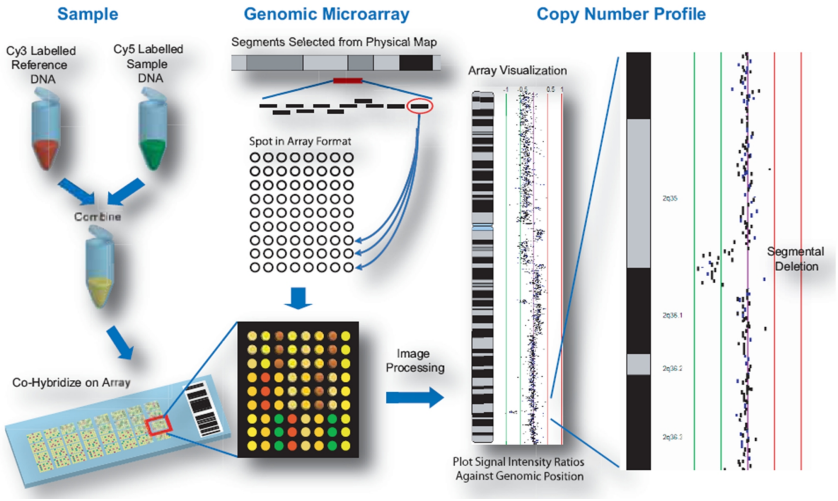


Figure 1.2: Schematic representation of the aCGH experimental protocol, extracted from Chari *et al* [18]. Sample and reference DNA are differentially fluorescently labeled, combined and co-hybridized on the array. The array is prepared by spotting probes selected from the physical map of the human genome. Fluorescence intensities are measured through image processing and a log ratio of intensities is produced for each probe on the array. One can then plot the log ratios as a function of physical location of the corresponding probe (shown right).

sider. Outliers can be explained as experimental noise due to measurement error or mismapped probes; very small real CNAs; or possibly CNVs. The characteristics of noisy signals, spatial correlation and outliers will be important when we develop our models in Chapters 2, 3 and 5.

Obviously, we will not know the labeling when the data is generated in an experiment. Since a typical study will involve tens of samples and each sample has 30,000 data points, manually labeling the data into regions of loss, neutral or gain is a labour intensive task that may be subject to investigator bias. We therefore turn to computational approaches to help consistently and accurately detect CNAs from aCGH.

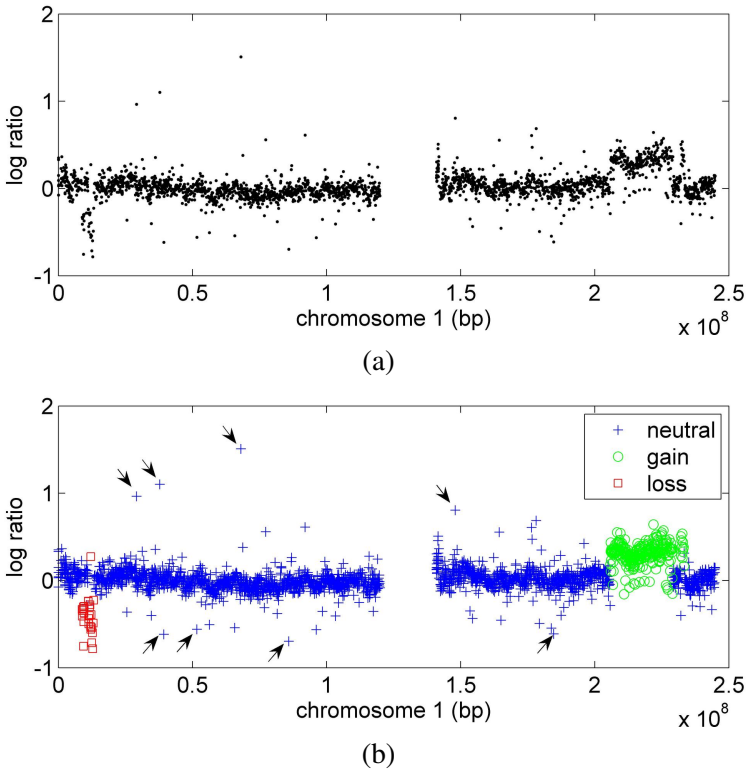


Figure 1.3: (a) Example aCGH data from DeLeeuw *et al* [19] for chromosome 1 of mantle cell lymphoma cell line HBL2. The horizontal axis is the physical chromosomal location of the probe and the vertical axis is the corresponding log ratio. (b) Same data as (a) with CNAs labeled by an expert. Blue crosses indicate neutral probes, green circles are gains and red squares are losses. Some outliers (chosen arbitrarily) are indicated by arrows.

1.3 Research goals

In this dissertation, we discuss three important research goals in the context of computationally detecting CNAs from aCGH data. The goals correspond to the key questions listed in Section 1.1.1. We develop statistical models to A) detect CNAs in a single aCGH sample, B) detect recurrent CNAs in multiple aCGH samples and C) cluster aCGH samples from a cohort into subgroups. These distinct computational problems are depicted schematically in Figure 1.4. The left column

shows the research goals, the right column shows the input and expected output for each goal. The major contributions of this dissertation are the specification of novel models and accompanying inference algorithms as solutions to each of these goals. We briefly outline the proposed computational problems and our solutions here, but refer the reader to the corresponding chapters and published work for details.

1.3.1 Goal A: inferring CNAs from a single aCGH experiment

The objective in Research goal A is to infer CNAs in a single aCGH experiment in order to delineate genomic regions of interest for further investigation. This work is outlined in Chapter 2. This task is by far the simplest of the three, but the framework introduced in its solution is critical to the solutions presented for the more complex goals. Recall Figure 1.3. This shows the raw aCGH data and corresponding CNAs inferred by a cytogeneticist. Our task is to replicate the 'calls' of the cytogeneticist using statistical models. As a solution, we present a novel hidden Markov model (HMM) to infer CNAs from an aCGH experiment. This work was originally published in Shah *et al* [20].

1.3.2 Goal B: detecting recurrent CNAs from multiple aCGH experiments

Research goal B (Figure 1.4B), is to infer recurrent CNAs from a *set* of aCGH experiments, whose corresponding patients form a phenotypic group (eg non-small cell lung cancer patients). The idea is to determine a pattern of CNAs, called a profile, that is common to the patients. The schematic in Figure 1.4B shows a toy example, where aCGH data from 5 patients is the input and the output is a probability curve showing where the recurrent gains and losses are most likely to be. As previously discussed, this gives insight into the molecular characteristics of the phenotype to further understand the disease. Our work on this problem is discussed in Chapter 3. The proposed solution is an extension of the HMM presented in Chapter 2 for goal A into a hierarchical HMM that models recurrent CNAs. This work was originally published in Shah *et al* [21].

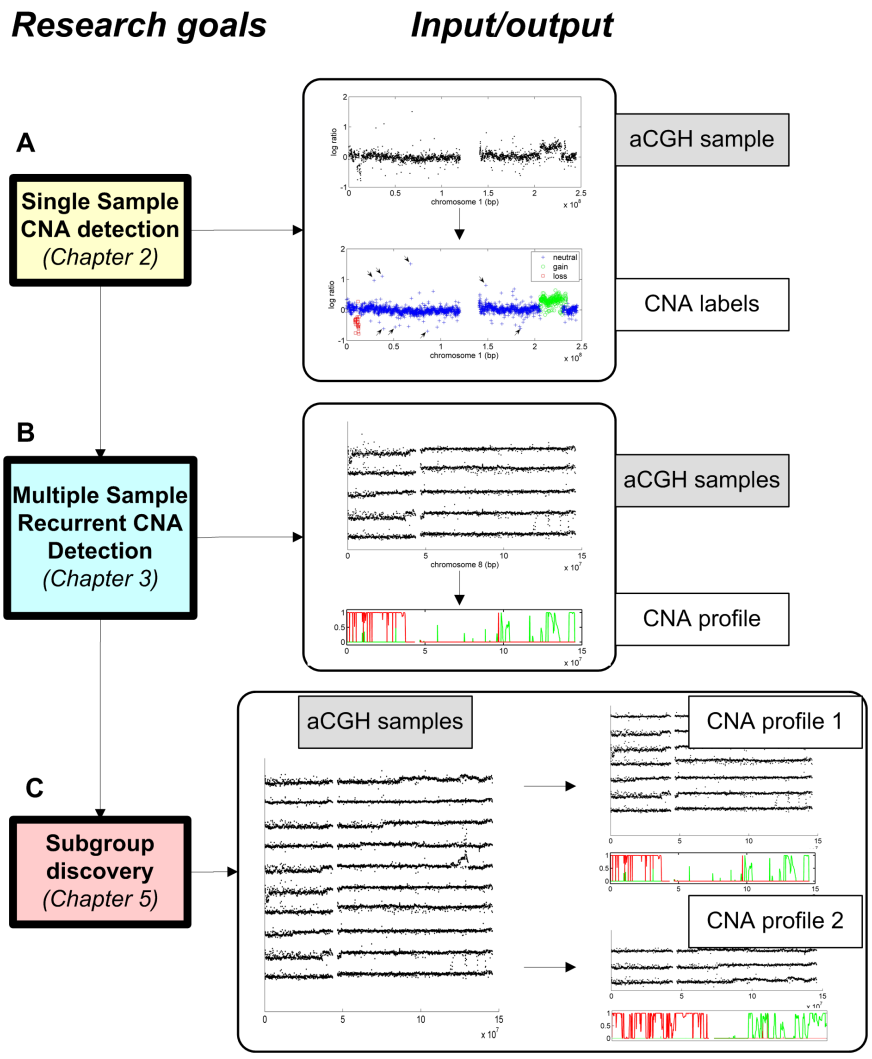


Figure 1.4: Schematic diagram of major research goals. Left column shows the goals indicating the corresponding chapters. Right column shows schematic diagrams of inputs (shaded rectangles) and outputs (unshaded rectangles). Arrows between boxes indicate dependencies. The more complex models depend on the simpler ones. In goal A, we want to classify each probe in the single sample input as loss, neutral or gain. In goal B, we want to infer a recurrent CNA profile from a set of samples. In goal C, we want to discover subgroups and their profiles in a set of samples.

1.3.3 Goal C: unsupervised clustering of aCGH experiments

Research goal C (Figure 1.4C) is to perform unsupervised clustering on a set of aCGH experiments where the cohort of patients is assumed to be composed of a fixed set of molecular subtypes. This is called subgroup discovery and has important application in suggesting multiple molecular mechanisms for acquiring disease characteristics and/or discovery of new molecular subtypes of disease. The toy example, shown in Figure 1.4C shows input data from 10 patients. The algorithm subgroups the patients into 2 groups (indicated by arrows) and computes the profiles for each group (labeled CNA profile 1 and 2). We present a novel model based clustering framework for this problem in Chapter 5.

1.4 Model-based approaches to analysis of aCGH data

Probabilistic graphical models (PGM) offer a robust and principled framework in which to model patterns in noisy and complex data [22]. This framework allows the expression of uncertain (noisy) data in terms of generative underlying probability distributions. PGMs represent probability distributions using graphs, where nodes are random variables in the system and edges represent probabilistic dependencies between nodes [23]. The observed data and the hidden underlying probability distributions can thus be represented in a unified framework (see for example Figures 2.4-2.7).

We take the approach of considering CNAs as latent patterns in the observed aCGH data. We adopt machine learning techniques and PGMs to infer the latent quantities from the observed data. These quantities are inferred with levels of uncertainty that are dependent on the noise characteristics of the data. With this approach in mind, we develop a comprehensive statistical modeling framework for aCGH data that is both flexible and modular. The basic idea is to specify models assumed to have generated the data, then estimate the parameters of the models so as to best explain what we observe. To this end, PGMs offer a convenient way in which to express hierarchical probabilistic models, for example Bayesian probability distributions with conjugate priors [23]. This is a critical aspect of our approach, as we leverage prior knowledge and intuition into our analysis in order to specify models that have intrinsic biological meaning, thus leading to interpretable

results for clinical investigators. In addition, we will show in Chapters 2, 3 and 5, the physical structure of the data can be conveniently represented using PGMs.

PGMs also allow us to build arbitrarily complex models using components of our framework to solve a given task. For each biological question involved in determining CNAs in a patient cohort, outlined above in Section 1.1.1, we propose novel and accurate model-based approaches using PGMs. As the problems increase in complexity (ie research goal C is more complex than research goal B, which in turn in more complex than research goal A), the solutions leverage the simpler models and simply incorporate them in a more complex structure. Moreover, PGMs provide a general "inference engine" [23] that can be applied to fitting arbitrarily complex models to data. We employ this theory in the development of our framework and inference routines.

1.5 Data sets

In order to see the utility, and to evaluate how well our solutions were working, we applied the methods related to our research goals to several real-world data sets generated by colleagues at the BC Cancer Agency. Table 1.1 lists the disease entity they represent, the number of aCGH experiments (cell lines or patients as the case may be), which research goal was applied, the Chapter that refers to the data set, and the status of the project. In all we tested and applied our models on seven different disease entities and a total of 339 aCGH samples.

The application ranged from cell-line data to clinical samples. A key point is that for some cell lines, we had ground truth data (CNAs were determined by manual analysis and some of those were verified by fluorescence in situ hybridization) and thus quantitative metrics could be computed to evaluate the accuracy of our approaches. In addition, the characteristics of cell line data is that they produce much cleaner signals than in clinical samples and therefore, made for very good initial benchmarking data sets on which to develop our algorithms. However, it was always our intention to evaluate how our algorithms would perform on previously unstudied clinical data sets for which no ground truth was available. The 106 aCGH samples generated for follicular lymphoma (the subject of Chapter 4), made for an ideal case study with the goal of revealing new science with respect

Table 1.1: Real-world aCGH data sets on which we have applied our models

Disease	#	Ref	Goal	Chap	status
mantle cell lymphoma	8 cell lines	[19]	A	2	complete
blastic-type lymphoma	11	[24]	A	2	complete
enteropathy T-cell lymphoma	30	[25]	A	2	complete
follicular lymphoma	106	[7]	A,C	4,5	complete
lung cancer	39 cell lines	[11]	B	3	complete
diffuse large B-cell lymphoma	92	n/a	A,B,C	5	ongoing
Hodgkin lymphoma	53	n/a	A,B,C	n/a	ongoing

to this disease. By working extensively with this data set, we were able to assess qualitatively how well our approaches were working on real data. This process required a focused collaboration with our colleagues that resulted in improving our models for robust application, enabling new and relevant clinical and biological inferences. In a similar vein, we remain engaged in two ongoing studies (bottom two rows of Table 1.1) in diffuse large B-cell lymphoma (DLBCL - 92 cases) and Hodgkin lymphoma (HL - 53 cases) where we continue to refine our models for CNA characterization of these diseases in a clinical setting.

In addition to the real-world data sets described above, we generated synthetic data for each goal in order to test the theoretical properties of our models. This enabled quantitative benchmarking against standard methods (see Chapters 2, 3 and 5 for details).

1.6 Dissertation outline

The rest of the dissertation is organized as follows: In Chapter 2, we demonstrate our approach to detecting CNAs from a single aCGH experiment. We outline two novel extensions to hidden Markov models that confer higher accuracy over standard methods, and introduce the statistical framework that forms the foundation for all work presented in this dissertation. Chapter 3 discusses our contribution to problem of detect recurrent CNAs in a set of aCGH experiments derived from a patient cohort. This work extends the models discussed in Chapter 2 to the multiple sample case. We derive a novel method that explicitly models driver and passenger CNAs, and thus successfully filters out the passengers, while reporting sparse profiles of

CNAs that represent putative driver alterations in the cohort. We demonstrate how our methods are more sensitive to finding signals in the data that may be lost using standard methods. Chapter 4 is an important chapter that describes the first high-resolution genomic profile for follicular lymphoma, that resulted in making refinements to our model presented in Chapter 2 for clinical application. We show how our methods led to the description of clinicopathologically significant CNAs that are now being rigourously pursued as candidate driver CNAs with prognostic relevance in this disease. We also seeded the idea of the clustering while working with this data that led to the concepts presented in Chapter 5. Chapter 5 introduces a novel model-based approach, based on a mixture of HMMs, for clustering aCGH data. This method outperforms partitioning and hierarchical clustering methods, and in an application to two rich clinical data sets demonstrates that it is capable of discovering clinically relevant molecular subtypes. In the FL cohort mentioned above, we show how our model successfully deals with heterogeneous molecular subtypes by stratifying patients into groups with prognostically distinct molecular profiles. Chapter 6 is a summary of our results and offers thoughts on future directions for the research presented herein. Each of the technical chapters 2-5 are intended to be self-contained and can be read independently of the others.

Chapter 2

Detecting CNAs from array CGH data

Summary

In this chapter, we consider the problem of inferring CNAs from a single aCGH experiment (research goal A). We begin by outlining the related work on this topic in Section 2.1 including smoothing, segmentation and state-space models; and illustrate potential limitations of previously published contributions. In Section 2.2, we introduce our contribution, an extension of a continuous emission hidden Markov model that explicitly models the statistical properties of aCGH data. We show in Section 2.3 how our novel HMM that borrows statistical strength across chromosomes for parameter estimation and explicit modeling of outliers outperforms standard baseline models using cell line data and synthetic data with ground truth [1]. In Section 2.4, we discuss the impact of this work to the bioinformatics, cancer and cytogenetics research communities and outline ideas for future directions.

[1] Some of the material in this chapter was previously published in: S P Shah, X Xuan, R J DeLeeuw, M Khojasteh, W L Lam, R Ng, and K P Murphy. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics, 22(14):431439, Jul 2006.

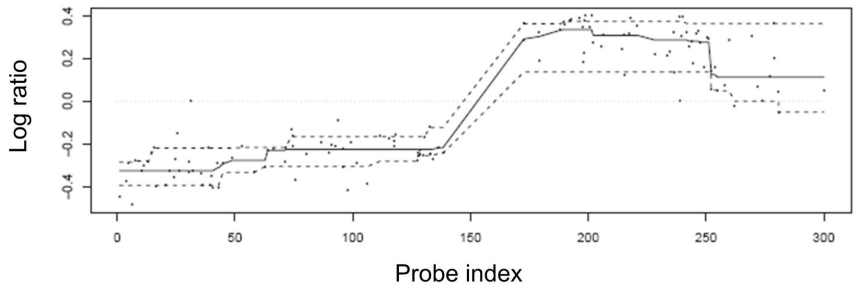


Figure 2.1: Output from Quantreg, a smoothing algorithm due to Eilers and Menenzes [26]. The output provides a smooth trace through the data, but does not provide any explicit predictive information about the location of the CNAs. Figure extracted from Eilers and Menenzes *et al* [26].

2.1 Related work: algorithms for single sample aCGH analysis

2.1.1 Notation and preliminaries

The analytical goal in this chapter is to infer CNAs from aCGH data derived from a single aCGH experiment. Let $y_{1:N_c} \in \mathbb{R}$ be the log ratios observed from the aCGH experiment, where $y_{1:N_c} = \{y_1, \ldots, y_{N_c}\}$. For a given probe t, y_t is noisily related to $\log \frac{I_t^s}{I_t^r}$ where I_t^s is the fluorescence intensity of the sample (or tumour) DNA at probe t and I_t^p is the fluorescence intensity of the reference (or normal) DNA at probe t. The intensity measurements are proportional to the DNA copy number number, which for the reference is usually 2. y_t is noisy due to technical errors and systematic variability arising from numerous sources and we assume the data has been normalised against these systematic biases[2]. We consider a single chromosome c at a time (this reflects the physical arrangement of DNA in the cell), with probes indexed from $(1,2,\ldots,N_c)$ where N_c is the number probes on the chromosome and the indices denote the probes' relative ordering by physical location on the chromosome. We wish to identify which probes in the data are

[2]Normalization is not the focus of the work presented in this thesis, but we refer the reader to Khojasteh *et al* [27] for detailed explanation of the sources of error and how they are corrected. This normalization method is used for all data presented herein.

likely to represent a CNA. We will examine how various authors in related work have modeled $y_{1:N_c}$ in order to interpret the data with respect to CNAs. In general, previous approaches can be characterised into four main categories: smoothing, segmentation, independently and identically distributed (IID) mixture models, and state space models. We will show the general characteristics of these methods and demonstrate their limitations. In Section 2.2, we propose improvements to these methods in order to gain accuracy in predicting CNAs.

2.1.2 Smoothing algorithms

We begin our discussion with smoothing algorithms. Smoothing approaches, such as [26, 28] generally model $y_{1:N_c}$ using regression to fit curves to the data. The output is a smooth trace $x_{1:N_c}$ through the data that is calculated by considering neighbouring probes. For example, Eilers and Menenzes [26] fit a curve to the data by minimizing:

$$J(x_{1:N_c}) = \sum_{t=1}^{N_c} |y_t - x_t| + \lambda \sum_{t=2}^{N_c} |x_t - x_{t-1}| \tag{2.1}$$

with respect to $x_{1:N_c}$. A typical example of smoothing output is shown in Figure 2.1 (extracted from Eilers and Menenzes [26]), depicting a smooth curve through the data. While this is helpful in that the data are denoised, smoothers do not provide explicit predictive information about which probes are CNAs - our primary objective in aCGH analysis. Smoothing algorithms are therefore used primarily as a data visualisation tool which must be subjectively interpreted to make real conclusions about the data. A somewhat more practical approach is realised by segmentation algorithms.

2.1.3 Segmentation algorithms

The bulk of the literature on aCGH analysis is related to segmentation algorithms. A segment is a contiguous set of probes assumed to share the same mean log ratio. Segmentation approaches partition the data into piecewise constant intervals by determining the segment boundaries, also referred to as chromosomal breakpoints. Segmenters therefore provide the users with a list of locations where the data is changing sharply. The breakpoints are usually determined so as to minimize the

within segment variation. Examples of segmentation algorithms include: DNACopy [29, 30], CGHSeg [31], aCGHSmooth [32] and GLAD [33]. Segmentation of aCGH data is also an example application of the multiple changepoint problem studied by Fearnhead [34]. Fearnhead's formulation of the problem models the data within a segment as follows:

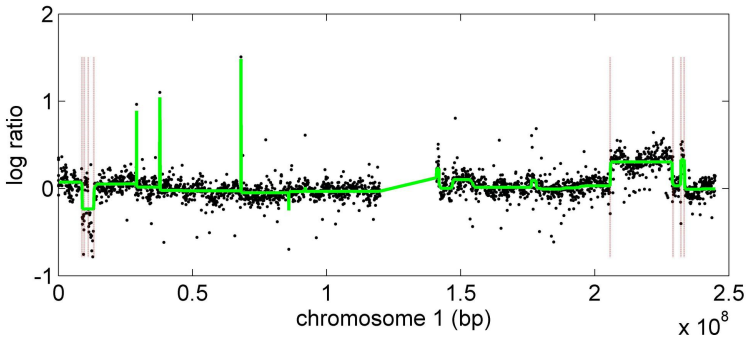
$$\vec{y}^i = \vec{G}^i \vec{\beta}^i + \varepsilon_i \tag{2.2}$$

where i represents the ith segment in the data, $\vec{G}^i$ is a matrix of basis functions for the ith segment under a polynomial piecewise constant assumption, $\vec{\beta}^i$ is the set of model parameters (eg mean for polynomial of order 1) for segment i and ε^i represents random Gaussian noise with mean 0 and segment-specific variance σ_i^2. Given this model, the task is to infer the number and positions of the changepoints.

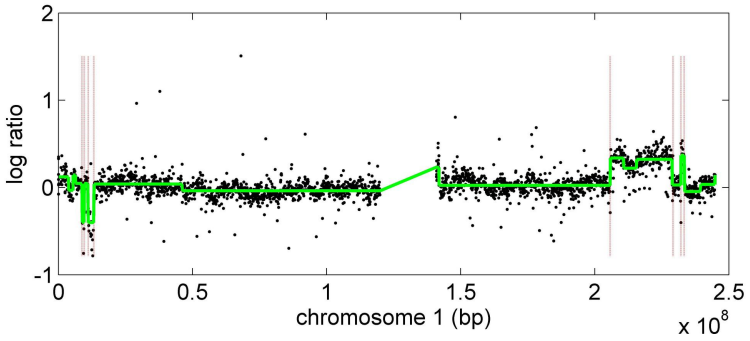
Figure 2.2 shows two examples of the output of segmentation using Fearnhead's algorithm [34] (a) and DNACopy (b)— considered to be the best segmentation algorithm in two separate evaluation studies [35, 36]. As Figure 2.2 illustrates, the drawback of segmentation is that there is no intrinsic biological meaning of the mean level of the segments. In fact, the data could be segmented into an arbitrary number of levels. Segmentation often produces an over-represented number of states in comparison to the biologically meaningful CNA states (note the single clone segments and numerous segments in the ground truth neutral regions). Post-processing is therefore required to infer the (loss, neutral, gain) CNA states. Algorithms such as GladMerge [33] and MergeLevels [36] are designed for this purpose. We have found that in general segmentation followed by post processing is susceptible to false positives. This is shown in Figure 2.3 which is the output of DNACopy (Figure 2.2(b)) post-processed with MergeLevels. From here onwards we will refer to this algorithm as DC+ML. We will demonstrate that it is more effective to jointly infer segments and changepoints simultaneously in Section 2.3.2.

2.1.4 IID Mixture models

A solution to the over-represented number of segments problem is to classify each probe into a fixed number of states K. We can assign biological meaning to the states, therefore the output is immediately interpretable. For example, IID Gaus-



(a) Fearnhead



(b) DNACopy

Figure 2.2: Example output of 2 segmentation algorithms on HBL2 chromosome 1: (a) Fearnhead and (b) DNACopy. The green horizontal lines indicate the mean level of the segments. The vertical lines indicate the ground truth changepoints. In both cases, the segmentation algorithms are predicting more segments than indicated by the ground truth.

sian mixture models (GMM) applied to aCGH by Hodgson *et al* [37] provide the desired output. We sketch this model as a directed graphical model (Bayesian network) in Figure 2.4. In GMMs for aCGH, the number of $K = 3$ states represent CNA loss, neutral and gain ($\{L,N,G\}$). GMMs introduce a set of $Z^c_{1:N_c}$ multinomial random variables, where $Z^c_t = k$ means probe t on chromosome c is in state k. The prior probability, $p(Z^c_t = k)$ is represented by $\pi_c(k)$. The key part of the model is a probabilistic dependency of y^c_t on Z^c_t:

$$p(y^c_t|Z^c_t = k) = \mathcal{N}(y^c_t|\mu^c_k, \sigma^c_k) \tag{2.3}$$

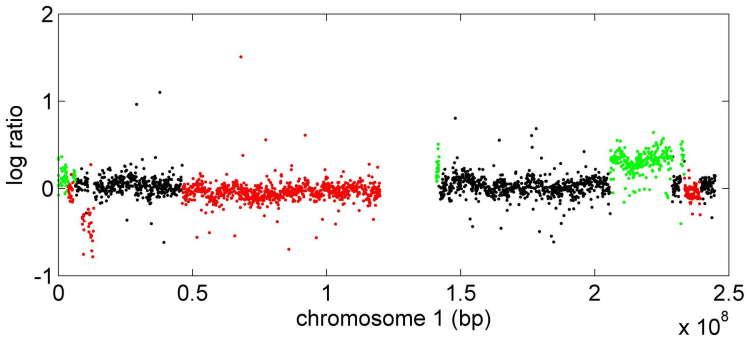


Figure 2.3: Example output of DNACopy + MergeLevels. The over-segmentation from DNACopy results in three false positive loss predictions in comparison to the ground truth.

Therefore, y_t^c is assumed to be generated by a state dependent Gaussian distribution[3]parameterized by μ_k^c, σ_k^c. The marginal distribution of y is therefore:

$$p(y_t^c) = \sum_{k=1}^{K} \pi_c(k) \mathcal{N}(y_t^c | \mu_k^c, \sigma_k^c) \tag{2.4}$$

a convex combination of the Gaussian emission densities, weighted by π_c. The parameters $(\pi_c, \mu_{1:K}^c, \sigma_{1:K}^c)$ can be fit to the data using maximum likelihood or MAP estimation in an expectation maximization (EM) framework (see Bishop (2006), ch 9 [22]). Since our goal is to infer the CNA state from the data, we can make use of Bayes' rule and calculate posterior probabilities $\gamma_t(k) = p(Z_t^c = k | y_{1:N_c}, \mu_k^c, \sigma_k^c)$:

$$\gamma_t(k) = \frac{\pi_c(k) \mathcal{N}(y_t | \mu_k^c, \sigma_k^c)}{\sum_l \pi_c(l) \mathcal{N}(y_t | \mu_l, \sigma_l)} \tag{2.5}$$

Typical output on HBL2 chromosome 1 (same data as in Figure 1.3) of the MAP method is shown in Figure 2.8 (top). The output indicates the posterior probability that each probe is a loss, neutral or gain, which is our desired output.

An important consideration in the MAP estimation is the use of conjugate prior

[3]This is an approximation that in our experience fit the data well in the vast majority of cases. In a small subset of the data, loss regions and gain regions showed slight skewness and thus alternative distributions such as Gamma might be used. This is a potential topic for future work.

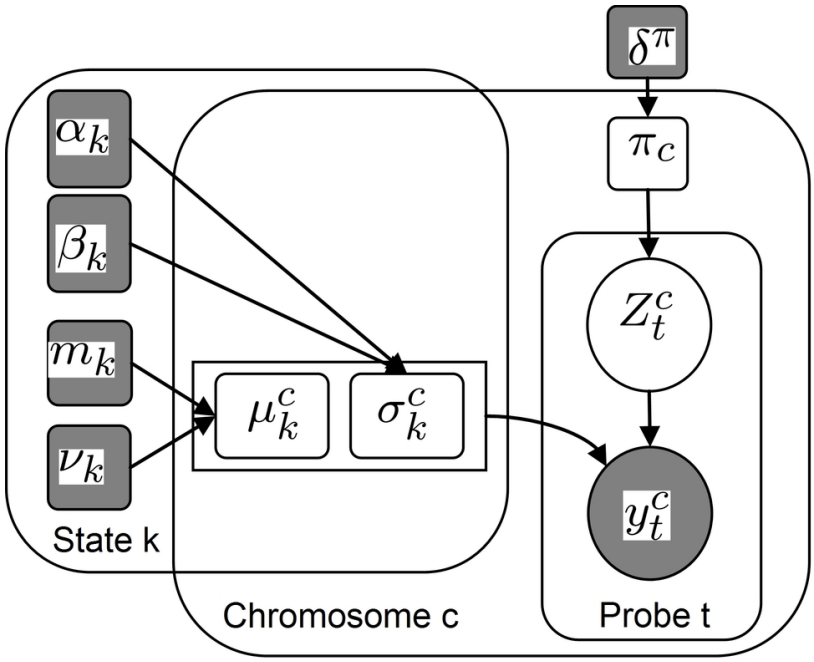


Figure 2.4: Probabilistic graphical model of the Gaussian mixture model (GMM). Square nodes are parameters or hyperparameters, round nodes are random variables. Shaded nodes are known quantities, unshaded nodes are unknown. The large rounded rectangles are called plates represent repetition of the contents inside. Arrows between nodes indicate probabilistic dependency between variables. We let c denote the chromosome and k represent the state. We show the generative mechanism for y_t^c as a state-dependent Gaussian emission density with parameters μ_k^c, σ_k^c indexed by the latent variables Z_t^c, representing CNA states. π_c is a multinomial distribution over the states and δ^π is a Dirichlet prior distribution over π_c. In this model, the latent state labels are independent, therefore spatial correlation is not modeled. Please refer to Figures 2.5, 2.6 and 2.7 for Markov extensions to this model that do model spatial correlation.

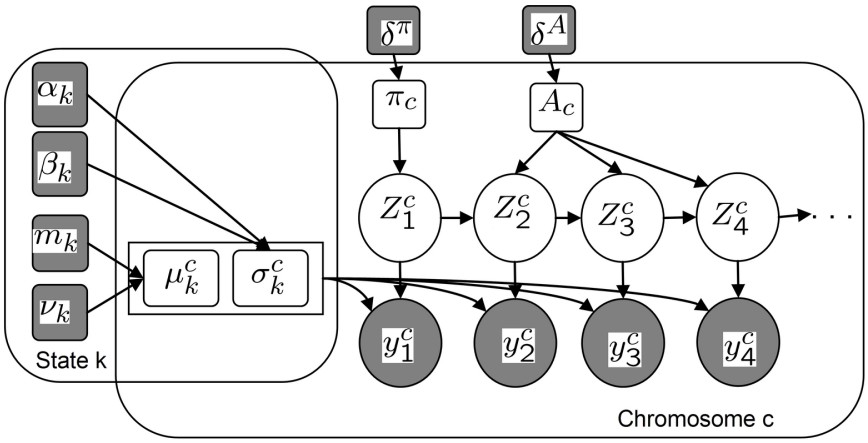


Figure 2.5: Graphical model of the continuous emission HMM (HMM-SC). See legend of Figure 2.4 for description of graphical model notation. In comparison to the GMM (Figure 2.4), Z_c^t are no longer plated for probe t. Instead Z_t^c is dependent on Z_{t-1}^c and a transition matrix A^c, thus modeling spatial correlation with Markov dynamics. In the HMM-SC model shows that parameters μ^c, σ^c, A^c are inside the chromosome plate, indicating they are chromosome specific.

distributions on the parameters of the model μ, σ (we drop the c superscript for brevity). These are modeled as follows:

$$\mu_k \sim \mathcal{N}(\mu_k | m_k, \sigma_k^2 \nu_k) \tag{2.6}$$

where m_k is the prior mean of μ_k and ν_k is the variance of the prior.

$$\sigma_k^{-2} \sim Gam(\sigma_k^{-2} | \alpha_k, \beta_k) \tag{2.7}$$

where *Gam* is the Gamma distribution and α_k, β_k are the shape and scale hyperparameters respectively. Setting the hyperparameters $(m_k, \nu_k, \alpha_k, \beta_k)$ is particularly important in such models because we wish to assign intrinsic meaning to k where $k = 1$ is loss, $k = 2$ is neutral and $k = 3$ is gain. Therefore, it is imperative that $\mu_1 < \mu_2 < \mu_3$ hold in order for the output to be interpretable. This condition enforces identifiability of the states and prevents the label switching problem [38].

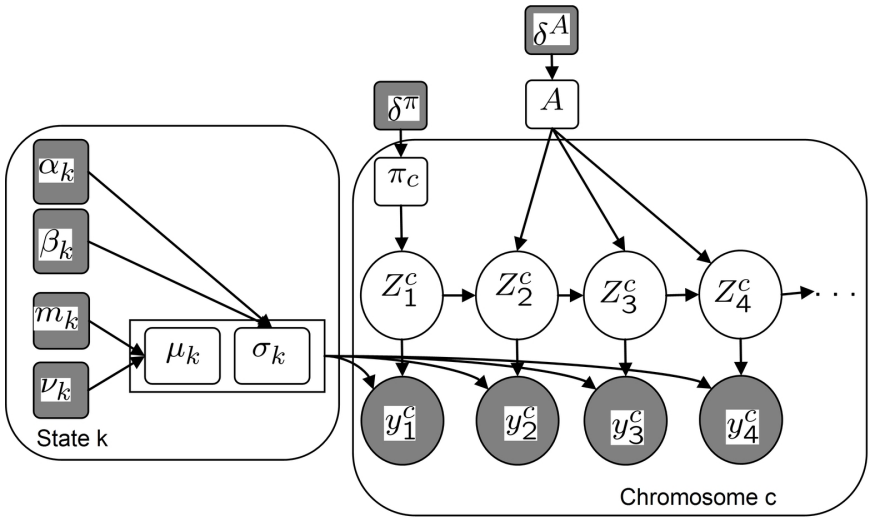


Figure 2.6: Graphical model of HMM-P. In comparison to Figure 2.5, we pull the μ, σ, A parameters outside the chromosome plate, indicating they are pooled (and thus, shared) across chromosomes, and we drop the c superscript. This leads to more accurate parameter estimation as the inference is informed by more data (all chromosomes vs one chromosome).

Maintaining identifiability can be accomplished through various techniques, one of which is to set strong priors on m_k by setting ν_k to be small (eg 10^{-4}). We will discuss this further in Section 2.2.3, but by way of introduction, we illustrate two separate runs of the GMM in Figure 2.9 with strong ($\nu_k = 10^{-4}$) (a) and weak ($\nu_k = 1$) (b) enforcement of $\mu_1 < \mu_2 < \mu_3$.

In addition to the label switching problem, the main limitation of GMMs is that each log ratio is assumed to be independent of all the others. Therefore spatial correlation is not considered in inferring the CNA states. Note that in Figure 2.8 (top) the data are coloured according to their vertical position, but not according to their horizontal position. In addition, as shown in Figure 2.10 the values of the converged parameters do not fit well to their empirical values (based on the ground truth labeling). We will show how this can be improved by modeling spatial correlation below.

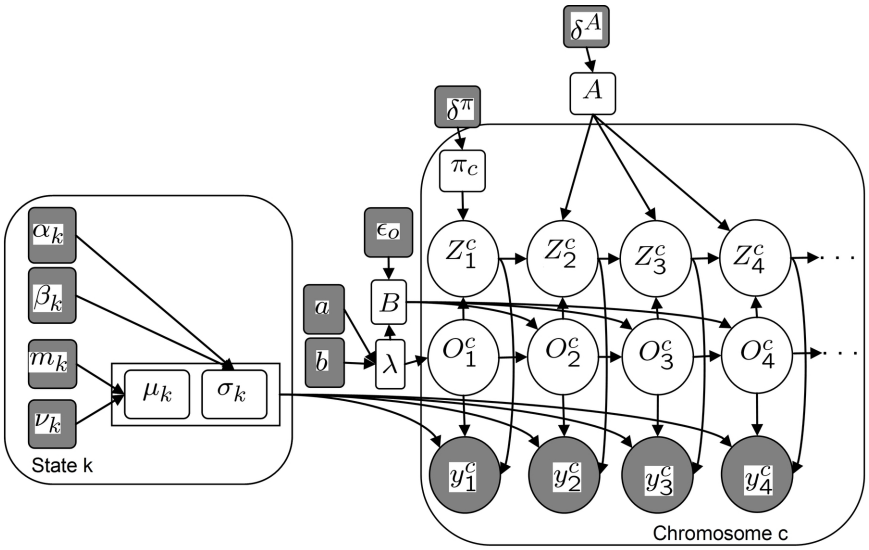


Figure 2.7: Graphical model of (HMM-R). HMM-R is similar to HMM-P (see Figure 2.6), but we add an extra process, O_t^c which models state-specific outliers. This is expected to reduce state transitions at outlying data points, thus reducing false positive CNA predictions. Please see text for details.

2.1.5 State space models

This class of models considers spatial correlation while classifying each probe into one of a fixed number of discrete classes or states. As in GMMs, we have the same semantic meaning of the states $\{L,N,G\}$. Useful and popular state space models for aCGH are hidden Markov models (HMM) [39]. HMMs are particularly appropriate as they attempt to simultaneously classify and segment the probes under a unified model.

Hidden Markov Models

Similar to the segmentation algorithms described above, HMMs segment the data into piecewise constant intervals, but instead of allowing an arbitrary number of levels, HMMs restrict the number of levels to a fixed number of $K = 3$ states as in GMMs. HMMs for aCGH have three important components: a latent sequence of

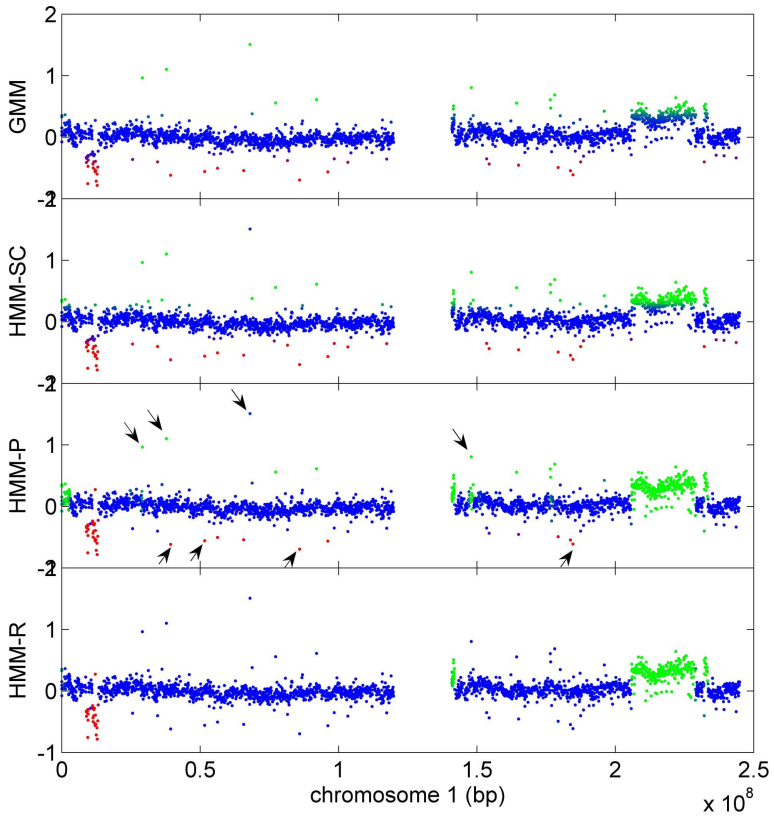
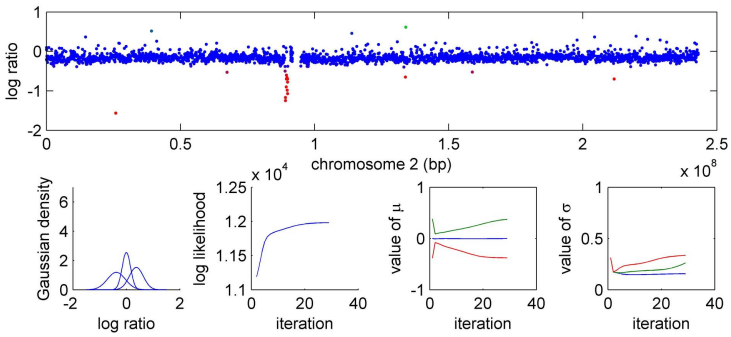
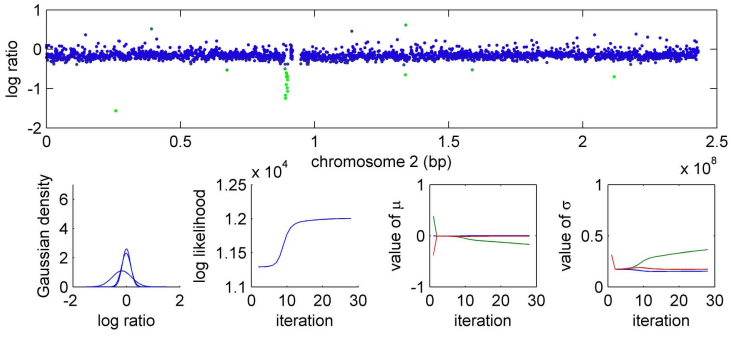


Figure 2.8: Comparison of $p(Z_t = k|y_{1:N}, \theta) = \gamma_t$ for the GMM, HMM-SC, HMM-P and HMM-R for chromosome 1 of HBL2. The data points are plotted using $[\gamma_t(L), \gamma_t(G), \gamma_t(N)]$ as a Red-Green-Blue colour vector for each probe. This allows a visual comparison to Figure 1.3 (b) and demonstrates that the HMM-R most closely resembles the ground truth. Systematic improvements over GMMs are achieved by HMM-SC, HMM-P and finally HMM-R. Arrow indicated in the plot for HMM-P are single clone predictions likely to be false positives. These are not predicted by HMM-R. This figure is best viewed in colour.



(a)



(b)

Figure 2.9: Illustration of the label switching problem on chromosome 2 of mantle cell lymphoma cell line NCEB1. (a), top shows $p(Z_t = k|\mu_k, \sigma_k, \pi) = \gamma_t(k)$ of the GMM using a strong prior where the probabilities are shown using $[\gamma_t(L), \gamma_t(G), \gamma_t(N)]$ an RGB colour vector with red=loss, green=gain and blue=neutral. The obvious loss near the centromeric end of the q-arm is correctly predicted as such. The bottom row of (a) from left to right show the Gaussian class conditional densities using the converged values of μ, σ; the log-likelihood plotted against iterations of the EM algorithm; the trace of μ against iterations of the EM algorithm (green is μ_3, blue is μ_2 and red is μ_1); and the trace of σ. The trace of μ indicates that $\mu_1 < \mu_2 < \mu_3$ is satisfied. (b) in contrast when a weak prior is used, label switching occurs and the converged value of $\mu_3 < \mu_1$. As a result, the loss region is erroneously predicted as a gain.

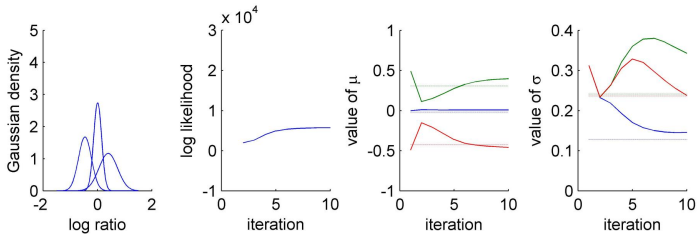


Figure 2.10: Convergence of model parameters for GMM. From left to right, the plots are the same as in bottom row of Figure 2.9. Note the problematic μ_G and σ_G parameters (two right-most plots) that do not fit their empirical values (based on ground truth labeling) well. The empirical estimates are denoted by horizontal lines. This results in many false negative predictions (see Figure 2.8 (top)).

discrete states, $Z^c_{1:N_c}$ for N_c probes on chromosome c; a set of emission densities conditional on Z^c_t:

$$p(y^c_t|Z^c_t = k) = \mathcal{N}(y^c_t|\mu^c_k, \sigma^c_k) \tag{2.8}$$

where $\mu^c_{1:K}, \sigma^c_{1:K}$ are specific to each chromosome, $K = 3$; and a K-by-K stochastic transition matrix A:

$$p(Z^c_t = j|Z^c_{t-1} = i) = A^c(i, j) \tag{2.9}$$

Z^c_t is a multinomial random variable where $Z^c_t \in \{L, N, G\}$. HMMs have the beneficial properties in that the hidden sequence of states $Z^c_{1:N}$ have biological interpretability (similar to GMMs) *and* they model spatial correlation by way of the state transition matrix A^c which encourages Z^c_t to be the same as its neighbours. The initial position is modeled as a stationary Dirichlet distribution $p(Z^c_1 = k) = \pi_c(k)$. A and π have standard Dirichlet conjugate priors parameterized by δ_A and δ_π respectively. A directed graphical model of the HMM, adapted for aCGH is depicted in Figure 2.5. The model depicts the probabilistic dependency of Z_t on Z_{t-1}, thus modeling the spatial correlation in the data. Please see [22] for details on directed graphical models and how they relate to state transition diagrams for HMMs.

Inference in HMMs

For each chromosome in turn, the inference goal of HMMs is to determine the most likely sequence of states $Z^c_{1:N}$ given $y^c_{1:N}$ and the model parameters $\theta^c = (\mu^c, \sigma^c, A^c, \pi_c)$ while simultaneously estimating θ^c. Similar to the GMM, this is accomplished in the EM framework. In the HMM case, $p(Z^c_t = k|y^c_{1:N_c}, \theta^c) = \gamma^c_t$ in the E-step is computed using the efficient $O(N_c)$ Forwards-Backwards algorithm [22]. θ^c is estimated in the M-step using standard MAP conjugate updating (similar to the GMM setting). We consider this the baseline HMM, upon which we make important extensions, described in Section 2.2. Full details of the inference algorithm are given in Bishop [22], chapter 13. We refer to this algorithm as HMM-SC (HMM, single chromosome) from here onwards. Figure 2.8 (HMM-SC) shows the output of HMM-SC on our running example chromosome 1 of HBL2. While the HMM is better than the GMM (more ground truth CNA probes are predicted), note that it still has false negatives. As in the GMM, this is perhaps due to the model parameters not matching their empirical values (based on the ground truth labeling) exactly (see Figure 2.11). In Section 2.2 we will show extensions to this baseline HMM are more accurate in predicting CNAs as they converge to model parameter values that are much closer to the empirical values.

Choosing the number of states in an HMM

One important issue with HMMs (and GMMs) is the requirement to choose the number of states. Thus far we have only considered 3-state models ($\{L, N, G\}$), however the biology may be more complicated than these models allow. If we consider the number of DNA copies of a normal person to be two, there may be one or two copies lost in a deletion. For gains, there may be arbitrarily many copies gained. Therefore to model the biology closely, we may require additional states in the HMM. Furthermore, in clinical samples, the ploidy (two-copy normal) assumption does not always hold [39]. Tumour genomes may be triploid or tetraploid which significantly affects the actual number of copies the samples can have. In addition, the tumour DNA sample is often contaminated with DNA from normal tissue that inherently affects the dynamic range of the observed logratios and may consequently make inference of discrete states more difficult. It is notable

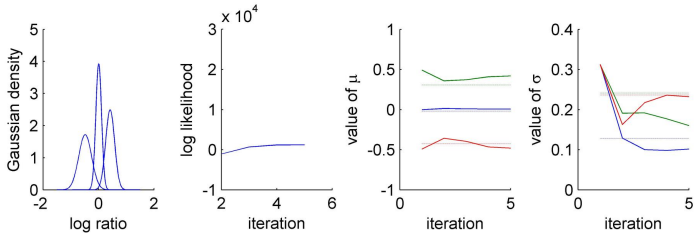


Figure 2.11: Convergence of model parameters for HMM-SC. From left to right, the plots are the same as in bottom row of Figure 2.9. Note that μ_G and σ_G parameters (two right-most plots) do not fit their empirical values (based on ground truth labeling) perfectly well. This results in many false negative predictions (see Figure 2.8 (2nd from top)). We will see in Section 2.2 how these parameter estimates can be improved for increased accuracy.

that three-state [40–42], four-state [20, 43], and six-state models [44] have been proposed. In addition, others have chosen the number of states based on penalised likelihood criteria such as AIC [39, 41]. We will demonstrate the effect of choosing different numbers of states in Section 2.3.3.

Other SSMs

Aside from HMMs, two other notable methods have been proposed that qualify as SSMs. Broet and Richardson [40] model spatial correlation using mixture model approach by way of a latent 1D Gaussian random field as opposed to a latent discrete 1D random field (ie HMM). Their approach produces posterior probabilities (analogous to γ_t as mixture weights of each probe belonging to each of three states ($\{L,N,G\}$) which can then be classified using thresholding or Bayes allocation rules. More recently, Shi *et al* [41] proposed a switching Kalman filter approach that can model spatial trends in the data that deviate from the piecewise constant assumption.

2.2 Methods: a novel HMM for inferring CNAs from aCGH data

In this section we outline three novel contributions we have made to modeling aCGH data with HMMs for both quantitative and qualitative improvement in CNA analysis. They include: i) improving HMM parameter estimation ii) modeling outliers and iii) automatic setting of hyperparameters. These are the main differences in our approach over other aCGH HMM methods such as as Fridlyand [39] and Guha [43].

2.2.1 Improving HMM parameter estimation by pooling

Thus far in our discussion, we outlined related work, all of which considered each chromosome in the data independently. In Shah *et al* [20], we showed that it is possible to pool the data across all the chromosomes when estimating certain parameters (μ, σ, A) of the model. This small extension to the HMM, which we call HMM-P (for pooled), results in considerable accuracy advantages. Based on the assumption that the mean log ratio levels for $\{L, N, G\}$ are consistent across chromosomes, the improved accuracy can be explained because the estimation of (μ, σ, A) can be guided by several fold more data points and thus borrow statistical strength from all chromosomes. Also, we often have the case where one or more chromosomes do not exhibit data that should belong to one of the states. If such a chromosome is treated independently, the parameter estimation for the 'missing' state defaults to the prior and is therefore not informed by any data. The HMM-P model is shown in Figure 2.6, which shows the (μ, σ, A) parameters outside of the chromosome plate, indicating that they are shared across chromosomes.

Note that Engler *et al* [42] proposed a pseudolikelihood SSM for the case where the data consists of a set of experiments (samples). Their method uses pooled estimates across chromosomes *and* samples. As we will discuss in Section 2.2.3, this may not be appropriate if the samples exhibit heterogeneous levels of $\{L, N, G\}$ due to differential mixtures of normal and tumour cells in the sample preparation, variability in the hybridization quality or different baseline ploidies [39].

The effect of the pooling procedure can be seen in Figure 2.8 (HMM-P) and Figure 2.12 which show a sharp reduction in false negatives and much more accu-

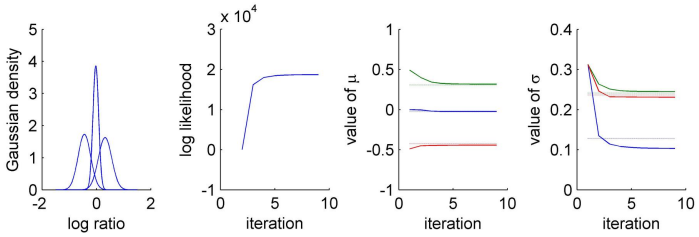


Figure 2.12: Convergence of model parameters for HMM-P. From left to right, the plots are the same as in bottom row of Figure 2.9. The parameter values converge more closely to the empirical values compared to HMM-SC (Figure 2.11) by considering all the data in the parameter estimation M-step. However, the convergence is still not exact.

rate parameter estimation compared to GMM and HMM-SC. The majority of the errors, (some of which are indicated by arrows in Figure 2.8) are due to misclassification of outliers. In the next section, we show how modeling outliers can further improve CNA detection accuracy and parameter estimation.

2.2.2 Modeling outliers

Thus far, we have considered the data as belonging to one of $\{L,N,G\}$ states. We call these states 'inlier' states. We propose to augment the state space of the model to consider an alternate generative mechanism for the data, the 'outlier' process. We call this new model HMM-R (for HMM robust). The motivation for this is given in the previous section in Figure 2.12.

We initially proposed a robust HMM capable of modeling outliers [20]. This was a simple approach that considered outliers from all the data (global outliers) with a single uniform distribution (simulated with a broad Gaussian) and inliers in the standard way as in HMM-P. Thus we modified the class conditional density as follows:

$$p(y_t|O_t,Z_t=k) = \begin{cases} \mathcal{N}(y_t|\mu_0,\sigma_0) & \text{if } O_t = 1 \\ \mathcal{N}(y_t|\mu_k,\sigma_k) & \text{if } O_t = 0 \end{cases} \tag{2.10}$$

Thus O_t acts like a "switching parent" variable, which selects between the outlier parameters μ_0,σ_0 or the inlier parameters, μ_k,σ_k. Examples of these distributions are shown in Figure 2.14 (b - top) for inliers and Figure 2.14 (b - middle) for

outliers. The outlier distribution approximates a uniform distribution.

We refined the outlier processing to consider outliers in the context of their neighbouring probes. We call these context-specific outliers. Figure 2.14 (a) illustrates the rationale behind this extension. The black arrows indicate two data points, one outlying from a neutral state and the other outlying from the loss state (red points). The outlying points both fall very close to the mean of the Gaussian of the gain state and therefore are susceptible to transitions to the gain state. Furthermore, they are not global outliers and so they would not be considered outliers by our previous method. Obviously, neither of these probes should be classified as a gain. To prevent misclassification of these probes, we introduce state-specific outlier density functions (one for each inlier state), shown in Figure 2.14 (a) capable of capturing locally outlying data points that would not be captured by our previous outlier distribution shown Figure 2.14 (b), middle. To achieve this, we introduce a binary switching (or Bernoulli) random variable O_t where $O_t = 1$ indicates that the log ratio for probe t is an outlier and $O_t = 0$ indicates that it is an inlier, to be modeled by the Z Markov process. We therefore modify the emission density as follows:

$$p(y_t|O_t, Z_t = k) = \begin{cases} \psi(y_t|\mu_k, \sigma_y) & \text{if } O_t = 1 \\ \mathcal{N}(y_t|\mu_k, \sigma_k) & \text{if } O_t = 0 \end{cases} \tag{2.11}$$

where

$$\psi(y_t|\mu_k, \sigma_y) = \begin{cases} \chi^{-2}(\mu_k - y_t|\sigma_y, \nu_0)/2 & \text{if } y_t \leq \mu_k \\ \chi^{-2}(y_t - \mu_k|\sigma_y, \nu_0)/2 & \text{if } y_t > \mu_k \end{cases} \tag{2.12}$$

We use the state mean of the outlier's neighbours to calculate a class conditional density that is based on a χ^{-2} distribution, but is symmetric. We denote this distribution as $\psi(y|\mu_k, \sigma_y)$ where μ_k is the mean of the state of the neighbours and σ_y is the global standard deviation of $y_{1:N}$. The distribution is symmetric about μ_k and integrates to 1. A clear example of the inlier and outlier densities for the loss state is shown in Figure 2.13. The shapes of all the emission densities superimposed are shown in Figure 2.14 (b) bottom.

The graphical model for HMM-R is shown in Figure 2.7. Table 2.1 contains the list of conditional probability distributions for the model. Outliers are depicted by O. As can be seen from the arrows pointing into Y, there are now two processes to generate the data: the inlier process modeled by Z and the outlier process

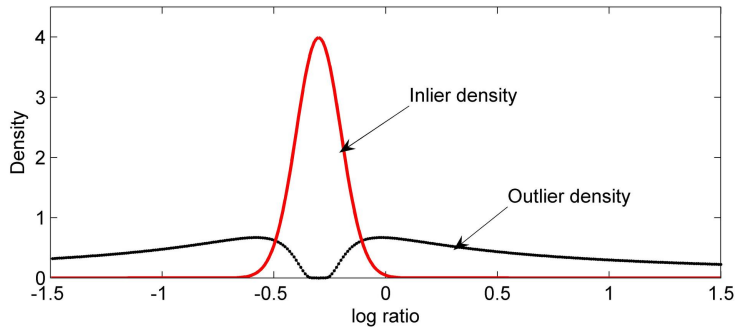
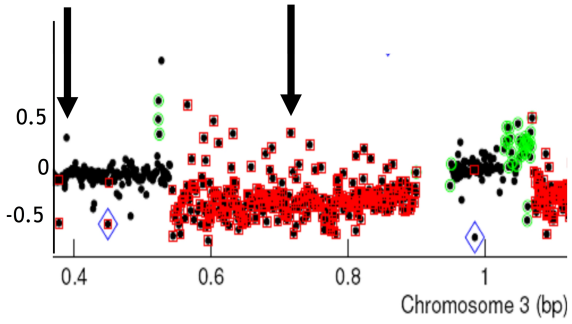


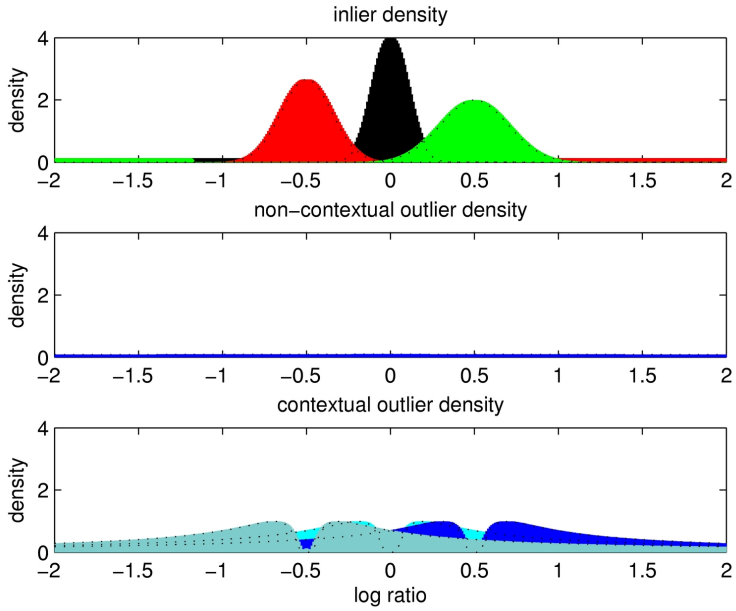
Figure 2.13: Inlier and outlier emission densities for HMM-R. We show the inlier density for the loss state (red curve), a standard Gaussian distribution, while it corresponding outlier density is depicted in black. The outlier density has a 'hole' centered at the mean of its corresponding inlier density, and has heavy tails that are better able to capture local outliers.

$$
\begin{array}{rcl}
p(A(i,.)|\delta^A) & \sim & Dir(A(i,.)|\delta^A) \\
p(\pi_c|\delta^\pi) & \sim & Dir(\pi_c|\delta^\pi) \\
p(Z_t=k|O_t=0,Z_{t-1}=j,A) & \sim & A(j,k) \\
p(Z_t=k|O_t=1,Z_{t-1}=j,A) & = & I(j=k) \\
p(O_t=1|O_{t-1}=1) & = & \varepsilon_o \\
p(O_t=1|O_{t-1}=0) & = & \lambda \\
p(\lambda|a,b) & \sim & Beta(\lambda|a,b) \\
p(y_t|O_t,Z_t=k) & \sim & \begin{cases} \psi(y_t|\mu_k,\sigma_y) & \text{if } O_t=1 \\ \mathcal{N}(y_t|\mu_k,\sigma_k) & \text{if } O_t=0 \end{cases} \\
p(\mu_k|m_k,v_k) & \sim & \mathcal{N}(\mu_k|m_k,\sigma_k^2 v_k) \\
p(\sigma_k^{-2}|\alpha_k,\beta_k) & \sim & Gam(\sigma_k^{-2}|\alpha_k,\beta_k)
\end{array}
$$

Table 2.1: Conditional probability distributions for HMM-R



(a)



(b)

Figure 2.14: (a) Motivation for contextual outliers. Plot shows ground truth losses as red squares, ground truth gains as green circles. The points delineated by the arrows are outlying from their neighbouring clones but lie close to the mean of the gain state and therefore the model is susceptible to transitions at these points. (b) model for inliers (top), global outliers (middle) used in Shah *et al* [20] and contextual outliers (bottom) where the distribution will have a much greater likelihood of capturing the points illustrated in (a).

modeled by O. The arrows between adjacent O nodes indicate that we enforce singleton outliers. As explained earlier, outliers may correspond to real CNAs, but the problem is that they are indistinguishable from experimental noise [45]. With high-dimensional arrays many investigators require CNAs to span at least 2 consecutive probes [45–47], so the singleton enforcement is well justified. Therefore, we consider O_{t-1} when evaluating $p(O_t)$ and only allow $O_{t-1} = 1$ and $O_t = 1$ with very low probability. The conditional probability distribution for $O_t = 1$ is:

$$p(O_t = 1|O_{t-1} = 1) = \varepsilon_o \tag{2.13}$$

$$p(O_t = 1|O_{t-1} = 0) = \lambda \tag{2.14}$$

where λ represents the prior probability of an outlier. λ has a conjugate Beta prior parameterized by (a,b). ε_o, $0 \leq \varepsilon_o << 1$, is the probability of having 2 consecutive outliers. These parameters are used to define the outlier transition matrix B:

$$B = \begin{pmatrix} 1-\lambda & \lambda \\ 1-\varepsilon_o & \varepsilon_o \end{pmatrix} \tag{2.15}$$

The set of parameters for HMM-R is consequently augmented to

$$\theta = (\mu, \sigma, A, \pi, B, \lambda).$$

To ensure that outliers are considered in the context of their neighbours, we set Z_t to Z_{t-1} if $O_t = 1$. We call this state latching. Therefore when determining Z_{t+1}, the inlier Markovian dynamics are maintained even though $O_t = 1$. This is similar to a 2nd order Markov process in that when $O_t = 1$, Z_{t+1} depends on Z_{t-1}, but because of the state latching condition, the computational cost during inference is not significantly increased. Note that when the model makes temporary "excursions" to the outlier state, it is not penalised by the state transition matrix A. Formally, the conditional probability distribution for $Z_t = k$ is given by:

$$p(Z_t = k|O_t = 0, Z_{t-1} = j, A) = A(j,k) \tag{2.16}$$

$$p(Z_t = k|O_t = 1, Z_{t-1} = j, A) = I(j = k) \tag{2.17}$$

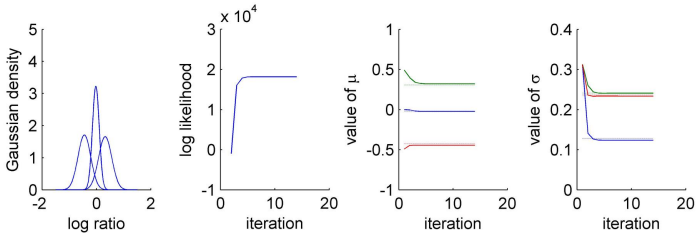


Figure 2.15: Convergence of model parameters for HMM-R. From left to right, the plots are the same as in bottom row of Figure 2.9. The parameter values converge more closely to the empirical values compared to HMM-P (Figure 2.12) by pooling all the data in the parameter estimation M-step.

HP	Description	Setting
m_L	Prior Gaussian mean on μ_L (loss state)	$-2\sigma_y$
m_N	Prior Gaussian mean on μ_N (neutral state)	0
m_G	Prior Gaussian mean on μ_G (gain state)	$2\sigma_y$
$v_{1:K}$	Prior Gaussian variance on $\mu_{1:K}$ (strong prior)	10^{-4}
$\alpha_{1:K}$	Prior shape parameter for Gamma prior on $\sigma_{1:K}^{-2}$	$10+\sigma_y$
$\beta_{1:K}$	Prior scale parameter for Gamma prior on $\sigma_{1:K}^{-2}$	1
a,b	Beta(a,b) prior on outlier	$(10^4\sigma_y, 10^5)$
δ^A	Dirichlet prior on A	$\begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}$
δ^π	Dirichlet prior on π_c	$(0.25, 0.5, 0.25)$

Table 2.2: Hyperparameters (HP), descriptions and settings for HMM-R

Processing outliers also allows the parameters in the model to be estimated more accurately, as shown in Figure 2.15. We will demonstrate how these important modifications yield better accuracy in Section 2.3.2.

2.2.3 Setting hyperparameters

The hyperparameters (the parameters of the conjugate priors, eg m, v) of the model need to be set at run time. The HMM-R hyperparameters include $(m_{1:K}, v_{1:K}, \delta_A, \delta_\pi, a, b)$ and are listed in Table 2.2. This leaves the user with many free parameters to set. This could be cumbersome when processing 10s or 100s of samples, given that

Algorithm 1 Expectation Maximization algorithm for HMM-R. We omit the π term for brevity and assume ε_o to be fixed. τ_c represents the $1:N_c$ indicies (t) of the probes on Chromosome c. Supporting functions are given in Algorithm 2. The input is the logratios, $Y_{1:N}$ and the output are the marginal posterior probabilities γ^Z and γ^O. ForwardsBackwards is described in Bishop [22].

$(m, \nu, \alpha, \beta, \delta_A, a, b)$ = setHyperparameters($Y_{1:N}$)
(μ, σ, A, B) = initialiseParameters($m, \nu, \alpha, \beta, a, b$)
for $iter = 1, 2, \ldots$ **do**
    /* Compute states (E step) */
    $b_{1:Nc}$ = makeLocalEvidence($Y_{1:N}, \mu_{1:K}, \sigma_{1:K}$)
    A^{ZO} = makeTransitionMatrix(A,B)
    $A', O' = 0$ /* reset pseudocounts */
    for $c = 1, 2, \ldots, C$ **do**
        $\gamma^{ZO}_{\tau_c}$ = ForwardsBackwards(A^{ZO}, b_{τ_c})
        $\gamma^{Z}_{\tau_c}$ = marginalise($\gamma^{ZO}_{\tau_c}$,O)
        $\gamma^{O}_{\tau_c}$ = marginalise($\gamma^{ZO}_{\tau_c}$,Z)
        $A' = A'$ + countTransitions($\gamma^{Z}_{\tau_c}, N_c, K$)
        $O' = O'$ + countTransitions($\gamma^{O}_{\tau_c}, N_c, 2$)
    end for
    /* Update parameter values (M step) */
    for k=1,...,K **do**
        /* Update emission density parameters */
        $n_k = \sum_t \gamma^Z(k)$
        $\bar{y}_k = \frac{1}{n_k} \sum_t \gamma^Z_t(k) y_t$
        $\bar{\nu}_k = \frac{1}{n_k} \sum_t \gamma^Z_t(k)(y_t - \bar{y}_k)^2$
        $\mu_k = \sigma^2_k m_k + n_k \nu_k \bar{y}_k$
        $\bar{\alpha}_k = \alpha_k + \frac{1}{2} n_k$
        $\bar{\beta}_k = \beta_k + \frac{1}{2} n_k \bar{\nu}_k$
        $\sigma_k^{-2} = \bar{\alpha}_k / \bar{\beta}_k$
    end for
    /* Update transition parameters */
    for j=1, ..., K **do**
        $A(j,k) = \frac{A'(j,k) + \delta_A(j,k)}{\sum_l A'(j,l) + \delta_A(j,l)}$
    end for
    /* Update outlier parameters */
    $\lambda = \frac{a + O'(inlier, outlier)}{b + a + N - 2}$
    $B = \begin{pmatrix} 1-\lambda & \lambda \\ 1-\varepsilon_o & \varepsilon_o \end{pmatrix}$
end for

Algorithm 2 Supporting functions for Algorithm 1

Function $(m, \nu, \alpha, \beta, \delta_A, a, b)$ = setHyperparameters($Y_{1:N}$)
σ_y = standardDeviation($Y_{1:N}$)
$m_L = -2\sigma_y$
$m_N = 0$
$m_G = 2\sigma_y$
$\nu_{1:K} = 0$
$\alpha_{1:K} = 10 + \sigma_y$
$\beta_{1:K} = 1$

Function (μ, σ, A, B) = initialiseParameters($m, \nu, \alpha, \beta, a, b, \delta^A, \varepsilon_0$)
$\mu_k = m_k$
$\sigma_k^{-2} = \frac{\alpha}{\beta}$
$A = \delta^A$
$\lambda = \frac{a}{a+b}$
$B = \begin{pmatrix} 1-\lambda & \lambda \\ 1-\varepsilon_o & \varepsilon_o \end{pmatrix}$

Function $b_{1:N}$ = makeLocalEvidence($Y_{1:N}, \mu_{1:K}, \sigma_{1:K}$)
$b_t(i,k) = \begin{cases} \psi(y_t|\mu_k, \sigma_y) & \text{if } i = 1 \\ \mathcal{N}(y_t|\mu_k, \sigma_k) & \text{if } i = 0 \end{cases}$

Function M^{AB} = makeTransitionMatrix(A,B)
$M^{AB} = A \times B$

Function (γ) = marginalise($p(X = x, Z = z)$,Z)
$\gamma_t = \sum_z p(X_t = x, Z_t = z)$

Function C = countTransitions(Z, N, K)
for i=1, . . . , K **do**
    for j=1, . . . , K **do**
        $C(i, j) = \sum_{t=2}^{N} p(Z_{t-1} = j, Z_t = i)$
    end for
end for

several characteristics in the data contribute to inter-sample variability in the log ratio levels of losses and gains and noise characteristics of the data. Examples include quality of the hybridization, the tumour/normal admixture of cells in the sample preparation and the ploidy of the tumour cells in the sample. Setting the hyperparameters therefore requires careful treatment. We devised a simple data-driven heuristic to set the hyperparameters. We calculated the standard deviation σ_y of the data and set the hyperparameters as follows:

$$m_L = -2\sigma_y, m_N = 0, m_G = 2\sigma_y \tag{2.18}$$

We set the mean of the neutral state to zero based on the assumption the data has been normalised (using a method such as Khojasteh *et al* [27]) so that the neutral state usually corresponds to a log ratio of 0. We set the mean of the aberrant states to reflect the typical deviations expected in this sample. This allows the model to adapt to automatically different noise levels. We set the prior variance on the mean to $\nu_k = 10^{-4}$. This was chosen to avoid the label switching problem (see Section 2.1.4) and to reflect that our method for choosing m was appropriate in the majority of the data we have observed. We set the shape parameter, α_k, of the prior Gamma distribution on σ^{-2} to $\alpha_k = 10 + \sigma_y$ to account for noise variability and $\beta_k = 1$, the scale parameter to reflect that this is a weak prior. The Beta prior on outliers are set as follows. $a = 10^4 \times \sigma_y, b = 10^5$ to scale the outlier probability to the overall noise in the data. We have found HMM-R to be robust to settings of δ_A and δ_π, the Dirichlet priors on A and π respectively. In practice we use a weak prior encouraging self transitions for δ_A and a uniform prior for δ_π.

This method produces good results for both cell line and clinical data (see Section 2.3.2) with no free parameters for the user to set. Of course the software allows advanced users to set the hyperparameters manually if desired. Note that given enough ground truth data, we could set the hyperparameters using empirical Bayes. It is not clear how much data is sufficient to do this robustly.

2.2.4 EM algorithm for HMM-R

The inference algorithm for HMM-R is similar to HMM-P, except we make adjustments for the extra variables O and λ. The full details of the algorithm are

sketched in Algorithm 1. The quantities of interest we which to infer are the probability probe t is in state k, $\gamma_t^Z = k$ and the probability that probe t is an outlier, γ_t^O. To begin we must initialize the parameters $\theta = (\mu, \sigma, A, \pi, \lambda, B)$. We use their expected values according to their priors:

$$\mu_k = m_k \tag{2.19}$$

$$\sigma_k^{-2} = \frac{\alpha_k}{\beta_k} \tag{2.20}$$

$$A = \delta_A \tag{2.21}$$

$$\pi = \delta_\pi \tag{2.22}$$

$$\lambda = \frac{a}{a+b} \tag{2.23}$$

$$B = \begin{pmatrix} 1-\lambda & \lambda \\ 1-\varepsilon_o & \varepsilon_o \end{pmatrix} \tag{2.24}$$

The E-step (Algorithm 1, line 4) is modified by grouping Z and O into a megavariable ZO with $2K$ states (where K is the number of CNA states). We collapse the 2 transition matrices A and B into $2K$-by-$2K$ transition matrix. We can then infer $\gamma_t^{ZO} = p(ZO_t|Y,\theta)$ with Forwards-Backwards. As in HMM-P, this is done for each chromosome independently. We marginalise out O to obtain:

$$\gamma_t^Z = \sum_{o=0,1} \gamma_t^{ZO}(o) \tag{2.25}$$

In the global outlier model, we ignored outliers in this step. In this model, since outliers are context-specific and have state meaning, we want to include them in the parameter estimation in the M-step. We marginalise out Z to obtain:

$$\gamma_t^O = \sum_{k=L,N,G} \gamma_t^{ZO}(k) \tag{2.26}$$

The parameters can then be updated in the M-step in the standard way as shown in Algorithm 1 beginning on line 15. In practice, if one wants to avoid interpreting the marginal probabilities given by Forwards-Backwards, a final run of the Viterbi

algorithm (see Bishop [22] for details) which gives the most probable sequence of states can be run once the model has converged.

2.2.5 Student-t emission model

If explicit modeling of outliers is not required, a class conditional Student-t distribution, parameterized by $(\mu_k, \lambda_k, \nu_k)$, the mean, precision and degrees of freedom, can be used in place of Equation 2.11. This results in:

$$p(y_t|Z_t = k) = St(Y_t|\mu_k, \lambda_k, \nu_k) \tag{2.27}$$

This emission density is robust to outliers and greatly simplifies inference by eliminating the outlier generative mechanism from the model. Thus the model is similar to HMM-P, but with a Student-t emission density rather than a Gaussian. This was also used in the models reported in Chapters 3 and 5.

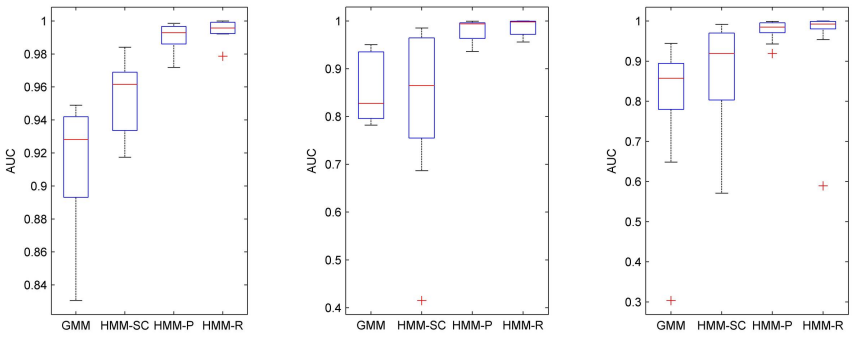
2.3 Results

2.3.1 Experiments on cell line and clinical data

We evaluated GMM, HMM-SC, HMM-P, HMM-R and DC+ML to assess which method was the most accurate at detecting CNAs. The algorithms were run on three data sets, all with ground truth annotation. The first set consisted of 8 mantle cell lymphoma (MCL) cell lines [19]. The remaining two were clinical samples: one set of 30 enteropathy T-cell lymphoma (ETL) samples [25] and one set of 11 blastic-type lymphoma (BL) (unpublished data). In all, there were 49 samples used in the evaluation.

Evaluation protocol

We used standard receiver-operator characteristic (ROC) curves to determine the accuracy of the methods. Each probe was given a binary label as a CNA (1) or not (0), based on the ground truth information. For GMM, HMM-SC, HMM-P, and HMM-R we calculated the true positive rate (TPR) as the proportion of CNA probes that were predicted as CNAs and false positive rate (FPR) as the proportion of pre-



(a) MCL (b) BL (c) ETL

Figure 2.16: Distribution of AUC for MCL (a) BL (b) and ETL (c) shown as box and whisker plots. The clear trand is that the best model is HMM-R, followed closely by HMM-P which is considerably better than HMM-SC which in turn is better than GMM.

dicted CNA probes that were not ground truth CNA probes. For GMM, HMM-SC, HMM-P and HMM-R, we calculated $p(CNA_t) = \gamma_t(L) + \gamma_t(G)$ and 'called' probe t a CNA if $p(CNA_t)$ exceeded a threshold τ. We then computed FPR and TPR for various values of τ, plotted TPR vs FPR and computed the area under the ROC curve (AUC) as a single measure of accuracy. Since DC+ML is non-probabilistic in its output, we plotted a single point on the ROC curves to compare it with the other methods.

2.3.2 Pooling and outlier processing lead to increased accuracy

Figure 2.16 shows distributions of AUC as box-and-whisker plots for the various models on the MCL, BL and ETL data sets. For the MCL data, HMM-R (0.99±0.00), HMM-P (0.99±0.00) were significantly more accurate (one-way ANOVA, $p = 3.9 \times 10^{-7}$) than both HMM-SC (0.95±0.01) and GMM (0.91±0.01) (mean AUC± stderr shown in parentheses). Similar results were seen in the BL data. HMM-R (0.99±0.01) and HMM-P (0.98±0.01) were significantly more accurate (one-way ANOVA, $p = 10^{-4}$) than both HMM-SC (0.82±0.05) and GMM (0.86±0.02). For ETL, HMM-R (0.98±0.01) and HMM-P (0.98±0.00) were significantly more accurate than GMM (0.83±0.02) (one-way ANOVA, p=5 $\times 10^{-10}$).

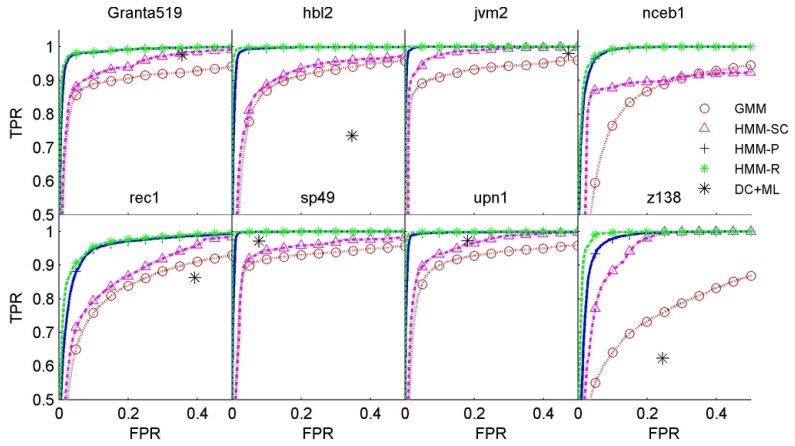


Figure 2.17: Receiver operator characteristic (ROC) curves for each sample in the MCL data. Vertical axis is TPR and horizontal axis is FPR. Results are shown for GMM (red circles), HMM-SC (pink triangles), HMM-P (blue crosses) HMM-R (green stars) and DC+ML (black stars). HMM-R and HMM-P are always above and to the left of the other algorithms. Some samples do not show a data point for DC+ML because it is off the scale.

The ROC plots for each sample are shown in Figures 2.17, 2.18 and 2.19 for MCL, BL and ETL. These plots allow us to compare performance of DC+ML (recall this is segmentation followed by post-processing). For nearly every one of the 49 samples, our 2 novel HMM variants (HMM-P, and HMM-R) outperform DC+ML and the other two standard models, GMM and HMM-SC. This is shown by their respective ROC curves always lying to the left and above the single data point for DC+ML, and the curves for GMM and HMM-SC. (Recall that DC+ML is non-probabilistic and therefore not amenable to full ROC curve analysis. Instead we compute a binary point estimate of FPR and TPR). These results systematically show the expected improvements obtained by modeling spatial correlation over the GMM (by HMM-SC); the improved parameter estimation by pooling (by HMM-P); and adding contextual outlier processing to the model (by HMM-R). Note that the results of the HMM-R are consistently very accurate, which, despite variability in the clinical data sets (BL, ETL), demonstrates that the hyperparameter setting

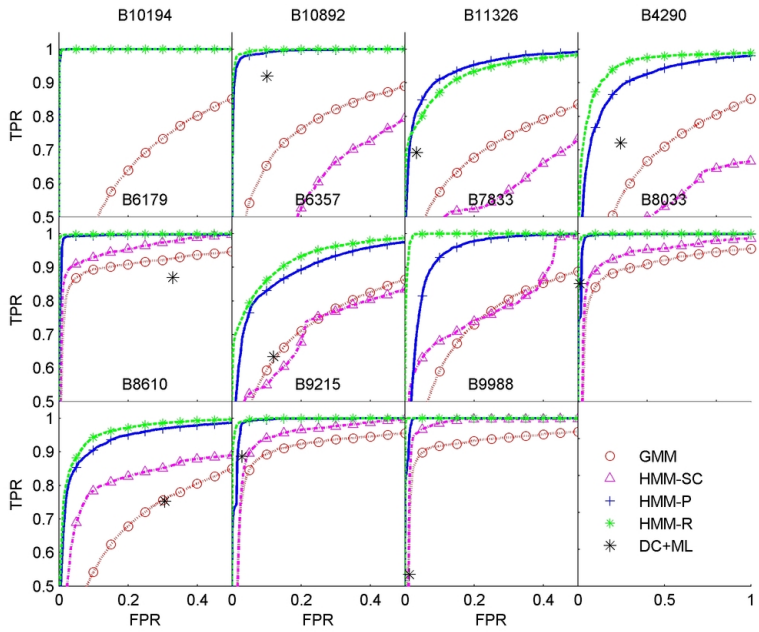


Figure 2.18: ROC plots for BL data. For legend and axes description, see Figure 2.17. HMM-R and HMM-P are better than the other algorithms. In general, the HMM-R curves are above and left of HMM-P.

method outlined in Section 2.2.3 is working well. In these experiments there were no free parameters, and all hyperparameters were set automatically based on the data.

2.3.3 3 state model works best

We investigated the effect of using 3, 4, 5, and 6 state models (HMM-R only) as there is little consensus in the literature on this issue (see Section 2.1.5). The 4-state model had an extra gain state, the 5-state model had 2 loss states, a neutral state and 2 gain states. The 6-state model had an additional gain state over the 5-state model (as proposed by van de Wiel *et al* [44]). Figure 2.20 shows that the 3 and 4 state models were the most accurate in all 3 data sets. Mean accuracy for MCL was 0.99 ± 0.00, 0.99 ± 0.00, 0.92 ± 0.04 and 0.91 ± 0.05 AUC respectively for the 3, 4, 5

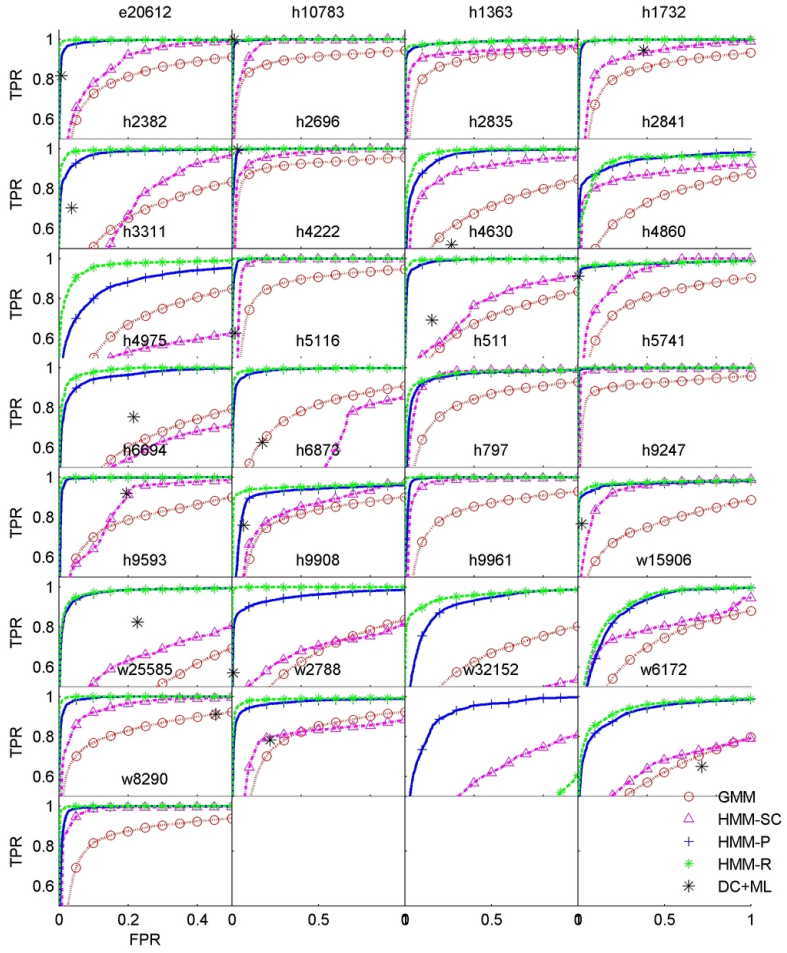
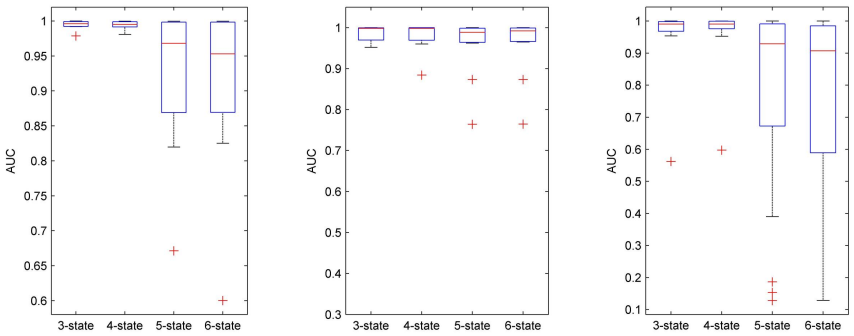


Figure 2.19: ROC plots for ETL data. For legend and axes description, see Figure 2.17. HMM-R and HMM-P are better than the other algorithms. In general, the HMM-R curves are above and left of HMM-P.



(a) MCL (b) BL (c) ETL

Figure 2.20: Distribution of AUC for 3, 4, 5 and 6 states of HMM-R for MCL (a), BL (b) and ETL (c). The 3 and 4 state models are consistently the best for all data sets.

and 6 state models. For BL, accuracy was 0.99±0.01, 0.98±0.01, 0.96±0.02 and 0.96±0.02 AUC. A similar trend was observed for ETL: 0.97±0.01, 0.97±0.01, 0.79±0.05 and 0.78±0.05 AUC for 3, 4, 5 and 6 state models. From these results, it is clear that the 3 and 4 state models are the most accurate and in the context of the HMM-R, additional states generally only hurt performance.

There are specific cases, however where the 3-state model missed ground truth losses. Consider the data from BL sample B11326 - see ROC curves in Figure 2.18, top row, 3rd from left). The results are not as accurate as many of the other samples. Chromosomes 1 and 2 for this sample are shown in Figure 2.21 (a) and (b) respectively. Figure 2.21 (c) shows the results with a 3-state model which misses many of the ground truth losses, while (d) shows results of a 5-state model (2 loss states, 1 neutral state and 2 gain states) which recovers the losses. In this specific case, there are 2 distinct loss levels. The 3-state model can only pick out one, while the 5-state model can accurately detect both. Note that identifiability is maintained by using strong priors to maintain the 'order' of the states (ie $\mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5$) as was shown previously for the 3-state model. This example demonstrates that occasionally it may be necessary to alter the number of states in the model. We implemented the number of states as a user settable parameter, therefore our software can easily adapt to special data sets with no adjustment.

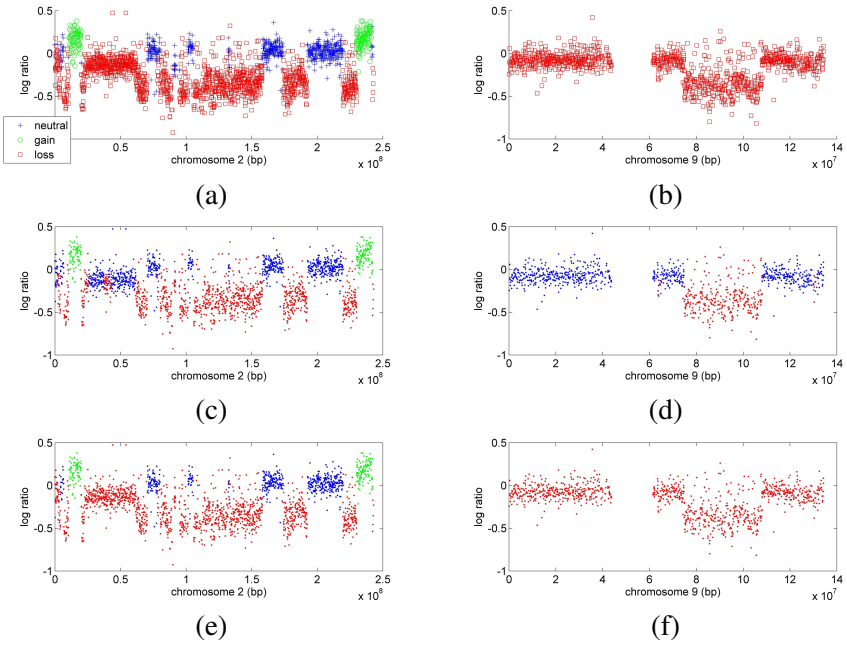


Figure 2.21: Example where 3-state model is not adequate. (a) and (b) show ground truth labeling of sample B11326 from the BL data. There are two distinct loss levels, likely due to single copy and two-copy deletions. The single copy deletions are not well modeled by the 3-state model (c,d), but are well-modeled by the 5-state model (e,f) which has 2 loss states.

2.4 Discussion

We described a novel approach to inferring CNAs from aCGH data using en extended HMM that leverages the statistical properties of aCGH data. We demonstrated systematically how our extensions of parameter estimation by pooling, contextual outlier processing and objective hyperparameter setting contribute to very accurate predictions of CNAs from single sample aCGH data and improved on standard approaches. This work has resulted in a solid theoretical model with accompanying implementation upon which to develop new ideas. The foundations laid in this chapter are built upon in subsequent chapters where we address research goals B and C: detection of recurrent CNAs from multiple aCGH experiments in Chapter 3 and clustering aCGH data in a population expected to be composed of

molecular subtypes in Chapter 5. In addition, we applied the HMM-R in a clinical setting. In collaboration with Dr. Doug Horsman's group, we generated aCGH data for 106 follicular lymphoma patients and the HMM-R was used as an analytical tool to detect prognostic markers in this disease. This is the subject of Chapter 4 and a recent clinically focused publication [7].

2.4.1 Impact

The work presented was among the first use of HMMs for aCGH analysis. Our original paper [20] has since been cited 19 times (Google scholar) and was included in a recent benchmarking study that ranked it in the highest performing group of algorithms [48]. Moreover, the HMM-R model is being used in several local cytogenetics research projects (Horsman lab, Dr. Randy Gascoyne lab and Dr. Wan Lam lab) at the BC Cancer Agency. The HMM-R method has also been used at the Ontario Cancer Institute in collaboration with Dr. Ming Tsao for a lung cancer project and the Children's and Women's hospital in collaboration with Dr. Patrice Eydoux in a study focusing on congenital CNAs in mental retardation. To promote usability by other researchers, we developed a distributable stand-alone software package with a user-friendly graphical user interface. This will be packaged with the SeeGH software [49] to support users of the SMRT array platform. The implementation of the HMM-R, written in MATLAB is available at: http://www.cs.ubc.ca/~sshah/acgh as part of a toolbox called CNA-HMMer . This package includes an implementation of the H-HMM algorithm we present in Chapter 3. In addition the Shah *et al* [20] was cited in the recently developed Matlab Bioinformatics toolbox demonstration on aCGH data analysis using HMMs.

2.4.2 Limitations and future work

Several assumptions are explicit in our model. First, we assume that the data can be classified into a fixed number of states that must be determined *a priori*. In some settings, this may be too restrictive. For example, high-level amplicons (gains of many copies) are of interest to identify as distinct from lower-level amplicons since the may represent targeted regions of the genome. Beal *et al* [50] present an intriguing paradigm in which the HMM formulation is assumed to have a countably

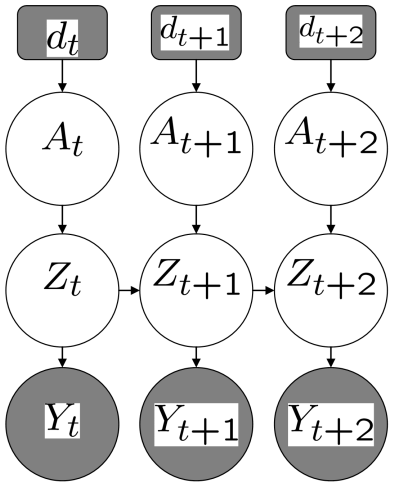


Figure 2.22: HMM with non-stationary transition matrix to account of unequal spacing of probes. In contrast to Figures 2.5-2.6, this model has a transition matrix at each probe t that is parameterized by d_t, the distance between $t-1$ and t. The greater the distance, the more likely a transition.

infinite number of states. Rather than optimize these parameters, they are integrated out using Dirichlet processes. Thus, they define a type of non-parametric HMM that does not require the specification of the number of states *a priori*. Extending and adapting our models using this technique is an open problem that may provide additional utility of our aCGH based HMM, especially when it is not clear how to set the number of states. Some post-processing would be necessary to make biological inferences from the potentially infinite number of states. Related to the number of states problem, Rueda *et al* [51] have made some progress in this area by using Bayesian model averaging in a reversible jump MCMC framework, thus expressing the model uncertainty in the number of states.

We assume that the probes are uniformly spaced across the chromosomes. While this is generally true in the SMRT arrays [16], complementary platforms such as Affymetrix SNP arrays have unequal spacing between probes. Colella *et al* [52] handle this situation by specifying a non-stationary transition matrix (one transition matrix per probe) assumed to be generated by a distance-based Dirichlet

prior. Thus, the transition matrix becomes:

$$p(Z_t = j|Z_{t-1} = i) = \begin{cases} \frac{\rho}{K-1} & \text{if } i \neq j \\ 1-\rho & \text{if } i = j \end{cases} \tag{2.28}$$

where

$$\rho = \frac{1}{2}(1 - e^{-\frac{d}{2L}}) \tag{2.29}$$

with d representing the distance between t and $t-1$, and L an expected length between transitions. The graphical model is shown in Figure 2.22. A_t represents the *location specific* transition matrix whose parameter settings are dependent on d. We have implemented this parameterization of the HMM into our model and found that for the data presented herein, the results are the same. However, application of our model to inference of copy number from SNP arrays (future work) should employ the distance based non-stationary transition matrix to compensate for unequal spacing of the probes. In addition, this model is easily extended to the multiple sample case (see Chapter 3). Stjernqvist *et al* [53] describe a continuous index HMM, whereby transitions are not specified at the level of probes, but at the level of nucleotides. This method is expected to perform better in the presence of unequal probe spacing and can 'interpolate' at a finer resolution where the state transitions are situated. Unfortunately, the inference algorithm of Stjernqvist *et al* may be computationally impractical as analysis of single chromosome took 25 CPU hours [53].

This concludes our chapter on single sample aCGH analysis. In the next chapter, we describe our work on research goal B: detecting recurrent CNAs from multiple aCGH experiments.

Chapter 3

Detecting driver CNAs from a set of aCGH experiments

3.1 Summary

In this chapter, we present a solution to the problem of inferring recurrent CNAs from a set of aCGH data (research goal B). The key contribution is that we demonstrate the benefit of using statistical models capable of inferring recurrent CNAs from the raw data directly (as opposed to the standard approach of using discretized data). We investigate three novel methods capable of leveraging the statistical strength present in the raw data and demonstrate that a model based on a hierarchical HMM performs best in a theoretical setting using simulated data and an applied setting using real data. We describe the biological motivation for this problem in Section 3.2. In Section 3.3, we develop the notation and formalise the computational problem. We then outline related work in this area (Section 3.3.2) noting that most of the described methods appeared after our contribution [21], but are relevant to the discussion [1] In Section 3.4 we describe the three novel approaches to the problem and demonstrate in Section 3.5 how our contributions confer sig-

[1] Some of the material in the related work section has been accepted for publication in: Shah SP. Computational methods for identification of recurrent copy number alteration patterns by array CGH. Cytogenetic and Genome Research. *In Press*. S. Karger AG, Basel

nificant advantages over baseline models [2]. The key contribution is a hierarchical HMM that explicitly models passenger and putative driver alterations, increasing specificity, and borrows statistical strength in the raw data across samples, increasing sensitivity. We conclude with a discussion of the limitations of our approaches and suggest future directions in Section 3.6.

3.2 Introduction to multiple sample analysis

A recurrent CNA in a cohort of patients is a CNA found at the same genomic location in multiple samples. Therefore, recurrent CNAs define a pattern that provides a molecular characterization of the cohort's phenotype, potentially identifying disrupted molecular processes, molecular targets for diagnosis, and development of novel therapeutics. Recent work has revealed previously undescribed recurrent CNAs that are implicated in cancer [8, 54], demonstrating that the catalogue of disease-related CNAs is far from complete. Generally, it is assumed that recurrent CNAs are evidence for so-called "driver" alterations, or alterations that are symptomatic and/or causative of the disease [12]. Indeed, some of these alterations are used for prognostic testing [55] and the development of diagnostic tools [56]. Furthermore, recurrent CNAs are thought to be selected for in the clonal evolution of a tumour and their study can suggest the presence of genes involved in disrupted cancer-related biochemical pathways [11][3]. Note that we make the explicit assumption that recurrent CNAs are merely *candidate* driver alterations that can be prioritized for functional studies to fully determine their roles. The "passenger" CNAs, in contrast, are those that are patient specific and are generally not shared across the population. They can be considered random effects and may result from acquired genomic instability, non-pathological copy number variations [3, 4] or other mechanisms that are not well-understood. Thus, separating driver CNAs from passenger CNAs is critical to reveal potential diagnostic/prognostic markers

[2] Some of the material in this chapter was previously published in: S P Shah, W L Lam, R T Ng, and K P Murphy. Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics, 23(13):450458, Jul 2007.

[3] We note parenthetically that the ability to detect driver alterations depends on the molecular homogeneity and composition of the patient cohort. If the cohort is heterogeneous (composed of several distinct molecular subtypes), important driver alterations of a rare subtype could be obscured by patterns from the remainder of the population. This topic is the focus of Chapter 5.

as well as therapeutic targets for improved clinical care and management of the disease [57].

In this Chapter, we describe the development of statistical models for the detection of recurrent CNAs across multiple aCGH experiments from a cohort of individuals. In Figure 3.1 we show aCGH data from a set of 8 mantle cell lymphoma cell lines, originally published in DeLeeuw *et al* [19]. (Note that routine studies often consist of 10s or 100s of cases, but we use this smaller example for illustrative purposes.) Recurrent CNAs, identified by visual inspection, are shown in the blue shaded areas. The problem of computationally detecting such recurrent CNAs is relatively under-represented in the bioinformatics literature. As such, the limitations of current algorithms in practice are not yet fully understood.

3.2.1 Statistical properties of the data

To further illustrate the complexity of this problem, we show examples from lung cancer cell lines in Figures 3.2-3.4 over small regions containing three different types of recurrent CNAs on chromosomes 8, 9 and 1. Figure 3.2 shows a recurrent CNA harbouring the *MYC* oncogene.

One common strategy to identify such a recurrent CNA is to first pre-process individual samples to make calls of losses and gains (using, for example HMM-R (see Chapter 2)), and then to infer recurrent CNAs using a threshold frequency of occurrence [12, 19, 47]. We call this process AF for alteration frequency (see Section 3.4 for details). While AF may detect signals as shown in Figure 3.2, pre-processing or discretizing the sequences separately may remove information by smoothing over short or low-amplification CNAs. However, by *jointly* considering all the data without pre-processing, we can borrow statistical strength [58] across the samples and identify locations where the signal is shared in the raw data. For example, in Figure 3.3, we show data at the locus containing an important NSCLC gene, carbonic anhydrase IX (CA9) [59, 60]. Logratios of probes overlapping the gene are shown as blue stars and are indicated with arrows. This shared CNA may be hard to detect using AF because when processing individual samples, single probe CNAs are often indistinguishable from experimental noise [45]. With high-dimensional arrays many investigators require CNAs to span at least 2 consecutive

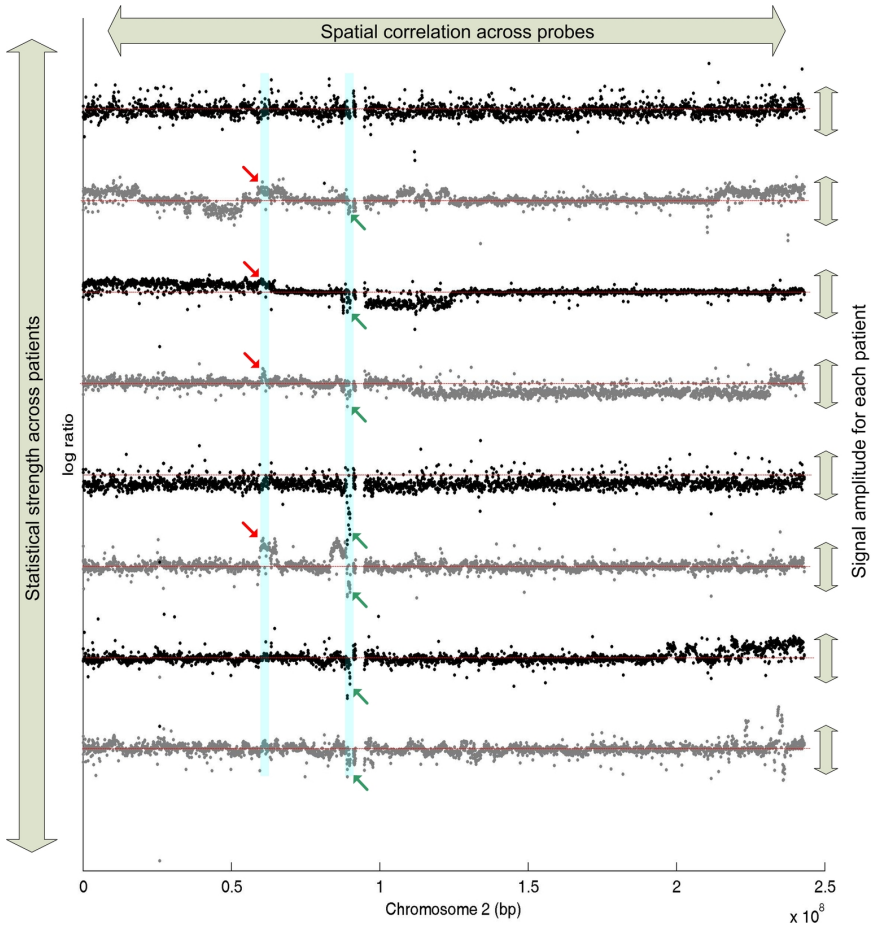


Figure 3.1: Example aCGH data from 8 mantle cell lymphoma cell lines [19] showing two examples of recurrent CNAs (shaded in blue) found on chromosome 2. Each horizontal set of dots represents the log ratios of a given cell line (or patient). The red dotted lines indicate the 0 log ratio (or expected neutral value). The probes that lie in the blue shaded areas (recurrent loss on the right, and recurrent gain on the left in four cell lines depicted by red arrows) comprise the desired output of an algorithm to detect recurrent CNAs. Note that for the recurrent CNAs, statistical strength across patients can be leveraged to detect them. Also note that CNAs tend to span regions of contiguous probes, thus spatial correlation across the chromosome should be leveraged. Finally the amplitude of the signal for each patient should be considered in the analysis.

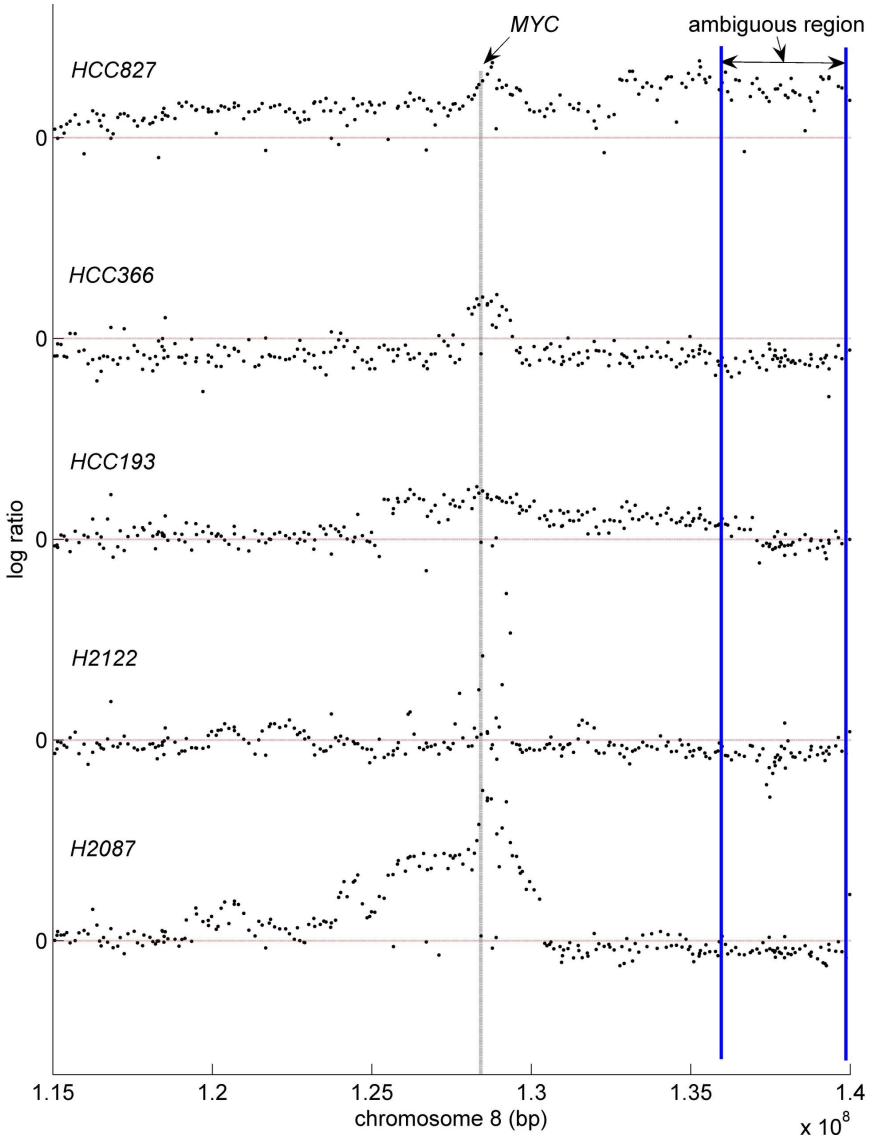


Figure 3.2: aCGH profiles for five NSCLC cell lines (labeled on the left) showing a high level shared amplification of a region spanning approx 3Mb containing the *MYC* oncogene on chromosome 8 (shown with vertical line). Horizontal red lines indicate the 0 log ratio level for each sample. Vertical grey lines indicate the position of a known gene of interest in NSCLC.

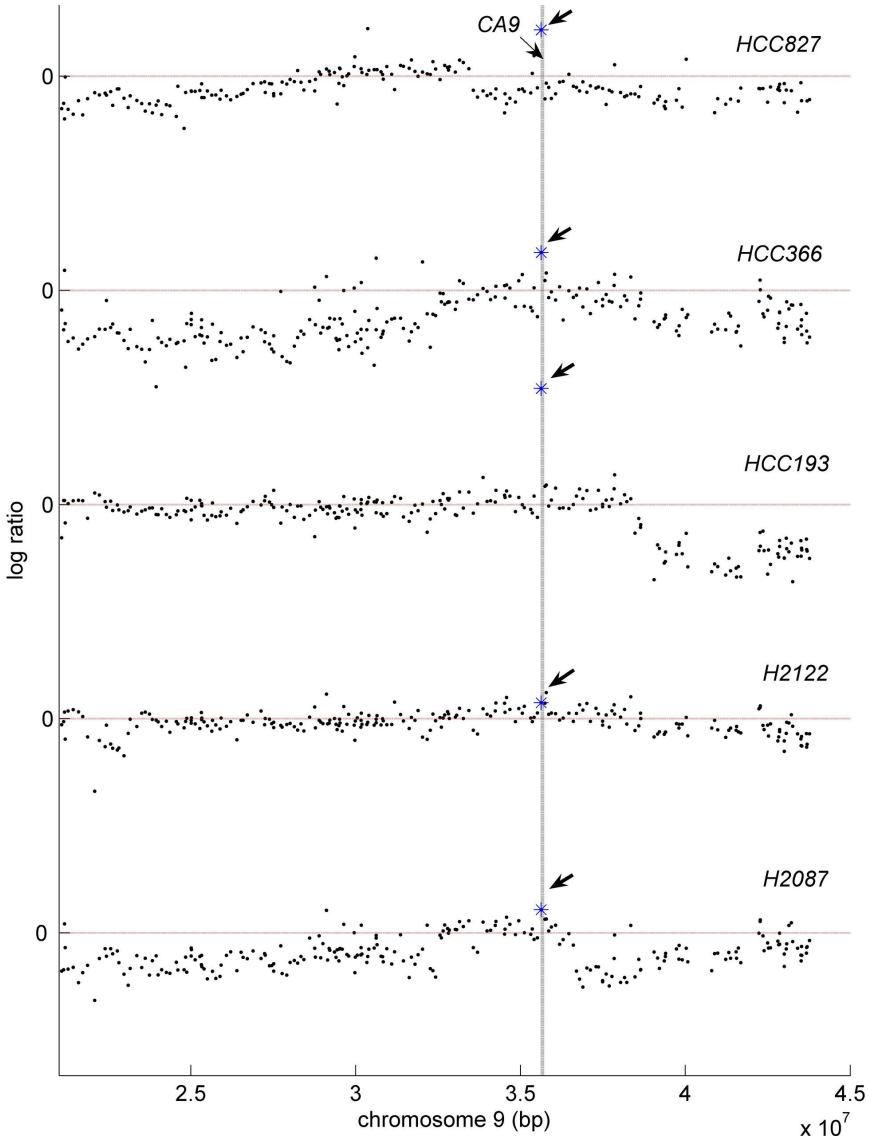


Figure 3.3: aCGH profiles for five NSCLC cell lines (labeled on the right) showing a single clone shared aberration at the CA9 locus on chromosome 9. Horizontal red lines indicate the 0 log ratio level for each sample. Vertical grey lines indicate the position of a known gene of interest in NSCLC.

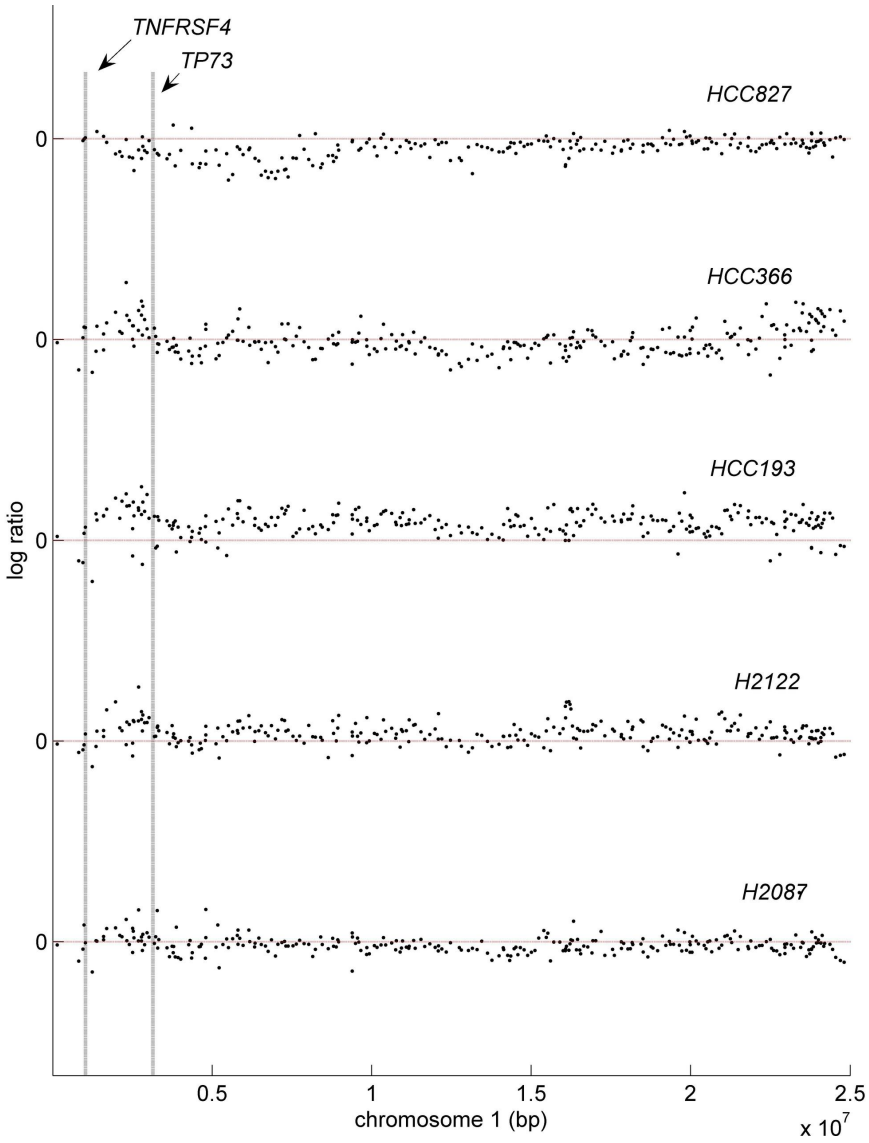


Figure 3.4: aCGH profiles for five NSCLC cell lines (labeled on the right) a low-level amplification on chromosome 1 including TNFRSF4 and TP73 - both implicated in NSCLC. This region is an example of a recurrent CNA that may be undetectable if each sample is pre-processed separately.

probes [45–47]. However, if a single probe CNA is shared across many samples, it may correspond to an important biological feature.

Figure 3.4 shows a third type of signal that is a low-level or subtle shared CNA. The region includes two known lung cancer related genes, *TNFRSF4* [61] and *TP73*. When compared to the *MYC* region in Figure 3.2, the level of amplification for *TNFRSF4* is much lower and is more difficult to distinguish from noise. However, cell lines H2122, HCC193 and HCC366 appear to share the low-level amplification. Furthermore, the *TP73* (a putative tumour suppressor involved in cell death [62]) loci exhibits low-level negative signals in three of the samples. These signals may be lost if each sample is pre-processed in isolation due to premature thresholding.

Most of the genome will not exhibit shared CNA patterns. Figure 3.2 (right end) shows a region from $\sim$135-140Mb (bounded by blue vertical lines) that is heterogeneous across the samples. One sample (HCC827) has an amplification while two are neutral (HCC193, H2087) and two are deletions (HCC366, H2122). This ambiguity in the signal across samples will be important when we develop our model in Section 3.4.

To address the goal of detecting recurrent CNAs from aCGH data, we present novel statistical models that extend the single sample hidden Markov model presented in Chapter 2 to the multiple sample case. We consider three different ways to do this. The first simply modifies the observation model of the HMM so that at each location, a vector of observations is generated, one per sample. We call the state sequence of the HMM the "master" sequence. It represents a classification of each probe location into a loss, neutral or gain state and hence it represents the canonical signal that encodes recurrent CNAs.

The second model augments this by allowing each observation in each sample to either be generated from the master sequence, or from its own private sequence. This allows for sample-specific random effects to be superimposed on the canonical signal. We demonstrate that this improves performance significantly. Finally, the third model augments the state space of the master sequence to allow undefined states, which represent locations which are ambiguous (such as the 135–140MB region in Figure 3.2). This allows the master to focus on the highly conserved regions, and to ignore heterogeneous locations. We will show that the resulting

output we infer is comparatively sparse, making it easier to create a short list of candidate locations for experimental follow up.

The remainder of this Chapter is organised as follows. We will outline notation and the computational problem as well as related work in Section 3.3. We describe our contributions to this problem in Section 3.4. In Section 3.5.1 we demonstrate our results on simulated data, where we know the ground truth, and in Section 3.5.2, we demonstrate results on well-studied lung cancer cell line data [11]. In Section 3.6 we summarize this chapter, discuss some limitations of our approach and suggest future research directions.

3.3 Related work

3.3.1 Notation and computational problem

The concept of separating putative driver alterations from passenger alterations is mirrored by the notion of computational dimensionality reduction and/or feature selection. If we consider the set of probes in the array as features, the task is to select a small number of features that are likely to represent recurrent CNAs, and thus a molecular profile of the disease. To help formalise this problem, we introduce notation in this section and define the computational problem of inferring recurrent CNAs from aCGH data. A schematic diagram of computational workflows to aid the reader is shown in Figure 3.3.1. The algorithms we will discuss in Section 3.3.2 differ in their steps taken to traverse this diagram, starting at the raw aCGH data in ending with recurrent CNAs. We will see how the paths through the workflow diagram confer certain advantages/disadvantages in prediction of recurrent CNA. The algorithms in the white box on the left operate on called or discrete data, while the algorithms in the white box on the right operated on raw data (see next section). The aCGH data consist of a log ratio $Y_t^p \in \mathbb{R}$ of hybridization intensity of tumour DNA vs normal DNA for each probe $t \in (1,\ldots,T)$ in the array and for each patient $p \in (1,\ldots,P)$ in the population. (Note that we have changed the notation to reflect multiple patients in the input and thus the c superscript for chromsome from Chapter 2 is replaced with a p for patient). $Y_{1:T}^{1:P} = \mathscr{D}$ thus represents the full data matrix and represents a noisy measurement of actual copy number. Note that we assume

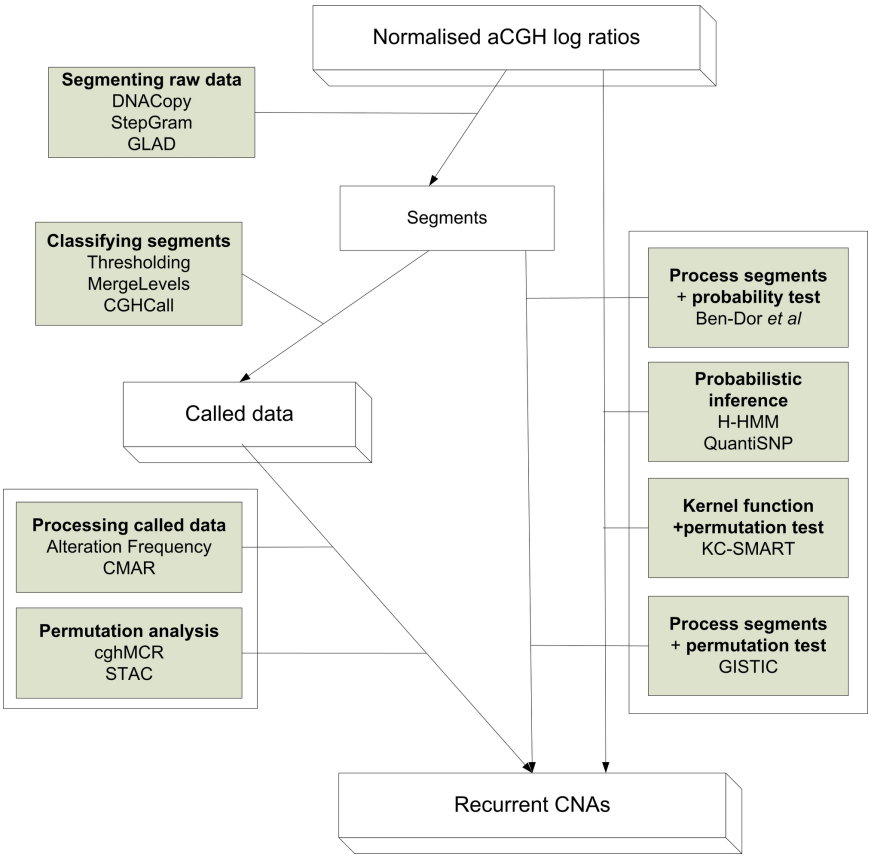


Figure 3.5: Workflows for inferring recurrent CNAs from aCGH data. We show the various steps used in predicting recurrent CNAs from aCGH data. The top part of the diagram shows the preprocessing steps some algorithms use to map raw data to called data. The algorithms in the white box on the left then process the called data to infer recurrent CNAs while those in the white box on the right process either continuous segmented data or the raw data directly. Also shown are which algorithms use probabilistic inference or permutation testing to produce their results. Please refer to Table 3.3.2 for availability of software.

that $Y_{1:T}^{1:P}$ is derived from a normalization step that has adjusted for technical biases and artifacts from the experimental protocol (see [63–65] for details).

A common step in analyzing aCGH data is to find a mapping $Y_{1:T}^{p} \rightarrow Z_{1:T}^{p}$ for each patient p where $Z_t^p = k$ represents a discrete copy number state, $k \in \{loss, neutral, gain\}$, to de-noise the data. These states correspond to: regions of genomic deletion; no change; and amplification, thus interpreting each probe's continuous log ratio with a biologically meaningful discrete label. As discussed in Chapter 2, to infer this mapping computationally is non-trivial (mainly due to sources of noise in the data) but has been extensively studied. We refer to the $Z_{1:T}^{1:P}$ matrix as called data and the $Y_{1:T}^{1:P}$ as raw data. The algorithms for inferring recurrent CNAs can be grouped to a large extent on whether they accept called or raw data as input. We will discuss the relative merits of the extant approaches and how the called or raw data can affect results. The output of all algorithms for detecting recurrent CNAs is a profile, which we represent by $\phi_{1:T}$. In some cases $\phi_{1:T}$ will represent a statistic or probability that a probe is recurrently altered, in other cases $\phi_{1:T}$ is simply a binary representation indicating presence or absence of a recurrently altered probe. It is assumed by most algorithms that recurrent CNAs span a relatively small fraction of the probes in the array.

3.3.2 Computational approaches for inferring recurrent CNAs

We divide the current approaches for inferring recurrent alterations into two categories: those that input the called data matrix $Z_{1:T}^{1:P}$ and those that input the raw data matrix $Y_{1:T}^{1:P}$. In general, there are three axes or dimensions across which the various approaches operate: i) the actual amplitude of the log-ratio signal contained in $Y_{1:T}^{1:P}$, ii) spatial correlation across probes (ie across columns in the data matrix) and iii) concurrence across the population (ie across rows of the data matrix). These dimensions are depicted with double-ended grey arrows on Figure 3.1. We will see how different algorithms exploit these characteristics. Note that an underlying assumption of some algorithms is that the patient cohort is relatively homogeneous. Preprocessing the data to separate the patients into subgroups with shared molecular patterns, or by known clinical subtype should be done if possible. We will discuss computational progress in this area in Chapter 5.

Algorithms for called data

The simplest algorithm for inferring recurrent CNAs from $Z_{1:T}^{1:P}$ is simply to compute a frequency of alterations for each probe such that:

$$\phi_t(k) = \frac{1}{P} \sum_{p=1}^{P} I(Z_t^p == k) \tag{3.1}$$

where $I(Z_t^p == k)$ is a function indicating that Z_t^p is in state k. Sets of probes from $\phi_t(loss)$ and $\phi_t(gain)$, are then selected based on frequency thresholding (for a recent example, see [66]). We refer to this method as *alteration frequency* (AF). While AF may be effective in some cases, it is limited in that it does not directly output meaningful statistics or probabilities to the investigator to quantify and thus compare the observed results.

Many authors treat the recurrent CNA problem as finding regions in the $Z_{1:T}^{1:P}$ matrix spanning contiguous set of probes in CNAs that maximally overlap across the patients. An example of this approach is reported in Aguirre *et al* [6] and available in the Bioconductor [67] package *cghMCR* (see Table 3.3.2 for availability). This method uses a step-wise approach and a permutation test to find recurrent CNAs based on a statistical score. The data are first segmented with DNACopy [29, 30]. Segments above an upper and lower user-settable threshold are labeled as CNAs resulting in the previously discussed $Y_{1:T}^p \rightarrow Z_{1:T}^p$ mapping. Highly altered CNAs are retained as important regions that define discrete locus boundaries. These regions are compared across patients to identify overlapping groups of positive or negative value segments. Minimal common regions (MCRs) are defined as regions having at least a user-defined recurrence rate across samples and where the median logratio for the probes with the segment across patients is above the 95% percentile in a permutation test. This method was the first to suggest a computational approach for identifying recurrent CNAs. Although it was shown to be effective in [6], it is somewhat ad-hoc and depends on number user-settable thresholds. One does not typically know how to correctly choose these thresholds, and a particular setting may not generalize well to other data sets.

A more mathematically motivated approach is presented by Rouveriol *et al* [68]. They present a formal definitional framework based on a binarized rep-

Algorithm	Input data	Availability
cghMCR [6]	Called	http://www.bioconductor.org
CMAR [68]	Called	from author
STAC [69]	Called	http://cbil.upenn.edu/STAC
CoCoA [70]	Raw	n/a
H-HMM [21]	Raw	http://www.cs.ubc.ca/∼sshah/acgh
KC-SMART [71]	Raw	n/a
QuantiSNP [52]	SNP	http://www.well.ox.ac.uk/QuantiSNP
GISTIC [72]	SNP	http://www.broad.mit.edu

Table 3.1: List of algorithms for recurrent CNAs, their data input and their availability if applicable.

resentation of the data (treating losses and gains separately and independently). The framework is used to develop two algorithms for discovering recurrent CNAs termed minimal altered region (MAR) and constrained minimal altered region (CMAR). A simplistic summary of the CMAR algorithm is that it searches for small rectangles of 1's in the input binary matrix, similar in concept to bi-clustering. Both are data mining methods based on finding closed constrained itemsets (sequences) in the binary matrix restricted to sequential data - a necessary extension due to the spatially ordered nature of the aCGH probes in the genome. The CMAR algorithm has a worst case running time of $O(T^2)$, which may limit its use to aCGH platforms with smaller numbers of probes.

Diskin *et al* [69] also input a binary matrix similar to Rouveriol *et al*. Their method, STAC, computes two complementary statistics for quantifying the likelihood of observed recurrent CNAs. The first estimates how often the observed frequency of an alteration would occur by chance. This is expected to uncover highly frequent alterations. The second is termed a footprint statistic, and is computed on the results of a greedy search strategy to find overlapping 'stacks' of alterations in the population. This is expected to detect recurrent CNAs that are low-frequency yet possibly of clinical importance. In both cases, permutation analysis is performed to assess the statistical significance of what is observed. The statistical output allows the prioritization for experimental follow up not possible in Rouveriol *et al* which produces binary output, however the permutation step may be impractical for very high resolution arrays.

cghMCR, CMAR and STAC all input called data. Working with called data

has its advantages in that with respect to some characteristics, the data are assumed to be de-noised. Thus, the specificity of predictions is expected to be high. However, in the next section we discuss how working with called data may limit the sensitivity of predictions in certain circumstances.

Algorithms for raw data

Thus far, we have considered algorithms that require discrete called data, or the $Z_{1:T}^{1:P}$ matrix, described above as input. Several authors: Lipson *et al* [73], Shah *et al* [21] (explained in Section 3.4), Klijn *et al* [71] and Ben-Dor *et al* [70] assert that inputting the raw data as input has advantages over called data.

Ben-Dor *et al* [70] argue that the amplitude of aberration should contribute to the inference of recurrent CNAs. Consider the case where the data are discretized into a small number of states (eg $\{loss, neutral, gain\}$), then (for example) high level amplicons, which arguably provide stronger evidence of being selected in the clonal evolution of the tumour, would contribute equally as a single copy gain to discovering recurrent CNAs. Furthermore, important high level amplicons may be infrequently targeted, making them harder to detect by methods discussed thus far. To leverage the amplitude of the signal, Ben-Dor *et al* use a statistical framework based on the concept of measuring probe penetrance. The approach begins by segmenting the raw data using continuous segmentation (using StepGram [73]), thus producing a intermediate form of the data that preserves amplitude, but is piecewise constant. Depending on the amplitude and the relative abundance of CNAs in the sample, a statistic is computed to quantify the significance of each CNA. More formally, given a region R spanning a putative CNA in a given patient p, the algorithm computes how many other regions of R's size in the continuous segmented data of p have at least the same average amplitude across patients. This computes a patient-specific score $s(p,R)$. Given P patients, the statistical significance of observing the data spanned by R in the whole population is given by an adjusted probability density function based on the Binomial distribution. This approach differs from the methods described in Section 3.3.2 in that it provides probabilistic output and models the signal amplitude across patients.

Klijn *et al* [71] suggest a method called KC-SMART, a locally weighted re-

gression algorithm based on kernel convolution to compute a smoothed estimate of the recurrent CNAs. The method also considers all three dimensions of the data: amplitude, spatial correlation and frequency of alteration. Using $Y_{1:T}^{1:P}$ as input, the data are separated into positive and negative log-ratios. The positive ratios are summed across patients and the negative ratios are summed across patients. These sums are used in computing the amplitude of a Gaussian kernel convolution function, whose values are then smoothed, providing a single estimate for the combined log ratios across the population at an arbitrary genomic position. Thus the profile is a smooth, continuous representation of the raw data matrix. Statistical significance of the amplitude of the peaks is assessed using permutation analysis. A key feature of the algorithm is that the width of the kernel is defined by the user, and thus can be tuned to find large recurrent CNAs and small recurrent CNAs.

3.3.3 Related algorithms for SNP arrays

Genotyping technology is also commonly used for copy number analysis. Single nucleotide polymorphism (SNP) chips can interrogate more than 1 million loci in the human genome in a one experiment and consequently robust computational approaches have been developed for their analysis. Although not explicitly designed for aCGH, the approaches described in this section are easily modified to use with aCGH data and thus are very relevant to the discussion.

Colella *et al* [52] suggest an HMM approach, QuantiSNP, with and emission model expressed in terms of the allele-specific intensities of the array. The hidden states in the model represent the combined copy number and genotype for each probe. The transition matrix between these states is non-stationary, and is computed using a distance based prior, accounting for unequal genomic spacing of the probes. Importantly, this method introduces Bayes factors for assessing significance levels for altered regions. These significance measures are computed on segments, whereas most HMMs for aCGH output likelihood of the best sequence, or probe-specific probabilities. For population-level analysis, the authors suggest placing a transition matrix at each probe that is jointly updated across patients. Thus the non-stationary transition matrix models recurrent CNAs by leveraging statistical strength across patients. We discuss this model further in Section 3.6.2.

Beroukhim *et al* [72] suggest a method called 'Genomic Identification of Significant Targets in Cancer' (GISTIC) based upon a step-wise workflow similar in spirit to KC-SMART, albeit with some notable differences. In Beroukhim *et al*, the amplitude of the logratios (inferred from the allele-specific probes) are summed across patients to compute a probe-level score. The probes are repeatedly permuted and scores are recalculated for each permutation. The probes with scores in the original data that occur rarely by chance are selected by thresholding. A novel contribution is that the data in significant 'regions' are post-processed to characterize 'peaks' as focal alterations (for example that span single genes), broad alterations (for example that span entire chromosomes), or overlapping peaks of both types. In contrast to Klijn *et al* where small and large peaks are found by iterative runs with different parameter settings, GISTIC explicitly classify large and small peaks in a single run. We have described published extant methods for recurrent CNA detection. We now describe our own contributions to this problem.

3.4 Methods

We implemented 4 models to study the problem of detection of recurrent CNAs from a set of aCGH data. We take a model-based approach that leverages our previous work described in Chapter 2. As mentioned above, a pattern consists (roughly speaking) of a list of locations which are highly conserved (either Loss, Neutral or Gain). The pattern can be represented as a "master" sequence of states $M_{1:T}$ where $M_t \in \{L,N,G\}$ is a multinomial random variable and $t \in (1,2,...,T)$. The locations where $M_{1:T} = L$ or $M_{1:T} = G$ are considered putative driver CNAs, and represent the output of our methods. Since we will often be uncertain about what the pattern should be at any given location, we will summarize our uncertainty using the (marginal) posterior distributions $\phi_t = p(M_t|\mathscr{D})$, which we call a *profile*. When we have data from different groups (as in our lung cancer data), we learn a different profile for each group, $\phi_t^g = p(M_t^g|\mathscr{D}^g)$. As shown in Section 3.5.2, we analyze four different phenotypic groups of lung cancer (ie $g = 1:4$). However, we will drop the g superscript for brevity. (Note that the problem of learning such groups from data is the subject of Chapter 5.)

Our task is related to learning profile HMMs for multiple sequence alignment

[74], but it is harder because the raw data is noisy and continuous-valued. Below, we describe four different approaches to the problem. The first is the method most widely used in current practice, and the remaining three are novel methods that we propose.

3.4.1 Alteration frequency (AF) model

As mentioned earlier, the simplest approach, AF, first processes each sample (or patient) $Y^p_{1:T}$ into a discrete sequence $Z^p_{1:T}$, where $Z^p_t \in \{L,G,N\}$. We chose the HMM-R method (see Chapter 2) for this implementation of AF to allow a more direct algorithmic comparison to the multiple sample HMMs we describe below. Note that other algorithms could be used for this step. For example, Coe *et al* [11] used aCGH-smooth [32] to pre-process the lung cancer data presented in Section 3.5.2. After preprocessing, we compute the empirical distribution over each state in each location to yield the profile $\phi_t = p(M_t|\mathscr{D})$, which can be represented as a $K \times T$ stochastic matrix, where $K = 3$ is the number of states, T is the length of the sequence, and each column sums to one. This can be further simplified to just compute the empirical probability of a recurrent CNA at each location, to yield a $1 \times T$ vector. The disadvantage of this method is that the mapping from Y^p to Z^p is done on each sample separately, so information cannot be shared across samples. Thus the method may smooth over important signals, as we will see.

3.4.2 Factored likelihood HMM (FL-HMM)

The second model, which we call "factored likelihood HMM" (FL-HMM), is a standard HMM model for $M_{1:T}$ (modeling the fact that CNAs tend to occur in runs), but where we modify the likelihood function to generate multiple samples instead of a single sample. Specifically, we assume the samples are conditionally independent given M_t and use a Gaussian observation model, yielding

$$p(Y_t^{1:P}|M_t = j) = \prod_{p=1}^{P} \mathscr{N}(Y_t^p|\mu_j^p, \sigma_j^p) \tag{3.2}$$

The observation model is a product over the emission densities of the samples, hence the term "factored likelihood". We have one mean and variance parameter

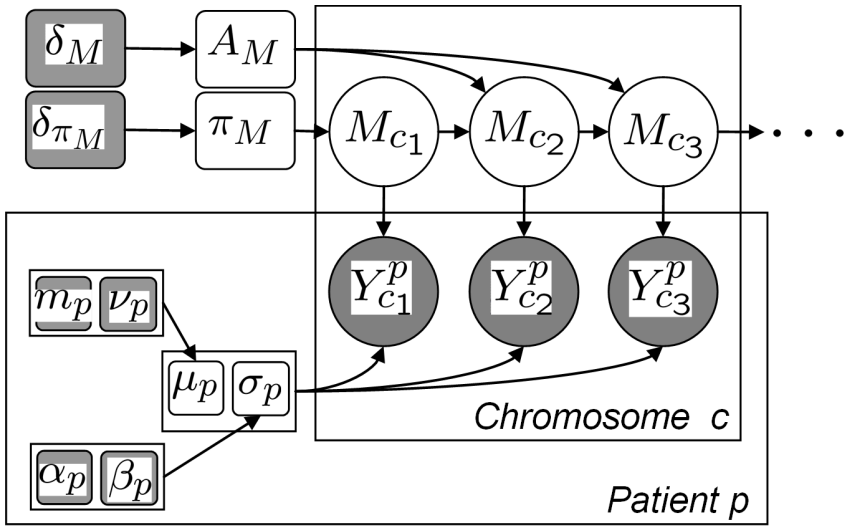


Figure 3.6: FLHMM model shown as a directed graphical model (Bayesian network). Circles represent random variables and rounded squares represent parameters. We only show the models for 3 probes, but in reality, the number of random variables is proportional to the number of probes on the chromosome, T_c. Unknown quantities are unshaded and observed quantities are shaded. $Y^p_{c,t}$ represents the observed log ratio of patient p in chromosome c at location t, $M_{c,t} \in \{L,N,G\}$ is the hidden master state. The shaded square nodes represent fixed hyper-parameters. Arrows between nodes indicate probabilistic dependencies. Boxes around variables are called "plates" and represent repetition of the contents inside. Thus we see that the observation parameters μ_p and σ_p are shared (tied) across chromosomes (since they are outside the c plate) but are specific to each sample (since they are inside the p plate), while the HMM parameters A_M, π_M are shared across chromosomes and samples.

for each of the 3 states of the HMM identical to the parameters described in Section 2.2. The mean and variance are patient specific, to model the fact that different samples often have quite different noise characteristics (see Section 2.2.3). In addition we pool the estimates of μ^p_j and σ^p_j for statistical strength as described in Section 2.2.1. The variable M_t has Markovian dynamics with transition matrix A_M, representing the probability of switching between the L/N/G states. The starting state distribution is denoted π_M. The model is shown as a directed graphical model

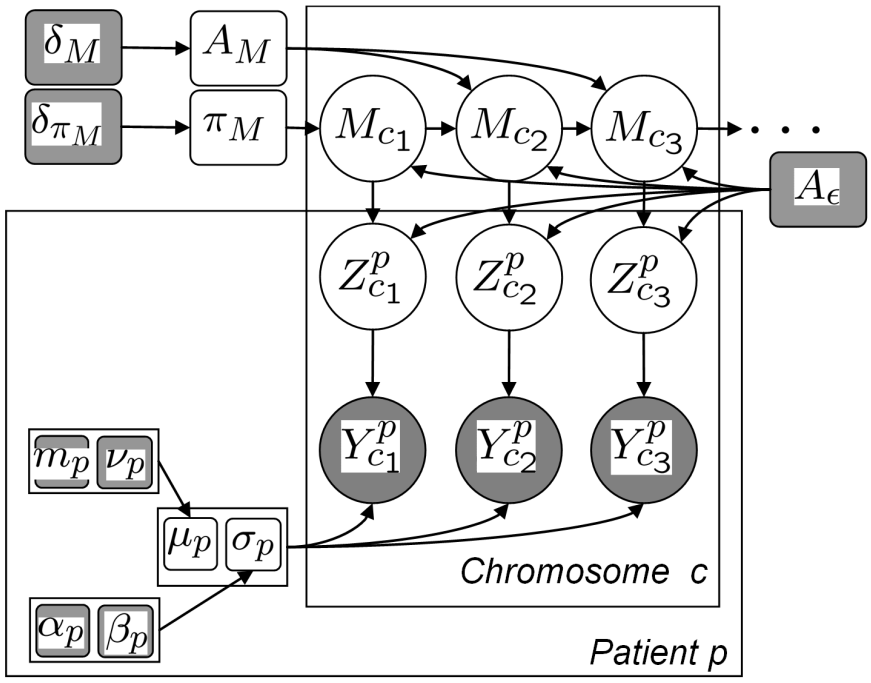


Figure 3.7: BFLHMM model shown as a directed graphical model (Bayesian network). Refer to caption for Figure 3.6 for details of the graphical model. $Y^p_{c,t}$ represents the observed log ratio of patient p in chromosome c at location t, $M_{c,t} \in \{L,N,G\}$ is the hidden master state. In comparison to Figure 3.6, here we introduce $Z^p_{c,t} \in \{L,N,G\}$, a hidden "slave" state expected to buffer the $M_{c,t}$ from large patient-specific deviations from neutral (please see text for details).

in Figure 3.6.

We add standard conjugate priors to all the parameters [58]. Specifically, for the multinomial distributions we use Dirichlet priors, $A_M \sim Dir(\delta_M)$ and $\pi_M \sim Dir(\delta_{\pi_M})$, where the matrix of pseudocounts δ_M encourages self-transitions, and δ_{π_M} encourages the neutral state. For the sample-specific emission density parameters, the priors, hyperparameter setting and initializations procedure is identical to the procedures outlined in Section 2.2 for HMM-R.

We use a Markov chain Monte Carlo (MCMC) inference algorithm to estimate the parameters of the model. This means that the parameters μ_k are sampled from

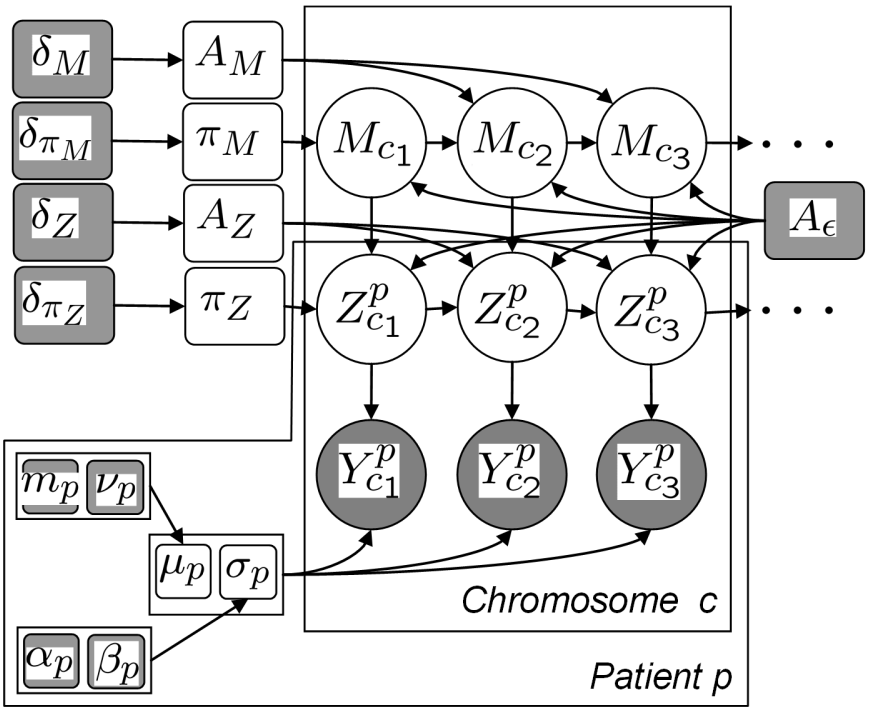


Figure 3.8: HHMM model shown as a directed graphical model (Bayesian networks). Refer to caption for Figure 3.6 for details of the graphical model. $Y^s_{c,t}$ represents the observed log ratio of sample s in chromosome c at location t, $M_{c,t}$ is the hidden master state and $Z^p_{c,t}$ is the hidden "slave" state. Here we augment the state space of $M_{ct} \in \{L, G, N, U\}$ and introduce Markovian dynamics on the Z^p_{ct} process (see the horizontal links and the new A_Z, π_Z parameters). Please refer to text for details.

their posterior distributions. This could be problematic in the rare case that the sampled estimate of μ_1 is greater than the sampled estimate of μ_3. Recall from Chapter 2, that we must ensure identifiability of the hidden states (i.e., to ensure state 1 means loss, 2 means neutral and 3 means gain). To do this, we use a truncated Gaussian on μ^p_j, to ensure $\mu^p_1 < \mu^p_2 < \mu^p_3$. The lower truncation bounds are set to $m^p_k - \sigma^p_y$ and the upper truncation bounds are set to $m^p_k + \sigma^p_y$. This is similar in nature to Guha *et al* [43]. Algorithm 3 shows the posterior distributions that each of the parameters are sampled from. Let $\theta = (\mu^{1:P}_{1:K}, \sigma^{1:P}_{1:K}, A_M, \pi_M)$ be all the parameters of the model. We can estimate the parameters of this model, $p(\theta|\mathscr{D})$,

using a Markov chain Monte Carlo (MCMC) algorithm called blocked Gibbs sampling [75]. This entails alternating between sampling $M_{1:T}$ as a block using the forwards-filtering backwards-sampling (FFBS) algorithm, and sampling the parameters individually conditioned on $M_{1:T}$ and the data: see Algorithm 3 for details. Alternatively, we can compute a point estimate, $\theta^{MAP} = \arg\max_\theta p(\mathscr{D}|\theta)p(\theta)$, using the EM algorithm. In either case, we initialize A_M with 0.9 on the diagonals and 0.1 spread over the remaining entries. We initialize π_M to favour neutral states. We can then run EM/ MCMC.

Note that we have also implemented a Student-t emission model in the EM setting with untruncated priors that equally well in practice and has fewer hyperparameters to set. This is because we can use untruncated, strong priors to control the identifiability problem (see Section 2.3.3) and in EM, the identifiability of the states is not subject to sampling. Moreover, EM converges much faster than MCMC resulting in dramatic runtime speedups, but as discussed below, it may result in less accurate predictions.

3.4.3 Buffered factored likelihood HMM (BFL-HMM)

The problem with the FL-HMM model is that M_t is summarizing the raw data $Y_t^{1:P}$. If any single sample at a given position has a large deviation from neutral, the master is likely to think that location is aberrated (because the neutral state cannot generate large aberrations). Thus large but rare deviations will be added to the profile. (This problem was also noticed by Lipson *et al* [73].) A simple fix to this is to add a "buffer" to each observation, $Z_t^p \in \{L,N,G\}$, which is responsible for generating the observation Y_t^p. Now the master will summarize these discrete states rather than the raw data. A key point is that in contrast to the AF model, we estimate Z and M simultaneously. See Figure 3.7.

In more detail, the BFL-HMM can be defined as follows. The "slave" Z_t^p processes are modeled as noisy versions of the master process: $p(Z_t^p = j|M_t = k) = A_\varepsilon(j,k)$, where

$$A_\varepsilon = \begin{pmatrix} \varepsilon & \frac{1-\varepsilon}{2} & \frac{1-\varepsilon}{2} \\ \frac{1-\varepsilon}{2} & \varepsilon & \frac{1-\varepsilon}{2} \\ \frac{1-\varepsilon}{2} & \frac{1-\varepsilon}{2} & \varepsilon \end{pmatrix} \tag{3.3}$$

Algorithm 3 Blocked Gibbs sampling algorithm for H-HMM. We omit the π_M and π_Z terms for brevity. FFBS stands for forwards-filtering backwards-sampling. The input is $y_{1:T}^{1:P}$, initial estimates for $Z_{1:T}^{1:P}, \mu^p \sigma^p$ and ε. The output is the marginal posterior probability $p(M_t|.)$

for $iter = 1, 2, \ldots$ **do**
  /* Sample states */
  for $t = 1 : T$ **do**
    $B(j,t) = \begin{cases} \prod_s A_\varepsilon(j, Z_t^p) & \text{if } j \in \{L, G, N\} \\ \prod_s A_Z(Z_{t-1}^p, Z_t^p) & \text{if } j = U \end{cases}$
  end for
  $M_{1:T} \sim FFBS(A_M, B^{1:T})$
  for $p = 1 : P$ **do**
    for $t = 1 : T$ **do**
      $B(j,t) = \mathcal{N}(y_t^p | \mu_j^p, \sigma_j^p)$
      $A_Z^t(i,j) = \begin{cases} A_\varepsilon(M_t, j) & \text{if } M_t \in \{L, G, N\} \\ A_Z(i,j) & \text{if } M_t = U \end{cases}$
    end for
    $Z_{1:T}^p \sim FFBS(A_Z^{1:T}, B^{1:T})$
  end for
  /* Sample parameters */
  $A_M \sim Dir(\delta_M + \sum_{c,t} I(M_{ct} = i, M_{c,t+1} = j))$
  $C_Z = \sum_{c,s,t} I(Z_{ct}^p = i, Z_{c,t+1}^p = j) I(M_t = U))$
  $A_Z \sim Dir(\delta_Z + C_Z)$
  for $s = 1 : P$ **do**
    for $j = 1 : K$ **do**
      $n_j^p = \sum_{c,t} I(Z_{ct}^p = j)$
      $\bar{y}_j^p = \frac{1}{n_j^p} \sum_{c,t} I(Z_{ct}^p = j) y_{ct}^p$
      $\bar{\lambda}_j^p = \frac{1}{n_j^p (v_j^p)^2 + (\sigma_j^p)^2}$
      $(\bar{\sigma_j^p})^{-2} = \frac{1}{(v_j^p)^2} + \frac{n_j^p}{(\sigma_j^p)^2}$
      $\mu_j^p \sim \mathcal{N}\left(\bar{\lambda}_j^p ((\sigma_j^p)^2 m_j^p + n_j^p (v_j^p)^2 \bar{y}_j^p), (\bar{\sigma_j^p})^2\right)$
      $\bar{\beta}_j^p = \frac{1}{2} \sum_{n=1}^{n_j^p} (I(Z_{ct}^p = j)(y_{ct}^p - \bar{y}_j^p))^2$
      $\lambda_j^p \sim Ga(\alpha_j^p + n_j^p/2, \beta_j^p + \bar{\beta}_j^p)$
    end for
  end for
end for

$$
\begin{array}{rcl}
p(A_M(i,.)|\delta_M) & \sim & Dir(A_M(i,.)|\delta_M) \\
p(\pi_M|\delta_{\pi_M}) & \sim & Dir(\pi_M|\delta_{\pi_M}) \\
p(M_t = j|M_{t-1} = i, A_M) & \sim & A_M(i,j) \\
p(A_Z(i,.)|\delta_Z) & \sim & Dir(A_z(i,.)|\delta_Z) \\
p(\pi_Z|\delta_{\pi_Z}) & \sim & Dir(\pi_Z|\delta_{\pi_Z}) \\
p(Z_t^p|Z_{t-1}^p, M_t, A_Z, A_\varepsilon) & \sim & \begin{cases} A_Z(Z_{t-1}^p, Z_t^p) & \text{if } M_t = U \\ A_\varepsilon(M_t, Z_t^p) & \text{if } M_t \in \{L,N,G\} \end{cases} \\
p(Y_t^p|Z_t^p = k, \mu^p, \sigma^p) & \sim & \mathcal{N}(Y_t^p|\mu_k^p, \sigma_k^p) \\
p(\mu_k^p|m_k^p, v_k^p) & \sim & \mathcal{N}(\mu_k^p|m_k^p, (\sigma_k^p)^2 v_k^p) \\
p((\sigma_k^p)^{-2}|\alpha_k^p, \beta_k^p) & \sim & Gam((\sigma_k^p)^{-2}|\alpha_k^p, \beta_k^p)
\end{array}
$$

Table 3.2: Conditional probability distributions for H-HMM

Here ε is the probability that the slave copies the master state. If we set $\varepsilon = 0$, the slaves never copy the master, so the posterior profile will equal the prior profile, i.e., we will not have learned anything, since M_t will be disconnected from the data Y_t^p. As we increase ε, each slave is influenced by the master with increased strength. Thus more of the samples will get reflected in the profile. If we set $\varepsilon = 1$, we are requiring that the slaves perfectly copy the master. This reduces to the FL-HMM model. In practice, we find it best to set $\varepsilon \sim 0.8$. See Section 3.5.2 for further discussion on the effect of ε. We can estimate the parameters in this model using MCMC or EM. We simply modify the algorithm to handle the fact that the observation model is now (a product of) a mixture of Gaussians, with mixing weights $p(Z_t^p = j|M_t = k)$.

3.4.4 Hierarchical HMM (H-HMM)

The problem with the BFL-HMM is that the slaves *have* to copy the master with probability ε at every location, even if this location is highly variable. We extend the model by adding an undefined (don't-care) state U to the master. Now if $M_t = U$, the slaves follow their own private Markovian dynamics, modeling local runs which are not shared (or assumed to be passengers), as shown in Figure 3.8. The

complete list of conditional probability distributions for this model is shown in Table 3.2. If $M_t \neq U$, they copy the master with probability ε as before:

$$p(Z_t^p | Z_{t-1}^p, M_t, A_Z, A_\varepsilon) = \begin{cases} A_Z(Z_{t-1}^p, Z_t^p) & \text{if } M_t = U \\ A_\varepsilon(M_t, Z_t^p) & \text{if } M_t \in \{L, N, G\} \end{cases}$$

The effect of this is that only highly conserved regions are stored in the profile; in the highly variable regions, the profile says "undefined". This makes the profile sparser, and easier to interpret. This is shown below in Figure 3.11. The degree of sparsity is controlled by ε: as we increase ε, the sparsity decreases, since more of the slaves influence the master (see Figure 3.13).

Estimating the parameters in this model is harder, since the Z^p chains become coupled due to the hidden common cause M (c.f., factorial HMMs [76]). However, conditioned on $M_{1:T}$, the $Z_{1:T}^p$ are independent and can be sampled in parallel using FFBS, so blocked Gibbs sampling is still easy. See Algorithm 3 for details. (Note that in the simpler BFL-HMM and FL-HMM models, we could integrate out the Z parameter since they are not Markovian, making the E-step easier.) An interesting feature of this model is that there are competing processes to explain the slaves. If the slave copies the master, its conditional probability distribution (CPD) is determined by A_ε, otherwise the CPD is determined by A_Z. Since A_Z is potentially estimated from a large subset of the data, it tends to converge to have diagonal values near 1. In contrast A_ε is fixed and therefore can be overwhelmed by the slave process. To avoid this, we use a strong prior on A_Z to discourage it from reaching near 1 on the diagonals, but still allowing it to be estimated from the data. This results in a 'fairer' competition between the A_Z process and the A_ε process.

We initialize the parameters as in the FL-HMM model. To initialize the states, we first sample each $Z_{1:T}^p$ using FFBS, with the master process turned off. We then initialize M_t to be the consensus majority state across $Z_t^{1:P}$, as in AF. As an alternative to MCMC, which is computationally demanding due to the number of MCMC samples required for convergence (1000s), we could use a Monte Carlo EM framework, where the E-step consists of alternately sampling $M_{1:T}|Z_{1:T}^{1:P}$, and $Z_{1:T}^{1:P}|M_{1:T}$ and the M-step consists of maximizing the parameters of the model given

$M_{1:T}$ and $Z_{1:T}^{1:p}$. Preliminary comparisons indicate that EM tends to converge faster but may give poorer results, perhaps because it is more prone to getting stuck in local optima.

3.4.5 Running time

Parameter estimation in all 4 models takes $O(T)$ time. This makes the technique scalable for use in high density oligonucleotide arrays or SNP arrays, frequently used for DNA copy number analysis, that may contain 500,000 or more probes per experiment. In practice the running time depends on the number of EM/MCMC iterations. For EM on the H-HMM model, we find the system converges within about 10 steps and takes about 90 minutes to learn a model from 20 samples with 32,000 probes each. (All experiments were performed in Matlab 7.2.0.294 (R2006a) on a Intel Xeon CPU @2.4GHz.) EM for the BFL-HMM and FL-HMM is much faster, since the E step can be performed exactly using the forwards-backwards algorithm, avoiding a Monte Carlo approximation.

3.5 Results

3.5.1 Quantitative results on synthetic data

Real data sets rarely have fully verified ground truth locations of recurrent CNAs. Thus, applying standard metrics to assess accuracy on real data is difficult. To overcome this, we created a synthetic data set derived from real data. We used eight mantle cell lymphoma samples originally published in Deleeuw *et al* [19] and used for a qualitative assessment in [68] and modified it to give us ground truth CNAs. We used the data for chromosome 20 (672 probes) which was reported to be relatively free of CNAs. We permuted the order of the data for each sample so as to remove any undetected shared signals that may be present across samples. We then inserted a recurrent CNA gain and a recurrent CNA loss at fixed positions of width w, in a fraction f of the samples. The clones within the region were shifted up/down (for gain/loss) by $\sigma_p \tau$ where p is one of the chosen patients, σ_p is the empirical variance of that patient's sample, and τ is the signal to noise ratio (SNR). Thus $\sigma_p \tau_p$ preserves the sample-specific heterogeneity of the noise. In or-

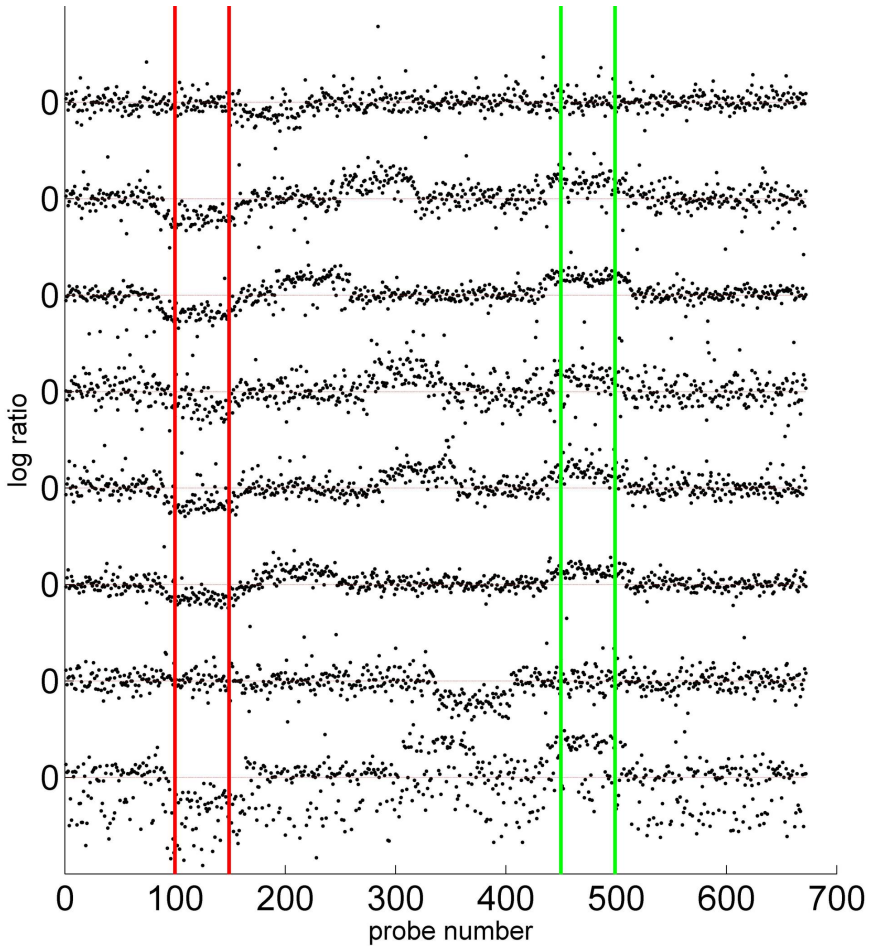
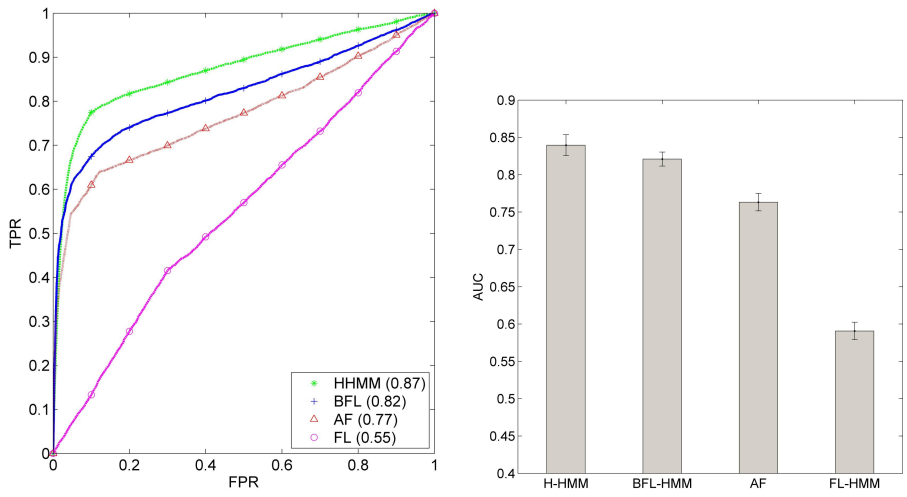


Figure 3.9: Example of the simulated data for $w = 50$, $\tau = 0.9$ and $f = 0.75$. Green lines (on the right) bound an inserted CNA gain, and red lines (on the left) bound an inserted CNA deletion.



(a) (b)

Figure 3.10: (a) ROC plot for the synthetic data for H-HMM (green stars), BFL-HMM (blue crosses), FL (red triangles) and AF (purple circles). TPR and FPR were calculated using results for all data, and therefore represent a summary of how the models compare over w, τ, f and ε. AUC for each model is indicated in brackets in the legend. H-HMM had the best performance overall (AUC=0.87), followed by BFL-HMM (AUC=0.82), AF (AUC=0.77) and FL (0.55). (b) Distributions of AUC for H-HMM, BFL-HMM, FL-HMM and AF over all values of w, τ, f and ε. H-HMM and BFL-HMM had statistically significantly better performance than AF and FL-HMM (one way ANOVA ($p \ll 0.01$). Whiskers indicate standard error bars. The mean and standard error AUC for the models was 0.84 ± 0.01 for H-HMM, 0.82 ± 0.01 for BFL-HMM, 0.76 ± 0.01 for AF and 0.59 ± 0.01 for FL-HMM.

der to "soften" the borders of the aberrations, we extended the borders by γ probes, where $\gamma \sim Gam(\alpha, 1)$ (α proportional to w - see text below). Here γ was sampled independently for each sample to ensure the exact borders of the aberrations were not shared. Finally, for each sample, we randomly sampled a location outside of the ground truth recurrent CNA and inserted a gain or loss (randomly chosen) of width w'. Figure 3.9 shows an example of the synthetic data for $w = 50$, $f = 0.75$, $\tau = 0.9$, $w' = 100$ ($\sim 15\%$ of the chromosome). The recurrent loss is at position

100-149 and the recurrent gain is at position 450-499. Comparing this figure to the real data in Figures 3.2-3.4, we see that the synthetic data is quite realistic and challenging.

We evaluated AF, FL-HMM, BFL-HMM and H-HMM on synthetic data for $w = (1, 10, 50)$, $f = (1/2, 3/4, 1)$, $\alpha = (1, 5, 10)$ and $\tau = (0.3, 0.6, 0.9, 1.2)$. For the BFL-HMM and the H-HMM, we set $\varepsilon = (0.8)$. Note for this large scale experiment we used (Monte Carlo) EM instead of MCMC for inference, to save time. However, preliminary results suggest that MCMC does work better, despite its increased cost.

Similar to experiments described in Section 2.3.1, we computed receiver operator characteristic (ROC) curves based on $p(M_t = A) = p(M_t = L) + p(M_t = G)$ where $p(M_t = A)$ is the probability that a recurrent CNA is predicted at position t. Using the ground truth labeling of the data, the false positive rate (FPR) is defined as $\frac{FP}{N}$ the number of probes incorrectly predicted as a CNA (FP) over the total number of non-CNA probes. The true positive rate (TPR) is defined as $\frac{TP}{P}$, the number of correctly predicted CNA probes (TP) over the true number of CNA probes. We plotted TPR vs FPR curves (varying a threshold on $p(M_t = A)$), and calculated area under this curve (AUC) as a measure of accuracy to test the effect over w,f,τ and ε across the various models.

Figure 3.10(a) shows a single summary ROC plot combining results for all values of w,f,τ and depicts the overall accuracy performance of the models. H-HMM had the highest accuracy (AUC=0.87) followed by BFL-HMM (AUC=0.82), AF (AUC=0.77) and FL (0.55). Figure 3.10(b) shows the mean AUC over for every setting of w,f,τ (repeated three times). The mean and standard error AUC for the models was 0.84 ± 0.01 for H-HMM, 0.82 ± 0.01 for BFL-HMM, 0.76 ± 0.01 for AF and 0.59 ± 0.01 for FL-HMM. H-HMM and BFL-HMM were significantly more accurate than AF and FL-HMM (one way ANOVA, $p \ll 0.01$). Although H-HMM had slightly higher mean of AUC than BFL-HMM, the result was not statistically significant. However, we show in the next section on lung cancer data that in practice, the H-HMM is considerably more useful to the investigator as it returns sparser, yet relevant predictions.

3.5.2 Qualitative results on lung cancer data

Ultimately we are interested in applying a model to aCGH data from clinically relevant samples. To compare the output characteristics of the various models, we ran the algorithms on aCGH samples from 39 well-studied lung cancer cell lines, originally published in [11, 47]. This data is particularly relevant since phenotype-specific patterns of recurrent CNAs have been experimentally validated. The samples can be subdivided into four groups: NSCLC Adenocarcinoma (NA), NSCLC Squamous cell carcinoma (NS), SCLC classical (SC) and SCLC variant (SV). Eighteen samples are NA, seven are NS, nine are SC and five are SV. This data has been rigorously studied and discordant shared patterns validated using PCR and gene expression have been identified across the major and minor groups [11, 47]. We fit separate profiles $\phi^g_{1:T}$, one per group, using each of the four models and we qualitatively assess the characteristics and biological relevance of the output, using results reported in [11] as a guide.

The experiments on synthetic data showed H-HMM and BFL-HMM are the best models. In this section we show how the explicit modeling of the ambiguity in the data by H-HMM displays a clear advantage over the other models. Recall that in Figures 3.2-3.4 we showed parts of chromosomes 8, 9, and 1 to illustrate different types of recurrent CNAs at important locations. Figures 3.11 and 3.12 show the output of H-HMM ($\varepsilon = 0.8$), BFL-HMM, FL-HMM and AF on the full chromosome 8 and the p-arm of chromosome 9. $p(M_t = gain)$ is plotted in green and $p(M_t = loss)$ is plotted in red. The clear trend is that H-HMM has sparser output and clearly predicts important regions in isolation. The sparsity is due to the output being dominated by positions where $p(M_t = U) = 1$. Note arrows at *MYC* and *CA9* for comparison to Figures 3.2 and 3.3. Considering Figure 3.11 in more detail, the p-arm (left) has a relatively high frequency of deletion and this is cleanly predicted by all models. In contrast, the centromeric half of the q-arm shows ambiguity in the AF plot. Both BFL-HMM and FL-HMM are unable to resolve the ambiguity as they are forced into a $\{L, N, G\}$ state, while the H-HMM can 'opt-out' of making a consensus prediction at these locations, choosing only to predict a CNA when the data cleanly support one (eg *MYC* locus). This illustrates the sparsity of the H-HMM compared to the other models.

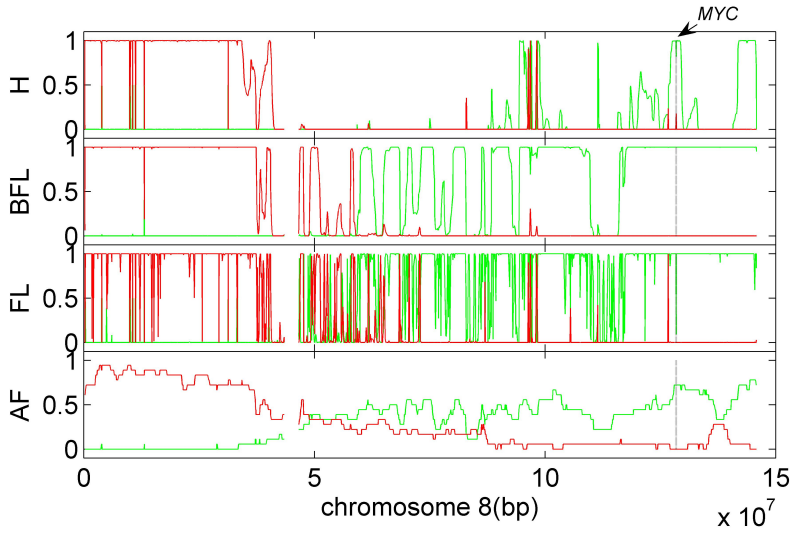


Figure 3.11: Output from top to bottom of H-HMM, BFL-HMM, FL-HMM and AF for the NA group, chromosome 8. The x-axis is the chromosomal position and the y-axis is predicted probability. Red plots indicate $p(M_t = L)$ and green plots indicate $p(M_t = G)$. Note the sparse, yet accurate predictions for the H-HMM at the MYC locus (recall Figure 3.4) and the p-arm loss prediction which recapitulates known results [47]. The other models either over-predict (BFL-HMM, FL-HMM) or under-predict (AF) the shared aberrations.

A similar result was seen for chromosome 1 at the *TPFRSF4* and *TP73* loci (see Figure 3.13 for results of H-HMM). Notice that BFL-HMM and FL-HMM also predict CNAs at these important genes. However, it is quite evident that they both over-predict, making it hard for an investigator to discern biologically relevant CNAs from spurious predictions. From Figure 3.11, we also see that AF has a peak at the *MYC* locus, but is unable to detect the recurrent CNA at *CA9* (Figure 3.12) with high frequency. In all 3 generative models, the signal is clearly predicted.

The combination of sparsity due to modeling ambiguity and the ability to tune ε allows the user to effectively set the false positive rate of the H-HMM. An example of the value of this is shown in Figure 3.13, displaying the results for group SC for various values of ε. The sparse output for $\varepsilon = 0.8$ reveals isolated peaks of high probability at locations of genes (*TNFRSF4, TP73, TNFRSF9, ZNF151, E2F2,*

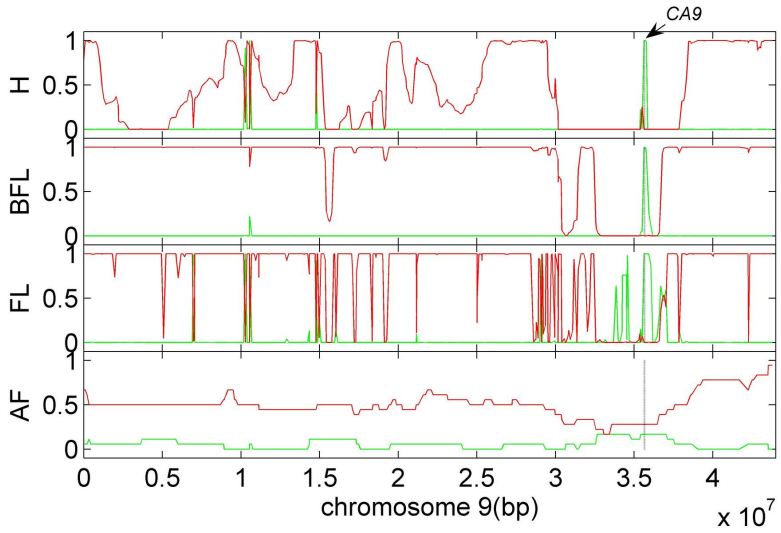


Figure 3.12: Output from top to bottom of H-HMM, BFL-HMM, FL-HMM and AF for the NA group for the p-arm of chromosome 9. (see Figure 3.11 for axes description). Similar to Figure 3.11, notice the sparse, yet accurate predictions for the H-HMM especially at the single probe *CA9* locus (recall Figure 3.3). The AF method does not predict CA9. BFL-HMM and FL-HMM both predict CA9, however they are over-predicting many other regions not likely to be shared CNAs.

FGR, EIF3S2, DMAP1, FUBP1, RAB13, HDGF, PPCC, NTRK1, TRAF5), whose expression is known to be altered in lung or other cancers. For example, *ZNF151* and *E2F2* were found to have copy number induced gene expression changes in [11]. Interestingly, the H-HMM predicts the *TP73* region as a narrow loss embedded within the gain region harbouring *TNFRSF4* shown in Figure 3.4. *TP73* was detected at only 22% frequency in AF and was not detected at all in BFL-HMM. Additional relatively narrow but high probability peaks correspond to the *EIF3S2* locus, which mediates the TGF-β pathway, *FUBP1* a transcriptional activator of *MYC* and the co-amplification of *TNFRSF4* and *TRAF5*, which are known interactors and activators in the NF-κB pathway [61]. These results are computational predictions, yet many provide compelling evidence that they merit experimental follow up.

To investigate whether H-HMM recapitulates the results in [11], we examined

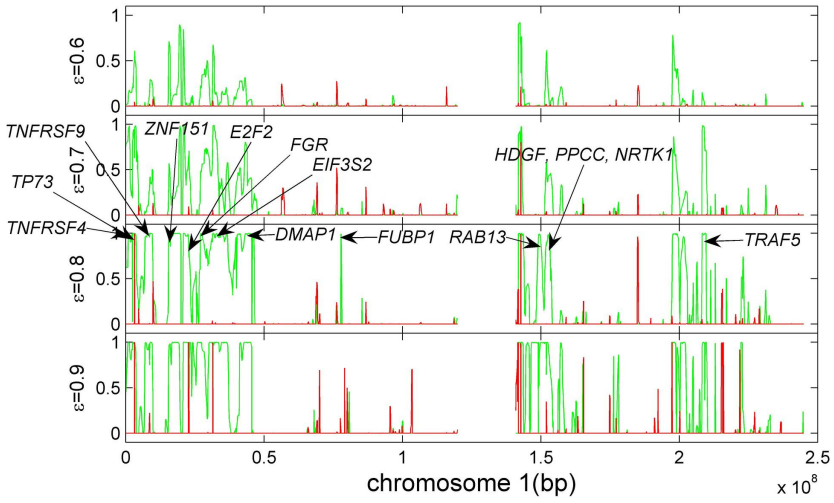


Figure 3.13: Output from H-HMM on chromosome 1 for different values of ε for SC group. For $\varepsilon \geq 0.8$, gain probability 'peaks' correspond to locations of several genes (annotated with arrow) implicated in lung or other cancers.

a subset of genes reported to be differentially disrupted in the two major groups, NSCLC and SCLC. These 22 genes are involved in key lung cancer pathways and therefore represent a highly relevant set of markers as a reference to assess our output. The H-HMM predicted shared aberrations in regions harbouring 14 of the 22 genes in at least 1 of the subgroups of NSCLC and SCLC. We counted a prediction if $p(M_t = L) > 0.5$, or $p(M_t = G) > 0.5$ for losses and gains respectively. The predicted genes included *STMN1, E2F2, SC, ZNF151, ID2, MAPK9, EGFR, CDK2NA, KNTC1, HMGB1, HSPH1, JJAZ1, NLK, JUNB, TIAM1, DSCAM*. Five of the regions were detected at $\varepsilon \leq 0.7$, eleven at $\varepsilon \leq 0.9$ and the remaining regions at $\varepsilon = 0.95$. This gives us a reasonable estimate for how to calibrate ε in order to predict relevant CNAs. The H-HMM did not predict recurrent CNAs harbouring the remaining genes *PRDM2, SOX11, MAP3K4, ING1, SMAD4, CCDC5, TCF4*.

We assessed if the H-HMM could determine differences in the profiles of the phenotypic groups (NA, NS, SC, SV), as this was part of the focus of the study of [11] and [47]. Figure 3.14 shows that the H-HMM produces very different pro-

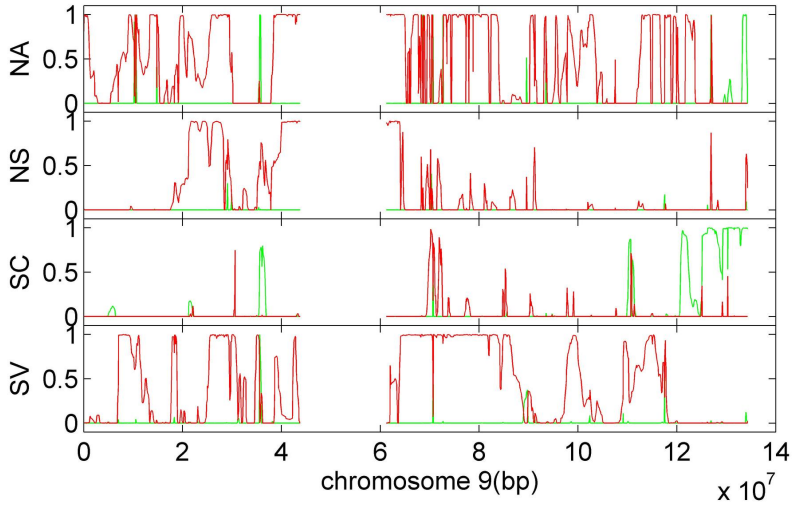


Figure 3.14: H-HMM output for chromosome 9 showing discordant patterns among the lung cancer groups (NA, NS, SC, SV).

files for chromosomes 9 across the different subgroups. This example chromosome was chosen as is was previously shown to have different patterns of CNA [11, 47]. Although anecdotal, our qualitative results give us confidence that the H-HMM is predicting biologically relevant recurrent CNAs. Combined with the result that the H-HMM is sparser in its output, we believe the H-HMM has the right characteristics of presenting biologically meaningful results to the investigator while maintaining a low false positive rate.

3.6 Discussion

We developed three novel methods that extend the single sample HMM for aCGH to the multiple sample case in order to infer recurrent CNAs. Our results indicate that the H-HMM, which simultaneously infers discrete labels for the samples and promotes sparsity by modeling ambiguity in the data is quantitatively and qualitatively better than simpler models and standard methods. In informal qualitative assessment we showed that the H-HMM produces meaningful biological output when compared to a list of experimentally validated genes. The H-HMM was able

to detect previously reported discordant patterns among the lung cancer groups - a key requirement to determine phenotype specific CNA patterns.

3.6.1 Other applications of H-HMM

Quality controls enforced in the procedure for generating aCGH data attempt to limit experiments that produce uninterpretable data due to noise. Occasionally, data is produced that pass through the quality control, but are still noisier than usual. In such cases, the experiment is often repeated, producing a replicate for a given sample. Replicates may also be produced by design. The H-HMM model is well suited to jointly analyze replicates from the same sample in order to infer a consensus CNA pattern in the presence of noise. In addition, our model could easily be extended to jointly analyze data from individuals generated from multiple platforms.

The output of the H-HMM has potential uses beyond identifying recurrent CNAs. As discussed earlier, it represents a sparse CNA profile of the cohort and is therefore a type of feature selection algorithm that identifies altered probes relevant to the disease entity under study. Therefore, these features could be used to train a classifier to recognize the phenotype associated with the cohort.

3.6.2 Limitations of H-HMM

The H-HMM has several limitations which need to be addressed. As mentioned earlier, there is competition between the slave process and the master process to explain the data, without restricting the slave transition matrix A_Z. In regions with very strong consensus, this model sometimes prefers to explain the data using the slave process rather than the master process. Qualitatively, the Student-t emission model seems to be more robust to this phenomenon. This remains an open problem and further study is necessary to evaluate the possible solutions.

Another limitation of H-HMM is that it is not likely to detect low frequency shared CNAs as the model forces the master to choose a discrete state. We propose that a discrete master sequence may be too limiting to detect low frequency alterations. We have addressed this in two important ways. First, we have adapted the model described in Colella *et al* [52] for aCGH data. This model was described for

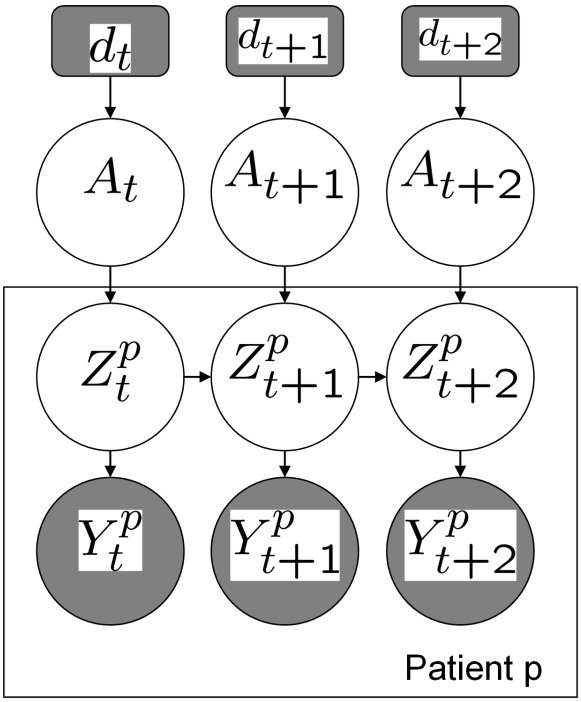


Figure 3.15: Multiple sample CMM. This is an extension of the model shown in Figure 2.22, where the non stationary transition matrices A_t are shared by all patients (A_t is outside p plate). Therefore, similar to the H-HMM, statistical strength can be borrowed across patients. However, the representation of the profile is continuous, rather than discrete, therefore low frequency events may be captured (see Figure 3.16, for example).

the single sample case in Chapter 2. Figure 3.15 shows the graphical model for the multiple sample case. We call this model the continuous master model (CMM). Figure 3.15 shows that the profile is represented by non-stationary, shared transition matrices A_t. The inference of A_t is dependent on all the patients, and thus, borrowing statistical strength across samples is still leveraged, but the master sequence is continuous, and capable of outputting low frequency events. Figure 3.16 shows a qualitative comparison of the H-HMM and CMM on the FL cohort. The CMM has desirable qualities in that it allows the investigator to quickly identify small regions of recurrent alteration.

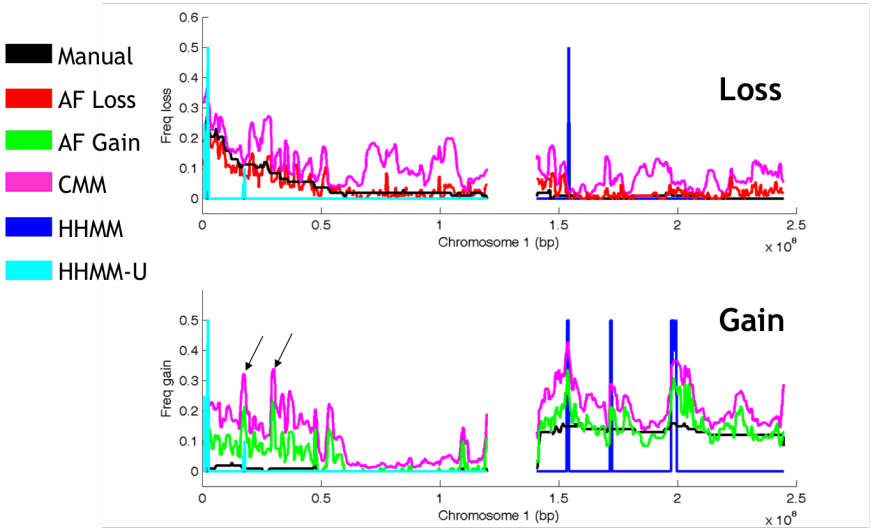


Figure 3.16: Comparison of recurrent CNA predictions of CMM and H-HMM on chromosome 1 for 106 follicular lymphoma patients (see Chapter 4). The top plot shown losses and the bottom plot shows gains. In this figure we superimpose the probability curves of frequency of manual calls (black), AF loss (red), AF gain (green), CMM (pink), H-HMM (blue) and H-HMM where $M_t = U$ (light blue). The characteristics of the output show that H-HMM is very sparse in comparison to AF and CMM, making the data easy to interpret. However, CMM (black arrows) detects signals that H-HMM does not. These may be important low frequency events missed by H-HMM.

In addition to the CMM, Chapter 5 discusses how infrequent, but important CNAs may be obscured by molecular heterogeneity in the data, and therefore missed by H-HMM. In Chapter 5, we describe a clustering approach for detecting putative driver CNAs in the presence of such heterogeneity.

Chapter 4

Case study: Genome-wide aCGH profiling of follicular lymphoma

4.1 Summary

In this chapter, we present a case study of determining CNAs from aCGH data [1]. This was a collaborative study between Dr. Douglas Horsman's group and myself. My role in the project was in designing and implementing the analysis strategy for this data. One important outcome is that we learned how to tune the parameters of the HMM-R for clinical application, and learned that indeed the model is performing very well on a real-world, previously unstudied data set. I performed all the bioinformatics related to this project including collaborative participation in the clinical data statistics. I am a co-first author on the accepted publication.

We generated aCGH data for 106 diagnostic biopsies of follicular lymphoma to characterize regional genomic imbalances. Using an analytical approach that defined regions of copy number change as intersections between visual annotations and the HMM-based algorithm, we identified 71 CNAs that were recurrent in 10% of cases. These ranged in size from 200 kb to 44 Mb, affecting chromosomes

[1] The material presented in this chapter has been accepted for publication: K-J Cheung, S P Shah, C Steidl, N Johnson, T Relander, A Telenius, B Lai, K P Murphy, W Lam, J M Connors, R T Ng, R D Gascoyne, and D E Horsman. Genome-wide profiling of follicular lymphoma by array comparative genomic hybridization reveals prognostically significant DNA copy number imbalances. Blood - in press, 2008

1, 5, 6, 7, 8, 10, 12, 17, 18, 19, and 22. We also demonstrated by cluster analysis that 46.2% of the 106 cases could be sub-grouped based on the presence of +1q, +6p/6q-, +7 or +18. Survival analysis showed that 21 of the 71 regions correlated significantly with inferior overall survival (OS). Of these 21 regions, 16 were independent predictors of OS using a multivariate Cox model that included the International Prognostic Index (IPI) Score. Two of these 16 regions (1p36.22-p36.33 and 6q21-q24.3) were also predictors of transformation risk and independent of IPI. These prognostic features may be useful to identify high-risk patients as candidates for risk-adapted therapies.

4.2 Introduction

Lymphoid malignancies account for 5% of cases of cancer in the U.S. and have continued to rise in frequency at 3-4% annually [77, 78]. Of the different types of indolent lymphoma, follicular lymphoma (FL) is most prevalent and has a variable clinical course with a median survival of 10 years [79]. While management strategies have changed, advanced-stage FL remains an incurable disease using conventional therapies [80]. Approximately 85% of FL is associated with a specific balanced translocation, t(14;18)(q32;q21), that leads to overexpression of the anti-apoptotic gene BCL2 due to its relocation in proximity to an IgH enhancer element [81–84]. This genetic abnormality alone, however, is unlikely to produce clinical FL, as BCL2 over-expressing transgenic mice do not develop lymphoma [85, 86] and t(14;18)-bearing lymphocytes have been frequently demonstrated in healthy individuals [87, 88].

If the pathogenesis of FL results from a sequential accumulation of genetic alterations [89], the analysis of early neoplastic lesions may define the critical events associated with the initial development and further progression. To the best of our knowledge, there have been 12 large studies reported in the Western Hemisphere in the last decade that have investigated chromosomal imbalances in FL using a combination of techniques including conventional karyotyping, comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) technology. The reported recurrent copy number alterations have consistently included losses of 1p32-36, 6q, 10q and 17p, and gains of 1q, 2p, 7, 9p, 12, 17q, 18q and

X [90–101]. The analysis by Hoglund *et al* [102] utilized computational analysis of a large number of published G-banded FL karyotypes to define early from late accruing genetic imbalances and demonstrated four putative pathways of clonal evolution in FL [102]. This karyotype-based study was hampered by the inherent inaccuracies of G-banding analysis and excluded from consideration all marker chromosomes and unbalanced chromosomal additions that are common features of FL karyotypes. Further examination of such complex karyotypes by multi-colour karyotyping may improve the definition of these recurrent aberrations [103], however, the metaphases typically obtained from short term lymph node cultures allow only for the detection of DNA imbalances that exceed 5-10 Mb in size and may represent only a fraction of the sideline diversity present in FL genomes. No studies to date have utilized a combination of high-resolution genomic analysis and a large FL cohort composed exclusively of diagnostic biopsies.

The advent of array comparative genomic hybridization (array CGH) technologies now provides the capability to detect subcytogenetic DNA copy number gains and losses. These techniques have led to improvements in the characterization of both acquired and inherited genetic abnormalities [104]. In this study we have applied whole genome tiling path BAC array CGH, with a $>$200 kb resolution for detection of copy number alterations in clinical specimens and a reported tolerance of up to 70% contamination by non-tumor cells [105], to a cohort of 106 FL diagnostic specimens with complete clinical information. We have generated a comprehensive profile of regional copy number imbalances with which to identify significant prognostic correlates in relation to both survival and transformation risk.

4.3 Materials and Methods

4.3.1 Patient materials

The 106 FL cases were selected from the Lymphoid Cancer Research Database of the British Columbia Cancer Agency (BCCA) in Vancouver, British Columbia, identified between 1987 and 1996 based on the availability of sufficient frozen diagnostic tumour material and information on clinical outcome. Importantly, these

cases were enriched in part for cases where two or more sequential specimens were available from the indolent phase or when transformation had occurred. Transformation was defined as either histologically proven (biopsy demonstrating large B-cell lymphoma), or clinically proven (one or more of the following: sudden rise in LDH to $>$ twice the normal level, rapid discordant localized nodal enlargement, and new unusual extranodal involvement of organs such as brain, lung and bone). The time to transformation was defined as the time from diagnosis to clinical or pathological endpoint described above. The International Prognostic Index (IPI) Score was used to risk-stratify these patients because information on the haemoglobin level and number of nodal sites were not available to generate a FLIPI score [106]. All cases were classified as FL based on the criteria defined by the World Health Organization classification of tumours of haematopoietic and lymphoid tissues [107]. Of the 106 cases, 20 have been included in previously reported studies [96, 102].

4.3.2 Cytogenetic analysis

Cytogenetic analysis of lymph node specimens was performed as previously described [96]. Fluorescence in situ hybridization (FISH) was performed using the LSI IGH/BCL2 probe according to the manufacturers protocol (Vysis, Downers Grove, IL, USA) to detect the presence of IGH/BCL2 genomic fusion. For validation of deletion of the 1p36.32 locus, the RP13-493G06 or RP11-756P03 BAC clones were selected from the array CGH profile and prepared for use as FISH probes as previously described, while BAC RP11-229M05 at 1q32.3 was used for copy number control [108]. For validation of the 6q23.3 locus deletion, the RP11-703G08 BAC was used, while RP11-516E15 at 6p12.3 served as copy number control. All BAC clones had previously been identity-verified by BAC-end sequencing and hybridized to normal metaphases to confirm the expected site of chromosomal localization. The frequency of false deletion for each BAC FISH probe was established by hybridization to normal lymphocyte cell suspensions and ranged from 0.5 to 3.0%. For the purpose of this study the cut-off value for true deletion was set at $>5\%$.

4.3.3 DNA extraction

Genomic DNA extraction was performed according to standard procedures using proteinase K digestion and fresh frozen tissue or cells stored at -80C. The DNA was further purified using the Gentra puregene tissue kit (Qiagen, Mississauga, Ontario).

4.3.4 Whole genome tiling path BAC array CGH

The sub-mega base resolution tiling array contains 26,819 BAC clones spotted in duplicate and covers >95% of the human genome [16]. Array CGH was performed as previously described [109]. The array slide was scanned using a charged-couple device camera system to capture the cyanine-3 and cyanine-5 channels (Applied Precision, Issiquah, WA). The images were then analyzed by SoftWoRx microarray analysis software (Applied Precision), followed by a stepwise normalization procedure [110]. Data were filtered based on both replicate standard deviation (data points with >0.1 standard deviation removed) and signal to noise ratio (data points with a signal to noise ratio <3 removed). Copy number alterations were visualized using the SeeGH software available at http://www.flintbox.ca/technology.asp?tech=FB312FB [111].

4.3.5 Computational analysis

Intersection analysis

Scoring of array CGH data was performed separately by two methods: visual analysis by a cytogeneticist (D.E.H.), using a criterion for an aberration defined as an apparent log ratio shift away from baseline in a minimum of three adjacent BACs (200kb or larger), and computational analysis by determining probability of aberration (loss, neutral, or gain) for each clone using the program CNA-HMMer v0.1 (available at http://www.cs.ubc.ca/~sshah/acgh/), which is based on a Hidden Markov Model (HMM) [20]. Only those alterations identified by both HMM and visual interpretation were accepted as true. We modified the emission model of the HMM described in Shah *et al* to be a mixture of Student-t distributions, achieving the equivalent robustness to outliers while producing output that was

more interpretable to the investigator [20]. In addition, this modification required fewer hyperparameters to be set, which were selected automatically using an 'empirical Bayes' type approach [21]. Concordance between the visual calls and the HMM predictions was assessed by calculating the area under the receiver operator characteristic (ROC) curve. ROC curves are a plot of the true positive rate (TPR - proportion of clones called as an aberration that were also predicted by the HMM) against the false positive rate (FPR - proportion of clones predicted as aberrant by the HMM that were not called visually). The area under the ROC curve (AUC) is a single measurement that represents the trade-off between TPR and FPR. AUC was calculated for each sample. The average AUC in this study was 0.93. A perfect AUC would be 1. All analyses were run using default settings.

Cluster analysis

Clustering of the 106 cases was performed using the K-medoids (also called partitioning around medoids) algorithm. The input data $X(i, j)$ represented the copy number of clone j in case i. Only clones that showed a 10% rate of recurrent loss or gain determined by intersection analysis were used for clustering. A Hamming distance function of a case to a medoid was used and the algorithm was run 1000 times using random initializations of the medoids. The run producing the lowest total distance of cases to their assigned medoids was reported. The number of clusters was chosen to be five based on the previous work by Hoglund *et al* [102].

Clinical correlations

The log-rank test using the Kaplan-Meier method was performed for univariate analysis assessing the prognostic significance of each of the 71 regional aberrations on survival and the risk of transformation. Each case was dichotomized as positive or negative for each of the 71 regions, where positive was defined as having at least one alteration in the region. The Cox proportional-hazards model was used to identify only those regions reaching significance independant of the IPI score. All clinical statistical data were computed using the SPSS version 11 software.

4.4 Results

4.4.1 Clinical data

The clinical, morphologic and cytogenetic information on the cohort is presented in Table 4.2. In brief, 56% were male and 44% were female with a median age of 53 years. Forty-two percent of the patients died, with a median overall survival time of 10.83 years after diagnosis. Overall, 50% of patients had developed transformed lymphoma over a median follow up time of 7.33 years. The median time to transformation was 6.61 years. The majority of patients who developed transformed lymphoma had biopsy proven transformation (64%). These patients had a similar clinical outcome as those whose transformed lymphoma was diagnosed on clinical grounds. Transformed lymphoma was the cause of death in 64% of patients, supporting the observation that transformation is an important cause of mortality in these patients and may be a confounding factor in assessing the risk of genetic alterations affecting survival in patients who develop transformed lymphoma.

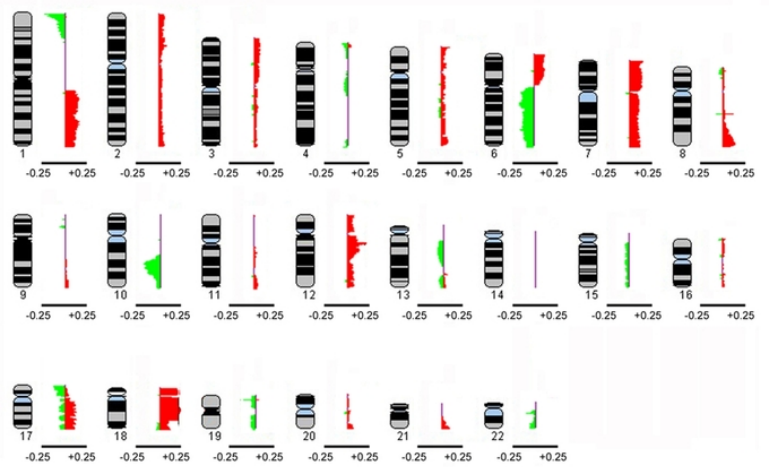
Treatment of these patients varied due to changes in era-specific approaches to management (Table 4.2). The effect of the addition of rituximab to standard chemotherapy could not be assessed reliably because of small numbers of patients (n=12; log-rank test on overall survival and transformation, p=0.7), recent incorporation of rituximab into primary treatment (after 2004) varying times of introduction (diagnosis, first progression, relapse, multiple relapses) and variable combination with standard agents (single agent or combination in multiple drug regimens).

4.4.2 Cytogenetic data

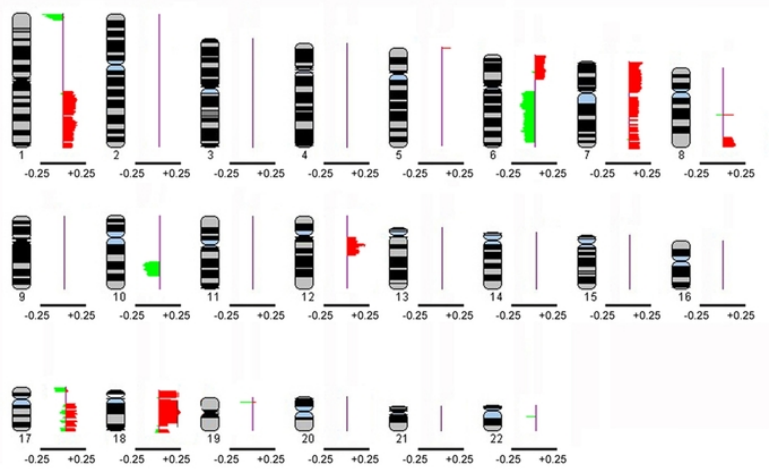
Ninety-three of the 106 cases had been studied by karyotype analysis and/or FISH using the IGH/BCL2 fusion probe. The t(14;18) or variant was present in 75 of these 93 cases (81%) but was absent in 18. Thirteen cases were not investigated by these techniques but had the standard morphologic features of FL.

Clinical characteristics	n=106 (%)	surv p	trans p
Median Age (yrs) = 53			
Male	59 (56)	0.4	0.7
Age $>$ 60	33 (31)	0.2	0.5
PS $>$ 1	13 (13)	0.003	0.3
LDH $>$ normal	24 (25)	0.008	0.1
Extranodal sties $>$ 1	13 (12)	0.4	0.6
Stage III/IV	74 (70)	0.01	0.1
IPI score:		0.003	0.02
0-1	34 (32)		
2-3	24 (32)		
4-5	6 (6)		
Diagnostic Pathology:			
FOLL1	63 (60)		
FOLL2	33 (31)		
FOLL3A	9 (9)		
Primary therapy:			
Observation	26 (24)		
Rad alone	15 (14)		
Single agent chemo	12 (11)		
Multi-agent chemo +/- rad	41 (39)		
Multi-agent chemo + rituximab	12 (11)	0.7	0.7
Outcome:			
Transformation:	53 (50)	0.02	
Biopsy proven	34 (64)	0.6	
Clinical	19 (36)		
Death: Unrelated	3 (7)		
From transformation	29 (64)		
From progressive indolent FL	13 (29)		
Median follow up alive = 7.33 yrs			
Median overall survival = 10.83 yrs			
Median time to transformation = 6.61 yrs			

Table 4.2: Patient characteristics of 106 FL specimens acquired at diagnosis. Abbreviations: PS: ECOG performance status; LDH, lactate dehydrogenase; IPI, international Prognostic Index; FOLL1 = follicular lymphoma grade 1, FOLL2 = grade 2, FOLL3A = grade 3A, rad = radiation, chemo = chemotherapy, FL = follicular lymphoma. surv p = logrank p-value (survival). trans p = logrank p-value (transformation)



(A)



(B)

Figure 4.1: Composite frequency ideogram plot of genome-wide copy number alterations in 106 diagnostic FL cases based on intersection analysis. (A) The frequencies of aberrations, represented by green signals for losses and red signals for gains, in the autosomes were derived from intersection analysis, where the union was taken between calls made visually by a cytogenetic pathologist and those determined by CNA-HMMer v0.1. (B) Composite frequency ideogram plot showing only those aberrations affecting 10% cases. The data were visualized using the SeeGH software. Genetic losses or gains are represented by green and red signals, respectively. The horizontal bar below each ideogram represents gain and loss frequencies of +0.25 and -0.25, respectively.

4.4.3 Profile of copy number alterations in FL

Each array CGH profile was annotated individually by visual inspection and by computation without knowledge of the associated karyotype. The individual profiles were combined to generate a genome-wide copy number profile of the 106 diagnostic FL specimens. Figure 4.1A represents a global composite profile ideogram of all aberrations affecting the 22 autosomes as determined by the intersection analysis. Figure 4.1B shows an ideogram of only those regions that were affected in $\geq$10% of cases. This 10% cutoff produced 71 altered regions ranging in size from 200 kb to 44 Mb. Overall, 97 of 106 cases (91.5%) had aberrations detected by array CGH with a median of 16.1% and a range from 0% to 32.2%. The most frequently altered region was band 1p36.22-p36.33 (11 Mb in size), showing 25.5% frequency of deletion. Table 4.3 provides details on the 71 regions of alteration.

A	B	C	D	E	I	J	M	N	O	P	Q
Chr	ID #	Start (bp)	End (bp)	Chr band	Size (bp)	Freq	US (pval)	CSI (pval)	UT (pval)	CTI (pval)	Genes
1p-	1	1120588	12511370	p36.22-p36.33	11390782	0.25	0.012	0.023	0.004	0.006	
1q+	2	144203231	149149179	q21.1-q25.1	4945948	0.13	NS		NS		
	3	149346998	171525137	q21.13-q25.1	22178139	0.16	NS		NS		
	4	173318349	186631007	q25.1-q31.1	13312658	0.14	0.039	NS	NS		
	5	192104220	215373858	q31.2-q41	23269638	0.17	NS		NS		
	6	216564705	223861673	q41-q42.12	7296968	0.12	0.049	0.041	NS		
	7	225671030	233337746	q42.13-q42.3	7666716	0.12	0.049	0.041	NS		
	8	236546601	244497402	q43-q44	7950801	0.11	0.014	0.018	NS		
	9	246412897	246741718	q44	328821	0.1	NS		NS		
5p+	10	568397	2059719	p15.33	1491322	0.1	0.003	0.002	NS		
6p+	11	101435	8166300	p24.3-p25.3	8064865	0.12	NS		0.01	0.023	
	12	9910293	15045693	p23-p24.3	5135400	0.11	NS		0.002	0.009	
	13	15322212	17396937	p22.3-p23	2074725	0.11	NS		0.002	0.009	
	14	18182851	19088319	p22.3	905468	0.1	NS		0	0.001	
	15	19741746	23462851	p22.3	3721105	0.1	NS		0	0.001	
	16	23588018	37906515	p21.2-p22.2	14318497	0.14	NS		0.006	0.009	
	17	39096919	42527171	p21.1-p21.2	3430252	0.11	NS		0.028	0.046	

Continued on next page...

A	B	C	D	E	I	J	M	N	O	P	Q
Chr	ID #	Start (bp)	End (bp)	Chr band	Size (bp)	Freq	US (pval)	CSI (pval)	UT (pval)	CTI (pval)	Genes
	18	42763847	44815780	p21.1	2051933	0.1	0.035	NS	0.008	0.007	
6q-	19	67856203	102485267	q12-q16.3	34629064	0.17	0.003	0.007	NS		
	20	104959688	145932243	q21-q24.3	40972555	0.17	0.016	0.038	0.002	0.001	
	21	146006611	158842383	q24.3-q25.3	12835772	0.12	NS		NS		
7p+	22	76475	9330205	p21.3-p22.3	9253730	0.16	0.007	NS	NS		
	23	11733679	12411228	p21.3	677549	0.11	0.021	NS	NS		
	24	13834918	57826849	p11.1-p21.2	43991931	0.15	0.049	NS	NS		
7q+	25	65183346	75792082	q11.21-q11.23	10608736	0.11	NS		NS		
	26	77519343	79287368	q21.11	1768025	0.11	NS		NS		
	27	79508881	81760593	q21.11	2251712	0.11	NS		NS		
	28	82203999	82396484	q21.11	192485	0.11	NS		NS		
	29	84076954	88755109	q21.11-q21.13	4678155	0.11	NS		NS		
	30	89067084	95807510	q21.13-q21.3	6740426	0.11	NS		NS		CYP51
	31	97314220	99254745	q21.3-q22.1	1940525	0.1	NS		NS		
	32	101780109	102036700	q22.1	256591	0.1	NS		NS		
	33	107062787	108434145	q22.3-q31.1	1371358	0.1	NS		NS		
	34	114602302	115391679	q31.2	789377	0.1	NS		NS		

Continued on next page...

A Chr	B ID #	C Start (bp)	D End (bp)	E Chr band	I Size (bp)	J Freq	M US (pval)	N CSI (pval)	O UT (pval)	P CTI (pval)	Q Genes
	35	115770526	117527587	q31.2-q31.31	1757061	0.1	NS		NS		
	36	119932154	124328532	q31.31-q31.33	4396378	0.1	NS		NS		
	37	124779872	128206833	q31.33-q32.1	3426961	0.11	NS		NS		
	38	131217474	135972389	q32.3-q33	4754915	0.11	NS		0.034	NS	
	39	136860444	140073345	q33-q34	3212901	0.11	NS		NS		
	40	140508716	144516051	q34-q35	4007335	0.12	NS		NS		
	41	147152251	151714055	q35-q36.1	4561804	0.12	NS		NS		
	42	151938143	158777885	q36.1-q36.3	6839742	0.12	NS		NS		
8q+	43	86699540	86900037	q21.2	207338	0.12	0.046	0.02	NS		
	44	127567638	144615332	q24.13-q24.3	17047694	0.16	NS		NS		KCNK9 NIBP PTK2/FAK PTP4A3
10q-	45	83519320	109743759	q23.1-q25.1	26224439	0.2	NS		NS		PTEN
12q+	46	39241592	40479848	q12	1238256	0.1	NS		NS		
	47	41164595	41849783	q12	685188	0.1	NS		NS		
	48	42875295	43431071	q12	555776	0.1	NS		NS		

Continued on next page...

A	B	C	D	E	I	J	M	N	O	P	Q
Chr	ID #	Start (bp)	End (bp)	Chr band	Size (bp)	Freq	US (pval)	CSI (pval)	UT (pval)	CTI (pval)	Genes
	49	45222971	47260459	q13.11	2037488	0.12	NS		NS		
	50	47315713	48862941	q13.11-q13.13	1547228	0.11	NS		NS		
	51	49766128	53769742	q13.13-q13.2	4003614	0.21	NS		0.038	NS	
	52	54124295	71253094	q13.2-q21.1	17128799	0.17	NS		NS		MDM2
17p-	53	433730	5025418	p13.2-p13.3	4591688	0.13	NS		NS		
	54	5418209	6976115	p13.1-p13.2	1557906	0.11	0.021	0.021	NS		
	55	7297479	7963869	p13.1	666390	0.1	0.019	0.01	NS		p53
	56	7873762	8172436	p13.1	298674	0.11	NS		NS		
	57	8633350	9331199	p13.1	697849	0.11	NS		NS		
17q+	58	28335370	34341092	q11.2-q12	6005722	0.12	0.003	0.017	NS		
	59	35115643	37009923	q12-q21.2	1894280	0.12	0.003	0.017	NS		
	60	40291911	49011357	q21.31-q22	8719446	0.13	0.012	0.037	NS		
	61	50468106	53853636	q22	3385530	0.12	NS		NS		ZNF161
	62	58325527	62428590	q23.2-q24.2	4103063	0.13	0.031	0.019	NS		
	63	64191239	69931286	q24.2-q25.1	5740047	0.12	0.007	0.007	NS		
	64	73418360	77524868	q25.3	4106508	0.11	0.021	0.017	NS		
18p+	65	35421	15060997	p11.21-p11.32	15025576	0.23	NS		NS		

Continued on next page...

A	B	C	D	E	I	J	M	N	O	P	Q
Chr	ID #	Start (bp)	End (bp)	Chr band	Size (bp)	Freq	US (pval)	CSI (pval)	UT (pval)	CTI (pval)	Genes
18q+	66	16793577	59123838	q11.1-q21.33	42330261	0.25	NS		NS		BCL2
	67	60284518	61729287	q22.1	1444769	0.1	NS		NS		
	68	63894258	64156108	q22.1	261850	0.1	NS		NS		
	69	70236233	76098439	q22.3-q23	5862206	0.12	NS		NS		
19p-	70	8566514	8784792	p13.2	218278	0.13	NS		NS		
22q-	71	21112875	21296725	q11.22	183850	0.1	NS		NS		

Table 4.3: Detailed information on the 71 regional aberrations affecting $\geq$10% of FL cases in intersection analysis (Data based on NCBI build 36.1). Columns are annotated with letters for easy reference in the text. Freq=Frequency of aberration, US=Univariate survival, CSI = Cox survival with IPI, UT = univariate transformation, CTI = Cox transformation with IPI

4.4.4 Association between copy number alterations and clinical parameters

When using the relative number of alterations per case to predict clinical outcome, defined as the number of altered BAC clones determined by intersection analysis divided by the total BAC clones (26,819) expressed as a percentage, we showed that there was no significant correlation between cases with 10% alterations and those with 10% in terms of overall survival and transformation risk (log-rank test, p=0.7 and p=0.06, respectively). However, significant correlation with overall survival and transformation risk was observed if cases with 5% alterations were compared with those containing 5% (log-rank test, p=0.02 and p=0.03, respectively).

Univariate analysis indicated that 21 regions correlated significantly with poor survival (Column M of Table 4.3), as did performance status, LDH, stage, and the IPI group (Table 4.2). Of these, 16 regions were independent of IPI in multivariate analysis (Column N of Table 4.3). Univariate analysis showed that 12 regions correlated significantly with risk of transformation (Column O of Table 4.3). The IPI group, but not the individual factors, was also predictive of transformation (Table 4.2). Ten of the 12 regions were identified as IPI-independent predictors of risk of transformation in multivariate analysis (Column P of Table 4.3). Del(1)(p36.22-p36.33) and del(6)(q21-q24.3 (identified by ID# 1 and 20, respectively) were not only associated with transformation and inferior outcome (see Figures 4.2 and 4.3) but were also IPI-independent predictors for both clinical variables (highlighted in grey in Table 4.3) and thus were selected as candidate regions for validation.

4.4.5 Validation of array CGH data

Two BAC clones, RP13-493G06 and RP11-756P03 that mapped to 1p36.32 and spaced 200 kb apart, were used for validation of the array CGH-detected 1p36 deletion. We performed FISH using these BAC clones on 10 selected cases. The RP11-229M05 probe at 1q32.3 was used as a control. Two of the 10 cases were determined by CGH intersection analysis to have no log ratio shift (no alteration) while 8 cases had evident deletions at 1p36.3 of variable size. The concordance rate between FISH and intersection analysis was 10 of 10 cases. Figure 4.4 shows the array CGH ideogram and FISH results for three representative cases, one without

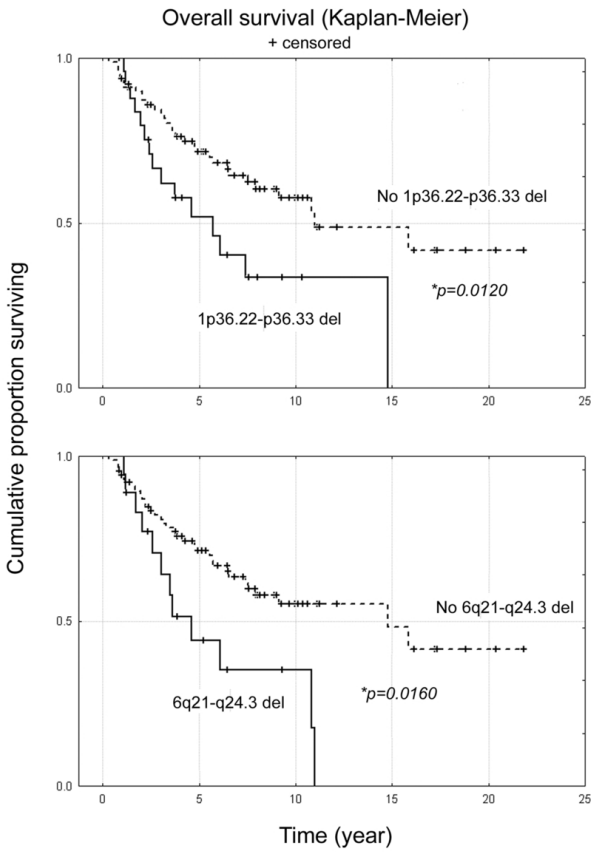


Figure 4.2: Kaplan-Meier survival of ID#1 [del(1)(p36.22-p36.33)] and #20 [del(6)(q21-q24.3)]. Log-rank test was performed to assess significance ($p \leq 0.05$).

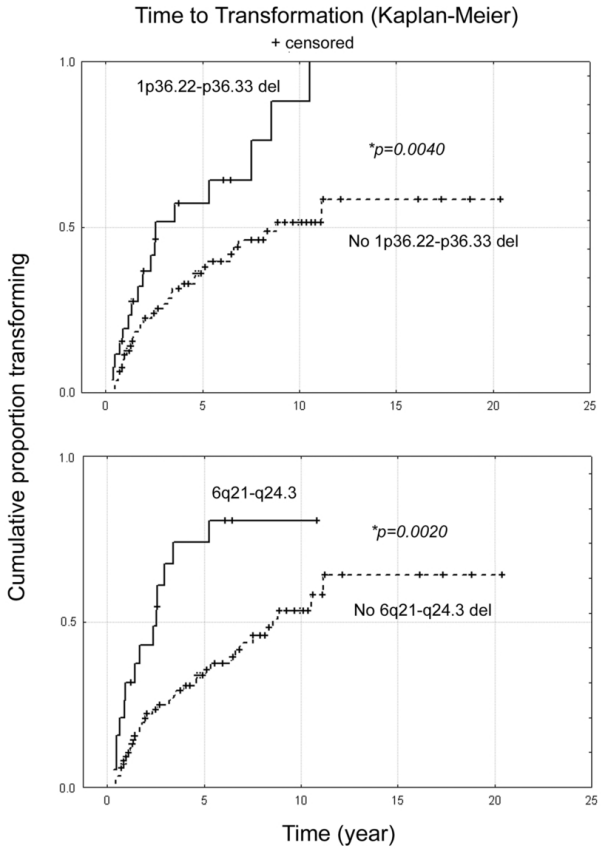


Figure 4.3: Kaplan-Meier time to transformation of ID#1 [del(1)(p36.22-p36.33)] and #20 [del(6)(q21-q24.3)]. Log-rank test was performed to assess significance ($p \leq 0.05$).

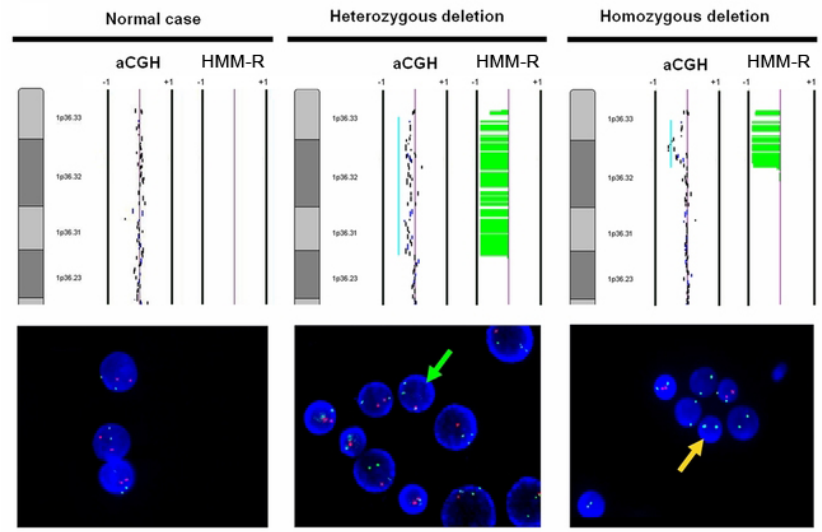


Figure 4.4: FISH validation of the 1p36.22-p36.33 region (ID# 1) which presented significant correlation with clinical outcome. aCGH (upper panel) and FISH (lower panel) demonstrating a case without deletion at 1p36.3 (normal), a case with heterozygous deletion at 1p36.3, and a case with homozygous deletion at 1p36.3. The 1p36.32 probe was labeled red while the control probe at 1q32 was labeled green. The upper panel for each case shows the HMM-R predictions.

a deletion, one with a heterozygous deletion and one with a homozygous deletion. As the region 6q21-q24.3 was too large (over 40 Mb) for case-specific FISH validation, we further narrowed this region by seeking areas of overlapping deletions affecting >15% of cases. Figure 4.5 shows the refinement of a broadly deleted region of the 6q arm (at the 10% cutoff level) to four small discrete regions of deletion (at the 15% cutoff level). The area that correlated significantly with survival and transformation risk corresponded to the single peak in band 6q23.3. The size of this peak was less than 300 kb and spanned from 93,765,93 to 94,111,251 bp (NCBI build 36.1). We used BAC clone RP11-703G08 for this region and RP11-516E15 from 6p12.3 as control for FISH analysis of 10 selected cases. Two of 10 cases were determined by intersection analysis to have no alteration while eight

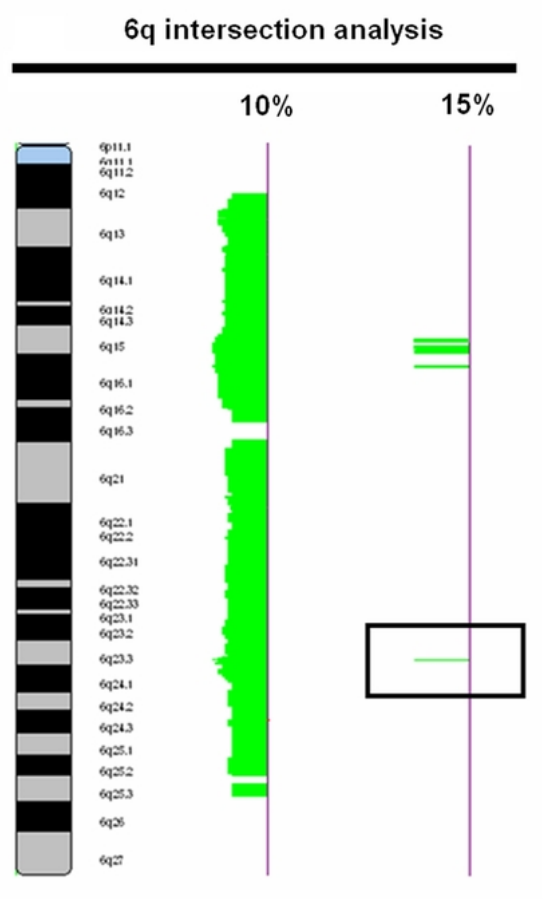


Figure 4.5: A composite array CGH ideogram profile of 6q alterations affecting $\geq$10% and $\geq$15% of FL cases was generated by intersection analysis. The alteration indicated by the black box in the 15% ideogram corresponds to the 6q23.3 region targeted for FISH validation.

cases had deletions by array CGH (as small as 810 kb) overlapping at 6q23.3. The concordance rate between FISH and intersection analysis was 10 of 10 cases. Figure 4.6 shows the array CGH ideogram data and FISH results for two representative cases, one with homozygous deletion and the other showing homozygous deletion at 6q23.3 with proximal and distal heterozygous deletion.

4.4.6 Correlation of array CGH findings with cytogenetic data

To illustrate the sensitivity of the array CGH platform compared to karyotype analysis, we examined the extent of correlation between array CGH and cytogenetic data in the 1p and 6q regions. Of the 27 cases with deletion of 1p36 detected by array CGH, 17 had karyotype data and of these only 7 (41.2%) showed an evident deletion or unbalanced translocation. Similarly, of the 22 cases with deletion of 6q detected by array CGH, 14 had karyotype data and 9 (64.3%) showed either whole chromosome loss, iso(6p) or deletion of 6q, whereas five cases showed normal 6q morphology.

4.4.7 Identification of high-level amplicons

In an attempt to identify high-level amplification in our cohort, we performed a simple computational thresholding approach where an amplicon was defined as 1) one that consisted of 3 or more contiguous BACs, 2) the log ratio of a BAC in an amplicon was at least 4 standard deviations above the mean log ratio of the sample, and 3) the frequency with which an amplicon occurred was at least 5% in order to minimize random aberrations in an individual case due to genomic instability, we found a high-level amplicon in 18q12.2 recurrent in 6.6% of cases (Table 4.4. By visual annotation using $>$1 log2 ratio shift as the definition of high-level amplification, 11 instances were found in four cases (Table 4.4).

4.4.8 Identification of secondary pathways

Based on a computational analysis of karyotype data, it was reported that dup(1q), del(6q), dup(7), and der(18) may constitute four distinct events arising secondary to t(14;18) in the early development of FL [102]. Using a clone-based approach (high resolution array CGH) and the application of a robust computational analysis,

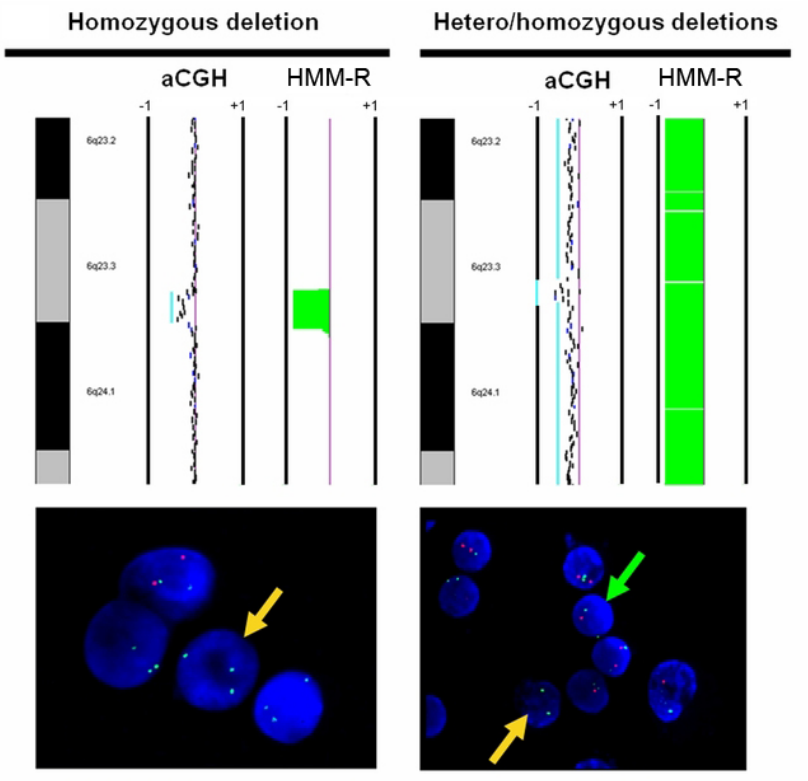


Figure 4.6: FISH validation of the 6q23.3 region (ID# 20) which presented significant correlation with clinical outcome. aCGH (upper panel) and FISH (lower panel) demonstrating a case with homozygous deletion and a case with homozygous deletion at 6q23.3 and proximal and distal heterozygous deletion. The 1p36.3 and 6q23.3 probes were labeled red while the control probe at 1q32.3 and 6p12.3 were labeled green. Green arrows in FISH indicate the presence of heterozygous deletion while yellow arrows indicate the presence of homozygous deletion. For array CGH, each dot represents a BAC clone and the light blue lines represent visual calling of aberrations. Loss is indicated by a shift to the left of center and gain by a shift to the right of center. Vertical lines are -1 and +1 scale bars of log2 ratios. In the upper panels, HMM-R predictions of the deletions are shown.

Region	DNA coordinates	Genes of interest	#
18q12.2	33040900-33600413	BRUNOL4	7
1p11.2-p12	119415209-120497765	NOTCH2	2
1q21.1-q22	143222462-153482569	PIAS3, BCL9, MCL1	
		IL6R, ADAM15, TNFAIP8L2	1
		mir-554, mir-190b, mir-92b	
1q23.2-q23.3	158419998-161795584	PEX19, DDR2, UHMK1	1
		mir-556	
1q23.3	159989344-161290912	DDR2, UHMK1, mir-556	1
12p11.21-p12.3	18053500-31175376	RERGL, PIK3C2G, KRAS	
		RASSF8, SSPN, mir-920	1
12q13.3-q21.1	56316396-72147648	OS9, TSPAN31, CDK4	
		RASSF3, TBK1, WIF1	
		DYRK2, IL22, MDM2	1
		RAP1B, YEATS4, FRS2	
		RAB21, mir-548c	
18q21.1	42698818-42831630		1
Xp11.4	37696318-40787875		1
Xp11.3	42402004-42902505		1
Xp11.1	57834572-58333582		1

Table 4.4: Detailed information on high-level amplicons in the 106 FL cohort (Data based on NCBI build 36.1)

we attempted to determine if similar pathway definitions could be obtained in our cohort. We first extracted the 4,912 BAC clones from the 71 regions of alterations and applied the k-medoids algorithm to the 106 cases to find clusters based on Hamming distance. Figure 4.7 presents a heat map where green signals indicate losses and red signals indicate gains. Of the 106 cases, 12 (11.3%) were clustered with dup(1q), 9 (8.5%) with dup(6p)/del(6q), 9 (8.5%) with dup(7), and 19 (17.9%) with dup(18). The remaining cases clustered into a group that exhibited no obvious pattern of alterations.

4.5 Discussion

The study reported here describes tiling path array CGH data for 106 cases of FL based exclusively on diagnostic biopsies. Seventy-one regions of alteration were identified to be recurrent in 10% or more cases. Of the 71 regions, 21 were shown to correlate significantly with inferior survival, however, only 16 were considered

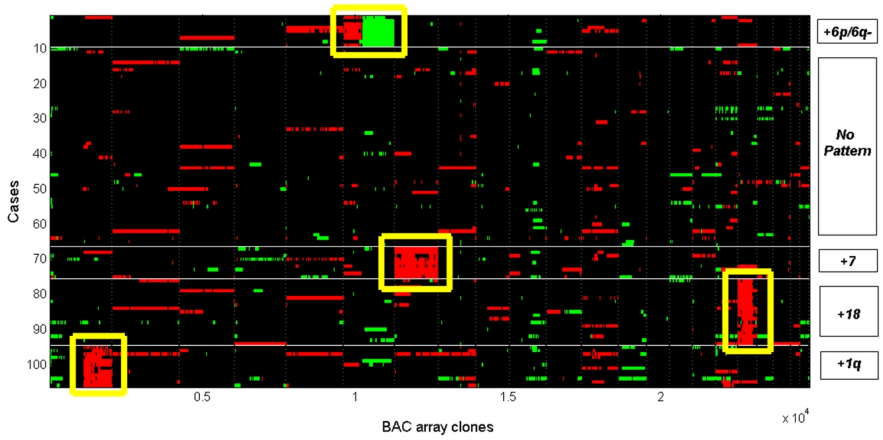


Figure 4.7: The K-medoids algorithm (where K=5) was applied to cluster both 106 cases (x-axis) and 4,912 BAC clones derived from 71 regions of aberrations. Four distinct clusters, +1q, +6p and 6q-, +7, and +18, were identified. The rest exhibited no obvious pattern of aberrations.

independent predictors from the IPI in a Cox multivariate model. Some of these regions, including deletion at 1p36.22-p36.33 (ID# 1), deletion at 6q21-q24.3 (ID# 20), gains at 17q11.2-q12 (ID# 55), 17q12-q21.2 (ID# 56), 17q21.31-q22 (ID# 57), and 17q24.2-q25.1 (ID# 60), provide a refinement of regions that were previously shown to be prognostic factors in overall survival [82, 112]. Our study also demonstrates that 12 of 71 regions were predictors of transformation risk in univariate analysis including: deletion of 1p36.22-p36.33 (ID# 1), gains of nearly the entire p arm of chromosome 6 (from p21.1 to p25.3, ID# 11-18), deletion of 6q21-q24.3 (ID# 20), gain of 7q32.3-q33 (ID# 38), and gain of 12q13.13-q13.2 (ID# 51). Other groups have reported CGH gains on both 6p and 7p to correlate with transformation from FL to DLBCL [93, 97], however, the 1p36 region has not been previously correlated with transformation.

While our investigation shows that two regions, deletions at 1p36.22-p36.33 and 6q21-q24.3, correlated with both inferior survival and higher transformation risk and were independent IPI prognostic predictors, many aberrant regions showed no positive correlation between these two clinical parameters. For instance, gains

of 6p were associated with higher transformation risk but had no effect on survival rate; likewise, gains of 5p and 17q that were associated with poor survival did not correlate with higher rate of transformation. Three possible explanations could be offered for these findings: 1) transformation and overall survival were not tightly linked in all cases, 2) some cases of FL behaved very aggressively, but did not show histological or clinical features used to define transformation in this study or, 3) the median follow-up was too short to appreciate an obvious relationship between transformation and overall survival in this cohort. This might explain why we did not observe a higher percentage of patients (42%) who had died during the observation period.

The most frequently altered region identified in this study was deletion of chromosome band 1p36.22-p36.33 (11 Mb in size) that occurred in 25.5% of cases. From the perspective of karyotype analysis, the faint subbands in 1p34 through 1p36 render cytogenetic analysis of 1p36 difficult and even relatively large deletions of this region may be overlooked, resulting in underreporting of deletions that could negatively affect previous correlations with prognosis. Of interest, a recent study by Ross *et al* found that 50% of 58 low-grade FL showed copy-neutral loss of heterozygosity (LOH) (also called acquired uniparental disomy (aUPD)) at 1p36, demonstrating that other mechanisms of gene inactivation may be involved at this and other sites [99]. Both deletions and copy-neutral LOH at the terminal portion of 1p have been implicated in other cancer types, including neuroblastoma, melanoma, germ cell tumors, lung cancer, and epithelial ovarian cancers, suggesting the presence of a tumor suppressor in the region [113]. A number of candidate genes reside in this region, including the cell cycle protein CDC2L1, the tumor necrosis factor (TNF) related receptor proteins such as TNFRSF9/14/18/25, the zinc finger transcription factor PRDM16, and the apoptotic factor DFFB. In related studies based on karyotype data we have observed that deletion of 1p36 occurs more frequently in transformed FL than in diagnostic cases (40% versus 24%, one-tailed p=0.0282) and that deletion of this region is seen in 50% of high-grade transformations with associated MYC translocations (manuscript in preparation). This latter association may in part explain the strong correlation of 1p36 with inferior survival and transformation observed in this study, as our cases may be enriched for patients that had experienced these events. It suggests that deletion of

1p36 may predispose patients to subsequent transformations with high proliferation rate, as described by Davies *et al* and Lossos *et al*, rather than lower proliferative transformations [114, 115].

Deletion of 6q is detected frequently in acute lymphoblastic leukemia, chronic lymphocytic leukemia, multiple myeloma, DLBCL and FL [108, 116]. Different studies have shown various regions to be involved, however, only one of these has focused specifically on FL in which a 2.3 Mb region of deletion was identified at 6q16.3 in 15% FL cases [108]. Our data show that nearly the entire 6q arm was involved in the majority of cases, though at the 15% cutoff level, only four very small regions of loss were defined. None of these regions overlap with that of the Henderson *et al* study [108]. The 500 kb deletion in 6q15 contains the CASP8 associated protein 2 involved in Fas-mediated apoptosis. Of particular interest is deletion of the 6q23.3 band that has been reported in 30-38% of ocular MALT lymphomas and FL [99, 116, 117]. Our array CGH data indicate that the 150 kb deletion at 6q23.3 affecting more than 15% of FL cases coincided with the TNFAIP3 (TNF- induced protein 3) gene. Deletion of this region was validated by FISH, suggesting that TNFAIP3 may be critical in FL development and/or progression. Furthermore, TNFAIP3 was implicated in a recent study based on correlation between genomic loss and gene expression, while it was unclear in another study as to whether TNFAIP3 or PERP (TP53 apoptosis effector) was implicated since sequencing analysis failed to uncover any mutations [99]. Deletion of TNFAIP3 can constitutively activate the NF-kB signaling pathway as it is a zinc finger protein inhibitor of NF-kB [118]. However, since deregulation of NF-kB appears to be uncommon in FL,46 other functions of TNFAIP3 may be responsible, or PERP may be involved since it lies only 200kb from TNFAIP3. Current evidence indicates that PERP induces TP53-mediated apoptosis and its deletion could lead to the promotion of tumor growth [119].

Although our cohort was partly enriched for diagnostic cases where a specimen was also available from a later transformation event, this selection did not significantly alter the expected median overall survival time (8-10 years versus our observed 10.83 years after diagnosis) and the expected median time to transformation (7 years vs our reported 6.61 years) for this disease. Based on the 3% annual transformation rate observed in FL [77], we would expect 22% of cases to have

transformed, whereas 50% of the study patients had transformed to DLBCL. This enrichment allowed us to detect genetic changes that are associated with transformation that would have been missed if too few patients in our study did not have that event. Given the high correlation between transformation and death, confirmation of the full clinical impact of cytogenetic alterations detected in this study should be validated on an unselected series of FL patients.

To exclude the possibility of random aberrant events, we also validated the copy number profile generated from the 106 FL diagnostic cases against an independent cohort of 37 FL (with similar age, grade and stage characteristics), 30 cases of mantle cell lymphoma (MCL), and 30 normal specimens. Our results indicated that 40 of the 71 (56.4%) aberrant regions, such as 1p36.22-p36.33 (ID# 1), 1q42.13-q42.3 (ID# 7) and 18p11.21-p11.32 (ID# 65), were unique to FL (Kruskal-Wallis test, $p<0.05$; data not shown) while others, such as 6p23-p24.3 (ID# 12) and 6q21-q24.3 (ID# 20), were shared by both FL and MCL. Overlapping regions between the normal controls, FL and MCL were also evident in regions such as 8q21.2 (ID# 43) and 5p15.33 (ID# 10) which represented copy number polymorphisms identified by this platform.

A number of CNVs as small as 80 to 200 kb can be detected by the array CGH platform [120] and may be evident in the global profile as discrete regions showing both duplication and deletion. Some CNVs may not exhibit this pattern and have been shown to be as large as a few megabases in size [121]. Since many CNV breakpoints cannot be precisely defined, and most importantly, that 58% of CNVs overlap with known RefSeq genes [121], we have elected not to filter these CNVs using any stringent criteria. A full understanding of the significance of the CNVs will require additional information on population-based frequencies, observed frequencies in specific types of lymphoma and the possible functional consequences exerted through associated genes, SNPs or other mechanisms.

Using only the percentage of alterations as a predictive measure of clinical outcome, we found that significant correlation with overall survival and transformation risk was present only if a criterion of 5% or more alterations (rather than 10%) was applied to dichotomize the cases. An explanation for this observation may be that as the criterion of percentage becomes too extreme (as in the example of 10% or more alterations), fewer and fewer cases will constitute one of the groups, thereby

significantly affecting statistical comparisons. Nevertheless, our findings were in general agreement with the commonly held notion that with increasing number of aberrations in the genome, prognosis is negatively affected [100].

This study has also shed further light on other recurrently altered genomic regions in FL. Potential genes that have been implicated in lymphomas are presented in Column Q of Table 4.3. For instance, the cytochrome gene CYP51 on 7q has been found overexpressed in lymphomas.16 In 8q, four candidate genes have been suggested: a potassium channel protein KCNK9, NF-KB-activating protein NIBP, the protein tyrosine kinase PTK2/FAK, and the protein tyrosine phosphatise PTP4A3.22 The PTEN phosphatase tumor suppressor at 10q23.1-q25.1 has been of constant interest [122]. MDM2 at 12q13.2-q21.1 has been reported to have altered expression that may negatively affect the stability of p53.51 Correlation between gene expression and genomic changes provided evidence that the Interleukin-3 zinc finger transcription factor ZNF161 at 17q23.2 may be involved [112]. BCL2 may be overexpressed as a result of extra copies of chromosome 18, especially in DLBCL [123], however, the bands 18q11-18q21.33 proximal to the BCL2 locus are also consistently over-represented in FL, implicating a gene proximal to the t(14;18) breakpoint in this amplification [96, 99].

Our search for high-level amplicons that occurred in at least 5% of cases led to the localization of a region in 18q22, which contains the BRUNOL4 gene. This gene belongs to a family of RNA-binding proteins involved in multiple aspects of RNA processing [124]. Its function in hematopoietic cells, however, has not been studied. Eleven other amplicons were found that occurred in less than 5% of cases. Viardot *et al* showed that among eight amplicons found in their 124 patients, their regions in 1q23-q25 and 12q13 overlapped with ours [100] as did the bands 8q24 and 12q13-q14 in Bentz *et al* study [91].

In an attempt to dissect the sequence of cytogenetic events occurring in the clonal evolution of FL, Hoglund *et al* conducted one of the first studies using published karyotype data to reconstruct the common pathways of clonal evolution secondary to the t(14;18). 26 Using principal component analysis to reduce data complexity for multivariate correlations, four major events consisting of dup(1q), dup(7), del(6q), and der(18) were identified to arise independently after the t(14;18). We attempted to utilize a clone-based approach to cluster both the 106 diagnostic

patients and 4,912 BAC clones extracted from the 71 regions of alteration. Using different methods and a mostly non-overlapping FL cohort, we have replicated the findings of the previous karyotype-based study. These data, in conjunction with previous findings, suggest that the early events of clonal progression in FL may evolve along a number of distinct pathways. These events may represent alternative critical steps following the primary event of BCL2 deregulation that are essential to the promotion of early clonal expansion, leading eventually to clinical manifestation of disease and transformation to more aggressive histologies. It is interesting to note that altogether 46.2% of our cases were represented in one of the four clustered groups, dup(1q), dup(6p)/del(6q), dup(7), and dup(18q), while the rest showed alterations that could not be explained by this approach. These may be cases where other types of biological mechanisms, such as copy-neutral LOH and/or methylation of genes of critical importance may be operative.

In conclusion, our data have confirmed and refined regions of aberrations found in previous findings and provided further insight into the distinct molecular pathways related to FL development using the clone-based cluster analysis. Most importantly, our study has identified deletion of 1p36 and 6q23 as significant prognostic indicators of clinical outcome. These correlations have been strengthened by the ability of high resolution analysis to detect submicroscopic deletions not previously detectable using other methods. The clinical relevance of these genetic alterations and their impact on disease progression will require additional studies of large patient cohorts, ideally managed with uniform therapy and lacking a selection bias.

Chapter 5

Model based clustering of aCGH data with simultaneous inference of calls, clusters and profiles

5.1 Summary

In this chapter, we investigate the problem of inferring molecular subtypes from aCGH data (Research goal C) [1]. We introduce a novel statistical framework, HMM-MIX that is capable of clustering patients in a cohort into distinct groups where patients assigned to a group share CNAs, defined by a group-specific profile. In an application of the model to clinical data derived from a cohort of 106 follicular lymphoma patients, HMM-Mix revealed subtypes that have previously undescribed prognostic significance. Moreover, in a cohort of 92 diffuse large B-cell lymphoma (DLBCL) patients, a novel subtype was discovered by HMM-Mix. In addition, we show that our model is significantly more accurate than simpler baseline models, including a published method tailored for aCGH, in a simulation study where subgroup assignments were known. The chapter is organised as follows: in Section 5.2, we introduce the biological and clinical motivation behind searching

[1] A version of this chapter has been submitted for publication: SP Shah, K-J Cheung, N Johnson, RD Gascoyne, DE Horsman, RT Ng and KP Murphy. Model based clustering of array CGH data. Proc Natl Acad Sci U S A. *In press.*

for molecular subtypes. In Section 5.3 we describe the HMM-Mix model, the simpler baselines and the data sets. Section 5.4 shows the results on the clinical data and the simulation study. In Section 5.5, we offer suggestions for future directions of this work.

5.2 Introduction

A critical complicating factor in the analysis of recurrent CNAs arises from the frequent observation that patient cohorts under study for a particular type of cancer exhibit heterogeneity in their molecular profiles. This has been demonstrated in breast [125], ovarian [126], and prostate cancers, as well as lymphomas [7, 102] among others, suggesting that the patients should be stratified into molecular subtypes where the patients within a group share a common group specific driver CNA profiles. This concept has been successfully applied many times over using gene expression data [125, 127], however it has been relatively under-studied in aCGH data.

Considering a cohort of patients as a composite of a fixed set of molecular subtypes has distinct advantages when determining recurrent CNAs. In the presence of large amounts of molecular heterogeneity, important driver alterations specific to a subgroup may be obscured and rendered indistinguishable from passenger alterations (or biological noise) in the data. By grouping or clustering the patients, within-group patterns are more likely to emerge, providing two advantages: groups of patients can be assessed for distinct clinical outcomes, and recurrent CNAs that might otherwise go undetected can be revealed. Subgrouping patients also has the potential of revealing CNAs that co-occur within a subtype and CNAs that are mutually exclusive between subtypes.

Importantly, it has been observed that definable molecular subtypes often correlate with clinical outcomes and in fact can, once identified, be considered as distinct disease entities [13] with different prognoses and/or response to therapy. Recent discovery of clinically relevant molecular subtypes by aCGH [8, 54] suggest that the inventory of CNA-derived molecular subtypes in cancer is not complete. Large scale projects such as the Cancer Genome Atlas Project [128] and the International Cancer Genome Consortium (ICGC: http://www.icgc.org) are now generating ge-

nomic array data sets from tumours from hundreds of patients for specific cancer types, thereby providing excellent potential for the discovery of new CNA-derived subtypes. In order to take full advantage of these data, robust and accurate computational algorithms for discovering molecular subgroups must be developed to keep pace with the data generation.

To address this goal, we introduce a novel statistical framework, *HMM-mix*, using model based clustering for inference of molecular subtypes from aCGH data. A key advantage of HMM-Mix is that it carries out *joint* inference of three quantities of interest: the discrete probe-level copy number calls, the assignment of patients to groups, and the profiles that define each group. This distinguishes HMM-Mix from standard methods that necessarily use a phased approach where these quantities need to be inferred in disjoint, independent steps. In addition, our approach performs simultaneous clustering and feature selection, similar in concept to Law *et al* [129], but with specific adaptation to aCGH data and built upon statistical frameworks introduced in our previous work in Chapters 2 and 3. Other approaches can employ feature weighting, however this can only be performed once at runtime. We demonstrate how the joint inference of the quantities of interest and simultaneous feature selection confer a significant performance advantage over both partitioning and hierarchical clustering methods [130] in a simulation study. More importantly, we show how the HMM-mix reveals clinically relevant subgroups in data derived from a cohort of 106 follicular lymphoma (FL) patients, originally reported in Cheung *et al* [7], and reveals previously unreported patterns of alteration in a cohort of 92 diffuse large B-cell lymphoma (DLBCL) patients that is the subject of a forthcoming manuscript [131].

5.3 Methods

Clustering algorithms can be divided into three categories: partitioning, hierarchical, and methods based on mixture models [129]. In this section, we introduce our model-based HMM-Mix framework, describe two partitioning algorithms we implemented for aCGH, and a hierarchical clustering algorithm designed for aCGH that was previously described in [130]. The latter approaches provide benchmarks against which we evaluate HMM-Mix.

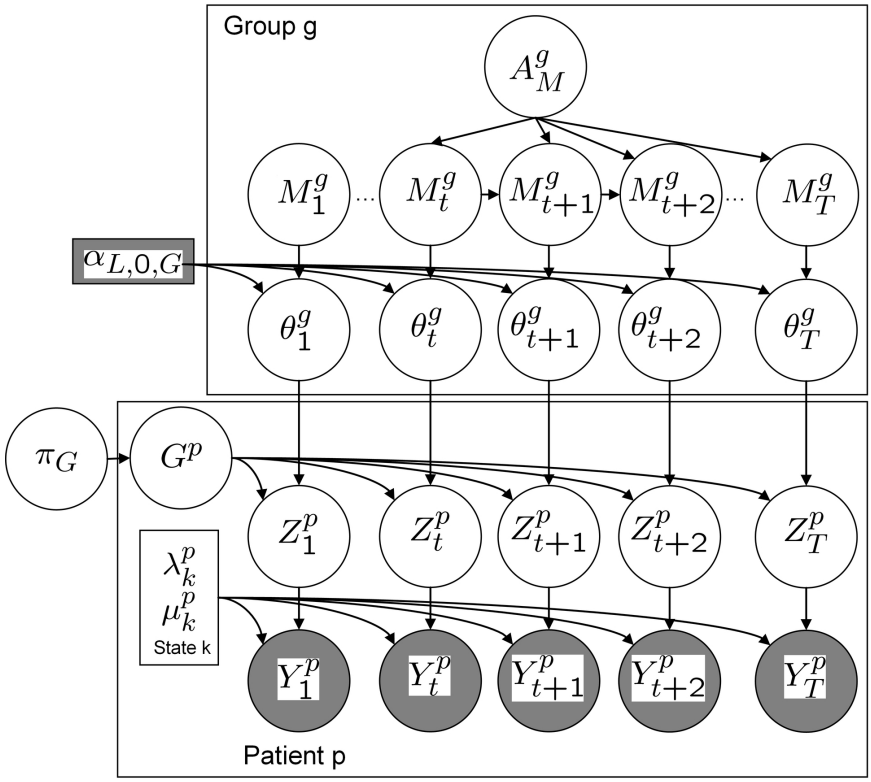


Figure 5.1: Proposed HMM-mix model for clustering aCGH data, represented as a directed graphical model [132]. Shaded nodes are observed/-fixed, unshaded nodes are hidden (unknown). The two boxes represent repetition over patients and groups. $Y_t^p \in \mathbb{R}$ is the observed aCGH logratio at probe t in patient p. $Z_t^p \in \{L,N,G\}$ is the discrete state, representing whether probe t is a loss, neutral or gain. Given $Z_t^p = k$, Y_t^p is assumed to be sampled from a class conditional Student-t distribution with parameters μ_k^p, λ_k^p and ν_k (not shown). $G^p \in \{1,\ldots,G\}$ is the group that patient p belongs to, which is sampled from a multinomial with parameter π_G. θ_t^g is the multinomial parameter over Z_t^p, which is sampled from a Dirichlet with parameter $\alpha_{M_t^g}$, where $M_t^g \in \{1,\ldots,C\}$ represents the state of the sparse profile for probe t in group g. A_M^g is the transition matrix for the profile model. Conditional probability distributions are shown in Table 5.1

5.3.1 The HMM-Mix model for clustering aCGH data

The principle idea in HMM-Mix is to extend the concept of explicitly modeling driver and passenger alterations described in Chapter 3 to the multi-group setting where we assume that the patient cohort is composed of distinct molecular subtypes (ie each patient belongs to one of $g \in (1,\ldots,G)$ groups). The goal is to simultaneously infer a profile that represents each group and discover the optimal stratification of the patients into the groups. The mixture model-based clustering concept is described in the general context in Raftery and Dean [133]. In our approach, the profile for each group is parameterized by a sparse HMM. Thus the data set as a whole is modeled as a mixture of G HMMs. We sketch the HMM-Mix model in Figure 5.1 as a graphical model. We represent the aCGH logratios as $Y_t^p \in \mathbb{R}$ for each probe $t \in (1,\ldots,T)$ in the array and for each patient $p \in (1,\ldots,P)$. Each probe maps to unique genomic coordinates and can therefore be positionally ordered along the chromosomes. $Y_{1:T}^{1:P}$ represents the full data matrix. For each datapoint we assume there is a discrete mapping from $Y_{1:T}^{1:P} \rightarrow Z_{1:T}^{1:P}$ where $Z_t^p \in k$ and k is a discrete copy number state $\in \{loss, neutral, gain\}$ as described in previous chapters [2]. The model assumes the observed data for each patient comes from a mixture of profiles, where each profile is modeled by a compound Dirichlet-Multinomial distribution represented by $\theta_{1:T}^g$ and $M_{1:T}^g$ (see below for details) and the group that each patient belongs to by a single multinomial $G^p = g$. We assume that a small set of probes represent the group-specific driver alterations and the rest of the probes can be explained by patient-specific passenger alterations. We now explain in detail how this is modeled.

Dirichlet Mixture Prior for sparse profiles

We assume the passenger probes are generated from a 'background' distribution, while the driver probes are generated from a group-specific 'foreground' distribution. Our framework is therefore designed to explicitly separate probes from the background from probes from the foreground. Moreover, we want to identify probes for which the patients in the group are highly biased towards gain or loss.

[2] We note that the model could easily be extended to accommodate more states, however for simplicity, we restrict ourselves to three states in this chapter.

Defining the set of probes that make up the foreground is therefore analogous to simultaneous feature selection and clustering as described by [129] for gene expression.

To accomplish this, we assume that θ_t^g is generated from a Dirichlet mixture prior $\alpha_{1:C}$. Dirichlet mixtures have been successfully used in related contexts, for example to model HMM-based profiles for protein families [134]. Here, we use an indicator variable $M_t^g = c$ to index the appropriate component of the mixture to be used to estimate θ_t^g, where $c \in \{L, 0, G\}$ and we fix the number of components, $C = 3$. Therefore, $p(\theta_t^g | M_t^g = c, \alpha_c)$ gives us an estimate of how likely the c^{th} component of the mixture was to have produced θ_t^g. This estimate is an evaluation of the Dirichlet density function at θ_t^g, with parameters $\alpha_c^{1:K}$. We set $\alpha_L = [a_L, 1, 1], \alpha_0 = [1, a_0, 1], \alpha_G = [1, 1, a_G]$, where $a_{L,0,G}$ represents how 'peaked' the Dirichlet distribution is for each of its components, and $a_L, a_G >> a_0 \geq 1$. α_0 can be considered a weak prior for the background distribution for passenger probes, where the patients in the group are heterogeneous or neutral in these locations. In contrast, α_L and α_G are priors for *loss* and *gain* probes respectively, where the strength of how uniform the columns are, depends on a. A key point is that we pool the data for all probes expected to be in the background in order to estimate the background distribution θ_0^g, while the data for the foreground distributions remains location specific. Having inferred $M_{1:T}^{1:G}$, we can then look at probes $\hat{t}$ for which $M_t^g \neq 0$ to obtain the probes of interest, and report these probes as a 'definition' of the profile for group g.

Note that θ has similar characteristics to the value ϕ output by AF discussed in Chapter 3. However a major difference is that θ is generative, whereas ϕ is not. Therefore, θ influences the Z calls which is an important feature in our model, as we will see.

HMM-Mix model specification

The model is a mixture of HMMs [135] with standard transition matrices, A^g, each a $C - by - C$ matrix that represents:

$$A^g(i, j) = p(M_t^g = j | M_{t-1}^g = i) \tag{5.1}$$

The Markovian dynamics on $M^g_{1:T}$ model the fact that the recurrent losses and gains will span sets of contiguous probes (see Chapter 3 for further details).

The emission model for HMM-Mix is a function of the observed data $Y^{1:P}_{1:T}$ and the current cluster responsibilities G^p. We use the conditional probability distributions of the model in order to compute standard quantities that can be used as input to the Forwards-Backwards framework for HMMs. Thus, the emission matrix $B^p_t(c,g)$ for patient p given its cluster assignment g and Dirichlet component $c \in \{L,0,G\}$ is defined as follows:

$$B^g_{p,t}(c) = p(Y^p_t | M^g_t = c, G^p = g, \alpha_c) \tag{5.2}$$

where

$$p(Y^p_t | M^g_t = c, G^p = g, \alpha_c) = \sum_k p(Y^p_t | Z^p_t = k) p(Z^p_t = k | M^g_t = c, G^p = g, \alpha_c) \tag{5.3}$$

$$p(Y^p_t | Z^p_t = k) = St(Y^p_t | \mu^p_k, \lambda^p_k, \nu_k) \tag{5.4}$$

where $\mu^p_{1:K}$ and $\lambda^p_{1:K}$ and $\nu_{1:K}$ (fixed) are patient specific parameters (mean, precision and degrees of freedom) of class conditional density Student-t distributions assumed to emit Y^p_t and

$$p(Z^p_t | M^g_t = c, G^p = g, \alpha_c) = \frac{\Gamma(\sum_k \alpha^k_c)}{\Gamma(1 + \sum_k \alpha^k_c)} \prod_{k=1}^{K} \frac{\Gamma(I(Z^p_t = k) + \alpha^k_c)}{\Gamma(\alpha^k_c)} \tag{5.5}$$

Note that we have integrated out θ as defined in Brown *et al* [134] where

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \tag{5.6}$$

Given the cluster assignments for all the data, the emission matrix for a given group g therefore becomes:

$$B^g_t(c) = \prod_p p(Y^p_t | M^g_t = c, G^p = g, \alpha_c)^{I(G^p = g)} \tag{5.7}$$

Given $B^g_{1:T}$ and A^g, the sequence $M^g_{1:T}$ can be computed using the standard forwards-backwards algorithm [22] (on a per chromosome basis) which returns the posterior marginal probabilities

$$\gamma^g_t(c) = p(M^g_t = c|Y^{1:P}_{1:T}, G^{1:P}, A^g) \tag{5.8}$$

of probe t for group g of Dirichlet component c having had generated the observed data. In order to infer the most likely sequence of Dirichlet components, we use the Viterbi algorithm.

Referring to the graphical model shown in Figure 5.1, we can read off the conditional independence properties of the variables in the Markov blanket (see [22]) of Z^p_t in order to compute the appropriate marginal probability:

$$p(Z^p_t = k|Y^p_t, G^p = g, \theta^g_t) = p(Z^p_t = k|\theta^g_t)p(Y^p_t|\mu^p_k, \lambda^p_k) \tag{5.9}$$

These quantities can in turn be used to update μ and λ as described in Archambeau [136].

We denote the weights of the mixture by π_G: a multinomial distribution where the components π^g_G model the proportion of the cohort assigned to group g, $0 \leq \pi^g_G \leq 1$ and $\sum \pi^g_G = 1$. Given π_G, we update the cluster assignments as follows:

$$p(G^p = g|Y^p_{1:T}, M^g_{1:T}, \alpha, \pi_G) = \frac{\pi^g_G \prod_t B^g_{p,t}(M^g_t)}{\sum_h \pi^h_G \prod_t B^h_{p,t}(M^h_t)} \tag{5.10}$$

It is also possible to update the hyperparameters of the Dirichlet prior α, depending on the data. This may be useful if we wish to have group-specific background distributions parameterized by α^g_0. Given $(M^g_{1:T}, Z^{1:P}_{1:T}, G^{1:P})$ we can calculate the counts $\vec{Z^g_c}$ where $G^p = g$ and $M^g_t = 0$. We can then estimate new parameters $\alpha^g_0|\vec{Z^g_c}$ using the iterative Newton-Rhapson method described in Minka [137].

Prior specification

All relevant parameters are assumed to be distributed according to their standard conjugate priors and are accordingly updated using *maximum a posteriori* (MAP) updating during inference. μ, λ are distributed according to a normal-gamma pa-

$p(M_t^g = j \mid M_{t-1}^g = i)$	$\sim$	$A^g(i,j)$
$p(\theta_t^g \mid M_t^g = c, \alpha)$	$\sim$	$Dir(\theta_t^g \mid \alpha_c^{1:K})$
$p(Z_t^p = k \mid G^p = g, \theta)$	$\sim$	$Mult(Z_t^p = k \mid \theta^g)$
$p(G^p = g \mid \pi_G)$	$\sim$	$Mult(G^p = g \mid \pi_G)$
$p(Y_t^p \mid Z_t^p = k, \mu_k^p, \lambda_k^p)$	$\sim$	$St(Y_t^p \mid \mu_k^p, \lambda_k^p, \nu_k)$
$p(\mu_k^p \mid m_k^p, \eta_k^p)$	$\sim$	$\mathcal{N}(\mu_k^p \mid m_k^p, \frac{\eta_k^p}{\lambda_k^p})$
$p(\lambda_k^p \mid S_k^p, \gamma_k^p)$	$\sim$	$Gam(\lambda_k^p \mid S_k^p, \gamma_k^p)$
$p(A^g(i,.) \mid \delta_A)$	$\sim$	$Dir(A^g(i,.) \mid \delta_A)$
$p(\pi_G \mid \delta_\pi)$	$\sim$	$Dir(\pi_G \mid \delta_\pi)$

Table 5.1: List of conditional probability distributions of HMM-Mix.

rameterized by (m, η) for μ and (S, γ) for λ and are set according to the description in Archambeault [136]. We assume the rows of A^g are distributed according to a Dirichlet prior, parameterized by δ_A, where we place emphasis on self-self transitions. Finally, we impose a flat Dirichlet prior, parameterize by δ_π, on π_G, assuming that all groups are equally likely *a priori*. All conditional probability distributions are given in Table 5.1.

Inference

We fit the model to the data using iterated conditional modes (ICM). Inference using EM is intractable in this model since the profile parameters are coupled. Moreover, an MCMC approach to infer full posterior distributions of the unknown quantities would be impractical due to the number of samples required. The quantities of interest: the cluster assignments, the profiles, the calls and the parameters are updated in sequence until convergence is reached. The runtime complexity of the algorithm is $O(T)$ but in practice is proportional to the number of groups G and patients P as well.

As in most model-based approaches, initializations of its parameters needs to be carefully considered. We initialize $\mu, \lambda, Z_{1:T}^{1:P}$ by fitting a modified version of the

Algorithm 4 Pseudocode for iterated conditional modes algorithm for HMM-Mix. Input: raw data $Y_{1:T}^{1:P}$, number of groups G and Dirichlet prior $\alpha_{loss,0,gain}$. Output: clusters $G^{1:P}$, calls $Z_{1:T}^{1:P}$ and profiles $M_{1:T}^{1:G}$. $K = 3$, the number of copy number states is fixed. N_g is the number of patients assigned to group g. Updating of the A_M^g transition matrices, π_M^g initial state distributions and μ, λ emission parameters are omitted for brevity. Supporting function listed in Algorithm 5

/* Initialise calls and emission model parameters */
for $p = 1, 2, \ldots, P$ **do**
    $m_{1:K}^p, \eta_{1:K}^p, S_{1:K}^p, \gamma_{1:K}^p$ = setHyperparameters($Y_{1:T}^p$)
    $Z_{1:T}^p, \mu_{1:K}^p, \lambda_{1:K}^p$ = HMM-R($Y_{1:T}^p, m_{1:K}^p, \eta_{1:K}^p, S_{1:P}^p, \gamma_{1:K}^p$)
end for
/* Initialise groups from calls */
$G^{1:P}$ = multipleRestartWKM($Z_{1:T}^{1:P}, G$) /* see Section 5.3.2 */
/* Initialise profiles from groups and calls */
for $g = 1, 2, \ldots, G$ **do**
    $\pi_G(g) = \frac{\delta_\pi(g)+N_g}{\sum_h \delta_\pi(h)+N_h}$
    $M_{1:T}^g$ = initialiseProfile($Z_{1:T}^{1:P}, G^{1:P}, g$)
end for
B = computeEmissionDensity($Y_{1:T}^{1:P}, M_{1:T}^{1:G}, G^{1:P}, \alpha_{1:C}$)
/* Begin ICM */
for $iter = 1, 2, \ldots, maxiter$ **do**
    /* Update the clusters */
    for $g = 1, \ldots, G$ **do**
        $G^p = argmax_g \frac{\pi_G^g \prod_t B_{p,t}^g(M_t^g)}{\sum_h \pi_G^h \prod_t B_{p,t}^h(M_t^h)}$
    end for
    /* Update the profiles */
    $M_{1:T}^g$ = Viterbi($B_{1:P,1:T}^g, A_M^g, \pi_M^g$)
    /* Update the calls */
    for $p = 1, \ldots, P$ **do**
        for $t = 1, \ldots, T$ **do**
            $Z_t^p = argmax_k \sum_g Mult(Z_t^p = k|\theta_t^g) St(Y_t^p|\mu_k^p, \lambda_k^p, \nu_k)^{I(G^p=g)}$
        end for
    end for
    /* Update μ, λ (omitted) and recompute emission density */
    B = computeEmissionDensity($Y_{1:T}^{1:P}, M_{1:T}^{1:G}, G^{1:P}, \alpha_{1:C}$)
end for

Algorithm 5 Supporting function for Algorithm 4

Function B = computeEmissionDensity($Y_{1:T}^{1:P}, M_{1:T}^{1:G}, G^{1:P}, \alpha_{1:C}$)
for $g = 1, \ldots, G$ **do**
  for $p = 1, \ldots, P$ **do**
    for $t = 1, \ldots, T$ **do**
      for $c = 1, \ldots, C$ **do**
        $B_{p,t}^{g}(c) = p(Y_t^p | M_t^g = c, G^p = g, \alpha_c)$ /* See equation 5.2 */
      end for
    end for
  end for
end for

single sample HMM (using a Student-t emission density rather than Gaussian) described in Shah *et al* [20] (see Chapter 2) to the data from each patient separately. The cluster groupings are initialized using WKM (see Section 5.3.2) with $Z_{1:T}^{1:P}$ as input. Figure 5.3.1 shows an example run of HMM-Mix initialised with WKM on synthetic data (described in Section 5.3.5). As is shown, in the figure, the initial clusters are a crude approximation to the truth. As the algorithm proceeds to convergence, the resulting clusters are far more accurate.

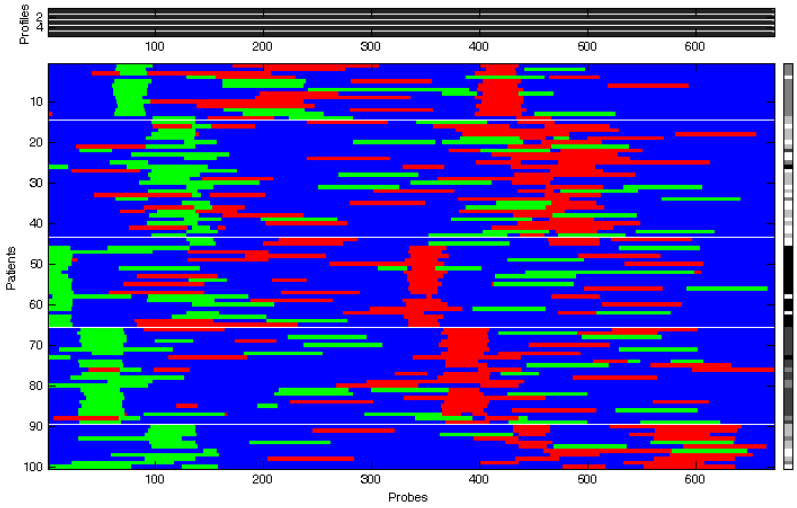
Given the initial cluster groupings and $Z_{1:T}^{1:P}$, $M_{1:T}^{g}$ is initialized by using the entropy, E_t^g for each probe for patients in group p:

$$E_t^g = -\sum_k f(Z_t = k, g) log(f(Z_t = k, g)) \tag{5.11}$$

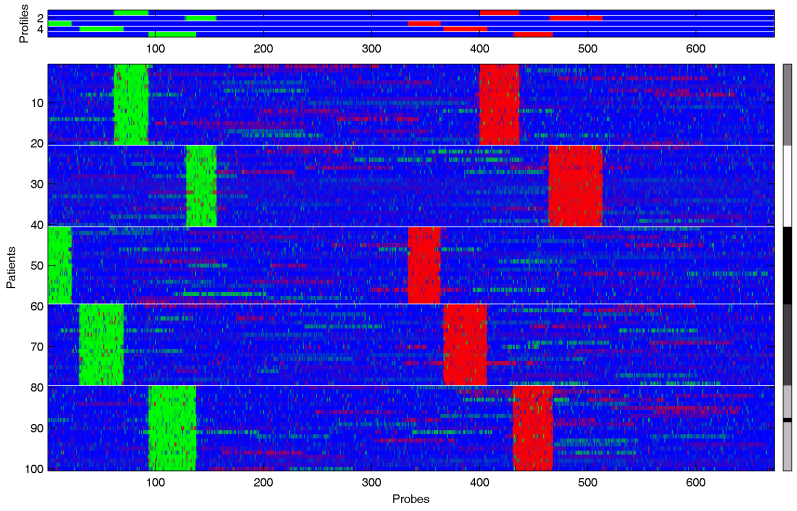
where $f(Z_t = k, g)$ is the normalized frequency of $Z_t = k$ for all patients in group g. Highly entropic probes are initialized to the background distribution, while positions relatively homogeneous for loss or gain are initialized accordingly.

5.3.2 Baseline algorithms

To compare HMM-Mix to partitioning based methods, we derived and implemented two K-medoids based algorithms. The first algorithm, KM, is based on a standard distance-based algorithm that outputs a clustering $G^p = g$ given the discrete call matrix $Z_{1:T}^{1:P}$ and the number of groups. G is chosen by the user, or it can be estimated heuristically from the data using the Silhouette coefficient [138] (see



WKM initialisation



HMM-Mix converged

Figure 5.2: WKM initialisation and convergence of HMM-Mix, showing how it dramatically improves the clustering in relation to the true class labels (shown right in each figure). The data is shown as a matrix of $Z_{1:T}^{1:P}$ with each row representing a patient and each column a probe. The colouring is representative of a call where $Z_p^t = loss$ is green, $Z_p^t = gain$ is red and neutral probes are blue. The horizontal white line separate the clusters. The ground truth classes are shown on the right as different grayscale shades. In a perfect clustering the grayscale shades would be arranged in uninterrupted blocks. In comparison to the WKM initialisation, the HMM-Mix converged estimate is far more accurate.

Section 5.3.4). We use Hamming distance to define the distance $d(i,j)$ between two patients i and j, as follows:

$$d(i,j) = \sum_{t=1}^{T} \delta(Z_t^i, Z_t^j) \tag{5.12}$$

where $\delta(Z_t^i, Z_t^j) = 0$ indicates that the discrete state at position t for patients i and j are equal. Otherwise $\delta(Z_t^i, Z_t^j)$ evaluates to 1. Based on an initial random assignment of the medoids to G distinct patients, the algorithm seeks to locally minimize the total distance of each data point to its assigned medoid. In an iterative framework, the algorithm searches for medoid assignments that reduce this total distance until a local minimum in the search space has been reached [3]. KM is prone to finding local minima and is sensitive to initializations, however the efficiency of the algorithm affords us the option to use a multiple restart framework. The two-step procedure of initializing the medoids randomly and running the algorithm to convergence is repeated 1000 times and the run producing the minimum total distance is kept.

Weighted K-medoids (WKM)

The KM algorithm described above treats all probes (features) equivalently when computing the distance function. Due to patient specific passenger CNAs that are likely not related to disease, we assume that only a small subset of features are important in determining the 'distance' between 2 patients with respect to the disease. We present a modified distance function that leverages the entropy of the features to compute the distance. The intuition is that probes spanning driver alterations that are different between subgroups will be more entropic than other probes. We show in Section 5.4 that this significantly improves performance. We calculate the entropy, H_t, of each probe as follows:

$$H_t = -\sum_{k} f(Z_t = k) log(f(Z_t = k)) \tag{5.13}$$

[3] Note that in general KM is far more efficient than the closely related k-means algorithm since the distance matrix $d(i,j)$ need only be computed once. In k-means, the means are updated to a new value at each iteration and the distance from each patient to each of the new means must be recalculated.

where $k \in \{loss, neutral, gain\}$ and $f(Z_t = k)$ is the (normalised) frequency probe t in state k taken over all patients where $\sum_{k=1}^{K} f(Z_t = k) = 1$. To minimize the effects of outliers from $E_{1:T}$, we apply a logistic transformation to obtain feature weights $W_{1:T}$ where:

$$W_t = \frac{1}{1 + e^{-\frac{H_t}{\alpha}}} \tag{5.14}$$

where we set $\alpha = 0.25$ is a smoothing constant chosen heuristically to avoid extreme weights[4] We modify the distance function between two patients i, j given in Equation 5.12 as follows:

$$d(i, j) = \sum_t W_t \, \delta(Z_t^i, Z_t^j) \tag{5.15}$$

Therefore the WKM algorithm is identical to the KM algorithm except that it uses Equation 5.15 in place of Equation 5.12 to compute distance.

Hierarchical clustering for aCGH

In recent work, van Wieringen *et al.* [130] introduce a system called "Weighted clustering of called array CGH data" (WECCA). This represents the first clustering approach to be tailored specifically to the aCGH data and is a specialized implementation of hierarchical agglomerative clustering. They define novel similarity (opposite of distance) and linkage metrics for hierarchical clustering that leverage the properties of the aCGH data. The authors also establish a weighted form of similarity, similar in spirit to the weighted-Hamming distance described above, although the weights are expected to be provided by the user, rather than empirically calculated.

5.3.3 Advantages of HMM-Mix over baseline methods

The HMM-Mix approach differs from KM, WKM and WECCA in three important ways: i) we use an adaptive feature selection approach where the features are selected simultaneously with the cluster assignments. WKM and WECCA both allow feature weighting, but this is only done once at runtime, preventing the fea-

[4] Without this smoothing transformation we found the dynamic range of the entropy measure was too great and did not lead to satisfactory results.

ture selection from being modified during the process of the clustering procedure; ii) HMM-Mix uses $Y_{1:T}^{1:P}$ as input. In previous work [21] (see also Chapter 3), we demonstrated that this confers an advantage over using $Z_{1:T}^{1:P}$ when inferring recurrent CNAs profiles since shared signals can be better elucidated by borrowing statistical strength present in the raw data. This has also been shown by Klijn *et al* [71]; iii) the HMM-Mix framework enables the updating of the discrete calls in the presence of the profiles by inferring the calls, the cluster assignments and the profiles jointly in a single inference routine. In contrast KM, WKM and WECCA all require discrete data as input, do not update the calls and do not produce profiles as output. For KM, WKM and WECCA, the process of inferring calls, cluster assignments and profiles is necessarily step-wise potentially leading to information loss progressing through the steps. Finally, HMM-Mix is a probabilistic model and thus, can be used without modification in a more complex setting.

5.3.4 Choosing the number of groups

A limitation in mixture model-based clustering is the requirement of specifying the number of groups *a priori*. In many contexts the investigator may not know how many groups to expect in the population. It may therefore be desirable to estimate the number of groups in an automated way. For distance based methods such as KM and WKM, the Silhouette coefficient computes a measure of clustering that considers both cohesion (how similar the points in a cluster are), and separation (how different the clusters are). For each data point, the Silhouette coefficient [139] is computed as:

$$s_i = (b_i - a_i)/max(a_i, b_i) \tag{5.16}$$

where a_i is the average distance to points in i's cluster and b_i is the minimum average distance to points in another cluster. Therefore, $-1 \leq s_i \leq 1$ where $s_i = 1$ is the optimal value (the case where $a_i = 0$). An overall measure of clustering is the average Silhouette coefficient:

$$S = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{5.17}$$

For two given clusterings where the number of groups was set to be $\hat{k}, k$, if $S_{\hat{k}} > S_k$ then that indicates $\hat{k}$ groups is better than k.

5.3.5 Data sets and evaluation protocol

Clinical data

We applied HMM-Mix to data derived from patient cohorts created to study i) follicular lymphoma (FL) (see Figure 5.3) and ii) diffuse large B-cell lymphoma (DLBCL) (see Figure 5.5). The FL data was previously reported in Cheung *et al* [7] (see also Chapter 4) and consists of 106 patients expected to fall into at least 4 genetic subtypes [102]. A characteristic of FL is that in some percentage of patients, the tumour undergoes a transformation to DLBCL, a more aggressive subtype of lymphoma where patients have poorer survival rates. Developing a prognostic CNA profile predictive of transformation is therefore of great interest for clinical management of FL. We measured the concordance of the predicted subgroups to this clinical data in an effort to discover prognostically relevant subgroups. The DLBCL data is the subject of a forthcoming manuscript [131] describing the aCGH findings in 92 patients with de novo DLBCL all treated uniformly with multi-agent chemotherapy (CHOP) and anti-CD20 monoclonal antibody rituximab. These largely represent consecutive cases from a population-based registry, but were enriched for primary treatment failures.

All clinical data were produced using the SMRT array platform [16] and contain approximately 27,000 probes per sample.

Simulated data

To test and compare performance of the various algorithms, where the true clustering was known, we generated simulated data. Based on real cell line data reported in DeLeeuw *et al* [19], we generated simulated data set in a manner similar to that described in Chapter 3. We performed 100 random draws (simulating patients) from the eight cell lines and extracted the 672 probes on chr 21 (chosen because it was reported to have relatively few alterations in the MCL cell lines). For each of the 100 patients, we shuffled the 672 clones and randomly assigned the pa-

tient to one of G groups. For each group, we preset coordinates of one recurrent gain and one recurrent loss. These group specific coordinates defined the profile for the group. The alterations were embedded into each patient's data at their group-specific coordinates with a small amount of randomly sampled noise added to the edges of the alterations to prevent perfectly overlapping recurrent alterations. Losses were logratio shifts of one standard deviation down and gains were shifts of one standard deviation up. Finally, for each patient, we randomly embedded alterations of length L at locations different than the group-specific alterations in order to simulate patient-specific 'passenger' alterations expected to be unrelated to the group profile. We created 10 replications with $G = 3, 5, 10$ and $L = 50, 75$ yielding 60 data sets.

Evaluation protocol

Given a data set for which the groupings are known, it is of interest to determine how well a given clustering algorithm reproduces the ground truth groupings. We refer to ground truth groupings as classes and predicted groupings (by any algorithm) as clusters. Consider a class matrix $C(i, j)$ where $C(i, j) = 1$ if datapoints i, j are in the same class and $C(i, j) = 0$ if they are in a different class. Similarly, consider a cluster matrix $P(i, j)$ where $P(i, j) = 1$ if i, j are predicted to be in the same cluster and $P(i, j) = 0$ if i, j are predicted to be in different clusters. To determine the clustering accuracy of the algorithms on the simulated data, we use the Jaccard coefficient. We compute the following quantities as described by Tan *et al* [139].

- f_{00}: the number of pairs of data points i, j for which both $C(i, j) = 0$ and $P(i, j) = 0$
- f_{01}: the number of pairs of data points i, j for which $C(i, j) = 0$ and $P(i, j) = 1$
- f_{10}: the number of pairs of data points i, j for which $C(i, j) = 1$ and $P(i, j) = 0$
- f_{11}: the number of pairs of data points i, j for which $C(i, j) = 1$ and $P(i, j) = 1$

The Jaccard coefficient is defined as:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \tag{5.18}$$

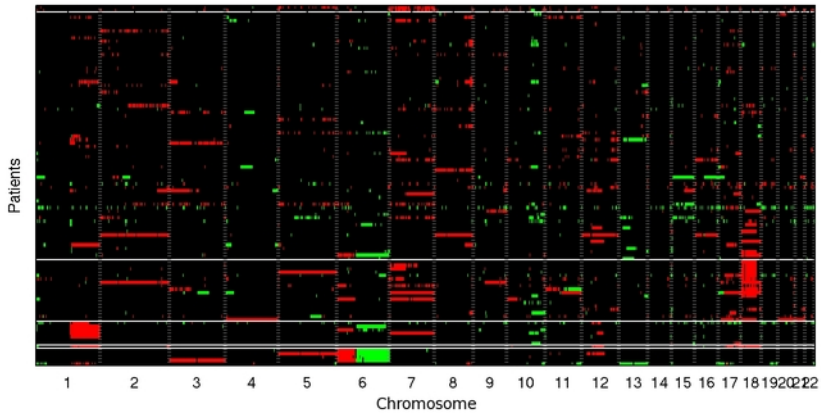
where $0 \leq J \leq 1$ and $J = 1$ indicates a perfect agreement of the clusters with classes.

5.4 Results

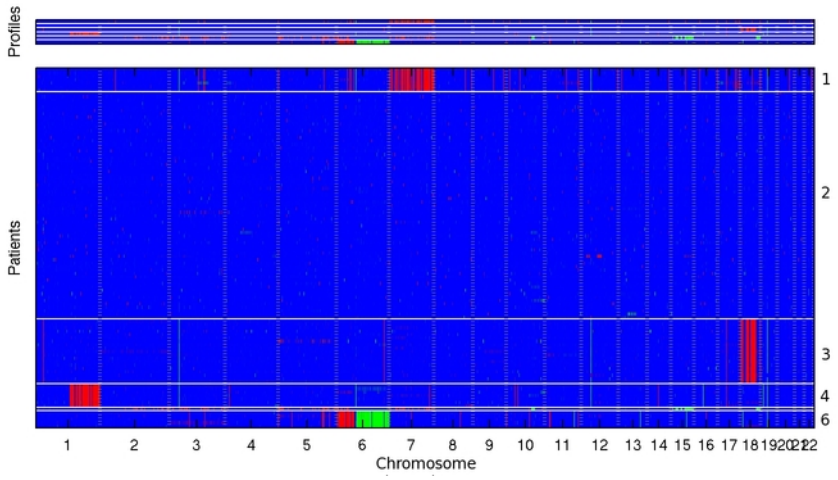
5.4.1 HMM-Mix discovers clinically relevant subgroups in FL data

We applied HMM-Mix to the FL cohort of 106 patients [7]. We initialized the model using WKM with 100 multiple restarts and we determined $G = 6$, the number of *a priori* groups using the maximum Silhouette coefficient over $G = (2, ..., 8)$. Figure 5.3 shows the WKM initializations and HMM-Mix converged clusters for the FL data. Figure 5.3(a) shows the initial $Z_{1:T}^{1:P}$ matrix where rows are patients and columns are probes. The rows are ordered according to their WKM cluster assignments. The green, red and black probes are predicted losses, gains and neutrals respectively. Figure 5.3C (bottom) shows the converged estimates of HMM-Mix where the rows have been ordered according to the HMM-Mix cluster assignments, and the data displayed are the re-estimated calls in the presence of the profiles. Figure 5.3(b) (top) shows the profiles of each group and it is clear that the re-estimated calls are heavily influenced by their corresponding profiles.

The results are characterized by the following groups (1): +7 (meaning gain of chromosome 7) (7 patients); (2): a 'null' group with no recurrent alterations (67 patients); (3): a group with +18 (19 patients); (4): a group with +1q and a small loss at 1p36 (7 patients); (5): a singleton outlier (1 patient); and (6): +6p/6q- (5 patients). Notably +1p, +6p/6q-,7+, and 18+ are known genetic pathways for acquired alterations in FL reported in Hoglund *et al*, detected using G-banded karyotyping [102]. Importantly, groups 1 and 6 had significantly reduced time to transformation (TTT), shown in Figure 5.4 (black and yellow Kaplan Meier curves respectively) by log-rank test ($p<0.01$), suggesting clinicopathologic significance of +7 and +6p/6q- as potential prognostic indicators for FL. Furthermore, we note that the clusters produced by HMM-Mix on the entire data set mirror those reported in



(a) FL WKM initializations



(b) FL HMM-Mix

Figure 5.3: Clustering of FL data showing the initial calls and WKM clusters (top) and the converged estimates of the calls, clusters and profiles by HMM-Mix (bottom). (a) The calls and clusters depicted as a heat map for WKM with G=6. The rows of the data indicate the patients and the columns indicate the probes. Red indicates gain, green loss and black neutral. The rows are ordered according to their assigned groups as predicted by WKM. (b) The posterior probability of the calls (where blue represents $p(Z_t^p = neutral)$), the clusters and the profiles (top) for the G=6 groups. In comparison to (a) the clusters are readily apparent from the data, they appear to be tighter and the re-estimated calls are clearly influenced by the profiles, resulting in far less noisy, and far more interpretable output. Importantly, 4 of the 6 groups (labeled on right) recapitulate the previously reported subtypes for FL. Group numbers that correspond to the time to transformation curves (Figure 5.4) are annotated on the right of (b). Groups 1 and 6 both had statistically significantly shorter time to transformation.

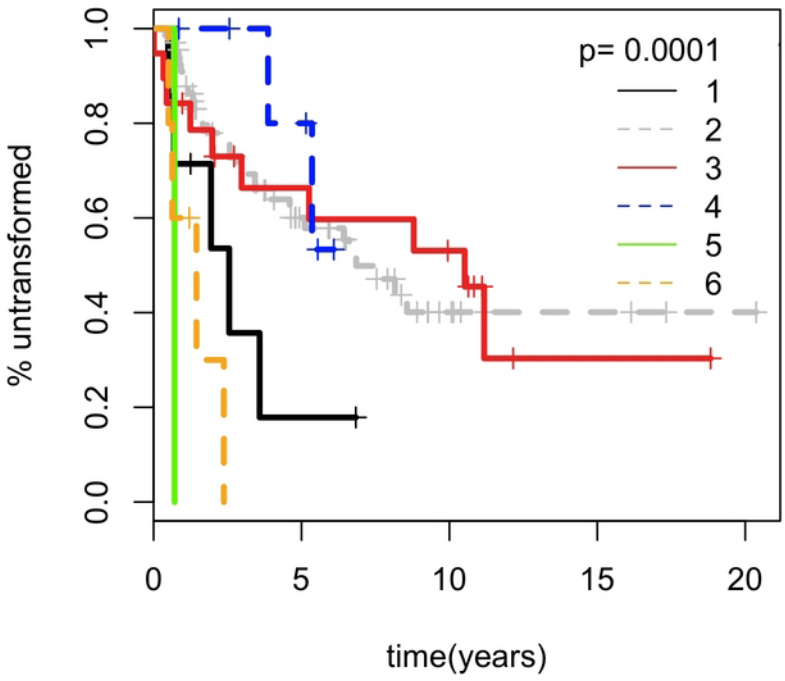
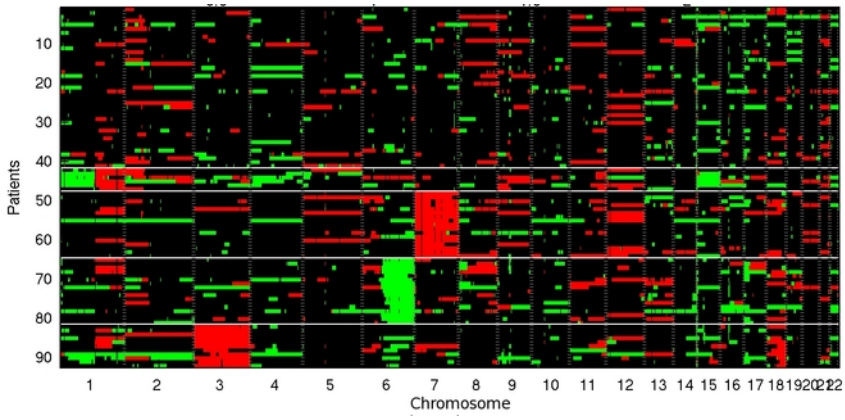


Figure 5.4: Time to transformation Kaplan-Meier curves for each group of patients as predicted by HMM-Mix for the FL cohort. Groups 1 and 6 (black and yellow) had significantly reduced time to transformation by log-rank test with 5 degrees of freedom. (The green curve corresponds to the singleton group shown in Figure 5.3). These correspond respectively to the groups characterized by 7+ and 6p-/6q+ (see Figure 5.3) and suggests that these recurrent CNAs confer inferior prognoses to the patients in these groups.

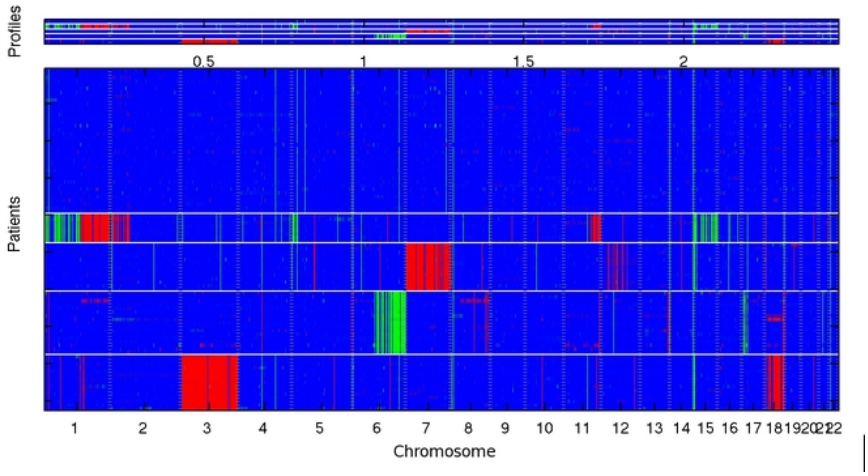
Cheung *et al* [7] where WKM was used in a supervised setting, where the $Z_{1:T}^{1:P}$ matrix was determined by manually curating computational predictions of CNAs and feature selection was performed by only including probes that exhibited a recurrent CNA in at least 10% of the patients (see Chapter 4 for details). In addition, the number of groups (5) in Cheung *et al* was chosen using supporting evidence from the literature, while in the HMM-Mix analysis, it was determined automatically from the data by Silhouette. HMM-Mix was therefore able to recapitulate clusters determined with significant manual interpretation of the data using a purely computational approach, and provide interpretable output to the investigator for further follow up. Finally, comparison between the WKM clusters and HMM-Mix clusters (Figure 5.3 (a) and (b)) shows that WKM only placed 2 patients in group 1 whereas HMM-Mix placed 7. Given that the outcome data (Figure 5.4) suggests that group 1 had prognostic significance, this result shows that despite what could be considered a 'poor' initializations, HMM-Mix was still able to identify the clinically relevant groups.

5.4.2 DLBCL data

Figure 5.5 shows the resultant subgroups from the 92 patients in the DLBCL cohort. Comparison of Figure 5.5(a) (initial calls and clusters) and (b) (HMM-Mix calls and clusters) shows that the algorithm is achieving the desired effect of focusing on putative driver or highly recurrent within-group alterations and ignoring non-recurrent passenger alterations, thus separating signal from noise. The data fell into 5 distinct groups characterized by a 'null' group with no discernible pattern and four groups characterized by: 1p-/+1q/+2p/+11q/15-, +7, 6q-, and +3/+18. The last group is a previously unreported pattern of alteration in DLBCL. Previous work had identified that both changes show increased frequency in the so-called activated B cell (ABC) subtype of DLBCL [140], but had not recognized that these two alterations travel together and may indeed define a unique molecular subgroup. This discovery merits further investigation as to its clinical significance in relation to survival and response to therapy.



(a) DLBCL WKM initialization



(b) DLBCL HMM-Mix

Figure 5.5: Clustering of 92 DLBCL profiles into 5 groups. Comparison between WKM initialization (a) and HMM-Mix (b) clearly shows HMM-Mix ability to reduce noise and report only highly conserved within-group patterns. The bottom cluster for HMM-Mix (b) shows a potentially novel subtype with gain of chr 3+/18+.

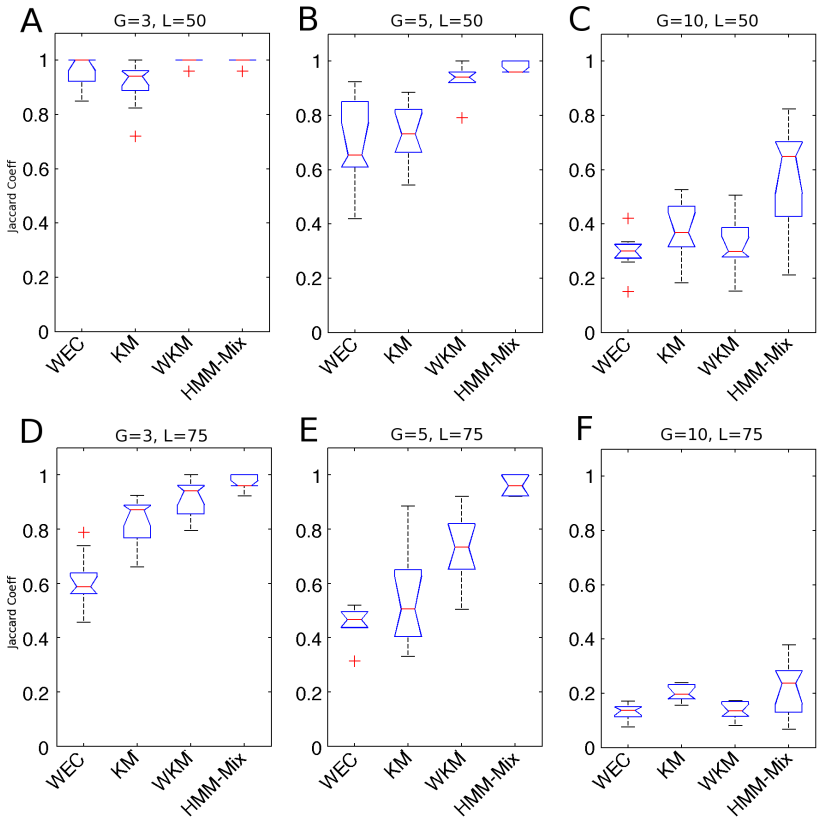


Figure 5.6: Distribution of accuracy of WECCA, KM, WKM and HMM-Mix for synthetic data generated with six different parameter settings. HMM-Mix was the most accurate for all six settings (see Table 5.2 for details). Each data set was composed of $P = 100$ patients with 672 probes each. From left to right there were $G = 3, 5, 10$ embedded groups in the data. The top row had randomly placed aberrations of $L = 50$ and the bottom row with $L = 75$. Distributions of Jaccard coefficient over 10 replicates of the G, L settings are shown as notched box plots where non-overlapping notches indicate statistical difference of the medians (red horizontal lines) with 95% confidence.

5.4.3 HMM-Mix more accurate in simulation study

Figure 5.6 shows the distribution of accuracy of WECCA, KM, WKM and HMM-Mix resulting from the simulation study. We show the Jaccard coefficient distributions over the 10 replicates of the G, L data generating parameters described above in Section 5.3.5. Table 5.2 contains the mean$\pm$ standard error for each of the datasets for the four methods. HMM-Mix showed the highest accuracy for all six settings. For number of groups $G = 3$ and passenger alteration length $L = 50$, HMM-Mix and WKM were more accurate than WECCA at recovering the ground truth classes and statistically more accurate than KM (one-way ANOVA, $p < 0.01$). For $G = 3$, $L = 75$ HMM-Mix was more accurate than WKM and statistically more accurate than both KM and WECCA ($p < 0.01$). For $G = 5$, $L = 50$, similar results were observed. HMM-Mix was more accurate than WKM and statistically more accurate than both KM and WECCA ($p < 0.01$). For $G = 5, L = 75$ and $G = 10, L = 50$ HMM-Mix was statistically more accurate than all other methods ($p < 0.01$). Finally, for $G = 10, L = 75$ all methods performed poorly, however HMM-Mix was still more accurate than KM and significantly more accurate ($p < 0.01$) than WKM and WECCA.

HMM-Mix was generally robust to the size of the randomly placed passenger alterations (see for example results for $G = 5, L = 50, 75$ shown in Figure 5.6 B and D). This suggests that the model is able to maintain its ability to detect group-specific alterations in the presence of additional noise created by larger passenger alterations. We also tested the robustness of HMM-Mix to initializations by comparing JC of the HMM-Mix predicted clusters when initialized by KM and WKM. We found that for $G = 3, 5$, $L = 50, 75$, converged results were nearly identical, despite the fact that WKM was significantly more accurate than KM for all four data sets. This suggests that in these settings, HMM-Mix is able to overcome a poor initialization, most likely due to its ability to re-estimate the calls and adapt the feature selection during inference. We suspect that these characteristics allow it to escape from local optima more readily than WKM approach that cannot re-estimate the calls and requires the feature selection to be fixed at runtime. Interestingly, HMM-Mix was considerably more sensitive to initialization in the $G = 10$ setting, indicating that in the presence of fewer datapoints per group, initialization

Table 5.2: Accuracy results for simulation study

Dataset	WECCA	KM	WKM	HMM-Mix	ANOVA P-val
G=3 L=50	0.959±0.018	0.916±0.027	0.996±0.004	0.996±0.004	4×10^{-3}
G=5 L=50	0.692±0.048	0.734±0.034	0.932±0.018	0.976±0.007	6×10^{-8}
G=10 L=50	0.296±0.022	0.375±0.033	0.317±0.031	0.580±0.065	7×10^{-5}
G=3 L=75	0.611±0.030	0.828±0.029	0.923±0.022	0.965±0.009	4×10^{-12}
G=5 L=75	0.460±0.019	0.548±0.057	0.730±0.043	0.964±0.011	6×10^{-11}
G=10 L=75	0.131±0.010	0.202±0.010	0.138±0.010	0.223±0.032	1×10^{-3}

is more important. We parenthetically note that we repeated the simulation experiment with $P = 500$ datapoints and noted higher accuracy in all settings, most notably $G = 10$ (data not shown).

5.5 Discussion and future work

The HMM-Mix model presented in this paper is effectively able to discover subgroups and profiles that define those subgroups given a set of aCGH data derived from a patient cohort. We showed the model's capability of finding clinically relevant subtypes in an FL cohort and a previously undescribed subtype in the DLBCL cohort. We demonstrated how the joint inference procedure of inferring copy number calls, cluster assignments and profiles, coupled with adaptive feature selection makes HMM-Mix significantly more accurate than partitioning and hierarchical clustering methods. Future work will entail further exploration of the 7+ and 6p-/6q+ subgroups detected in the FL cohort for prognostic significance for TTT and determining clinical relevance of the DLBCL subgroups we reported. Extension of HMM-Mix to high density SNP arrays (eg Affymetrix 6.0) will be of interest as patterns of both genotype and copy number can be elucidated. HMM-based models for SNP arrays introduced in Colella *et al* [52] and Scharpf *et al* [141] will be investigated for extension to the clustering setting using the HMM-Mix framework introduced here.

Chapter 6

Conclusion

The goals of this dissertation were to solve three important and challenging problems in array CGH data analysis, namely the inference of DNA copy number alterations (CNAs) in data derived from a cohort of patients. We developed and applied model based analytical approaches for the detection of CNAs in a single aCGH experiment, recurrent CNAs in multiple experiments, and subgroup discovery in cohorts exhibiting molecular heterogeneity. For each of these tasks, we improved on standard and/or available methods, producing state-of-the-art solutions. Our solutions combine to form a robust and comprehensive statistical framework based on principled probabilistic graphical models and machine learning techniques for processing aCGH data. Working together with clinical and biomedical researchers, we developed and applied our methods to real-world settings, leading to new insights in 2 types of lymphoma and the first high-resolution description of CNAs in follicular lymphoma.

6.1 Summary of contributions

6.1.1 Robust HMM for single sample aCGH analysis

We developed a novel continuous emission robust HMM (HMM-R) tailored specifically to aCGH. Using a fully Bayesian representation of hidden copy number states with implicit biological meaning, we proposed an improved parameter es-

timation method during inference that leveraged statistical strength present across chromosomes, leading to significantly more accurate results. In addition, we specified a model to explicitly model outliers in the data, preventing spurious singleton-probe CNA predictions. Finally, we developed a heuristic, data-driven method to set the hyperparameters and initialize our model and demonstrate highly accurate results on real data with virtually no free parameters for the user to set. This work forms the backbone of our framework and was leveraged extensively in our other contributions.

6.1.2 Inferring recurrent CNAs from a set of aCGH data

We extended the HMM for single sample analysis to the problem of detecting recurrent CNAs from a set of aCGH data. We developed a hierarchical HMM (H-HMM) that explicitly models passenger and driver CNAs as separate generative processes. This novel idea results in sparse and specific predictions of recurrent CNAs and produces output that is more favourable to the investigator than baseline models. In addition, our model borrows statistical strength present in the raw data across patients resulting in increased sensitivity over baseline models. The output of our model is a profile that represents the putative important alterations in the cohort.

6.1.3 Model-based clustering of aCGH data

We used the idea of modelling passengers and drivers and extended its application to a multi-group setting in model-based approach to cluster patients into subgroups called HMM-Mix. We proposed a model that simultaneously clusters patients into subgroups, infers the profiles and re-estimates the CNA calls in the presence of the profiles. We showed how this inference technique improves on baseline models that perform these steps as discrete and disjoint steps. Furthermore, we perform adaptive feature selection in our method capable of modifying feature selection during inference of the clusters. This allows the model to focus on highly conserved CNAs in the group-specific profiles yielding more accurate clustering of the data.

6.1.4 Genome-wide profiling of follicular lymphoma

We applied HMM-R and HMM-Mix to data derived from a cohort of 106 follicular lymphoma patients. Our analysis produced the first genome-wide, high resolution molecular portrait of this disease. We found prognostically significant CNAs, now under investigation for clinical relevance in more focused studies. Experimental validation of HMM-R predictions led to confirmation of 8/8 genomic deletions in the 1p36 region of chromosome 1. 2/2 controls that were predicted as normal also validated. These results showed that our methods are working well in a clinical setting. HMM-Mix was able to reproduce known molecular subtypes and 2 clinically relevant subtypes whose patients exhibited reduced time to transformation to a more aggressive form of lymphoma.

6.2 Future work

Our contributions form a solid foundation we can build upon for more complex and sophisticated bioinformatics problems. For example, while it has been suggested that gene expression is often related to copy number [12], the relationship between the two is complex and not yet fully understood [142]. While numerous matched datasets of aCGH and gene expression arrays now exist, development of principled methods for their joint analysis is an open problem. Extending our framework for application to the problem of jointly analyzing gene expression and copy number would be a reasonable place to begin.

In addition to copy number alterations which result in modified DNA structure and sequence of the tumour genome, epigenomic changes due to chromatin remodelling and altered methylation patterns play a key role in the gene expression patterns of tumours [143]. As genome wide assays produce data measuring epigenetic changes, these data will need to be integrated with copy number and gene expression data in order to obtain a comprehensive spectrum of the mutational changes in a tumour. Developing effective analytical solutions to integrating these data will be necessary in order to derive knowledge of which genes/biochemical pathways are disrupted in disease.

As previously discussed, SNP genotyping arrays are in widespread use for studying genetic diseases [144]. As shown by Colella *et al* [52] and Beroukhim

et al [72], sophisticated models have been proposed for recurrent CNA detection. However, to our knowledge, model-based clustering methods have not. Extending our HMM-Mix framework to SNP arrays is therefore an open problem to pursue.

Next generation sequencing (NGS) technology is providing a sea change in the field of genomics. NGS has emerged as an effective tool to sequence genomes at relatively low cost for fairly high coverage. In addition to detecting SNPs, sequence mutations, and small insertions and deletions, the nature of the data enables sequence coverage to be used as proxy for copy number changes [145]. This translates into nucleotide resolution breakpoint detection and a digital, direct measure of DNA copy number. Moreover, paired-end technology allows the detection of copy number balanced changes such as inversions and translocations that are not detectable using aCGH. We have begun to apply our models on output from this data with good success. However, the nuances needed to properly tailor the models for the most effective results are not yet understood. This is another open problem we are actively working on.

6.3 Concluding thoughts

The fields of computational biology and cancer biology are now intricately linked. As we try and decipher the important molecular events in the progression of cancer, principled analytical approaches must be applied to the data to not only detect relevant events, but to generate new hypotheses for follow up. As we showed in the FL and DLBCL cohorts, new candidate biomarkers with prognostic significance were revealed, generated via predictions by methods presented in this dissertation. As these candidates are pursued, additional data requiring rigorous analyses are produced and the cycle begins anew. It is our hope that our models are continuously applied to new data sets enabling investigators to generate new hypotheses related to copy number changes in human diseases. Much work remains in determining a catalogue of "driver" alterations in cancer. We can only hope that the work presented herein will provide useful tools to accelerate this process.

Bibliography

[1] C Lee, A J Iafrate, and A R Brothman. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet*, 39(7 Suppl):48–54, Jul 2007. → pages 1

[2] R A Weinberg. *The biology of cancer.* Garland Science, Taylor and Francis Group, 2007. → pages 1, 2, 3

[3] R Redon, S Ishikawa, K R Fitch, L Feuk, G H Perry, T D Andrews, H Fiegler, M H Shapero, A R Carson, W Chen, E K Cho, S Dallaire, J L Freeman, J R González, M Gratacòs, J Huang, D Kalaitzopoulos, D Komura, J R MacDonald, C R Marshall, R Mei, L Montgomery, K Nishimura, K Okamura, F Shen, M J Somerville, J Tchinda, A Valsesia, C Woodwark, F Yang, J Zhang, T Zerjal, J Zhang, L Armengol, D F Conrad, X Estivill, C Tyler-Smith, N P Carter, H Aburatani, C Lee, K W Jones, S W Scherer, and M E Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, Nov 2006. → pages 1, 50

[4] K K Wong, R J deLeeuw, N S Dosanjh, L R Kimm, Z Cheng, D E Horsman, C MacAulay, R T Ng, C J Brown, E E Eichler, and W L Lam. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet*, 80(1):91–104, Jan 2007. → pages 1, 50

[5] D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000. → pages 2

[6] A J Aguirre, C Brennan, G Bailey, R Sinha, B Feng, C Leo, Y Zhang, J Zhang, J D Gans, N Bardeesy, C Cauwels, C Cordon-Cardo, M S Redston, R A DePinho, and L Chin. High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A*, 101(24):9067–9072, Jun 2004. → pages 2, 60, 61

[7] K-J Cheung, S P Shah, C Steidl, N Johnson, T Relander, A Telenius, B Lai, K P Murphy, W Lam, J M Connors, R T Ng, R D Gascoyne, and D E

Horsman. Genome-wide profiling of follicular lymphoma by array comparative genomic hybridization reveals prognostically significant DNA copy number imbalances. *Blood - in press*, 2008. → pages 11, 46, 115, 116, 129, 131, 134

[8] S F Chin, A E Teschendorff, J C Marioni, Y Wang, N L Barbosa-Morais, N P Thorne, J L Costa, S E Pinder, M A van de Wiel, A R Green, I O Ellis, P L Porter, S Tavaré, J D Brenton, B Ylstra, and C Caldas. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol*, 8(10):R215, 2007. → pages 50, 115

[9] G Tonon, K K Wong, G Maulik, C Brennan, B Feng, Y Zhang, D B Khatry, A Protopopov, M J You, A J Aguirre, E S Martin, Z Yang, H Ji, L Chin, and R A Depinho. High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A*, 102(27):9625–9630, Jul 2005. → pages 2

[10] C Greenman, P Stephens, R Smith, G L Dalgliesh, C Hunter, G Bignell, H Davies, J Teague, A Butler, C Stevens, S Edkins, S O'Meara, I Vastrik, E E Schmidt, T Avis, S Barthorpe, G Bhamra, G Buck, B Choudhury, J Clements, J Cole, E Dicks, S Forbes, K Gray, K Halliday, R Harrison, K Hills, J Hinton, A Jenkinson, D Jones, A Menzies, T Mironenko, J Perry, K Raine, D Richardson, R Shepherd, A Small, C Tofts, J Varian, T Webb, S West, S Widaa, A Yates, D P Cahill, D N Louis, P Goldstraw, A G Nicholson, F Brasseur, L Looijenga, B L Weber, Y E Chiew, A DeFazio, M F Greaves, A R Green, P Campbell, E Birney, D F Easton, G Chenevix-Trench, M H Tan, S K Khoo, B T Teh, S T Yuen, S Y Leung, R Wooster, P A Futreal, and M R Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, Mar 2007. → pages 2

[11] B P Coe, W W Lockwood, L Girard, R Chari, C Macaulay, S Lam, A F Gazdar, J D Minna, and W L Lam. Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br J Cancer*, 94(12):1927–1935, Jun 2006. → pages 3, 11, 50, 57, 65, 77, 79, 80, 81

[12] J R Pollack, T Sorlie, C M Perou, C A Rees, S S Jeffrey, P E Lonning, R Tibshirani, D Botstein, A L Borresen-Dale, and P O Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–12968, Oct 2002. → pages 3, 50, 51, 141

[13] T Sorlie. Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur J Cancer*, 40(18):2667–2675, Dec 2004. → pages 3, 115

[14] A Rosenwald, G Wright, K Leroy, X Yu, P Gaulard, R D Gascoyne, W C Chan, T Zhao, C Haioun, T C Greiner, D D Weisenburger, J C Lynch, J Vose, J O Armitage, E B Smeland, S Kvaloy, H Holte, J Delabie, E Campo, E Montserrat, A Lopez-Guillermo, G Ott, H K Muller-Hermelink, J M Connors, R Braziel, T M Grogan, R I Fisher, T P Miller, M LeBlanc, M Chiorazzi, H Zhao, L Yang, J Powell, W H Wilson, E S Jaffe, R Simon, R D Klausner, and L M Staudt. Molecular diagnosis of primary mediastinal b cell lymphoma identifies a clinically favorable subgroup of diffuse large b cell lymphoma related to hodgkin lymphoma. *J Exp Med*, 198(6):851–862, Sep 2003. → pages 3

[15] M J van de Vijver, Y D He, L J van't Veer, H Dai, A A Hart, D W Voskuil, G J Schreiber, J L Peterse, C Roberts, M J Marton, M Parrish, D Atsma, A Witteveen, A Glas, L Delahaye, T van der Velde, H Bartelink, S Rodenhuis, E T Rutgers, S H Friend, and R Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, Dec 2002. → pages 3

[16] AS. Ishkanian, CA. Malloff, SK. Watson, RJ. DeLeeuw, B. Chi, BP. Coe, A. Snijders, DG. Albertson, D. Pinkel, MA. Marra, V. Ling, C. MacAulay, and WL. Lam. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, 36(3):299–303, Mar 2004. → pages 4, 47, 89, 129

[17] D. Pinkel and DG. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, 37 Suppl:11–17, Jun 2005. → pages 4

[18] Chari R, Lockwood W W, and Lam W L. Computational methods for the analysis of array comparative genomic hybridization. *Cancer Informatics*, 2, 2006. → pages 5

[19] R J de Leeuw, J J Davies, A Rosenwald, G Bebb, R D Gascoyne, M J Dyer, L M Staudt, J A Martinez-Climent, and W L Lam. Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, 13(17):1827–1837, Sep 2004. → pages 4, 6, 11, 39, 51, 52, 73, 129

[20] S P Shah, X Xuan, R J DeLeeuw, M Khojasteh, W L Lam, R Ng, and K P Murphy. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, 22(14):431–439, Jul 2006. → pages 7, 27, 28, 29, 32, 46, 89, 90, 124

[21] S P Shah, W L Lam, R T Ng, and K P Murphy. Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, 23(13):450–458, Jul 2007. → pages 7, 49, 61, 62, 90, 128

[22] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. → pages 9, 18, 25, 26, 35, 39, 121

[23] MI Jordan. Graphical models. *Statistical Science (Special Issue on Bayesian Statistics)*, 19:140–155, 2004. → pages 9, 10

[24] L R Kimm, R J deLeeuw, K J Savage, A Rosenwald, E Campo, J Delabie, G Ott, H K Muller-Hermelink, E S Jaffe, L M Rimsza, D D Weisenburger, W C Chan, L M Staudt, J M Connors, R D Gascoyne, and W L Lam. Frequent occurrence of deletions in primary mediastinal b-cell lymphoma. *Genes Chromosomes Cancer*, 46(12):1090–1097, Dec 2007. → pages 11

[25] R J Deleeuw, A Zettl, E Klinker, E Haralambieva, M Trottier, R Chari, Y Ge, R D Gascoyne, A Chott, H K Müller-Hermelink, and W L Lam. Whole-genome analysis and HLA genotyping of enteropathy-type T-cell lymphoma reveals 2 distinct lymphoma subtypes. *Gastroenterology*, 132(5):1902–1911, May 2007. → pages 11, 39

[26] PH. Eilers and RX. de Menezes. Quantile smoothing of array CGH data. *Bioinformatics*, 21(7):1146–1153, Apr 2005. → pages 14, 15

[27] M Khojasteh, W L Lam, R K Ward, and C MacAulay. A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics*, 6:274–274, Nov 2005. → pages 14, 37

[28] L. Hsu, SG. Self, D. Grove, T. Randolph, K. Wang, JJ. Delrow, L. Loo, and P. Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, Apr 2005. → pages 15

[29] AB Olshen, ES Venkatraman, R Lucito, and M Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, Oct 2004. → pages 16, 60

[30] E S Venkatraman and A B Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, Mar 2007. → pages 16, 60

[31] F Picard, S Robin, M Lavielle, C Vaisse, and JJ Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1):27–27, Feb 2005. → pages 16

[32] K Jong, E Marchiori, G Meijer, A V Vaart, and B Ylstra. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, 20(18):3636–3637, Dec 2004. → pages 16, 65

[33] P. Hupé, N. Stransky, JP. 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, Dec 2004. → pages 16

[34] Paul Fearnhead. Exact and Efficient Bayesian Inference for Multiple Changepoint problems. *Statistics and Computing*, 16:203–213, 2006. → pages 16

[35] WR Lai, MD Johnson, R Kucherlapati, and PJ Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763–3770, Oct 2005. → pages 16

[36] H Willenbrock and J Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, Nov 2005. → pages 16

[37] 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, Dec 2001. → pages 17

[38] M. Stephens. Dealing with label-switching in mixture models. *J. Royal Statistical Society, Series B*, 62:795–809, 2000. → pages 20

[39] J. Fridlyand, A. Snijders, D. Pinkel, D. Albertson, and A. Jain. Hidden Markov Models approach to the analysis of array CGH data. *Journal of Multivariate Statistics*, 90:132–153, 2004. → pages 22, 26, 27, 28

[40] P Broët and S Richardson. Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, Feb 2006. → pages 27

[41] Y Shi, F Guo, W Wu, and EP Xing. GIMscan: A new statistical method for analyzing whole-genome array CGH data. In *Proceedings of RECOMB*, 2007. → pages 27

[42] D A Engler, G Mohapatra, D N Louis, and R A Betensky. A Pseudolikelihood Approach for Simultaneous Analysis of Array Comparative Genomic Hybridizations (aCGH). *Biostatistics*, 7(3):399–421, Jan 2006. → pages 27, 28

[43] S Guha, Y Li, and D Neuberg. Bayesian Hidden Markov Modeling of Array CGH Data. Technical report, Harvard School of Public Health, 2006. → pages 27, 28, 68

[44] M A van de Wiel, K I Kim, S J Vosse, W N van Wieringen, S M Wilting, and B Ylstra. Cghcall: calling aberrations for array cgh tumor profiles. *Bioinformatics*, 23(7):892–894, Apr 2007. → pages 27, 42

[45] J A Veltman and B B de Vries. Diagnostic genome profiling: unbiased whole genome or targeted analysis? *J Mol Diagn*, 8(5):534–537, Nov 2006. → pages 33, 51, 56

[46] C Baldwin, C Garnis, L Zhang, M P Rosin, and W L Lam. Multiple microalterations detected at high frequency in oral cancer. *Cancer Res*, 65(17):7561–7567, Sep 2005. → pages

[47] C Garnis, W W Lockwood, E Vucic, Y Ge, L Girard, J D Minna, A F Gazdar, S Lam, C Macaulay, and W L Lam. High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *Int J Cancer*, 118(6):1556–1564, Mar 2006. → pages 33, 51, 56, 77, 78, 80, 81

[48] R Pique-Regi, J Monso-Varona, A Ortega, R C Seeger, T J Triche, and S Asgharzadeh. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, 24(3):309–318, Feb 2008. → pages 46

[49] B Chi, R J DeLeeuw, B P Coe, C MacAulay, and W L Lam. Seegh–a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, 5:13–13, Feb 2004. → pages 46

[50] MJ Beal, Z Ghahramani, and CE Rasmussen. "the infinite hidden markov model". In *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, 2002. MIT Press. → pages 46

[51] O M Rueda and R Díaz-Uriarte. Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput Biol*, 3(6):e112, Jun 2007. → pages 47

[52] S Colella, C Yau, J M Taylor, G Mirza, H Butler, P Clouston, A S Bassett, A Seller, C C Holmes, and J Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*, 35(6):2013–2025, Mar 2007. → pages 47, 61, 63, 82, 138, 141

[53] S Stjernqvist, T Rydén, M Sköld, and J Staaf. Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, 23(8):1006–1014, Apr 2007. → pages 48

[54] N Hosoya, M Sanada, Y Nannya, K Nakazaki, L Wang, A Hangaishi, M Kurokawa, S Chiba, and S Ogawa. Genomewide screening of DNA copy number changes in chronic myelogenous leukemia with the use of high-resolution array-based comparative genomic hybridization. *Genes Chromosomes Cancer*, 45(5):482–494, May 2006. → pages 50, 115

[55] H Yaziji, L C Goldstein, T S Barry, R Werling, H Hwang, G K Ellis, J R Gralow, R B Livingston, and A M Gown. HER-2 testing in breast cancer using parallel tissue-based methods. *JAMA*, 291(16):1972–1977, Apr 2004. → pages 50

[56] C Schwaenen, M Nessling, S Wessendorf, T Salvi, G Wrobel, B Radlwimmer, H A Kestler, C Haslinger, S Stilgenbauer, H Döhner, M Bentz, and P Lichter. Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. *Proc Natl Acad Sci U S A*, 101(4):1039–1044, Jan 2004. → pages 50

[57] L Chin and J Gray. Translating insights from the cancer genome into clinical practice. *Nature*, 242:553–563, Apr 2008. → pages 51

[58] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 2004. 2nd edition. → pages 51, 67

[59] S J Kim, Z N Rabbani, R T Vollmer, E G Schreiber, E Oosterwijk, M W Dewhirst, Z Vujaskovic, and M J Kelley. Carbonic anhydrase IX in early-stage non-small cell lung cancer. *Clin Cancer Res*, 10(23):7925–7933, Dec 2004. → pages 51

[60] D E Swinson, J L Jones, D Richardson, C Wykoff, H Turley, J Pastorek, N Taub, A L Harris, and K J O'Byrne. Carbonic anhydrase IX expression, a novel surrogate marker of tumor hypoxia, is associated with a poor prognosis in non-small-cell lung cancer. *J Clin Oncol*, 21(3):473–482, Feb 2003. → pages 51

[61] S Kawamata, T Hori, A Imura, A Takaori-Kondo, and T Uchiyama. Activation of OX40 signal transduction pathways leads to tumor necrosis factor receptor-associated factor (TRAF) 2- and TRAF5-mediated NF-kappaB activation. *J Biol Chem*, 273(10):5808–5814, Mar 1998. → pages 56, 79

[62] D Alarcon-Vargas, S Y Fuchs, S Deb, and Z Ronai. p73 transcriptional activity increases upon cooperation between its spliced forms. *Oncogene*, 19(6):831–835, Feb 2000. → pages 56

[63] M Khojasteh, W L Lam, R K Ward, and C MacAulay. A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics*, 6:274–274, Nov 2005. → pages 59

[64] J C Marioni, N P Thorne, A Valsesia, T Fitzgerald, R Redon, H Fiegler, T D Andrews, B E Stranger, A G Lynch, E T Dermitzakis, N P Carter, S Tavaré, and M E Hurles. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol*, 8(10):R228, Oct 2007. → pages

[65] P Neuvial, P Hupé, I Brito, S Liva, E Manié, C Brennetot, F Radvanyi, A Aurias, and E Barillot. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7:264–264, 2006. → pages 59

[66] A Idbaih, Y Marie, C Lucchesi, G Pierron, E Manié, V Raynal, V Mosseri, K Hoang-Xuan, M Kujas, I Brito, K Mokhtari, M Sanson, E Barillot, A Aurias, J Y Delattre, and O Delattre. BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. *Int J Cancer*, 122(8):1778–1786, Apr 2008. → pages 60

[67] R C Gentleman, V J Carey, D M Bates, B Bolstad, M Dettling, S Dudoit, B Ellis, L Gautier, Y Ge, J Gentry, K Hornik, T Hothorn, W Huber, S Iacus, R Irizarry, F Leisch, C Li, M Maechler, A J Rossini, G Sawitzki, C Smith, G Smyth, L Tierney, J Y Yang, and J Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004. → pages 60

[68] C Rouveirol, N Stransky, P Hupé, P L Rosa, E Viara, E Barillot, and F Radvanyi. Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, 22(7):849–856, Apr 2006. → pages 60, 61, 73

[69] S J Diskin, T Eck, J Greshock, Y P Mosse, T Naylor, C J Stoeckert, B L Weber, J M Maris, and G R Grant. STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res*, 16(9):1149–1158, Sep 2006. → pages 61

[70] A Ben-Dor, D Lipson, A Tsalenko, M Reimers, L O Baumbusch, M T Barrett, J N Weinstein, A-L Brresen-Dale, and Z Yakhini. Framework for Identifying Common Aberrations in DNA Copy Number Data. In *Research in Computational Molecular Biology*, volume 4453 of *Lecture Notes in Computer Science*, pages 122–136. Springer, 2007. → pages 61, 62

[71] C Klijn, H Holstege, J de Ridder, X Liu, M Reinders, J Jonkers, and L Wessels. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res*, 36(2):e13, Feb 2008. → pages 61, 62, 128

[72] R Beroukhim, G Getz, L Nghiemphu, J Barretina, T Hsueh, D Linhart, I Vivanco, J C Lee, J H Huang, S Alexander, J Du, T Kau, R K Thomas, K Shah, H Soto, S Perner, J Prensner, R M Debiasi, F Demichelis, C Hatton, M A Rubin, L A Garraway, S F Nelson, L Liau, P S Mischel, T F Cloughesy, M Meyerson, T A Golub, E S Lander, I K Mellinghoff, and W R Sellers. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A*, 104(50):20007–20012, Dec 2007. → pages 61, 64, 142

[73] D Lipson, Y Aumann, A Ben-Dor, N Linial, and Z Yakhini. Efficient calculation of interval scores for DNA copy number data analysis. *J Comput Biol*, 13(2):215–228, Mar 2006. → pages 62, 69

[74] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998. → pages 65

[75] S Scott. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 2002. → pages 69

[76] Z. Ghahramani and M. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997. → pages 72

[77] A Al-Tourah, M Chhanabhai, K Gill, P Hoskins, R Klasa, C Paltiel, K Savage, L Sehn, T Shenkier, R Gascoyne, and J Connors. Incidence, predictive factors and outcome of transformed lymphoma: a population-based study from british columbia. In *Annals of Oncology. 9th International Conference on Malignant Lymphoma*, volume 64, 2005. → pages 86, 110

[78] A Jemal, R Siegel, E Ward, T Murray, J Xu, and M J Thun. Cancer statistics, 2007. *CA Cancer J Clin*, 57(1):43–66, Jan-Feb 2007. → pages 86

[79] E S Jaffe, N L Harris, H Stein, and J W Vardiman. *Pathology & Genetics of Tumours of Haematopoietics and Lymphoid tissues: World Health Organization Classification of Tumours*. IARC Press, France: Lyon, 2001. → pages 86

[80] T Y Kang, L A Rybicki, B J Bolwell, S G Thakkar, S Brown, R Dean, M A Sekeres, A Advani, R Sobecks, M Kalaycio, B Pohlman, and J W Sweetenham. Effect of prior rituximab on high-dose therapy and autologous stem cell transplantation in follicular lymphoma. *Bone Marrow Transplant*, 40(10):973–978, Nov 2007. → pages 86

[81] W B Graninger, M Seto, B Boutain, P Goldman, and S J Korsmeyer. Expression of bcl-2 and bcl-2-ig fusion transcripts in normal and neoplastic cells. *J Clin Invest*, 80(5):1512–1515, Nov 1987. → pages 86

[82] H Tilly, A Rossi, A Stamatoullas, B Lenormand, C Bigorgne, A Kunlin, M Monconduit, and C Bastard. Prognostic value of chromosomal abnormalities in follicular lymphoma. *Blood*, 84:1043–1049, Aug 1994. → pages 108

[83] Y Tsujimoto, L R Finger, J Yunis, P C Nowell, and C M Croce. Cloning of the chromosome breakpoint of neoplastic b cells with the t(14;18) chromosome translocation. *Science*, 226:1097–1099, 1984. → pages

[84] J J Yunis, G Frizzera, M M Oken, J McKenna, A Theologides, and M Arnesen. Multiple recurrent genomic defects in follicular lymphoma. a possible model for cancer. *N Engl J Med*, 316(2):79–84, Jan 1987. → pages 86

[85] T J McDonnell, N Deane, F M Platt, G Nunez, U Jaeger, J P McKearn, and S J Korsmeyer. bcl-2-immunoglobulin transgenic mice demonstrate extended b cell survival and follicular lymphoproliferation. *Cell*, 57(1):79–88, Apr 1989. → pages 86

[86] T J McDonnell and Korsmeyer S J. Progression from lymphoid hyperplasia to high-grade malignant lymphoma in mice transgenic for the t(14; 18). *Nature*, 349(6306):254–6, Jan 1991. → pages 86

[87] G Dölken, G Illerhaus, C Hirt, and R Mertelsmann. Bcl-2/jh rearrangements in circulating b cells of healthy blood donors and patients with nonmalignant diseases. *J Clin Oncol*, 14(4):1333–1344, Apr 1996. → pages 86

[88] J Limpens, R Stad, C Vos, C de Vlaam, D de Jong, G J van Ommen, E Schuuring, and P M Kluin. Lymphoma-associated translocation t(14;18) in blood b cells of normal individuals. *Blood*, 85(9):2528–2536, May 1995. → pages 86

[89] B Vogelstein and K W Kinzler. Cancer genes and the pathways they control. *Nat Med*, 10(8):789–799, Aug 2004. → pages 86

[90] H Avet-Loiseau, M Vigier, A Moreau, M P Mellerin, F Gaillard, J L Harousseau, R Bataille, and N Milpied. Comparative genomic hybridization detects genomic abnormalities in 80% of follicular lymphomas. *Br J Haematol*, 97(1):119–122, Apr 1997. → pages 87

[91] M Bentz, C A Werner, H Döhner, S Joos, T F Barth, R Siebert, M Schröder, S Stilgenbauer, K Fischer, P Möller, and P Lichter. High incidence of chromosomal imbalances and gene amplifications in the classical follicular variant of follicle center lymphoma. *Blood*, 88(4):1437–1444, Aug 1996. → pages 112

[92] M Berglund, G Enblad, U Thunberg, R M Amini, C Sundström, G Roos, M Erlanson, R Rosenquist, C Larsson, and S Lagercrantz. Genomic imbalances during transformation from follicular lymphoma to diffuse large b-cell lymphoma. *Mod Pathol*, 20(1):63–75, Jan 2007. → pages

[93] R Boonstra, A Bosga-Bouwer, M Mastik, E Haralambieva, J Conradie, E van den Berg, A van den Berg, and S Poppema. Identification of chromosomal copy number changes associated with transformation of follicular lymphoma to diffuse large b-cell lymphoma. *Hum Pathol*, 34(9):915–923, Sep 2003. → pages 108

[94] J R Cook, S Shekhter-Levin, and S H Swerdlow. Utility of routine classical cytogenetic studies in the evaluation of suspected lymphomas: results of 279 consecutive lymph node/extranodal tissue biopsies. *Am J Clin Pathol*, 121(6):826–835, Jun 2004. → pages

[95] J Fitzgibbon, S Iqbal, A Davies, D O'shea, E Carlotti, T Chaplin, J Matthews, M Raghavan, A Norton, T A Lister, and B D Young. Genome-wide detection of recurring sites of uniparental disomy in follicular and transformed follicular lymphoma. *Leukemia*, 21(7):1514–1520, Jul 2007. → pages

[96] D E Horsman, J M Connors, T Pantzar, and R D Gascoyne. Analysis of secondary chromosomal alterations in 165 cases of follicular lymphoma with t(14;18). *Genes Chromosomes Cancer*, 30(4):375–82, Apr 2001. → pages 88, 112

[97] R E Hough, J R Goepel, H E Alcock, B W Hancock, P C Lorigan, and D W Hammond. Copy number gain at 12q12-14 may be important in the transformation from follicular lymphoma to diffuse large b cell lymphoma. *Br J Cancer*, 84(4):499–503, Feb 2001. → pages 108

[98] A N Mohamed, M Palutke, L Eisenberg, and A Al-Katib. Chromosomal analyses of 52 cases of follicular lymphoma with t(14;18), including blastic/blastoid variant. *Cancer Genet Cytogenet*, 126(1):45–51, Apr 2001. → pages

[99] C W Ross, P D Ouillette, C M Saddler, K A Shedden, and S N Malek. Comprehensive analysis of copy number and allele status identifies multiple chromosome defects underlying follicular lymphoma pathogenesis. *Clin Cancer Res*, 13(16):4777–4785, Aug 2007. → pages 109, 110, 112

[100] A Viardot, P Möller, J Högel, K Werner, G Mechtersheimer, A D Ho, G Ott, T F Barth, R Siebert, S Gesk, B Schlegelberger, H Döhner, and M Bentz. Clinicopathologic correlations of genomic gains and losses in follicular lymphoma. *J Clin Oncol*, 20(23):4523–4530, Dec 2002. → pages 112

[101] X Zhang, S Karnan, H Tagawa, R Suzuki, S Tsuzuki, Y Hosokawa, Y Morishima, S Nakamura, and M Seto. Comparison of genetic aberrations in cd10+ diffused large b-cell lymphoma and follicular lymphoma by comparative genomic hybridization and tissue-fluorescence in situ hybridization. *Cancer Sci*, 95(10):809–814, Oct 2004. → pages 87

[102] M Höglund, L Sehn, J M Connors, R D Gascoyne, R Siebert, T Säll, F Mitelman, and D E Horsman. Identification of cytogenetic subgroups and karyotypic pathways of clonal evolution in follicular lymphomas. *Genes Chromosomes Cancer*, 39(3):195–204, Mar 2004. → pages 87, 88, 90, 105, 115, 129, 131

[103] V S Lestou, R D Gascoyne, C Salski, J M Connors, and D E Horsman. Uncovering novel inter- and intrachromosomal chromosome 1 aberrations in follicular lymphomas by using an innovative multicolor banding technique. *Genes Chromosomes Cancer*, 34(2):201–210, Jun 2002. → pages 87

[104] B S Emanuel and S C Saitta. From microscopes to microarrays: dissecting recurrent chromosomal rearrangements. *Nat Rev Genet*, 8(11):869–883, Nov 2007. → pages 87

[105] C Garnis, B P Coe, S L Lam, C MacAulay, and W L Lam. High-resolution array cgh increases heterogeneity tolerance in the analysis of clinical samples. *Genomics*, 85(6):790–793, Jun 2005. → pages 87

[106] P Solal-Céligny, P Roy, P Colombat, J White, J O Armitage, R Arranz-Saez, W Y Au, M Bellei, P Brice, D Caballero, B Coiffier, E Conde-Garcia, C Doyen, M Federico, R I Fisher, J F Garcia-Conde, C Guglielmi, A Hagenbeek, C Haïoun, M LeBlanc, A T Lister, A Lopez-Guillermo, P McLaughlin, N Milpied, P Morel, N Mounier, S J Proctor, A Rohatiner, P Smith, P Soubeyran, H Tilly, U Vitolo, P L Zinzani, E Zucca, and E Montserrat. Follicular lymphoma international prognostic index. *Blood*, 104(5):1258–1265, Sep 2004. → pages 88

[107] N L Harris, E S Jaffe, J Diebold, G Flandrin, H K Muller-Hermelink, J Vardiman, T A Lister, and C D Bloomfield. World health organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the clinical advisory committee meeting-airlie house, virginia, november 1997. *J Clin Oncol*, 17(12):3835–3849, Dec 1999. → pages 88

[108] L J Henderson, I Okamoto, V S Lestou, O Ludkovski, M Robichaud, M Chhanabhai, R D Gascoyne, R J Klasa, J M Connors, M A Marra, D E Horsman, and W L Lam. Delineation of a minimal region of deletion at 6q16.3 in follicular lymphoma and construction of a bacterial artificial chromosome contig spanning a 6-megabase region of 6q16-q21. *Genes Chromosomes Cancer*, 40(1):60–5, May 2004. → pages 88, 110

[109] R J de Leeuw, J J Davies, A Rosenwald, G Bebb, R D Gascoyne, M J Dyer, L M Staudt, J A Martinez-Climent, and W L Lam. Comprehensive whole genome array cgh profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, 13(17):1827–1837, Sep 2004. → pages 89

[110] M Khojasteh, W L Lam, R K Ward, and C MacAulay. A stepwise framework for the normalization of array cgh data. *BMC Bioinformatics*, 6:274–274, 2005. → pages 89

[111] B Chi, R J DeLeeuw, B P Coe, C MacAulay, and W L Lam. Seegh–a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, 5:13–13, Feb 2004. → pages 89

[112] V S Lestou, R D Gascoyne, L Sehn, O Ludkovski, M Chhanabhai, R J Klasa, H Husson, A S Freedman, J M Connors, and D E Horsman. Multicolour fluorescence in situ hybridization analysis of t(14;18)-positive follicular lymphoma and correlation with gene expression data and clinical outcome. *Br J Haematol*, 122(5):745–759, Sep 2003. → pages 108, 112

[113] A Rajgopal, I M Carr, J P Leek, D Hodge, S M Bell, P Roberts, K Horgan, D T Bonthron, P J Selby, A F Markham, and K A MacLennan. Detection by fluorescence in situ hybridization of microdeletions at 1p36 in lymphomas, unidentified on cytogenetic analysis. *Cancer Genet Cytogenet*, 142(1):46–50, Apr 2003. → pages 109

[114] A J Davies, A Rosenwald, G Wright, A Lee, K W Last, D D Weisenburger, W C Chan, J Delabie, R M Braziel, E Campo, R D Gascoyne, E S Jaffe, K Muller-Hermelink, G Ott, M Calaminici, A J Norton, L K Goff, J Fitzgibbon, L M Staudt, and T Andrew Lister. Transformation of follicular lymphoma to diffuse large b-cell lymphoma proceeds by distinct oncogenic mechanisms. *Br J Haematol*, 136(2):286–293, Jan 2007. → pages 110

[115] I S Lossos, A A Alizadeh, M Diehn, R Warnke, Y Thorstenson, P J Oefner, P O Brown, D Botstein, and R Levy. Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. *Proc Natl Acad Sci U S A*, 99(13):8886–8891, Jun 2002. → pages 110

[116] K Honma, S Tsuzuki, M Nakagawa, S Karnan, Y Aizawa, W S Kim, Y D Kim, Y H Ko, and M Seto. Tnfaip3 is the target gene of chromosome band 6q23.3-q24.1 loss in ocular adnexal marginal zone b cell lymphoma. *Genes Chromosomes Cancer*, 47(1):1–7, Jan 2008. → pages 110

[117] W S Kim, K Honma, S Karnan, H Tagawa, Y D Kim, Y L Oh, M Seto, and Y H Ko. Genome-wide array-based comparative genomic hybridization of ocular marginal zone b cell lymphoma: comparison with pulmonary and nodal marginal zone b cell lymphoma. *Genes Chromosomes Cancer*, 46(8):776–783, Aug 2007. → pages 110

[118] I E Wertz, K M O'Rourke, H Zhou, M Eby, L Aravind, S Seshagiri, P Wu, C Wiesmann, R Baker, D L Boone, A Ma, E V Koonin, and V M Dixit. De-ubiquitination and ubiquitin ligase domains of a20 downregulate nf-kappab signalling. *Nature*, 430(7000):694–699, Aug 2004. → pages 110

[119] R A Ihrie and L D Attardi. Perp-etrating p53-dependent apoptosis. *Cell Cycle*, 3(3):267–269, Mar 2004. → pages 110

[120] K K Wong, R J deLeeuw, N S Dosanjh, L R Kimm, Z Cheng, D E Horsman, C MacAulay, R T Ng, C J Brown, E E Eichler, and W L Lam. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet*, 80(1):91–104, Jan 2007. → pages 111

[121] R Redon, S Ishikawa, K R Fitch, L Feuk, G H Perry, T D Andrews, H Fiegler, M H Shapero, A R Carson, W Chen, E K Cho, S Dallaire, J L Freeman, J R González, M Gratacòs, J Huang, D Kalaitzopoulos, D Komura, J R MacDonald, C R Marshall, R Mei, L Montgomery, K Nishimura, K Okamura, F Shen, M J Somerville, J Tchinda, A Valsesia, C Woodwark, F Yang, J Zhang, T Zerjal, J Zhang, L Armengol, D F Conrad, X Estivill, C Tyler-Smith, N P Carter, H Aburatani, C Lee, K W Jones, S W Scherer, and M E Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, Nov 2006. → pages 111

[122] JJ. Davies, IM. Wilson, and WL. Lam. Array CGH technologies and their applications to cancer genomes. *Chromosome Res*, 13(3):237–248, 2005. → pages 112

[123] E Galteland, E A Sivertsen, D H Svendsrud, L Smedshammer, S H Kresse, L A Meza-Zepeda, O Myklebost, Z Suo, D Mu, P M Deangelis, and T Stokke. Translocation t(14;18) and gain of chromosome 18/bcl2: effects on bcl2 expression and apoptosis in b-cell non-hodgkin's lymphomas. *Leukemia*, 19(12):2313–2323, Dec 2005. → pages 112

[124] Y Yang, CL Mahaffey, N Brub, TP Maddatu, GA Cox, and WN Frankel. Complex seizure disorder caused by brunol4 deficiency in mice. *PLoS Genet.*, 3(7):e124, Jul 2007. → pages 112

[125] C M Perou, T Sorlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lonning, A L Borresen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000. → pages 115

[126] L Khalique, A Ayhan, M E Weale, I J Jacobs, S J Ramus, and S A Gayther. Genetic intra-tumour heterogeneity in epithelial ovarian cancer and its implications for molecular diagnosis of tumours. *J Pathol*, 211(3):286–295, Feb 2007. → pages 115

[127] G Wright, B Tan, A Rosenwald, E H Hurt, A Wiestner, and L M Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma. *Proc Natl Acad Sci U S A*, 100(17):9991–9996, Aug 2003. → pages 115

[128] F S Collins and A D Barker. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am*, 296(3):50–57, Mar 2007. → pages 115

[129] M H C Law, M A T Figueiredo, and A K Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1154–1166, 2004. → pages 116, 119

[130] W N van Wieringen and M A van de Wiel. Nonparametric Testing for DNA Copy Number Induced Differential mRNA Gene Expression. *Biometrics*, May 2008. → pages 116, 127

[131] N Johnson, S P Shah, deLeeuw R J, and Gascoyne R D. Specific patterns of copy number alterations in r-chop treatment failures in dlbcl. Forthcoming. → pages 116, 129

[132] WR Gilks, S Richardson, and DJ Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996. → pages 117

[133] A E Raftery and N Dean. Variable selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101:168–178, 2006. → pages 118

[134] M. P. Brown, R. Hughey, A. Krogh, I. S. Mian, K. Sjölander, and D. Haussler. "using dirichlet mixtures priors to derive hidden markov models for protein families". In *ismb*, pages 47–55, 1993. → pages 119, 120

[135] Padhraic Smyth. Clustering sequences with hidden markov models. In *Advances in Neural Information Processing*, 1997. → pages 119

[136] C Archambeau. *Probabilistic Models in Noisy Environments - And their Application to a Visual Prosthesis for the Blind*. PhD thesis, Universit catholique de Louvain, 2005. → pages 121, 122

[137] T Minka. Estimating a dirichlet distribution. Technical report, Microsoft Research, 2000. → pages 121

[138] van der Laanm M J, Pollard K S, and Bryan J. A New Partitioning Around Medoids Algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003. → pages 124

[139] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. → pages 128, 130

[140] S Bea, A Zettl, G Wright, I Salaverria, P Jehn, V Moreno, C Burek, G Ott, X Puig, L Yang, A Lopez-Guillermo, W C Chan, T C Greiner, D D Weisenburger, J O Armitage, R D Gascoyne, J M Connors, T M Grogan, R Braziel, R I Fisher, E B Smeland, S Kvaloy, H Holte, J Delabie, R Simon, J Powell, W H Wilson, E S Jaffe, E Montserrat, H K Muller-Hermelink, L M Staudt, E Campo, and A Rosenwald. Diffuse large b-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood*, 106(9):3183–3190, Nov 2005. → pages 134

[141] Scharpf R B, Parmigiani G, Pevsner J, and Ruczinski I. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *The Annals of Applied Statistics*, 2(2):687–713, 2008. → pages 138

[142] H Lee, S W Kong, and P J Park. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics*, 24(7):889–896, Apr 2008. → pages 141

[143] P A Jones and S B Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, Feb 2007. → pages 141

[144] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007. → pages 141

[145] P J Campbell, P J Stephens, E D Pleasance, S O'Meara, H Li, T Santarius, L A Stebbings, C Leroy, S Edkins, C Hardy, J W Teague, A Menzies, I Goodhead, D J Turner, C M Clee, M A Quail, A Cox, C Brown, R Durbin, M E Hurles, P A Edwards, G R Bignell, M R Stratton, and P A Futreal. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–729, Jun 2008. → pages 142