Inferring tumor evolution using computational phylogenetics

Ayshwarya Subramanian

August 19, 2013

Department of Biological Sciences
Carnegie Mellon University
Pittsburgh, PA 15213

A Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Thesis Committee:

Russell Schwartz, Chair
Stanley Shackney
Robert Murphy
Carl Kingsford

Abstract

Cancer research has made tremendous progress in understanding the basic biology of tumors. One of the key insights that has informed work in this area is the recognition that a tumor is an evolutionary system, in which individual cells undergo a process of rapid mutation and selection leading to a progression in phenotypes and, typically, aggressiveness of the tumor. Tumor phylogenetics is a strategy for interpreting the evolution of tumors using computer algorithms for phylogenetics, i.e., the inference of evolutionary trees. The approach takes advantage of a large body of phylogenetic theory and algorithms, developed primarily for inferring evolution among species, to interpret complex tumor data sets as evidence for evolutionary processes. The result is a tumor phylogeny, or phylogenetic tree, a reconstruction of the sequences of mutations that cells within a tumor or class of tumors accumulate over the course of their progression. The goals of finding such trees are to better interpret heterogeneity within and among tumors, identify and classify tumor subtypes with possible underlying mechanisms of action, learn markers of progression for key steps in tumor evolution, and enable predictive modeling of likely tumor progression steps that may ultimately assist in diagnosis and treatment.

In this dissertation, we discuss a computational framework for reconstructing phylogenies from genome-scale tumor array and sequencing data. We first present a novel phylogenetic pipeline for building tumor phylogenies from whole-genome copy number variation data. The steps included computational unmixing for resolving heterogeneity in genomic data from tumors, a statistical method for progression marker discovery, a statistical method for data discretization, application of character-based phylogeny reconstruction, and analyses of the resulting trees to draw biological significance. We then describe HMM-CNA, an improved model for discovering progression markers from cohorts of patient tumor copy number data that are especially relevant for phylogeny reconstruction via a custom multi-sample Hidden Markov model (HMM). We next present a novel strategy for phylogeny building from single cell sequencing data by inferring features that can accurately capture the composition of the individual genome sequences and distinguish among stages of tumor progression. We demonstrate these contributions on both simulated and human breast tumor biopsy and cell line data assuming a maximum parsimony model of evolution. Finally, we discuss future directions for building a more realistic model of tumor evolution by integrating patterns in genome structural changes with the functional elements they encode. We close with a discussion of recent research, current trends, and challenges and opportunities facing the field.

Acknowledgments

I wish to express my foremost gratitude to my PhD advisor Dr.Russell Schwartz. In addition to being a great research guide and mentor, I have also learnt many valuable life lessons from him. The Schwartz lab has been an enabling environment which has helped me grow in many ways as a person. I could not have asked for a better advisor and I feel truly honored to be his student. I thank my collaborator and thesis committee member Dr. Stanley Shackney for his mentorship and guidance. I also thank my other committee members Dr. Robert Murphy and Dr. Carl Kingsford for their time and feedback. I am grateful to my collaborator Dr. Adrian Lee for being generous with his time and research inputs. I thank the Department of Biological Sciences and my funding sources for supporting my doctoral work.

I thank Ena, Kristen, Emily, Carol, Donya, Cynthia and Al from the Department of Biological Sciences, and Nichole and Thom from the Lane Center for Computational Biology for their assistance at various stages of this work. I thank my lab-mates, collaborators, co-authors, classmates and mentors for their valuable time and insights. I express my deepest gratitude to everyone who has contributed both directly and indirectly to the completion of this dissertation. I especially express my love and heartfelt gratitude to my parents and sister for everything.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

In this chapter, we present the scientific question of enquiry pursued in this dissertation, the thesis statement and highlight the major contributions made. We also provide an outline for the rest of the dissertation.

Cancer research has made tremendous progress in understanding the basic biology of tumors. One of the key insights that has informed work in this area is the recognition that a tumor is an evolutionary system, in which individual cells undergo a process of rapid mutation and selection leading to a progression in phenotypes and, typically, aggressiveness of the tumor [1]. The study of tumor evolution is concerned with tracing these specific changes in time and space and their underlying causes and mechanisms of action, as observed in the transformation of healthy body tissue to benign, and progressively, malignant and aggressive, metastatic tissue types. Most of these changes occur primarily in the genomes of the tissue cells and then transition into functional changes in patterns of gene expression and translation, which are then observed as systemic changes in phenotypes.

Tumor evolution has been studied at various levels of observation: systemic physiological determinants in the cancer patients, tissue-level changes in tumor biopsy morphology and changes in molecular levels of protein, RNA and DNA, to name a few. Each level of study has promoted the formulation of specific theories and the integration of such knowledge is often necessary to present a complete picture of tumor evolution. Regardless of the research dimension, the key question confronting tumor evolution is:

What is the sequence of events underlying the onset and progression of tumorigenesis?

The answer to this question is not straightforward and has been tackled using various approaches to collect data spatio-temporally with a view to infer a generalized model of tumor evolution. The presence of heterogeneous or multiple cell populations resulting from non-homogenous changes in the genome of tissue cells, makes the task of building a generalized model of tumor evolution more challenging as tumors in patients with similar overall phenotypes may have different underlying molecular mechanisms of action. This intra-tumoral [2] and

inter-tumoral [3] heterogeneity pose unique challenges and the goal of most current research is to delineate certain general underlying sequences common to all tumors, which can complement more personalized and tumor-specific progression pathways. Advances in genomic methods including array based technologies and most recently, single nucleus sequencing have allowed access to high resolution and abundant data at the DNA, RNA and protein levels. Further several additional attributes of prognosis, treatment, outcome as well as demographics have provided additional heterogeneous data types. Other meta-data include reports from pathologists, outcomes from Immunohistochemistry (IHC) tests, genetic association tests and Fluorescent in-situ hybridization (FISH) experiments. More recently, the availability of paired data samples of normal and tumor from the same patient, regional sections from a single tumor biopsy as well as the availability of both metastatic and primary tumor samples from the same patient have provided several possibilities in data collection and experiment design. In addition to heterogeneity, the randomness in mutability can lead to noise recognized as passenger aberrations, which co-exist with the actual, or driver aberrations driving the progression of the tumor.

1.1 Thesis statement and contributions

In this thesis, we tackle the problem of delineating tumor evolution using methods and tools in computational phylogenetics. We claim the following statement:

The representation of tumor progression as an ordered set of changes in a phylogeny or evolutionary tree can both reveal the driver progression pathways and infer missing progression states. The goals of finding such trees are to better interpret heterogeneity within and among tumors, identify and classify tumor subtypes, learn markers of progression for key steps in tumor evolution, and enable predictive modeling of stages in tumor progression that may ultimately assist in diagnosis and treatment.

In order to substantiate the claim, we present the following contributions:

1. A novel phylogenetic pipeline that can build tumor phylogenies from whole genome data
   We have designed and implemented a novel phylogenetic pipeline for building tumor evolutionary trees or phylogenies from whole genome copy number variation data. This pipeline provides a step-by-step procedure for turning whole genome tumor profiles into phylogenies with biologically meaningful information. The steps included an optional computational unmixing to infer heterogeneity, a statistical method for progression-marker discovery, a statistical method for data discretization, application of character-based

phylogeny reconstruction, and analyses of the resulting trees to draw biological significance.

2. Improved method for phylogenetic marker detection and calling
   Implemented an improved method HMM-CNA for discovering progression markers from cohorts of patient tumor copy number data that are especially relevant for phylogeny reconstruction via a custom multi-sample Hidden Markov Model (HMM). HMM-CNA improves upon the state-of-the-art with respect to speed of computation, accurate noise inference and the ability to analyze multiple samples at once.

3. A novel approach to reconstructing phylogenetic trees from sequencing data
   We have reconstructed evolutionary trees from single-cell sequencing data of tumor cells by designing novel features that can accurately capture the composition of the individual genome sequences. These features can be used to both build distance-based phylogenies as well as accurately classify primary and metastatic tumor data.

1.2 Organization of the thesis

In Chapter 2, we provide a general introduction to the topic of tumor evolution and phylogenetics and introduce relevant concepts and terms. We will also survey various existing approaches to studying tumor progression.
In Chapter 3, we introduce the various tumor genomics data types under study and the issue of tumor heterogeneity. We will also present an unmixing approach to resolve such heterogeneity with results and analyses.
In Chapter 4, we present a general tumor phylogenetic pipeline with a special application to unmixed tumor copy number data. The pipeline includes separate steps for inferring heterogeneity, identifying markers of progression, discretizing such data markers to get phylogenetically amenable data types and the application of tree reconstruction algorithms. We also provide a brief description of the analyses of such phylogenetic trees.
In Chapter 5, we present HMM-CNA, a novel approach to inferring phylogenetically informative markers of progression that is an improvement over the general method described previously. We elucidate two key applications for unmixed data and larger raw datasets.
In Chapter 6, we present an application of the pipeline to breast tumor cell line structural rearrangement data.
In Chapter 7, we detail novel strategies for building phylogenetic trees from whole genome sequencing data.
In Chapter 8, we summarize the major conclusions of the thesis and speculate on future directions.

methylation

Chapter 2

Principles of Tumor Evolutionary phylogenetics

In this chapter, we provide background on the key principles and methods of tumor evolution and tumor phylogenetics and survey seminal results in the field. The contents are adapted from the book chapter [4].

The National Cancer Institute Surveillance Epidemiology and End Results (SEER) [5] projects that 1,660,290 individuals will be diagnosed with cancer in 2013 in the US alone. There are more than 100 types of cancers based on their tissue of origin. According to the National Cancer Institute, cancer maybe defined as a set of diseases in which abnormal cells grow in an uncontrolled manner and are capable of invading normal, healthy cells. We define normal or healthy cells as cells which are capable of maintaining basic housekeeping functions like balanced growth, death, metabolism and homeostasis. Abnormal cells may then be defined as those cells which have lost a sense of control or balance of these basic housekeeping functions. As a result of this loss of balance, they present "abnormal" traits. Cancerous or tumor cells are recognized by their acquisition of a set of "abnormal" traits which make them a threat to the normal functioning of the local organ systems and eventually, the survival of the organism on the whole. Hanahan and Weinberg [6] studied these abnormal traits of tumorous cells and classified them into six major categories of cell function: sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis and resisting cell death. Thus, cancerous cells present these six hallmarks of cancer.

These hallmarks, brought about by changes in cellular functions, can be traced back to changes or aberrations in the genome which in turn get translated to changes in the cellular regulatory system, transcriptome or patterns of gene expression, epigenome or patterns of chromatin modifications and proteome or patterns of protein translation. The cause for changes in the genome maybe hereditary or genetic or environmental. Hereditary factors include inheritance of recessive mutations with functional effects or a predisposition for further accumulation of lethal mutations. Genetic changes include random mutations induced by errors in DNA replication due to aging and senescence. The mutations maybe induced by environmental factors, also called carcinogens like UV radiation, chemicals like nicotine, pollutants, viruses etc. In this thesis, we focus on the story of tumorige-

nesis from the onset of cancer-driving mutations in healthy cells and do not delve into causative mechanisms.

Cancer research has made tremendous progress in understanding the basic biology of tumors. One of the key insights that has informed work in this area is the recognition that a tumor is an evolutionary system, in which individual cells undergo a process of rapid mutation and selection leading to a progression in phenotypes and, typically, aggressiveness of the tumor [1]. As the data available for understanding tumors has grown in size and complexity, largely due to next-generation sequencing (NGS), it has increasingly become apparent that an understanding of the process of tumor evolution is necessary to make sense of these data. A more detailed understanding of the common pathways of tumor evolution could not only shed light on the basic mechanistic events driving tumor development but also provide powerful evidence to guide treatment strategies and predict progression. Phylogenetic algorithms have provided one powerful set of tools for drawing meaningful inferences from complex tumor data sets in terms of models of tumor evolution.

2.1 Understanding Tumors as Evolutionary Systems

Tumor evolution is generally understood to be a process of progressive acquisition of genetic or epigenetic abnormalities [7]. Our understanding of the details of this process has become more nuanced over time alongside our ability to profile tumors at the genetic level. Early mathematical models [8] in the field suggested that tumor progression was driven by a series of DNA mutations, an insight that helped lay the basis for the more nuanced two-hit hypothesis [9] of the requirement of pairs of mutations for tumor formation and later more sophisticated models of accumulated genetic abnormalities [10]. More recent tumor progression models have been increasingly informed by population genetic theory. A defining characteristic of tumors is hypermutability [11], providing a source of high diversity in genotype and phenotype, which may be reinforced by external environmental or hereditary factors [12], but is now generally understood that the tumor genome evolves following the principles of Darwinian selection [7, 13]. A framework for understanding these selective pressures was developed by Hanahan and Weinberg [6], who sought to categorize the specific functions for which tumor evolution selects. Work continues on more detailed mathematical models of tumors as evolutionary systems (cf., [14, 15, 16]).

An important consequence of this process of rapid evolution and selection is high heterogeneity both within and between tumors. Tumors are not homogeneous masses of cells but rather contain many distinct cell populations representing different stages or directions of evolution within single tumors [17, 18]. Debate exists

within the field [19] as to whether this diversity primarily reflects evolution within tumors themselves (the clonal evolution hypothesis) or evolution within a separate population of progenitor stemlike cells (the cancer stem cell hypothesis). At the same time, the variability of this process from patient to patient also produces high heterogeneity between tumors. Genomic technologies have revealed, however, that even clinically indistinguishable tumors may show very different gene expression profiles [20, 21, 22] potentially predictive of different prognoses or responses to treatment [23, 24]. This recognition of distinct molecular mechanisms underlying diverse subtypes of tumors has in turn helped spawn the notion of targeted therapeutics [25], drugs that target defining abnormalities of specific tumor subtypes and, in the process, more selectively target tumor cells with reduced toxicity for healthy tissue (cf., [26]).

Much of the research challenge in oncology today concerns the difficulty of finding meaningful information to guide the development of diagnostics and therapeutics in the face of the enormous diversity within and between tumors. Identifying the meaningful features defining subtypes of tumors is ultimately a question of characterizing major pathways of evolution by which these subtypes develop. Distinguishing mutations causal for tumor progression that may make good therapeutic targets (known as drivers) from those that result at random from tumor hypermutability (known as passengers) is ultimately a question of distinguishing those mutations under selective pressure from those that are not. These are extremely challenging problems for a field that is suddenly awash in data, data that is complicated by the enormous diversity cell-to-cell and tumor-to-tumor as a direct result of these same evolutionary processes we seek to understand.

2.2 Phylogenetics Basics

Phylogenetics provides one answer to the problem of how the field of cancer research can draw meaningful inferences of the underlying evolutionary process of cancer from the large, highly heterogeneous data sets that confront the field. To help frame that discussion, we first provide a brief introduction to the field of phylogenetics, including the basic concepts and terminology needed for the remainder of this thesis.

A phylogeny is a representation of the evolutionary history of a set of species or organisms and their common ancestors. At the most basic level, a phylogeny is a tree, defined by a set of nodes (typically representing the species, or taxa , under study) and a set of edges connecting pairs of nodes (typically representing ancestral relationships between nearest relatives in the tree). Fig. 2.1 provides examples of hypothetical phylogenies. A tree may have a single node known as

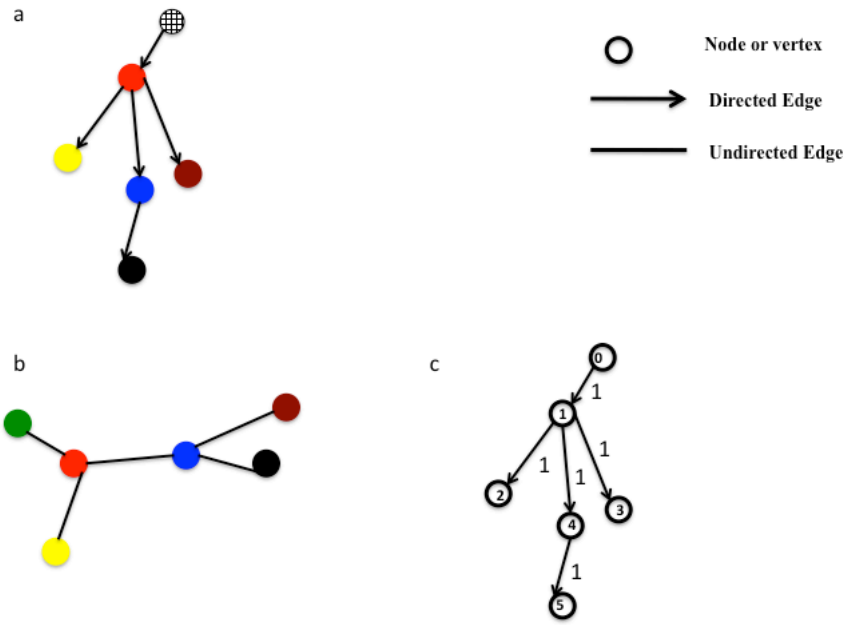


Figure 2.1: Examples of phylogenies: A phylogeny is an acyclic graph consisting of nodes or vertices connected by edges. A phylogeny can have one of two forms: (a) rooted, where the checkered node is the root and all nodes below it are children, and (b) unrooted, where there is no root and hence no direction in the phylogeny. The edges may be weighted or unweighted. (c) An example of a weighted phylogeny where each edge has weight 1.

the root , identified as the common ancestor of all other nodes in the tree. For any pair of nodes sharing an edge, the node that is higher in the tree (closer to the root) is called a parent node and that lower in the tree (farther from the root) is called the child. Nodes with no children are known as leaves and commonly represent observed members of a species.

Phylogenetics is a branch of computational biology that arose to infer phylogenies for collections of organisms. In order to infer a phylogeny, three components are required: (1) data describing evolutionary differences between the taxa under consideration, (2) an underlying model of evolution, and (3) an algorithm, or computational procedure, for building a tree given the data and model of evolution. Many variations exist on all three components of the phylogenetic inference process.

At a high level, phylogeny inference approaches are generally split into two categories depending on the kind of data they use: distance-based, where data

is assumed to describe a measure of distance between or similarity of taxa and the goal is generally to find a tree consistent with pairwise distance data, or character-based, where data is assumed to be a discrete array of evolutionary features characterizing the taxa (e.g., specific DNA bases found in some conserved region of all taxa) and the goal is generally to find a tree describing specific evolutionary changes (mutations) along tree edges, explaining how the observed taxa might have evolved from a common ancestor.

Models of evolution typically follow from the type of data available. The simplest use the principle of parsimony, i.e., that simpler trees are more likely to be correct. For distance-based trees, parsimony leads to a model called minimum evolution [27], in which one seeks to minimize the total amount of evolutionary distance in the tree. In the character-based case, it is known as maximum parsimony, in which one seeks to minimize numbers or weights of mutations across the tree [28]. Parsimony models are often deprecated in favor of probabilistic models capturing more detailed quantitative representations of a model of evolution. Most commonly used are maximum-likelihood (ML) models, in which one seeks a tree that maximizes the probability of the tree having produced the observed data [29]. A more sophisticated alternative is Bayesian modeling [30], which uses a similar sort of probabilistic model but samples over possible ranges of model parameters to provide a more nuanced picture of uncertainty in the phylogeny inference.

Just as models must be designed to suit the available data, algorithms are determined by both data and models. Phylogenetic inference in all common models is computationally intractable, meaning that there are no known computer algorithms that can reliably find the most plausible phylogeny for a given data set and evolutionary model. In practice, then, a variety of heuristic methods are usually used. For distance-based phylogenetics, more tractable simplifications of the major models are generally used, such as the neighbor-joining approximation to minimum evolution[31]. Character-based models commonly use variants of Markov chain Monte Carlo (MCMC) methods, an approach for randomly searching sets of possible trees [32]. Trade-offs are often required between model realism and tractability, with ML often favored as a good compromise between more realistic but computationally difficult Bayesian models and more tractable but less realistic parsimony models. Many software packages are available for the major phylogeny algorithms, however, that make it easy to access effective phylogenetics without needing a deep understanding of the algorithmic theory behind it. Popular codes include PHYLIP [33], PAUP[34], RAxML [35], and MRBAYES[36].

2.3 Adapting Phylogenetic Methods to Tumor Evolution The application of phylogenetics to tumor data is a promising approach to detailing the multiple interacting events underlying tumor progression. It relies on the key observations

that cancer is at its heart an evolutionary phenomenon and a tumor is an evolving system, suggesting that computational methods for reconstructing evolutionary systems should in principle provide a way to reconstruct processes of tumor evolution (Fig. 2.2). The field of tumor phylogenetics was introduced by Desper et al. in 1999 in a pioneering paper [37] setting out the concept of tumor phylogenies, also called oncogenetic trees. Since this seminal work, methods for tumor phylogenetics have been extended to many data sources, evolutionary models, and algorithms. Fig. 2.3 summarizes the basic steps of a generic tumor phylogeny pipeline, and Fig. 2.5 summarizes the major contributions in the field. Fig. 2.4 provides a simple example of the possible output of such a tumor phylogeny inference, in this case an unrooted neighbor-joining tumor phylogeny built from a set of primary breast tumor samples with multiple sections from each of several individual biopsies [38].

2.3 Tumor Phylogenetics

2.3.1 Phylogenetics at the Tissue Level

The first approaches to tumor phylogenetics relied on a model of treating distinct tumors as species and seeking a tree among the tumors. Although tumors in distinct patients are not literally descended from common ancestors, they nonetheless can be explained by descent from common ancestors provided they undergo similar pathways of evolution. By observing how different tumors group into a tree, one can in principle identify tumors with similar molecular bases, identify sequences of events not apparent from examination of individual tumors, and project early decision points in the evolution of tumors.
The earliest approaches depended on pre-genomic methods for assaying tumor state. The original work of Desper et al. [37], for example, used comparative genomic hybridization (CGH) data probing copy number gains and losses of large genomic regions preselected for their relevance to tumor progression. Similar coarse-grained karyotyping approaches were employed by a variety of other studies on tumor phylogenetics [39, 40, 41] More sophisticated, but still essentially pre-genomic, approaches emerged to allow a more nuanced portrait of tumor evolution, for example, through the use of microsatellite typing to detect allelic imbalances inaccessible to CGH methods [42]. Other studies have also explored how multiple heterogeneous types of marker could be combined in a single tumor phylogenetic profile, e.g., combining point mutations of target genes, epigenetic methylation markers, and microsatellite instability into a unified model of progression [43].

Several computational approaches were attempted for analyzing these pre-

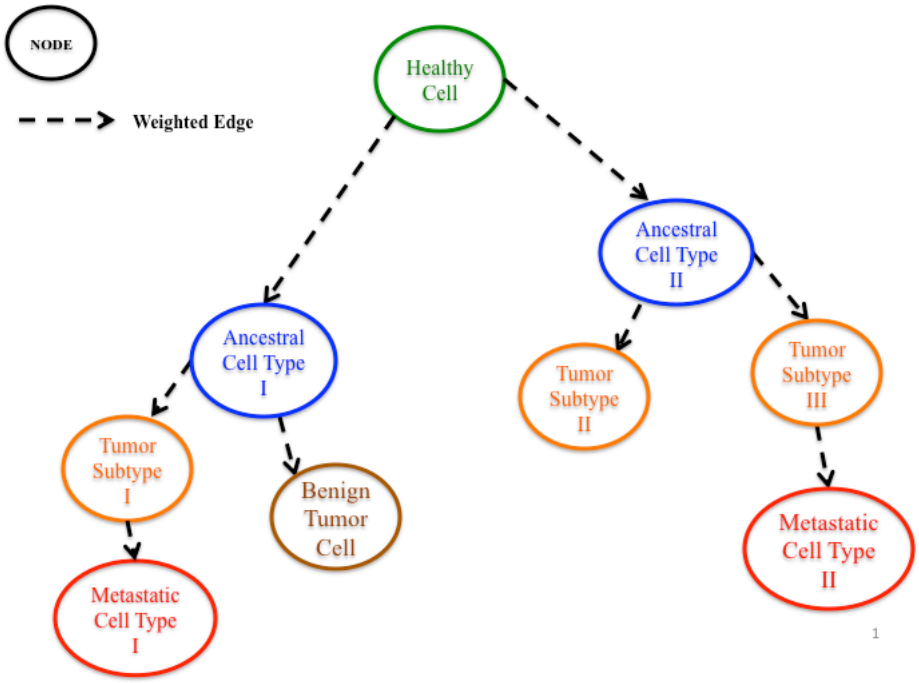


Figure 2.2: A tumor phylogeny represents a model of possible trajectories of evolution through discrete stages of progression as a normal healthy tissue transforms into malignant tissue with increasing degrees of aggression. Here, the ovals represent nodes of a rooted phylogeny and the dashed arrows represent weighted edges. The weights can be changes in mutations or copy number variations or structural variations. The green circle, which is the root, represents a healthy cell and it can acquire genomic changes to attain either ancestral cell type I or II. Ancestral cell type I can further mutate to either acquire the status of tumor subtype I or become a benign tumor cell. Tumor subtype I can in turn mutate to acquire an aggressive metastatic cell type I. Similarly, ancestral cell type II can acquire distinct mutations to become either tumor subtype II or III. Tumor subtype III is more aggressive and can mutate to become metastatic cell type II.

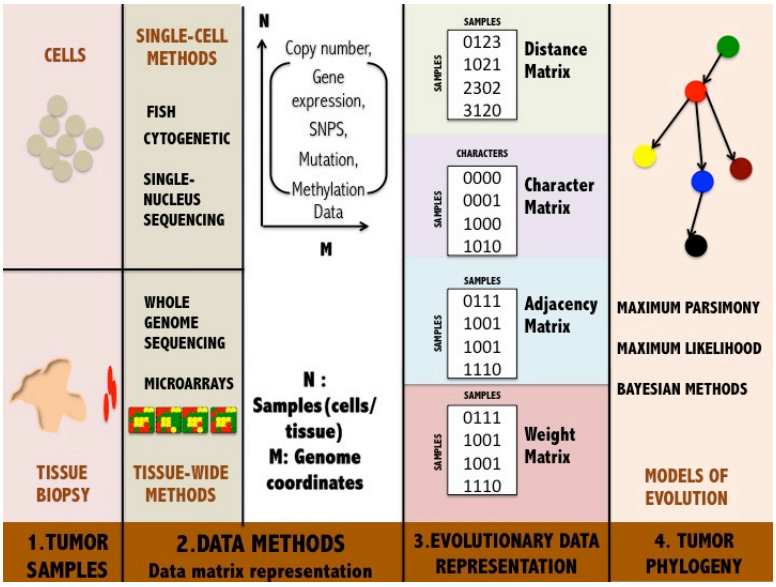


Figure 2.3: A general tumor phylogeny pipeline. In Step 1, tumor samples are collected from patients. The samples are organized either at the single-cell resolution or the whole tissue resolution as biopsies or blood samples. In Step 2, evolutionary data is obtained from the raw tumor tissue sample using either single-cell based methods like FISH or single-nucleus sequencing or tissue-wide methods like microarray experiments or next-generation sequencing. In Step 3, the evolutionary data is processed into a phylogeny data matrix. The data format may vary depending on whether distance-based or character-based methods are used for phylogenetics. In Step 4, a tumor phylogeny inference algorithm supporting the underlying model of evolution is applied to generate the tumor phylogeny.

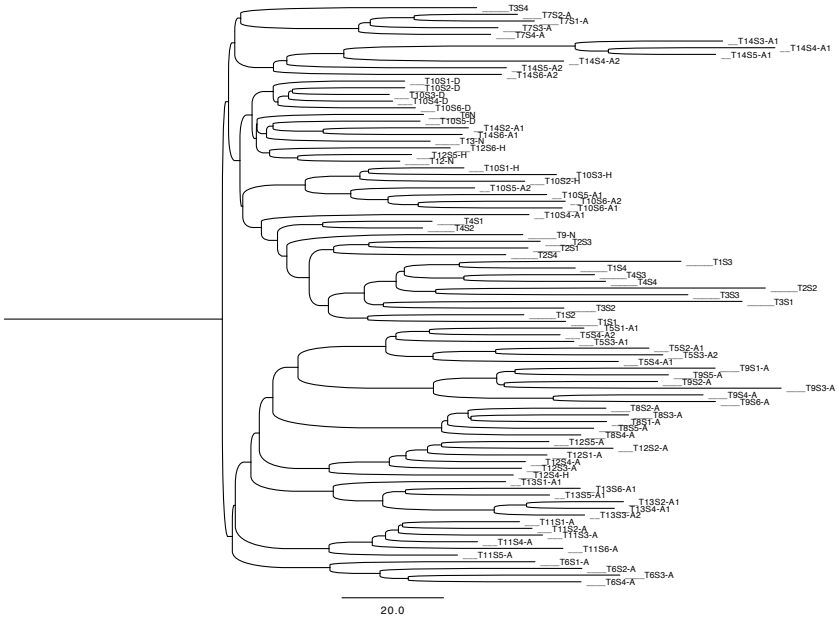


Figure 2.4: An example tumor phylogeny. The data consisted of log DNA copy number ratios from aCGH experiments on 87 tumor sections derived from 14 primary ductal breast tumors [64]. Phylogenies were built using the phylogeny reconstruction software PHYLIP to build an unrooted neighbor-joining tree using Euclidean distances between log copy number vectors. In the tree, leaves are labeled by section (S) number and tumor (T) sample number (e.g., T1S1 indicates section 1 from tumor 1). N stands for normal and A stands for aneuploidy based on additional information on ploidy available in the primary reference for the data.

Approach No	Tumor data retrieval technology	Molecular Data Type	Evolutionary Data type	Model of evolution	Phylogeny Algorithm	Reference
1.	Tissue wide: CGH	Chromosome copy number values	Matrix of weights (joint probabilities)	Oncogenetic Tree model	Chow-Liu	Desper et al. 1999, Simon et al. 2000, Jiang et al. 2000, Huang et a 2002
2.	Tissue wide: CGH	Chromosome copy number values	Matrix of weights (joint probabilities)	Minimum Evolution	Fitch and Neighbor(PHYLIP)	Desper et al. 2000
3.	Tissue wide: CGH	Chromosome Break-point data	Matrix of patterns of correlations with noise correction	Oncogenetic Tree model	Graph model	Szabo et al. 2002
4.	Tissue-wide: Microarray	Gene expression data	Distance-matrix	Minimum evolution	Neighbor-joining, Weighted Least Squares, FastME	Desper et al. 2004
5.	Tissue-wide: Microarray	Gene expression data	Distance-matrix	Minimum evolution	Neighbor-joining, Weighted Least Squares, FastME	Desper et al. 2004
6.	Single cell: cytogenetic data	Copy number, structural rearrangement	Matrix of statistical dependencies	Maximum Likelihood Estimation	Variant of Felsenstein 1981.	von Heydebreck et al. 2004
7.	Tissue-wide: CGH, Loss of Heterozygosity (LOH) scores	Copy number	Character –based feature matrix	K – Mutagenic Trees mixture model	Mtreemix	Beerenwinkel et al. 2005, Rahnenfuhrer et al. 2005
8.	Tissue-wide	Micro-satellite data	Matrix of probabilistic weights	Minimum evolution	Maximum weight branching algorithm, distance based	Chen et al. 2006
9.	Single-cell : FISH	Copy number	Character-based	Minimum evolution	Minimum Spanning Tree inference	Pennington et al. 2006
10.	Tissue Wide: Microarray	Gene Expression	Character-based using polarity assessment	Maximum Parsimony	Heuristic-based (PHYLIP)	Abu-Asab et al. 2008
11.	Tissue wide: CGH	Chromosome copy number values	Matrix of pairwise correlations	Oncogenetic Tree model with error correction	Chow-Liu	Sweeney et al. 2009
12.	Tissue-wide: microarray	Gene expression	Distance-based	Maximum likelihood Bayesian model	Minimum Spanning Tree inference	Shackney and Schwartz. 2010
13.	Regionally sectioned tissue-wide: SPP	Copy Number	Distance-based	Minimum evolution	Neighbor-joining	Navin et al. 2010
14.	Tissue-wide: Microarray	Differential Gene Expression	Distance-based	Minimum evolution	Neighbor-joining	Riester et al. 2010
15.	Tissue-wide: aCGH	Copy Number	Character-based	Maximum Parsimony	Heuristic-based (PAUP)	Subramanian et al. 2011
16.	Single-cell genome sequencing	Copy number	Distance based	Minimum evolution.	Neighbor-joining	Navin et al. 2012
17.	Genome and exome sequencing of regional sections	Mutation profile, copy number	Character-based	Based on clonal ordering	Empirically derived	Gerlinger et al. 2012
18.	Deep sequencing of reads	Copy Number	Character-based	Graph -theoretical	Allelic and connectivity graphs	Greenman et al. 2012

Figure 2.5: Summary of major contributions in the field of tumor phylogenetics. This table lists the various tumor phylogenetic approaches sorted by categories of tumor data retrieval technology, the molecular data type, the evolutionary data types, the model of evolution, the phylogeny algorithms employed and the corresponding publications.

genomic data types. The earliest work on oncogenetic trees [37, 39] used a custom character-based likelihood model and adapted a classic combinatorial optimization algorithm [44]to find trees under this model. A number of theoretical results later improved on this approach, for example, by extending it to error-prone data [45, 46] and by developing more sophisticated probabilistic [47] and combinatorial [48] frameworks in which to model patterns of correlations among frequent mutations. An alternative approach using distance-based methods was soon developed to account for dependencies between mutational events and to bring to bear many sophisticated algorithms already available for distance-based phylogeny inference [39, 40, 49]. A significant advance in tumor phylogenetics came with the appearance of whole-genome data sets, predominantly microarrays, which allowed much larger marker sets as well as de novo discovery of significant progression markers. Expression microarrays were adopted for this task [50], with phylogenetic methods providing a generalization of clustering methods and then being applied for tumor classification [51, 52, 53] to give a finer-scale hierarchical classification as well as predictions of early stages along progression of distinct subtypes. A variety of other similar array data types have been used for studies of tumor classification, including SNP arrays [54], methylation arrays [55] and aCGH arrays [56].

The move to whole-genome data sets had important implications for algorithms for phylogenetics, primarily by encouraging a shift to distance-based methods better able to deal with the much larger numbers of markers available [50]. Additional algorithmic work has aimed to adapt phylogenetic approaches more specifically to tumor development. One novel problem is proper normalization of evolutionary distances, since large changes in global expression patterns do not necessarily correspond to large numbers of mutations, a problem partially addressed by measuring evolutionary distance by changes at the level of inferred expression modules rather than genes [57] and by establishing novel tumor evolutionary distance scales based on degrees of differentiation [58]. Another problem has been rooting tumor phylogenies, addressed through novel strategies for outgrouping, a common phylogenetic technique for determining the root of a tree by adding a species distant from those under study [59, 60].

2.3.2 Phylogenetics at the Cellular Level

An alternative approach to tumor phylogenetics arose from a somewhat different conception of the problem focused on building phylogenies of single tumors based on cell-to-cell heterogeneity within single tumors. The major source of such data for this strategy has been fluorescence in situ hybridization (FISH), generally used to probe copy numbers of specific genes, genomic regions, or chromosomes of interest. Such data can provide a detailed and accurate profile of copy number

variations cell-to-cell in a tumor, allowing one to reconstruct profiles of fine steps in tumor development that would not be seen at the population resolution. Such data is, however, limited to small numbers of markers per cell, ranging from two in the earliest studies [17, 61] to eight probes per cell in the most recent [62]. Superficially, the algorithmic problem of phylogenetics from FISH data is similar to that of phylogenetics on other forms of discrete copy number data. Algorithms for this kind of data used similar forms of character-based phylogenetics to those of pre-genomic copy number studies on whole tumors, although with special handling of some kinds of evolutionary mechanism characteristic of tumor evolution, such as aneuploidy [63, 64]. In addition, the cell-level resolution resulted in treating inference of phylogenies within single patients as a separate problem from synthesizing phylogenies from distinct tumors into a model of common trends across a population. While the former was solved by phylogenetic tree-building algorithms, the latter required a simple variant of ideas from two other forms of common phylogenetic algorithm: supertree algorithms [65], which join simple trees into more complex models, and consensus tree algorithms [66], which identify common features among a set of trees.

These single-cell approaches provide a way around a major challenge facing tissue-level methods: turning intratumor heterogeneity from a confounding factor into a valuable source of information about likely pathways of fine-scale progression. They are, however, severely limited by the difficulty of probing more than a few markers of state in single cells. Such a limitation precludes discovery of novel markers and limits the complexity of models one can build, a significant disadvantage relative to the genomic methods available for tissue-level analysis.

2.3.3 Bridging the Gap Between Tissue and Cell Data

Tissue-level and cell-level analyses each bring advantages and disadvantages, prompting a search for new methods capable of bringing the benefits of both. Some experimental strategies were developed to better control for intratumor heterogeneity in phylogenetic studies. One strategy is to use cultured cell lines rather than primary tumors [67], providing a more uniform set of primary data from which to perform phylogenetics, although at the cost of having a data that may be poorly representative of the diversity and heterogeneity found within and between primary tumors. Two important steps were made in an influential paper that advanced the importance of tumor heterogeneity to understanding progression [38]: more precise subsampling of biopsies by microdissection and the use of cell sorting to subdivide tumor cell populations by ploidy, each allowing generation of more uniform cell populations for phylogenetic analysis. An alternative approach to bridging cell- and tissue-level analyses was computational inference of uniform populations from heterogeneous samples. A theoretical basis for such analysis was developed based

on earlier pre-genomic assays of tumors by computationally modeling tumor data as products not of pure evolutionary trees but as mixtures of distinct samples from trees, known as mutagenic tree models [68, 69, 70] suitable for relatively small numbers of markers. These methods have been applied, for example, to reconstruct progression pathways from karyotyping data on meningiomas [71]. Mixture modeling emerged as a computational strategy for genome-scale analysis, in which one attempts to explain multiple raw array data sets as mixtures of cell types [72, 73] to be inferred computationally, an approach similar to one previously used to control for contamination of tumor data by normal cells [74] and applied for discovery of likely sites of origin of tumors [75, 76] Further algorithmic advances adapted these initially distance-based methods into character-based by segmenting inferred mixture components [77, 78], enabling both phylogenetic inference and discovery of phylogenetically informative markers from whole-genome array data.

These hybrid approaches represented a step forward for tumor phylogenetic studies, enabling in principle the simultaneous study of whole-genome marker sets and cellular-scale evolutionary processes. In practice, however, highly noisy data sets, limited amounts of data, and uncertainty introduced at every stage of computational processing hinder the discovery of detailed, accurate models of tumor phylogenetics. In the face of high heterogeneity both within and between tumors, new advances were needed in data generation and data processing to build accurate and detailed models of tumor evolution.

2.3.4 Profiling Tumors by NGS

Next-generation sequencing offers the promise of massive throughput data, high resolution, and extensive detail. Large amounts of NGS data have been gathered in a series of major patient studies [79, 80] while continuing improvements in sequencing technology are making it easier and cheaper to incorporate NGS data into more patient studies. NGS data provides a way to mitigate several challenges facing earlier studies by allowing typing and discovery of markers at arbitrary resolution, providing more accurate quantitation of copy numbers, and allowing measurement of many kinds of data (copy number, mutation, structural rearrangement, and epigenetic) from the same basic technology. At the same time, NGS data create substantial new problems for phylogenetic analysis due to computational and statistical challenges of handling the much larger marker sets NGS produces and the difficulty of interpreting some forms of NGS data (e.g., structural rearrangements) phylogenetically.

Despite these challenges, phylogenetics from NGS data is not in principle substantially different from that with pre-NGS genomic data types. Additional

preprocessing is, however, required. Typically, a first step is mapping NGS data to a reference genome. Data spanning different kinds of variation will then require custom preprocessing. The most common approach for dealing with NGS copy number data has been to bin the DNA sequence data into windows and then determine copy numbers for the windows. Several algorithms have been published that delve deeper into the nuances of segmenting the genome data [81, 82]. For structural rearrangement information, paired-end sequencing is commonly needed to reveal sources of novel chimeric sequences [83, 84, 85]. Processing mutation data follows the model of haplotype inference in population data where after alignment to the genome, genotypes are inferred and variants identified [86, 87]. RNA-seq follows a similar approach of measuring counts at individual base locations to estimate fine-scale expression levels [88, 89, 90].

2.3.5 Phylogenetics on Whole Tumor Sequencing

Surprisingly, few NGS tumor phylogenetic studies have yet been published given the amount of data now available. NGS has, however, already enabled some important new directions in tumor phylogenetics. For example, NGS has made it possible to reconstruct rearrangement phylogenies [91], revealing an incredibly complex and diverse landscape of rearrangements across tumor types that was invisible to prior genomic methods. As NGS has helped reveal the extent of genomic aberrations, it has also made apparent the need to deal with intratumoral heterogeneity. Advances in profiling tumor state from the pre-NGS era have thus been adapted and extended for NGS data. Similar regional sectioning and ploidy profiling techniques to those that began to reveal tumor heterogeneity at the whole-genome level in the pre-NGS era [68] have been developed for NGS studies. An important recent study in NGS tumor phylogenetics [18] showed some of the power of these methods through NSG studies of regionally sectioned renal carcinomas, revealing a complex landscape characterized by high intratumor heterogeneity. Clonal ordering [92], a simple model of tumor evolution, makes it possible to reconstruct evidence for multiclonal progression pathways from such data. Such regional profiling studies provide substantial power to characterize heterogeneity beyond that visible at the whole tumor level but nonetheless provide only a partial solution to the problem of separating heterogeneous signals in even small subregions of tumors. Single-cell studies [61, 93] have shown that far greater heterogeneity is present than can be resolved by even a fine-grained microdissection.

A key advance for tumor studies then has been the development of single-cell sequencing. By sequencing individual cells from single tumors, it becomes possible to systematically assay large numbers of markers of varying types within single tumors. The first approaches to this task made it possible to reconstruct

likely tumor phylogenies largely from relatively coarse-grained data on aneuploidy of individual cells [94]. Further studies have extended the range and quality of the data, for example, by allowing one to recover nearly complete spectra of point mutations from which it is possible to reconstruct detailed models of progression of single tumors [95, 96]. Single-cell sequencing technology has been applied in other contexts also to single cell expression profiling [97], although not yet to our knowledge to tumor phylogenetics.

Despite the huge promise of single-cell sequencing, there remain substantial technological gaps before we can truly characterize tumor genomes accurately at the single-cell level. A key problem is the need for single-cell genome amplification, which introduces biases, both systematic and random, in measured copy numbers, and potentially nonspecific amplification of contaminants and introduction of chimeric sequences that might be confused with true translocations [98]. Technological improvements such as multidisplacement amplification (MDA) [99, 100] are reducing the systematic bias in amplification over early whole-genome amplification (WGA) approaches [94], but cannot overcome the inherent randomness of the process. Furthermore, adequate scalable computational methods are still lacking for key problems of interpreting single-cell NGS data, such as reconstructing detailed genomic rearrangement phylogenies.

2.4 Conclusions

Phylogenetics provides a crucial set of tools for making sense of the evolutionary processes that underlie tumor development in the face of the flood of data that NGS has unleashed, but major challenges remain to realize its potential. Cancer biology, like much of biological and medical research, has transformed in a span of a few years to a data-driven field dependent on computational algorithms for managing data and drawing meaningful inferences from it. Phylogenetic theory provides a framework and a set of models and algorithms well suited to understanding what is at its heart an evolutionary system. This framework has advanced alongside technologies for probing tumor state, helping us to assemble profiles of the common ways tumors develop across patient populations. With the NGS era, these computation methods have never been more necessary. However, major challenges confront both sequencing technologies and phylogenetic inferences if we are to truly reconstruct in detail the common evolutionary trajectories that underlie tumor development across cancers, subtypes, and patients.

Chapter 3

Resolving tumor heterogeneity via computational unmixing

In this chapter, we discuss a computational unmixing method for inferring heterogeneity in whole genome tumor data. We analyze the resulting heterogeneous cell types and propose approaches for validating the same by comparing with Fluorescent in-situ hybridization (FISH) data. This work was done in collaboration with David Tolliver, Charalampous Tsourakakis, Stanley Shackney and Russell Schwartz. Specifically, David Tolliver designed and implemented the robust unmixing optimization scheme and performed experiments on synthetic data. Ayshwarya Subramanian designed and performed experiments on real data, analyzed the results, and designed and implemented a scheme for comparison of the unmixing results with FISH data. Stanley Shackney and Russell Schwartz were involved in posing the heterogeneity inference problem and in the overall design and analysis of the unmixing scheme. Partial contents of this chapter are adapted from [73].

Genomic studies have dramatically improved our understanding of the biology of tumor formation and treatment. In part this has been accomplished by harnessing tools that profile the genes and proteins in tumor cells, revealing previously indistinguishable tumor sub-types that are likely to exhibit distinct sensitivities to treatment methods [23, 20, 22, 101]. As these tumor sub-types are uncovered, it becomes possible to develop novel therapeutics more specifically targeted to the particular genetic defects that cause each cancer [25, 102, 103]. While recent advances have had a profound impact on our understanding of tumor biology, the limits of our understanding of the molecular nature of cancer obstruct the burgeoning efforts in "targeted therapeutics" development. These limitations are apparent in the high failure rate of the discovery pipeline for novel cancer therapeutics [104] as well as in the continuing difficulty of predicting which patients will respond to a given therapeutic. A striking example is the fact that traztuzumab, the targeted therapeutic developed to treat HER2-amplified breast cancers, is ineffective in many patients who have HER2-overexpressing tumors and yet effective in some who do not [105]. Furthermore, sub-types typically remain poorly defined — *e.g.*, the "basal-like" breast cancer sub-type, for which different studies have inferred very distinct genetic signatures [20, 22, 24] — and yet many patients do not fall into any known sub-type. Our belief, then, is that clinical treatment of cancer will reap considerable benefit from the identification of new cancer sub-types and genetic signatures.

One promising approach for better elucidating the common mutational patterns by which tumors develop is to recognize that tumor development is an evolutionary process and apply phylogenetic methods to tumor data to reveal these evolutionary relationships. Much of the work on tumor evolution models flows from the seminal efforts of [37] on inferring *oncogenetic trees* from comparative genomic hybridization (aCGH) profiles of tumor cells. A strength in this model stems from the extraction of ancestral structure from many probe sites per tumor, potentially utilizing measurements of the expression or copy number changes across the entire genome. However, this comes at the cost of overlooking the diversity of cell populations within tumors, which can provide important clues to tumor progression but are conflated with one another in tissue-wide assays like aCGH.

3.1 Inferring tumor heterogeneity using geometric unmixing

The cell-by-cell approaches, such as [64, 106], use this heterogeneity information but at the cost of allowing only a small number of probes per cell. In recent work, Schwartz and Shackney [72] proposed bridging the gap between these two methodologies by computationally inferring cell populations from tissue-wide gene expression samples. This inference was accomplished through "geometric unmixing," a mathematical formalism of the problem of separating components of mixed samples in which each observation is presumed to be an unknown convex combination[1] of several hidden fundamental components. Other approaches to inferring common pathways include mixture models of oncogenetic trees [69], PCA-based methods [107], conjunctive Bayesian networks [108] and clustering [109].

Unmixing falls into the class of methods that seek to recover a set of pure sources from a set of mixed observations. Analogous problems have been coined "the cocktail problem," "blind source separation," and "component analysis" and various communities have formalized a menagerie of models with distinct statistical assumptions. In a broad sense, the classical approach of principal component analysis (PCA) [110] seeks to factor the data under the constraint that, collectively, the fundamental components form an orthonormal system. Independent component analysis (ICA) [111] seeks a set of statistically independent fundamental components. These methods, and their ilk, have been extended to represent non-linear data distributions through the use of kernel methods (see [112, 113] for details), which often confound modeling with black-box data transformations. Both PCA and ICA break down as pure source separators when the sources exhibit

[1] A point p is a convex combination combination of basis points $v_0, ..., v_k$ if and only if the constraints $p = \sum_{i=0}^{k} \alpha_i v_i$, $\sum_i \alpha_i = 1$ and $\forall i \ : \ \alpha_i \geq 0$ obtain. The fractions α_i determine a mixture over the basis points $\{v_i\}$ that produce the location p.

a modest degree of correlation. Collectively, these methods place strong independence constraints on the fundamental components that are unlikely to hold for tumor samples, where we expect components to correspond to closely related cell states.

The structure of our present inference problem, that of extracting multiple correlated fundamental components, has motivated the development of new methods for unmixing genetic data. Similar unmixing methods were first developed for tumor samples by Billheimer and colleagues [74] to improve the power of statistical tests on tumor samples in the presence of contaminating stromal cells. Similarly, a hidden Markov model approach to unmixing was developed by Lamy *et al.* [114] to correct for stromal contamination in DNA copy number data. These recent advances demonstrate the feasibility of unmixing-based approaches for separating cell sub-populations in tumor data. Outside the bioinformatics community, geometric unmixing has been successfully applied in the geo-sciences [115] and in hyper-spectral image analysis [116].

The recent work by [72] applied the hard geometric unmixing model (see §3.1.1.1) to gene expression data with the goal of recovering expression signatures of tumor cell sub-types, with the specific goal of facilitating phylogenetic analysis of tumors. The results showed promise in identifying meaningful sub-populations and improving phylogenetic inferences. They were, however, hampered by limitations of the hard geometric approach, particularly the sensitivity to experimental error and outlier data points caused by the simplex fitting approach. An example of simplex fitting in the plane is shown in Figure 3.1, illustrating why the strict containment model used in [115, 116, 72] is extremely sensitive to noise in the data. In the present work we introduce a soft geometric unmixing model (see §3.1.1.2) for tumor mixture separation, which relaxes the requirement for strict containment using a fitting criterion that is robust to noisy measurements. We develop a formalization of the problem and derive an efficient gradient-based optimization method. We develop this method specifically for analyzing tissue-wide DNA copy number data as assessed by array comparative genomic hybridization (aCGH) data. We demonstrate the value of the soft unmixing model by comparison to a hard unmixing method on synthetic and real aCGH data. We apply our method to an aCGH data set taken from [38] and show that the method identifies state sets corresponding to known sub-types consistent with much of the analysis performed by the authors.

The data are assumed to be given as g genes sampled in s tumors or tumor sections. The samples are collected in a matrix, $M \in \Re^{g \times s}$, in which each row corresponds to an estimate of gene copy number across the sample population obtained with aCGH. The data in M are processed as raw or baseline normalized raw input, rather than as $\log$ ratios. The "unmixing" model, described below, asserts that each sample m_i, a column of M, be well approximated by a convex combination of a fixed set of $C = [c_0|...|c_k]$ of $k+1$ unobserved basis distributions over the gene measurements. Further, the observed measurements are assumed

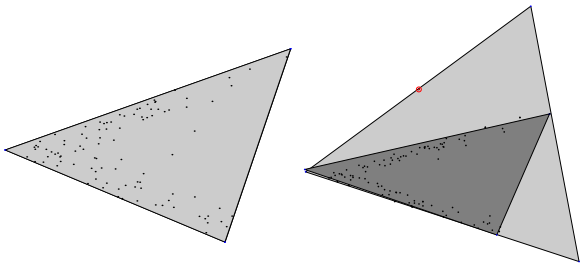


Figure 3.1: *Left:* The minimum area fit of a simplex containing the sample points in the plane (shown in black) using the program in §3.1.1.1. On noiseless data, hard geometric unmixing recovers the locations of the fundamental components at the vertices. *Right:* However, the containment simplex is highly sensitive to noise and outliers in the data. A single outlier, circled above, radically changes the shape of the containment simplex fit (light gray above). In turn, this changes the estimates of basis distributions used to unmix the data. We mitigate this short coming by developing a soft geometric unmixing model (see §3.1.1.2) that is comparatively robust to noise. The soft fit (shown dark gray) is geometrically very close to the generating sources as seen on the left.

to be perturbed by additive noise in the $\log$ domain, *i.e.*:

$$m_i = b^{\log_b(CF_i)+\eta}$$

where F_i is the vector of coefficients for the convex combination of the $(k+1)$ basis distributions and η is additive zero mode *i.i.d.* noise.

3.1.1 Algorithms and Assumptions

Given the data model above, the inference procedure seeks to recover the $k+1$ distributions over gene-copy number or expression that "unmix" the data. The procedure contains three primary stages:

1. Compute a reduced representation x_i for each sample m_i,
2. Estimate the basis distributions K_{min} in the reduced coordinates and the mixture fractions F,
3. Map the reduced coordinates K_{min} back into the "gene space" recovering C.

The second step in the method is performed by optimizing the objective in §3.1.1.1 or the robust problem formulation in §3.1.1.2.

Obtaining the reduced representation

We begin our calculations by projecting the data into a k dimension vector space (i.e., the intrinsic dimensionality of a $(k+1)-$vertex simplex). We accomplish this using principal components analysis (PCA) [110], which decomposes the input matrix M into a set of orthogonal basis vectors of maximum variance and retain only the k components of highest variance. PCA transforms the $g \times s$ measurement matrix M into a linear combination $XV + A$, where V is a matrix of the principal components of M, X provides a representation of each input sample as a linear combination of the components of V, and A is a $k \times s$ matrix in which each row contains s copies of the mean value of the corresponding row of M. The matrix X thus provides a reduced-dimension representation of M, and becomes the input to the sample mixture identification method in Stage 2. V and A are retained to allow us to later contruct estimated aCGH vectors corresponding to the inferred mixture components in the original dimension g.

Assuming the generative model of the data above, PCA typically recovers a sensible reduced representation, as low magnitude log additive noise induces "shot-noise" behavior in the subspace containing the simplex with small perturbations in the orthogonal complement subspace. An illustration of this stage of our algorithm can be found in Figure 3.2.

Sample mixture identification

Stage 2 invokes either a hard geometric unmixing method that seeks the minimum volume simplex enclosing the input point set X (Program 3.1) or a soft geometric unmixing method that fits a simplex to the points balancing the desire for a compact simplex with that for containment of the input point set (Program 3.2). For this purpose, we place a prior over simplices, preferring those with small volume that fit or enclose the point set of X. This prior captures the intuition that the most plausible set of components explaining a given data set are those that can explain as much as possible of the observed data while leaving in the simplex as little empty volume, corresponding to mixtures that could be but are not observed, as possible.

Upon completion, Stage 2 obtains estimates of the vertex locations K_{min}, representing the inferred cell types from the aCGH data in reduced coordinates, and a set of mixture fractions describing the amount of each observed tumor sample attributed to each mixture component. The mixture fractions are encoded in a $(k+1) \times s$ matrix F, in which each column corresponds to the inferred mixture fractions of one observed tumor sample and each row corresponds to the amount of a single component attributed to all tumor samples. We define F_{ij} to be the fraction of component i assigned to tumor sample j and F_j to be vector of all mixture fractions assigned to a given tumor sample j. To ensure that that the observations are modeled as convex combinations of the basis vertices, we require that $F\mathbf{1} = 1$.

Cell type identification

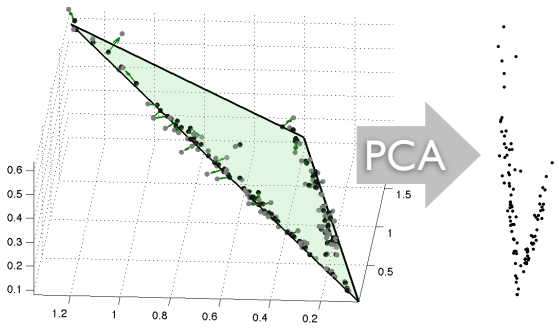


Figure 3.2: An illustration of the reduced coordinates under the unmixing hypothesis: points (show in gray) sampled from the $3-$simplex embedded are $\Re^3$ and then perturbed by log-normal noise, producing points shown in black with sample correspondence given the green arrows. Note that the dominant subspace remains in the planar variation induced by the simplex, and a 2D reduced representation for simplex fitting is thus sufficient.

The reduced coordinate components from Stage 2, K_{min}, are projected up to a $g \times (k+1)$ matrix C in which each column corresponds to one of the $k+1$ inferred components and each row corresponds to the approximate copy number of a single gene in a component. We perform this transformation using the matrices V and A produced by PCA in Stage 1 with the formula $C = V^T K_{min} + A$, augmenting the average to $k+1$ columns.

Finally the complete inference procedure is summarized in the following pseudocode:

Given tumor sample matrix M, the desired number of mixture components k, and the strength of the volume prior γ:

1. Factor the sample matrix M such that $M^T = XV + A$
2. Produce the reduced $k-$dimensional representation by retaining the top k components in X
3. Minimize Program 3.1, obtaining an estimate of the simplex K^0_{min}
4. Minimize Program 3.2 starting at K^0_{min}, obtaining K_{min} and F
5. Obtain the centers C in gene space as $C = A + V^T K_{min}$

3.1.1.1 Hard Geometric Unmixing

Hard geometric unmixing is equivalent to finding a minimum volume $(k+1)-$simplex containing a set of s points $\{X\}$ in $\Re^k$. A non-linear program for hard geometric

unmixing can be written as follows:

$$\begin{aligned} \underset{K}{min} &: \log \texttt{vol}(K) \\ \forall i &: x_i = KF_i \\ \forall F_i &: F_i^T \mathbf{1} = 1,\ F_i \succeq 0 \end{aligned} \tag{3.1}$$

where $\log \texttt{vol}$ measures the volume of simplex defined by the vertices $K \doteq [v_0|...|v_k]$ and $F \succeq 0$ requires that $\forall_{ij}.\ F_{ij} \geq 0$. Collectively, the constraints ensure that each point be expressed exactly as a unique convex combination of the vertices. Exact nonnegative matrix factorization (NNMF), see [117], can be seen as a relaxation of hard geometric unmixing. Exact NNMF retains the top two constraints while omitting the constraint that the columns of F sum to unity – thus admitting all positive combinations rather than the restriction to convex combinations as is the case for geometric unmixing.

Approximate and exponential-time exact minimizers are available for Program 3.1, in our experiments we use the approach of [116], which sacrifices some measure of accuracy for efficiency.

3.1.1.2 Soft Geometric Unmixing

Estimates of the target distributions, derived from the fundamental components (simplex vertices), produced by hard geometric unmixing are sensitive to the wide-spectrum noise and outliers characteristic of log-additive noise (*i.e.*, multiplicative noise in the linear domain). The robust formulation below tolerates noise in the sample measurements m_i and subsequently in the reduced representations x_i, improving the stability of these estimates. The sensitivity of hard geometric unmixing is illustrated in Figure 3.1. The motivation for soft geometric unmixing is to provide some tolerance to experimental error and outliers by relaxing the constraints in Program 3.1 allowing points to lie outside the boundary of the simplex fit to the data. We extend Program 3.1 to provide a robust formulation as follows:

$$\begin{aligned} \underset{K}{min} &: \sum_{i=1}^{s} |x_i - KF_i|_p + \gamma \log \texttt{vol}(K) \\ \forall F_i &: F_i^T \mathbf{1} = 1,\ F_i \succeq 0 \end{aligned} \tag{3.2}$$

where the term $|x_i - KF_i|_p$ penalizes the imprecise fit of the simplex to the data and γ establishes the strength of the minimum-volume prior. Optimization of Program 3.2 is seeded with an estimate produced from Program 3.1 and refined using MATLAB's *fminsearch* with analytical derivatives for the $\log \texttt{vol}$ term and an LP-step that determines mixtures components F_i and the distance to the boundary for each point outside the simplex.

We observe that when taken as whole, Program 2 can be interpreted as the

negative log likelihood of a Bayesian model of signal formation. In the case of array CGH data, we choose $p = 1$ (*i.e.*, optimizing relative to an ℓ_1 norm), as we observe that the errors may be induced by outliers and the ℓ_1 norm would provide a relatively modest penalty for a few points far from the simplex. From the Bayesian perspective, this is equivalent to relaxing the noise model to assume *i.i.d.* heavy-tailed additive noise. To mitigate some of the more pernicious effects of log-normal noise, we also apply a total variation-like smoother to aCGH data in our experiments. Additionally, the method can be readily extended to weighted norms if an explicit outlier model is available.

3.1.1.3 Analysis & Efficiency

The hard geometric unmixing problem in §3.1.1.1 is a non-convex objective in the present parameterization, and was shown by [118] to be NP-hard when $k + 1 \geq \log(s)$. For the special case of minimum volume tetrahedra ($k = 3$), [119] demonstrated an exact algorithm with time complexity $\Theta(s^4)$ and a $(1+\varepsilon)$ approximate method with complexity $O(s + 1/\varepsilon^6)$. Below, we examine the present definition and show that Programs 3.1 and 3.2 have structural properties that may exploited to construct efficient gradient based methods that seek local minima. Such gradient methods can be applied in lieu of or after heuristic or approximate combinatorial methods for minimizing Program 3.1, such as [115, 116] or the $(1+\varepsilon)$ method of [119] for simplexes in $\Re^3$.

We begin by studying the volume penalization term as it appears in both procedures. The volume of a convex body is well known (see [120]) to be a log concave function. In the case of a simplex, analytic partial derivatives with respect to vertex position can used to speed the estimation of the minimum volume configuration K_{min}. The volume of a simplex, represented by the vertex matrix $K = [v_0|...|v_k]$, can be calculated as:

$$\texttt{vol}(K) = c_k \cdot \texttt{det}\left(\Gamma^T K K^T \Gamma\right)^{1/2} = c_k \cdot \texttt{det}\ Q \tag{3.3}$$

where c_k is the volume of the unit simplex defined on $k+1$ points and Γ is a fixed vertex-edge incidence matrix such that $\Gamma^T K = [v_1 - v_0|...|v_k - v_0]$. The matrix Q is an inner product matrix over the vectors from the special vertex v_0 to each of the remaining k vertices. In the case where the simplex K is non-degenerate, these vectors form a linearly independent set and Q is positive definite (PD). While the determinant is log concave over PD matrices, our parameterization is linear over the matrices K, not Q. Thus it is possible to generate a degenerate simplex when interpolating between two non-degenerate simplexes K and K'. For example, let K define a triangle with two vertices on the $y-$axis and produce a new simplex K' by reflecting the triangle K across the $y-$axis. The curve $K(\alpha) = \alpha K + (1-\alpha)K'$ linearly interpolates between the two. Clearly, when $\alpha = 1/2$, all three vertices of $K(\alpha)$ are co-linear and thus the matrix Q is not full

rank and the determinant vanishes. However, in the case of small perturbations, we can expect the simplexes to remain non-degenerate.

To derive the partial derivative, we begin by substituting the determinant formulation into our volume penalization and arrive at the following calculation:

$$\begin{aligned} \log \texttt{vol}(K) &= \log c_k + \frac{1}{2} \log \texttt{det} Q \\ &\propto \log \prod_{d=1}^{k} \lambda_d(Q) = \sum_{d=1}^{k} \lambda_d(Q) \end{aligned}$$

therefore the gradient of $\log \texttt{vol}(K)$ is given by

$$\frac{\partial \log \texttt{vol(K)}}{\partial K_{ij}} = \sum_{d=1}^{k} \frac{\partial}{\partial K_{ij}} \lambda_d = \sum_{d=1}^{k} z_d^T (\Gamma^T E_{ij} E_{ij}^T \Gamma) z_d$$

where the eigenvector z_d satisfies the equality $Q z_d = \lambda_d z_d$ and E_{ij} is the indicator matrix for the entry ij. To minimize the volume, we move the vertices along the paths specified by the negative $\log$ gradient of the current simplex volume. The Hessian is derived by an analogous computation, making Newton's method for Program 3.1, with log barriers over the equality and inequality constraints, a possible optimization strategy.

Soft geometric unmixing (Program 2) trades the equality constraints in Program 1 for a convex, but non-differentiable term, in the objective function $\sum_{i=1}^{s} |x_i - KF_i|_p$ for $p|1 \leq p \leq 2$. Intuitively – points inside the simplex have no impact on the cost of the fit. However, over the course of the optimization, as the shape of the simplex changes points move from the interior to the exterior, at which time they incur a cost. To determine this cost, we solve the nonnegative least squares problem for each mixture fraction F_i, $min_F \ : \ (KF_i - x_i)^T (KF_i - x_i)$. This step simultaneously solves for the mixture fraction, and for exterior points, the distance to the simplex is determined. The simplex is then shifted under a standard shrinkage method based on these distances.

We evaluated our methods using synthetic experiments, allowing us to assess two properties of robust unmixing 1) the fidelity with which endmembers (subtypes) are identified and 2) the relative effect of noise on hard versus robust unmixing. We then evaluate the robust method on a real world aCGH data set published by [38] in which ground truth is not available, but for which we uncover much the structure reported by the authors.

3.1.2 Methods: Synthetic Experiments

To test the algorithms given in §5.2 we simulated data using a biologically plausible model of ad-mixtures. Simulated data provides a quantitative means of evaluation

as ground truth is available for both the components C and the mixture fractions F_i associated with each measurement in the synthetic design matrix M. The tests evaluate and compare hard geometric unmixing §3.1.1.1 and soft geometric unmixing §3.1.1.2 in the presence of varying levels of log-additive Gaussian noise and varying k. By applying additive Gaussian noise in the log domain we simulate the heteroscedasticity characteristic of CGH measurements (*i.e.* higher variance with larger magnitude measurements). By varying k, the dimensionality of the simplex used to fit the data, we assess the algorithmic sensitivity to this parameter as well as that to γ governing the strength of the volume prior in Program 2. The sample generation process consists of three major steps: 1) mixture fraction generation (determining the ratio of sub-types present in a sample), 2) end-member (*i.e.* sub-type) generation and 3) the sample perturbation by additive noise in the log-ratio domain.

3.1.2.1 Mixture Sampler

Samples over mixture fractions were generated in a manner analogous to Polya's Urn Process, in which previously sampled simplicial components (*e.g.*, line segments, triangles, tetrahedra) are more likely to be sampled again. This sampling mechanism produces data distributions that are similar to those we see in low dimensional projections of aCGH data when compared against purely uniform samples over mixtures. An example of a low dimensional sample set and the simplex that was used to generate the points is shown in Figure 3.3.

To generate the mixture fractions F_i for the i^{th} sample, the individual components in C^{true} are sampled without replacement from a dynamic tree model. Each node in the tree contains a dynamic distribution over the remaining components, each of which is initialized to the uniform distribution. We then sample s mixtures by choosing an initial component according to the root's component distribution and proceed down the tree. As a tree-node is reached, its component distribution is updated to reflect the frequency with which its children are drawn. To generate the i^{th} sample, the fractional values F_i are initialized to zero. As sample generation proceeds, the currently selected component C_j updates the mixture as $Fij \sim \texttt{uniform}[(1/2)f_p^j, 1]$ where f_p^j is the frequency of j's parent node. For the i^{th} mixture, this process terminates when the condition $1 \leq \sum_{j=1}^{k+1} F_{ij}$ holds. Therefore, samples generated by long paths in the tree will tend to be homogenous combinations of the components C^{true}, where as short paths will produce lower dimensional substructures. At the end of the process, the matrix of fractions F is re-normalized so that the mixtures associated with each sample sum to unity. This defines a mixture F_i^{true} for each sample – *i.e.* the convex combination over fundamental components generating the sample point.

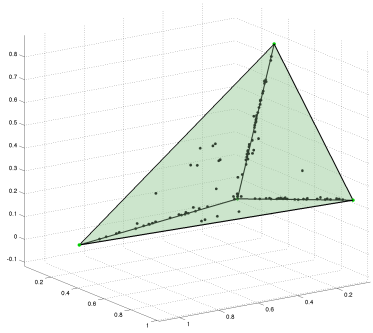


Figure 3.3: An example sample set generated for §3.1.2.2 shown in the "intrinsic dimensions" of the model. Note that sample points cleave to the lower dimensional substructure (edges) of the simplex.

3.1.2.2 Geometric Sampling of End-members & Noise

To determine the location of the end-members we specify an extrinsic dimension (number of genes) g, and an intrinsic dimension k (requiring $k+1$ components). We then simulate $k+1$ components by constructing a $g \times (k+1)$ matrix C^{true} of fundamental components in which each column is an end-member (*i.e.* subtype) and each row is the copy number of one hypothetical gene, sampled from the unit Gaussian distribution and rounded to the nearest integer. Samples m_i, corresponding to the columns of the data matrix M, are then given by:

$$m_i = 2^{log_2\left(C^{true}F_i^{true}\right)+\frac{1}{2}\sigma\eta} \tag{3.4}$$

where $\eta \sim \texttt{normal}(0,1)$ and the mixture fractions F_i^{true} were obtained as in §3.1.2.1.

3.1.2.3 Evaluation

We follow Schwartz and Shackney [72] in assessing the quality of the unmixing methods by independently measuring the accuracy of inferring the components and the mixture fractions. We first match inferred mixture components to true mixture components by performing a maximum weighted bipartite matching of columns between C^{true} and the inferred components C^e, weighted by negative Euclidean distance. We will now assume that the estimates have been permuted according to this matching and continue. We then assess the quality of the mixture component identification by the root mean square distance over all entries of all components between the matched columns of the two C matrices:

$$\texttt{error} = \frac{1}{g(k+1)}\left|\left|C^{true} - C^e\right|\right|_F^2 \tag{3.5}$$

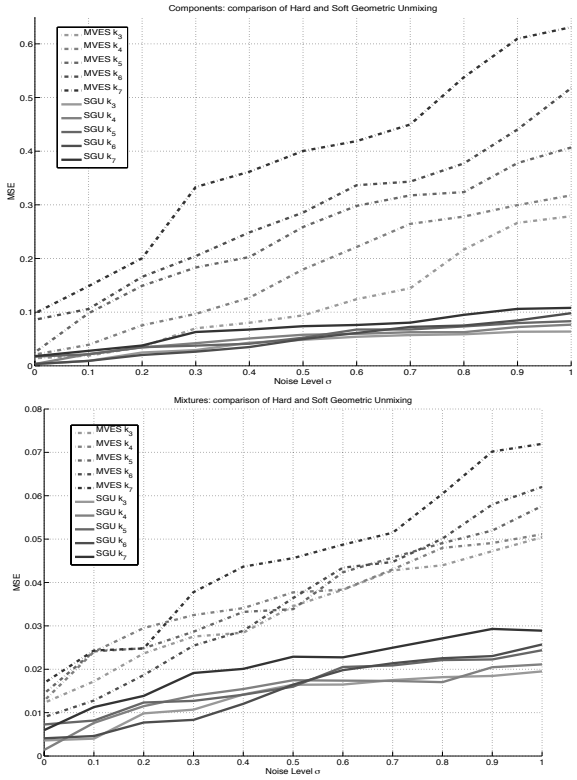


Figure 3.4: Left: mean squared error for the component reconstruction comparing Hard Geometric Unmixing (MVES: [116]) and Soft Geometric Unmixing (SGU) introduced in §3.1.1.2 for the experiment described in §3.1.2.2 with variable γ. The plot demonstrates that robust unmixing more accurately reconstructs the ground truth centers relative to hard unmixing in the presence of noise. Right: mean squared error for mixture reconstruction comparing MVES and SGU.

where $||A||_F = \sqrt{\sum_{ij} a_{ij}^2}$ denotes the Frobenius norm of the matrix A.

We similarly assess the quality of the mixture fractions by the root mean square distance between F^{true} and the inferred fractions F^e over all genes and samples:

$$\texttt{error} = \frac{1}{g(k+1)} \left|\left|F^{true} - F^e\right|\right|_F^2 . \tag{3.6}$$

This process was performed for $s = 100$ and $d = 10000$ to approximate a realistic tumor expression data set and evaluated for $k = 3$ to $k = 7$ and for $\sigma = \{0, 0.1, 0.2, ..., 1.0\}$, with ten repetitions per parameter.

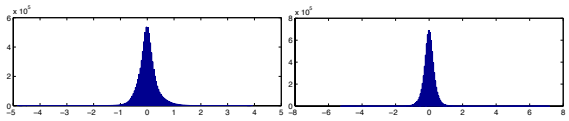


Figure 3.5: Empirical motivation for the $\ell_1 - \ell_1$−total variation functional for smoothing CGH data. The left plot shows the histogram of values found in the CGH data obtained from the [38] data set. The distribution is well fit by the high kurtosis Laplacian distribution in lieu of a Gaussian. The right plot shows the distribution of differences along the probe array values. As with the values distribution, these frequencies exhibit high kurtosis.

3.1.3 Results: Synthetic Data

The results for the synthetic experiment are summarized in Figure 3.4. The figure shows the trends in MSE for hard geometric unmixing §3.1.1.1 and soft geometric unmixing §3.1.1.2 on the synthetic data described above. As hard geometric unmixing requires that each sample lie inside the fit simplex, as noise levels increase (larger σ), the fit becomes increasingly inaccurate. Further, the method MVES deteriorates to some degree as order k of the simplex increases. However, soft geometric unmixing degrades more gracefully in the presence of noise if an estimate of the noise level is available with ± 0.1 in our current model. The trend of soft unmixing exhibiting lower error and better scaling in k than hard unmixing holds for both components and mixture fractions, although components exhibit a higher average degree of variability due to the scale of the synthetic measurements when compared to the mixture fractions.

3.1.4 Array Comparative Gene Hybridization (aCGH) Data

We further illustrate the performance of our methods on a publicly available primary Ductal Breast Cancer aCGH Dataset furnished with [38]. This dataset is of interest in that each tumor sample has been sectored multiple times during biopsy which is ideal for understanding the substructure of the tumor population. The data consists of 87 aCGH profiles from 14 tumors run on a high-density ROMA platform with 83055 probes. Profiles are derived from 4-6 sectors per tumor, with samples for tumors 5-14 sub-partitioned by cell sorting according to total DNA content, and with healthy control samples for tumors 6, 9, 12, and 13. For full details, the reader is referred to Navin et al. [38]. The processed data consists of $\log_{10}$ ratios and which were exponentiated prior to the PCA step (Stage 1) of the method.

3.1.4.1 Preprocessing

To mitigate the effects of sensor noise on the geometric inference problem we apply a total variation (TV) functional to the raw log-domain data. The $\ell_1 -$

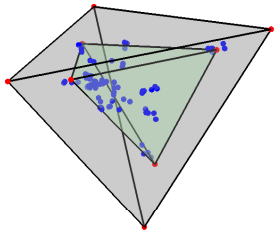


Figure 3.6: The simplex fit to the CGH data samples from [38] ductal data set in $\Re^3$. The gray tetrahedron was return by the optimization of Program 3.1 and the green tetrahedron was returned by the robust unmixing routine.

ℓ_1-TV minimization is equivalent to a penalized projection onto the over-complete Harr basis preserving a larger degree of the signal variation when compared to discretization methods (*e.g* [121, 122]) that employ aggressive priors over the data distribution. The procedure seeks a smooth instance x of the observed signal s by optimizing the following functional:

$$\min_{x} \; : \; \sum_{i=1}^{g} |x_i - s_i|_1 + \lambda \sum_{i=1}^{g-1} |x_i - x_{i+1}|_1 \tag{3.7}$$

The functional 3.7 is convex and can be solved readily using Newton's method with log-barrier functions ([120]). The solution x can be taken as the maximum likelihood estimate of a Bayesian model of CGH data formation. That is, the above is the negative log-likelihood of a simple Bayesian model of signal formation. The measurements $\hat{x}_i$ are assumed to be perturbed by the *i.i.d.* Laplacian noise and the changes along the probe array are assumed to be sparse. Recall that the Laplacian distribution is defined as $Pr(x) = \frac{1}{z} \exp \frac{-|x|}{a}$. In all experiments the strength of the prior λ was set to $\lambda = 10$. The data fit this model well as illustrated in Figure 3.5. The dimension of the reduced representation k, fixing the number of fundamental components, was determined using the eigengap heuristic during the PCA computation (Stage 1). This rule ceases computing additional principal components when the difference in variances jumps above threshold.

3.1.4.2 Unmixing Analysis and Validation

The raw data was preprocessed as described above and a simplex was fit to the reduced coordinate representation using the soft geometric unmixing method (see §3.1.1.2). A three dimensional visualization of the resulting fit is shown for the [38] data set in Figure 3.6. To assess the performance with increasing dimensionality, we ran experiments for polytope dimensionality k ranging from 3 to 9.Following the eigen-gap heuristic we chose to analyze the results for $k = 6$. The γ value was picked according to the estimated noise level in the aCGH dataset

and scaled relative to the unit simplex volume (here, $\gamma = 100$). The estimated 6 components/simplex vertices/pure cancer types are labeled $C_1, C_2, ..., C_6$.

Figure 3.7 shows mixture fraction assignments for the aCGH data for $k = 6$. While there is typically a non-zero amount of each component in each sample due to imprecision in assignments, the results nonetheless show distinct subsets of tumors favoring different mixture compositions and with tumor cells clearly differentiated from healthy control samples. The relative consistency within versus between tumors provides a secondary validation that soft unmixing is effective at robustly assigning mixture fractions to tumor samples despite noise inherent to the assay and that produced by subsampling cell populations. It is also consistent with observations of Navin *et al.*

It is not possible to know with certainty the true cell components or mixture fractions of the real data, but we can validate the biological plausibility of our results by examining known sites of amplification in the inferred components. We selected fourteen benchmark loci frequently amplified in breast cancers through manual literature search. Table 3.1 lists the chosen benchmarks and the components exhibiting at least 2-fold amplification of each. Figure 3.8 visualizes the results, plotting relative amplification of each component as a function of genomic coordinate and highlighting the locations of the benchmark markers. Thirteen of the fourteen benchmark loci exhibit amplification for a subset of the components, although often at minimal levels. The components also show amplification of many other sites not in our benchmark set, but we cannot definitively determine which are true sites of amplification and which are false positives. We further tested for amplification of seven loci reported as amplified by Navin et al. [38] specifically in the tumors examined here and found that six of the seven are specifically amplified in one of our inferred components: PPP1R12A (C_2), KRAS (C_2), CDC6 (C_2), RARA (C_2), EFNA5 (C_2), PTPN1 (C_3), and LPXN (not detected). Our method did not infer a component corresponding to normal diploid cells as one might expect due to stromal contamination. This failure may reflect a bias introduced by the dataset, in which many samples were cell sorted to specifically select aneuploid cell fractions, or could reflect an inherent bias of the method towards more distinct components, which would tend to favor components with large amplifications.

We repeated these analyses for the hard unmixing with a higher amplification threshold due to the noise levels in the centers. It detected amplification at 11 of the 14 loci, with spurious inferences of deletion at four of the 11. For the seven sites reported in Navin *et al.*, hard unmixing identified five (failing to identify EFNA5 or LPXN) and again made spurious inferences of deletions for three of these sites, an artifact the soft unmixing eliminates. The full results are provided in supplementary section S1. The results suggest that hard unmixing produces less precise fits of simplexes to the true data.

We can also provide a secondary analysis based on Navin *et al.*'s central result that the tumors can be partitioned into monogenomic (those appearing to

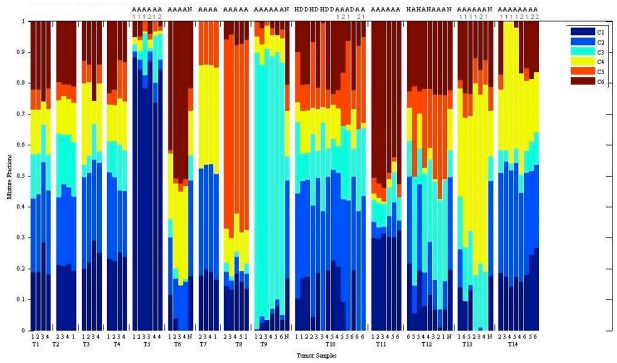


Figure 3.7: Inferred mixture fractions for six-component soft geometric unmixing applied to breast cancer aCGH data. Data is grouped by tumor, with multiple sectors per tumor placed side-by-side. Columns are annotated below by sector or N for normal control and above by cell sorting fraction (D for diploid, H for hypodiploid, A for aneuploid, and A1/A2 for subsets of aneupoloid) where cell sorting was used.

show essentially a single genotype) and polygenomic (those that appear to contain multiple tumor subpopulations). We test for monogeniety in mixture fractions by finding the minimum correlation coefficient between mixture fractions of consecutive tumor sectors (ignoring normal controls) maximized over all permutations of the sectors. Those tumors with correlations above the mean over all tumors (0.69) were considered monogenomic and the remainder polygenomic. Navin *et al.* assign $\{1, 2, 6, 7, 9, 11\}$ as monogenomic and $\{3, 4, 5, 8, 10, 12, 13, 14\}$ and polygenomic. Our tests classify $\{1, 2, 5, 6, 7, 8, 11\}$ as monogenomic and $\{3, 4, 10, 12, 13, 14\}$ as polygenomic, disagreeing only in tumors 5 and 8. Our methods are thus effective at identifying true intratumor heterogeneity in almost all cases without introducing spurious heterogeneity. By contrast, hard unmixing identifies only tumors 5 and 8 as polygenomic, generally obscuring true heterogeneity in the tumors (see supplementary section S1).

Our long-term goal in this work is not just to identify sub-types, but to describe the evolutionary relationships among them. We have no empirical basis for validating any such predictions at the moment but nonetheless consider the problem informally here for illustrative purposes. To explore the question of possible ancestral relationships among components, we manually examined the most pronounced regions of shared gain across component. Figure 3.9 shows a condensed view of the six components highlighting several regions of shared amplification between components. The left half of the image shows components 3, 5, and 1, revealing a region of shared gain across all three components at 9p21 (labeled B). Components 5 and 1 share an additional amplification at 1q21 (labeled A). Components 1 and 5 have distinct but nearby amplifications on chromosome 17,

Marker	Locus	Component	Marker	Locus	Component
MUC1	1q21	C1,C4	BRCA2	13q12.3	C5
PIK3CA	3q26.3	C3,C6	ESR2	14q23	C1
ESR1	6q25.1	C4	BRCA1,	17q21	C5,C6
EGFR	7p12	C5	ERBB2		
c-MYC	8q24	C1,C3,C5	STAT5A,	17q11.2	C5
PTEN	10p23	none	STAT5B		
PGR	14q23.2	C6	GRB7	17q12	C6
CCND1	11q13	C4	CEA	19q13.2	C6

Table 3.1: Benchmark set of breast cancer markers selected for validation of real data, annotated by gene name, genomic locus, and the set of components exhibiting amplification at the given marker.

with component 1 exhibiting amplification at 17q12 (labeled D) and component 5 at 17q21 (labeled C). We can interpret these images to suggest a possible evolutionary scenario: component 3 initially acquires an amplification at 9p21 (the locus of the gene CDKN2B/p15INK4b), an unobserved descendent of component 3 acquires secondary amplification at 1q21 (the locus of MUC1), and this descendent then diverges into components 1 and 5 through acquisition of independent abnormalities at 17q12 (site of PGAP3) or 17q21 (site of HER2). The right side of the figure similarly shows some sharing of sites of amplification between components 2, 4, and 6, although the amplified regions do not lead to so simple an evolutionary interpretation. The figure is consistent with the notion that component 2 is ancestral to 4, with component 2 acquiring a mutation at 5q21 (site of APC/MCC) and component 4 inheriting that mutation but adding an additional one at 17q21. We would then infer that the amplification at the HER2 locus arose independently in component 6, as well as in component 5. The figure thus suggests the possibility that the HER2-amplifying breast cancer sub-type may arise from multiple distinct ancestral backgrounds in different tumors. While we cannot evaluate the accuracy of these evolutionary scenarios, they nonetheless provides an illustration of how the output of this method is intended to be used to make inferences of evolutionary pathways of tumor states.

3.2 Validation on FISH Data

3.2.1 Introduction

Fluorescent in-situ hybridization or FISH is a commonly used experimental technique for identifying features of genomes including ploidy and chromosomal arrangements. The procedure involves hybridizing DNA probes to sections of sample genome DNA for labeling specific chromosomes, genes or other genetic markers.

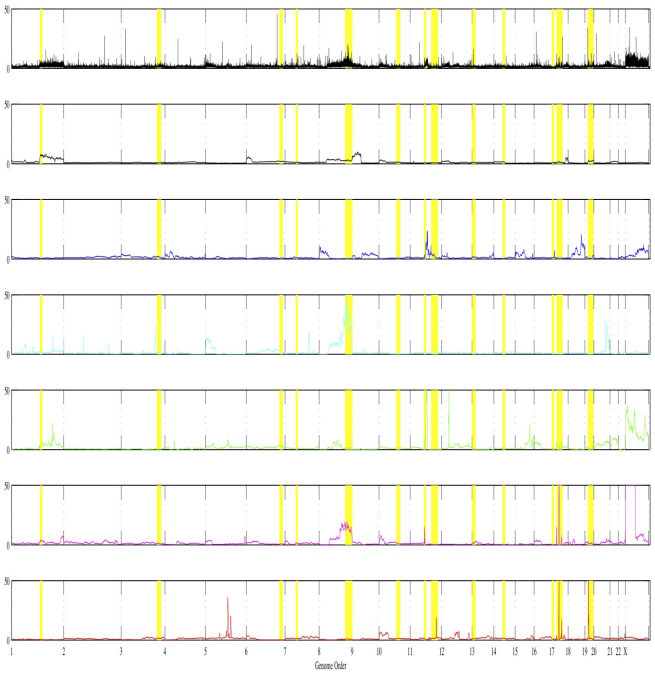


Figure 3.8: Copy numbers of inferred components versus genomic position. The average of all input arrays (top) is shown for comparison, with the six components below. Benchmarks loci are indicated by yellow vertical bars.

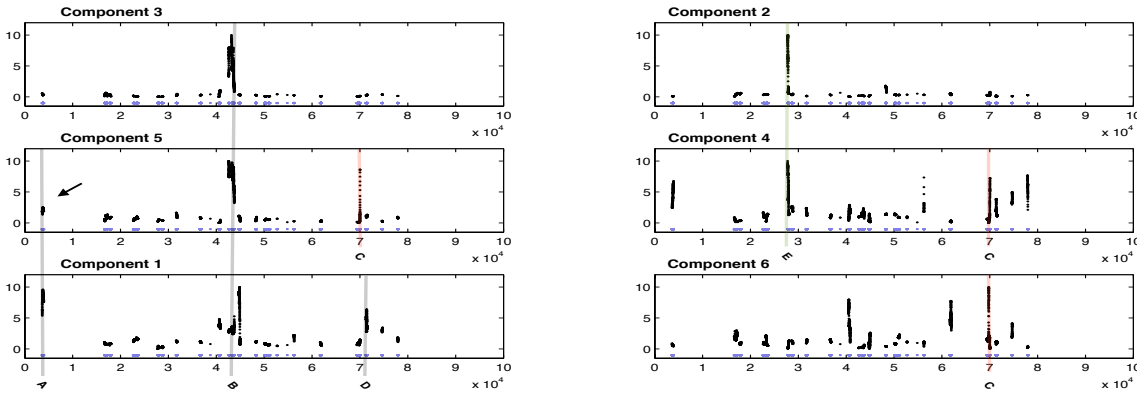


Figure 3.9: Plot of amplification per probe highlighting regions of shared amplification across components. The lower (blue) dots mark the location of the collected cancer benchmarks set. Bars highlight specific markers of high shared amplification for discussion in the text. *Above:* **A**: 1q21 (site of MUC1), **B**: 9p21 (site of CDKN2B), **C**: 7q21 (site of HER2), **D**: 17q12 (site of PGAP3), **E**: 5q21 (site of APC/MCC).

The probes fluoresce on hybridization and the number of hybridized probes for a certain genetic marker can be counted using microscopy. These counts inturn can be used to determine amplifications, deletions and rearrangements of genetic markers of interest. Probes binding to centromeres can be used to determine changes in ploidy. FISH is an active medical diagnostic procedure for detecting various chromosomal anomalies(E.g. trisomy, solid tumors, micro deletions). In cancer research, FISH has been used to study structural rearrangements in the genome of tumor cells including changes in overall cell ploidy.

3.2.2 Data

We use data derived from FISH experiments on paired breast tumor samples from primary and metastatic regions in 13 patients [62]. The primary tumor stage is called DCIS(Ductal carcinoma in-situ) and the advanced tumor stage is IDC(Infiltrating ductal carcinoma). The FISH experiments used a maximum of 11 probes : 8 gene markers and 1 marker for ploidy and 1 or 2 centromere markers. Several cells are studied from each of the 26 breast tumor samples. Each representative cell in a sample can be represented by a data vector of counts for each probe where each count represents the number of times each probe was observed in the given cell in that sample. An additional data element counts the number of such cells in the sample with the same pattern of probe counts to avoid repeat observations.

The data matrix derived from the experiments consists of 12-13 columns, 2-3 columns with counts of each ploidy/chromosome centromere probe found in each cell of each patient, 8 columns with counts of each gene marker and a last column with the number of cells showing the given count distribution. There are as many rows as there are representative cells examined for the patient.

For comparison with the aCGH data, the aCGH probes closest in genome location to the FISH probes were selected from the unmixed component profiles. Ploidy and centromere probes were not included in the comparison as the corresponding information was not available in the aCGH data. The profile of copy number ratios for these markers probes are shown in Figure 3.10.

3.2.3 Methods

To allow for meaningful comparison, we binarize the data. The FISH data was reduced to binary format by reducing any ploidy value of diploid or below to the state 0 and any amplified ploidy states to the value 1. For our analysis, we pooled the data across all patient samples. This resulted in a total of 134 unique cell states across all samples.

The unmixed data was reduced to binary format by rejecting a hypothesis test that the data are drawn from a Gaussian centered at mean 1 and empirical variance across all probes and components. The 6 components are represented as in

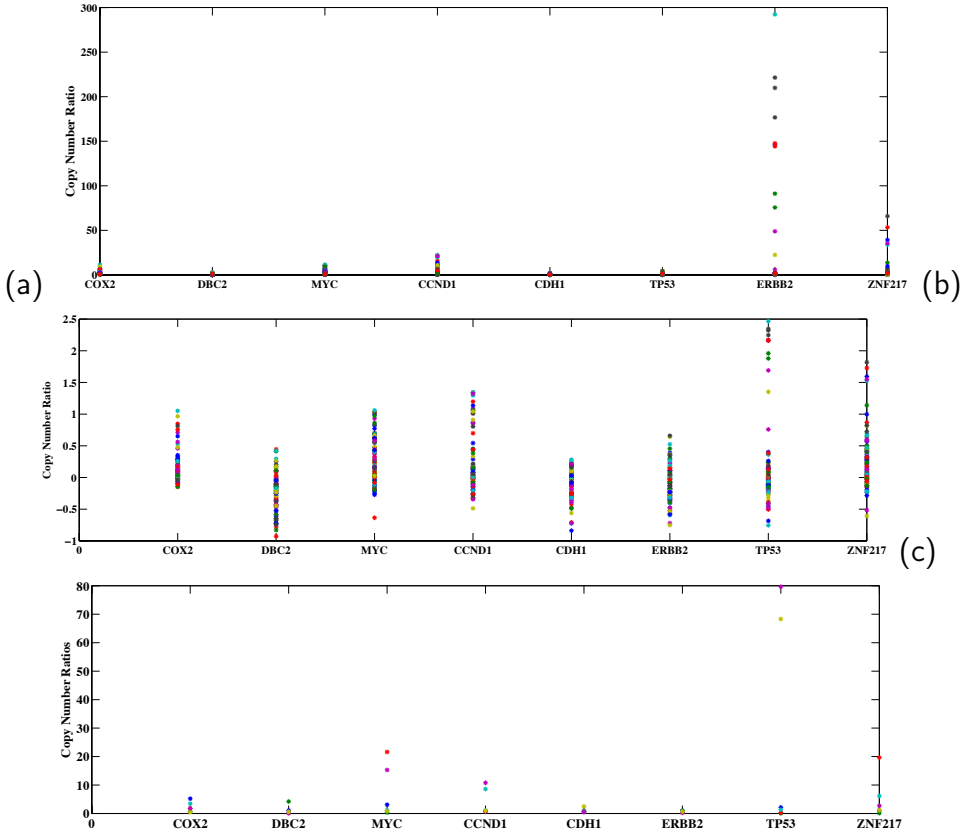


Figure 3.10: Copy number ratios for the 8 cytogenetic markers (d) Raw Navin aCGH Data (e) Log Navin aCGH Data (f) Real Unmixed Navin Data.

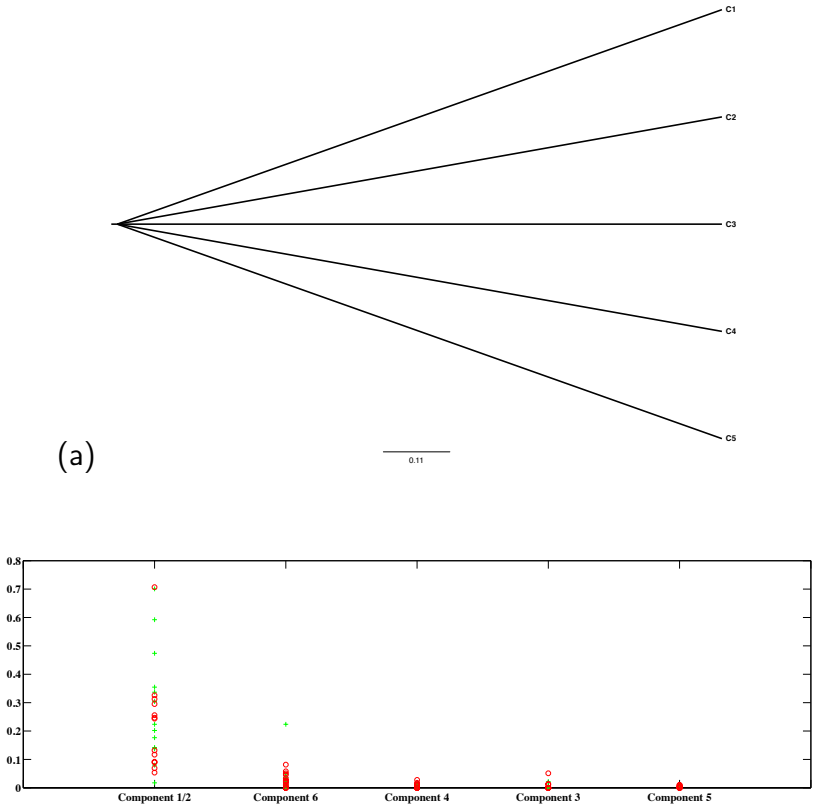


Figure 3.11: Green is DCIS and red is IDC

the table below:

COX2 DBC2 MYC CCND1 CDH1 ERBB2 TP53 ZNF217
Component 1 and 2 : 0 0 0 0 0 0 0 0
Component 3 : 0 0 1 0 0 0 0 1
Component 4 : 0 0 0 1 0 0 0 1
Component 5 : 0 0 1 1 0 0 1 0
Component 6 : 0 0 0 0 0 0 1 0

There are 5 unique states as both components 1 and 2 represent the normal state. Consensus maximum parsimony trees were built by bootstrapping over 10000 replicates for the unmixed data.

The FISH cell states were then checked for matches with the Navin unmixed cell components. The frequencies of occurrence of the tumor cell components in the FISH data samples are shown in the figure below.

The combined data cloud can be visualized in the PC space as below.

We came up with a statistical test for similarity between the FISH cell states

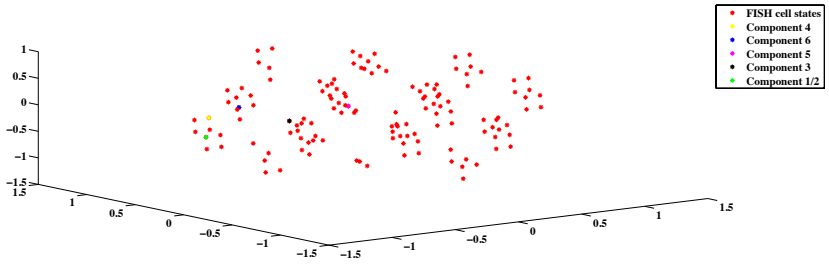


Figure 3.12: FISH Data cell types in PC space. The green, black, yellow and blue dots are components 1/2, 3, 4 and 6 respectively. Comparison of inferred single cells from mixture modeling to true single cell data by FISH. Points represent individual breast cancer cells assayed by FISH (red dots) and mixture components (non-red dots) plotted in a space of three principle components of the full data. The plot shows point clusters corresponding to frequently observed breast cancer cell states, with mixture components found in a subset of these clusters.

and the unmixed components. There are a total of 2367 cells among the 26 samples. Each cell is represented by a binary vector of size 8 for the 8 gene probes as described above. The similarity between a cell and a component is measured by the hamming distance between the two. The test statistic obtained by summing over the minimum distances between a cell and its closest component across all cells was 4998. This was compared to the distribution of the test statistic obtained when the cells were obtained by randomly sampling assuming independence of probes and biased by the empirical frequencies of finding amplifications at each probe. Over 1000 trials, the average distance was got to be 3879.9 with a standard deviation of 41.79. The probability that the value of 4998 is drawn from such a distribution is < 0.001.

3.3 Conclusion

We have developed a novel method for unmixing aCGH data to infer copy number profiles of distinct cells states from tumor samples. The method uses "soft geometric unmixing" to provide superior tolerance to experimental noise and outliers compared to the prior work. We have further developed an efficient gradient-based optimization algorithm for this objective function. We have shown through tests on simulated data that the soft unmixing approach dramatically improves accuracy of inference of components and mixture fractions in the presence of high noise or large component numbers relative to a hard unmixing method. We have further verified, with application to a set of real aCGH data from breast cancer patients, that the method is effective at separating components corresponding to distinct subsets of known breast cancer markers. The specific patterns of gain and loss in

the components are suggestive of patterns of evolution among the tumor types. The work thus demonstrates the potential of tumor sample unmixing applied to aCGH data to infer copy number profiles of cell populations from heterogenous tumor samples. In addition to facilitating studies of tumor evolution, the methods may have value to many other applications of mixture separation from noisy data. We also introduce an approach to compare the resulting virtual unmixed cell components with real FISH cells. The results suggest that while the virtual single-cell mixture components do match major cell populations supported by true single cell FISH data, the FISH data also reveal much greater complexity than can be inferred through the mixture approach, helping to motivate the shift in focus to phylogenetics of direct cell-level data proposed here.

Chapter 4

Inference of tumor phylogenies from genomic assays on heterogeneous samples

In the Chapter, we describe a general pipeline for tree-building from genome-scale tumor data. The contents are adapted from the conference proceedings [123] and the extended journal paper[77]

Tumorigenesis can in principle result from many combinations of mutations, but only a few roughly equivalent sequences of mutations, or "progression pathways," seem to account for most human tumors. Phylogenetics provides a promising way to identify common progression pathways and markers of those pathways. This approach, however, can be confounded by the high heterogeneity within and between tumors, which makes it difficult to identify conserved progression stages or organize them into robust progression pathways. To tackle this problem, we previously developed methods for inferring progression stages from heterogeneous tumor profiles through computational unmixing. In this chapter, we develop a novel pipeline for building trees of tumor evolution from the unmixed tumor data. The pipeline implements a statistical approach for identifying robust progression markers from unmixed tumor data and calling those markers in inferred cell states. The result is a set of phylogenetic characters and their assignments in progression states to which we apply maximum parsimony phylogenetic inference to infer tumor progression pathways. We demonstrate the full pipeline on simulated and real comparative genomic hybridization (CGH) data, validating its effectiveness and making novel predictions of major progression pathways and ancestral cell states in breast cancers.

4.1 Introduction

The application of genomic technologies to cancers has revealed that patients with tumors that appear indistinguishable to the clinician may have completely different causes at the molecular level [20, 23] resulting in very different prognoses [24] and responses to possible treatments [22]. Nonetheless, most human cancers seem to follow a relatively small number of progression pathways [23, 20, 124], each

characterized by an approximately equivalent sequence of mutations. This observation is key to the success of targeted therapeutics, a groundbreaking approach to cancer treatment in which drugs are developed to treat specific molecular abnormalities shared by large subgroups of patients [25]. By identifying common progression pathways and characterizing their conserved features, it is hoped that we can find new subgroups of patients who will respond to a common treatment, identify the specific abnormalities that will provide effective therapeutic targets for those subgroups, and develop clinically useful diagnostic tests to identify new patients in those subgroups. There are considerable practical challenges to each of these steps, however.

One of the significant challenges to identifying and characterizing progression pathways is the heterogeneity of cancers both within and between patients [104]. Any two patients, even with a common progression pathway, will exhibit many differences in the details of the causal mutations along that pathway, as well as in the assortment of random passenger mutations distinct to each patient that do not contribute to their pathology [125]. Even within a single patient, a tumor will generally be highly heterogeneous, with genetically distinct cell populations corresponding to different stages along the progression of their tumor and possibly even different branches along those progression pathways within a single tumor [126]. This heterogeneity is problematic for methods for profiling tumor states, since there is at present no technology to determine the genetic states of single cells at a genomic scale. Genome-wide methods for tumor profiling — such as expression microarrays, RNA-seq, or array comparative genomic hybridization (aCGH) — necessarily mix contributions from many discrete cell types. This mixing would be expected to result in a conflation of distinct states along a progression pathway, obscuring characteristics of individual subpopulations of cells and hiding the discrete steps in progression that may provide clinically valuable markers of early stages in progression or important clues to major decision points in a tumor's evolution. This heterogeneity is particularly challenging to phylogenetic approaches to inferring tumor progression [37], which depend on our ability to at least approximately identify discrete steps in tumor evolution and can benefit greatly from information about ancestral states and the combinations of states present in distinct tumor samples [72].

There are various ways to approach the problem of heterogeneity in tumor phylogeny inference. One approach is to use alternative technologies designed to profile single cells as a way of directly observing discrete states within tumors. This approach has been successfully used for tumor phylogeny inference from single cell fluorescent in situ hybridization (FISH) data [17, 63]. Using single-cell assays has substantial drawbacks, however, because single-cell technologies can profile only a few preselected markers per cell. An alternative is to separate cells into approximately homogeneous populations prior to applying genomic methods, as was done recently by [38], who used a combination of microdissection and post-dissection cell sorting to separate discrete sub-populations of cells prior to

whole-genome DNA copy number profiling by aCGH. A third alternative, used in the present work, is to apply genomic technologies to heterogeneous samples but attempt to computationally separate distinct cell populations from the outputs of these samples. Such computational unmixing methods have been previously used in tumor analysis to correct for stromal contamination of tumor cells [74] and have been useful to similar applications of evolutionary inference from heterogeneous samples, such as in reconstructing evolutionary steps in viral quasispecies [127].

In previous work, we proposed the use of such unmixing methods for identifying cell states for phylogeny inference [72] and demonstrated their ability to separate biologically meaningful tumor cell populations from expression microarray data [72] and aCGH data [73]. In this chapter, we build on that prior work by developing a pipeline for converting inferred cell profiles into phylogenetic trees describing likely stages of tumor progression and common progression pathways by which they evolve. This pipeline implements four distinct steps. The first applies our prior unmixing model [73] to infer profiles of major progression steps from heterogeneous tumor data. The second step uses a novel statistical test to identify amplified genomic regions that can serve as markers of progression. The third step then uses a second statistical approach to call these markers as amplified or non-amplified in individual inferred cell states, creating a matrix of phylogenetic states suitable for character-based phylogenetic inference. The fourth step then applies maximum parsimony phylogeny inference to the resulting data to identify likely progression trees, labeled by changes in the marker set inferred in step two. These progression trees establish a model of tumor evolution identifying discrete steps of progression among these markers and possible ancestral stages of tumor progression not directly apparent from the identified components. Validation on simulated data demonstrates the effectiveness of the method at identifying markers, assigning them to progression states, and inferring trees from those states. Application to real breast cancer CGH data results in a phylogeny that recapitulates key features of our current understanding of major breast cancer progression pathways while elaborating in several potentially significant ways. The work represents, to our knowledge, the first use of character-based phylogenetic inference for similar whole-genome tumor profiles, providing advantages over prior distance-bsaed approaches in identifying likely markers and describing specific mutations that may underlie key steps in tumor progression.

In the remainder of this chapter, we present our method and an application to a publicly available aCGH data set. In section 2, we describe our overall phylogenetic inference pipeline and the novel computational and statistical methods developed for it. In section 3, we provide details on specific use of the methods developed here and their application to the analysis of the breast tumor aCGH data of [38]. In section 4, we present the results, identifying a set of phylogenetic markers and a resulting tumor phylogeny. In section 5, we discuss the biological significance of the results, examining both their concordance with prior literature and interesting novel predictions of the methods. Finally, in section 6, we consider avenues for

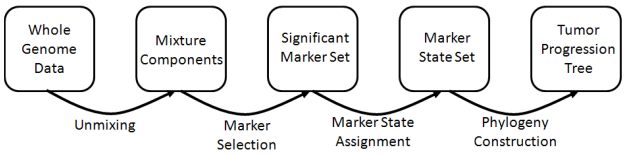


Figure 4.1: Workflow diagram summarizing the major steps in our unmixing-based phylogenetic analysis pipeline.

future work.

4.2 Materials and Methods

4.2.1 Algorithms

At a high level, our method consists of an analysis pipeline to convert raw data on profiles of heterogeneous tumor samples into phylogenetic inferences on computationally inferred profiles of discrete cell states. While the method can in principle work with any technology for profiling tumor state, we assume in the present work that we are specifically using aCGH data describing DNA copy numbers at a discrete genome-wide probe set. The data are assumed to be in the form of copy numbers of n probes in m tumors or tumor sections. These data are assumed to be raw or baseline normalized raw input, rather than the conventional $\log$ ratios.

The overall analysis pipeline is summarized in Figure 4.1. The pipeline consists of the following steps:

1. Computational unmixing of raw aCGH data to infer aCGH profiles of well-populated tumor states,
2. Identification of significantly amplified marker regions of the genome from the component aCGH data,
3. Assignment of marker states to components,
4. Phylogenetic inference on cell states to produce an inferred progression tree.

The individual steps of this analysis are as described below.

Unmixing Analysis

Our phylogenetic approach assumes data has been separated into mixture components. We initially accomplished this assignment using an unmixing method previously developed by our group [73] based on an interpretation of the problem

as that of fitting a simplex to an observed set of data points, where simplex vertices will then correspond to inferred components of the mixture. The method is based on prior work by Ehrlich and [115] adapted to better handle the high dimension and noise level characteristic of genomic data. We have since updated that method to use non-negative matrix factorization (NNMF) [117] to eliminate the possibility of negative copy number values and other artifacts that can induce in the code.

Figure 4.2 illustrates the unmixing procedure. We first preprocess the data by applying L1-L1 total variation denoising to the raw aCGH profiles. In the initial method, we then use principal components analysis (PCA) to convert aCGH profiles of tumor samples to points in a low-dimensional space. The aCGH profiles are then explained as mixtures drawn from a set of common cell types by fitting a simplex to the point set, with some allowance for noise in the data. Any point in the simplex can then be explained as a linear combination of the vertices of the simplex. These vertex points are interpreted as the cell types from which each tumor sample is generated and can be projected back into the original dimension of the aCGH array to construct virtual aCGH profiles of the inferred cell types. The outputs of the method are an inferred set of mixture components, identifying a projected copy number of each cell type at each probe, and a set of mixture fractions, explaining each observed tumor sample as a sum of fractional contributions of cell types. The mixture components can be represented as a matrix C in which each entry c_{ij} describes the inferred copy number of component or cell type i at aCGH probe j. For the present pipeline, we use only the component matrix C and discard the mixture fractions. Space does not permit a detailed description of the method, so we refer the reader to [72] for a more thorough description of our general unmixing strategy for tumor phylogenetics and to [73] for a detailed discussion of the specific noise-tolerant unmixing algorithm used in our primary results here. Our most recent algorithm functions identically except that initially dimensionality reduction is accomplished by NNMF rather than PCA and an additional non-negativity constraint is imposed during the optimization of components C.

The primary results below are based on components previously determined in Tolliver et al. [73] by the PCA-based method, although the improved method is applied to develop components from simulated data and from a secondary breast cancer data set to provide additional points of comparison.

Identification of Amplified Genomic Regions

Once we have the inferred components, it is next necessary to identify markers for tracking phylogenetic state. For aCGH data, we seek genomic regions that are amplified in subsets of tumors. We focus on amplifications due to a technical limitation of the unmixing approach. Unmixing is performed in the linear, rather than log, domain and a deletion represents only a small linear change in copy

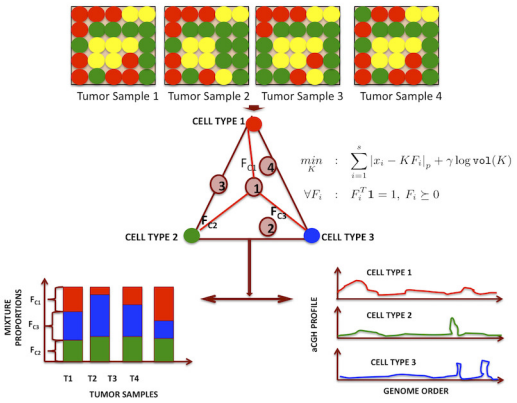


Figure 4.2: Illustration of the unmixing approach. Tumor samples T1–T4 are assayed by aCGH, generating genome-wide copy number profiles. The aCGH profiles are intepreted as points in a space (two-dimensional in the example) and are unmixed by fitting a simplex to the point set (a 3-simplex, or triangle, in the example). The vertices of the simplex represent inferences of three cell types (1, 2 and 3) from which T1–T4 can be explained. These vertices are then projected back to the dimension of the aCGH arrays to construct virtual aCGH profiles of the inferred cell types. The outputs are these virtual aCGH profiles and the inferred fractional amount of each cell type in each tumor sample.

number, so we expect the method to have poor sensitivity to deletions. Given the high variability from probe to probe in the data, it is necessary to use a statistically robust test for amplification. To accomplish this, we developed a test designed to test for significant amplification of a window of w contiguous probes across the m components.

We assume Gaussian noise in the data, thus modeling each individual probe as drawn from a Gaussian distribution with mean 1 (corresponding to diploid DNA). The variance is assumed to be the empirically measured variance, σ^2, across all probes in all components. We then seek to reject the hypothesis that the collection of $w \times m$ probes under consideration were drawn from the corresponding Gaussian. For this purpose, we take as our statistic the sum of squares of Z-scores of the probe values:

$$X_k = \sum_{i=1}^{m} \sum_{j=k}^{k+w-1} \left(\frac{c_{ij} - 1}{\sigma} \right)^2$$

where k is the index of the first probe in the window. Under the null hypothesis, this statistic would be expected to be chi-square distributed with $w \times m$ degrees of freedom. We thus test for significant amplification with a one-sided chi-square significance test for the appropriate degrees of freedom.

We apply this test to sliding windows of probes of fixed width w across the

genome. After identification of discrete amplified windows, we apply a post-processing step to collapse any overlapping amplified windows into a single larger window and treated the union of probes in all overlapping significant windows as the marker for subsequent analysis.

We would normally expect the detected regions to be a subset of those one would find by performing a comparable statistical test on the raw aCGH measurements rather than the inferred components, as we would expect that features that are not robust to a significant fraction of samples will be interpreted as noise and suppressed at the unmixing step.

The scan for significant windows was done through custom Matlab code using the `chi2cdf` function for chi-square significance testing.

Assignment of Marker States to Components

After identifying a set of markers, we next need to determine the states of those markers in each inferred cell component. For this purpose, we again treat the problem as that of attempting to reject the hypothesis that the individual copy numbers are drawn from a Gaussian of mean 1 and variance corresponding to the empirically measured variance across all probes. For each component i and marker j, we compute the mean copy number over all probes in the given marker for the given component:

$$\mu_{ij} = \frac{1}{b_j - a_j + 1} \sum_{k=a_j}^{b_j} c_{ik}$$

where a_j is the leftmost probe index and b_j the rightmost probe index for marker j. We then evaluate the single-sided p-value for the hypothesis that μ_{ij} is drawn from a Gaussian with mean 1 and variance $\sigma^2/(b_j - a_j + 1)$, where σ^2 is again the empirical variance across all probes in all components. We implicitly build in a prior probability that any given marker is not amplified in any given component by using a p-value cutoff of 0.001 for calling a probe amplified. The result of this analysis is an assigned state (amplified or not amplified) for each component at each phylogenetic marker. These values can be represented as an $m + 1 \times k$ matrix P of phylogenetic markers, where element p_{ij} is a binary value indicating whether marker j is amplified or not amplified in component i.

Custom Matlab code was used to assign phylogenetic states to each component at each marker using the `normcdf` function.

Phylogeny Construction

The matrix of phylogenetic marker states P produced in the previous step serves as the input to a character-based phylogenetic inference. Given the lack of any sound empirical basis for setting parameters for a Bayesian or maximum likelihood

method, we favor use of a simpler parsimony method and therefore treat tumor phylogeny inference as the problem of finding a maximum parsimony Steiner tree [128] in which the observed components are leaves of the tree. For similar reasons, we do not weight markers, treating gain or loss of any marker as equally likely and seeking a minimum weight Steiner tree capable of explaining the data. The actual phylogeny construction is accomplished with PAUP [34] (Portable version 4.0b10 for Unix). The program was run with the Maximum Parsimony Optimality Criterion using heuristic search for 10 repetitions, random sequence addition, and the Tree Bisection Reconnection option for swapping. Trees were visualized with GraphViz [129].

4.2.2 Computational Analysis

Simulated Data: As a first validation, we applied our methods to a set of simulated aCGH data to specifically test the effectiveness of our method at identifying markers, grouping them into components, and properly placing the components in a phylogenetic tree. We simulated data for a single hypothetical chromosome of 1000 probes, assuming cell states evolve according to a binary tree from an initially diploid root state. We then assumed each of the edges would contribute a single mutation, represented as a segment of 11 consecutive probes with amplification level 20 placed uniformly at random on the simulated chromosome, rejecting placements that would place segments less than 10 probes away from another segment. We then drew 200 simulated tumor samples from this tree of components by choosing a single node at random from the tree and using all nodes on the path from the root to chosen node as the mixture components of that sample. We chose mixture fractions for the components in each simulated tumor sample by choosing uniform random weights for each component assigned to the sample and normalizing by the sum of these weights to derive fractional contributions of each component to each mixture fraction. Finally, we add simulated Gaussian noise to each probe value for each simulated tumor sample with mean zero and standard deviation set to 0.05, 0.10, 0.15, and 0.20 in separate experiments. We then applied the NNMF-based unmixing algorithm with regularization parameter 100 and the analysis pipeline described above using a p-value cutoff of 10^{-6} and window size of 5 for marker identification.

We measured accuracy based on amplified segments correctly identified, components correctly identified, and tree edges correctly identified. We first assessed the fraction of the amplified segments correctly identified during marker selection for each scenario. Next, we computed the fraction of components correctly identified, with an assignment judged correct if it was assigned the same state as the true component for all markers that were correctly identified in the previous step. Finally, we assessed the fraction of tree edges correctly identified among those subdividing nodes correctly identified in the previous step. A tree edge was considered correct if it subdivided the node set identically in the inferred tree and

in the true tree when collapsed to the subset of nodes identified correctly during the marker assignment step. All three analyses were repeated for $k = 4 - 7$ components for each of the four noise levels.

Real Data: Our primary analysis consisted of application of our method to a set of previously identified mixture components derived in [73] using a publicly available set of aCGH data from sectioned primary ductal breast tumors [38]. This dataset was selected because the sectioning and cell sorting approach developed by Navin et al. was specifically chosen to facilitate phylogenetic inference and provides additional data on intra-tumor heterogeneity useful in validating the methods. The raw data comprises 87 tumor sectors obtained from 14 ductal breast cancer tumors run on a high-density ROMA platform with 83,055 probes. We confined our analysis to the twenty-two autosomal chromosomes, reducing the dataset to 78,874 probes.

The raw aCGH data was preprocessed and unmixed as described in our prior work [73]. As before, data was converted from log to linear domain, denoised with a total variation denoising, and unmixed to generate components. Six components were chosen, as described in the prior work, based on an analysis of the eigen-decomposition of the data. The resulting components are the same as those described in that prior paper and we refer the reader there for detailed information on the unmixing method and its application to this data set.

Phylogenetic markers were determined from the resulting component matrix as described in Approach. We used a window size of $w = 20$ for the initial sliding-window scan of the genome. The p-value threshold for each window in isolation was set to 10^{-8} to account for Bonferroni correction for the 78,855 sliding windows of size 20 possible for the 78,874 probes. This threshold corresponds to a corrected p-value threshold of 7.9×10^{-4}. After collapsing overlapping windows, we found a total of 27 phylogenetic marker regions significantly amplified across samples. In order to investigate the possible biological significance of these markers, we identified all genes overlapping the probe set for each marker region using the UCSC Genome Browser [130] applied to the human reference genome build 17 (NCBI35). We use NCBI build 35, rather than a more recent build, to conform to the aCGH platform specifications. We further attempted to identify any genes with a known association with cancer by manually examining Online Mendelian Inheritance in Man (OMIM) [131] entries for all genes overlapping the probes, specifically noting those with a prior association with cancers in general or breast cancers specifically.

Application to an Independent Data Set: As a secondary validation of our approach, we applied it to a second set of mixture components derived from a second publicly available second breast cancer aCGH dataset [132] consisting of 44 predominantly advanced primary breast tumors and 10 breast cancer cell lines. The dataset consists of 59 samples and 6691 probes each corresponding to a single gene, making it substantially lower in resolution than the Navin et al. dataset. We ran our recent NNMF-based unmixing method with TV denoising

regularization parameter 6 and, unmixing parameters $k = 6$ components and $\gamma = 100$ regularization, with window size 20 and bonferroni corrected p-value cutoff 1.7×10^{-7}. While the lower resolution of the data prevented direct comparison to the Navin et al. results, we evaluated the method based on its ability to identify four markers (on 1q, 8q, 17q and 20q) specifically cited by the authors of the study as well as others that showed up as important markers in the analysis of the Navin et al. data.

4.3 Results and Discussion

4.3.1 Results

Simulated Data: Figure 4.3 summarizes results on the simulated data. Surprisingly, marker-level accuracy generally improves with increasing component numbers but appears relatively insensitive to noise level over the ranges examined here. The average accuracy across all scenarios is 79.2%. No false positives markers were detected in any of the simulations. Component-level accuracy shows a more complicated profile, with generally worse performance for larger numbers of components at any given noise level. Analysis of specific identified components suggests a common error is the identification of more than one inferred component closely corresponding to a single true component, leading to other true components getting omitted from the data. The overall average accuracy in component assignment is 72.8% over all scenarios. The accuracy of tree edges in partitioning the identified components is 100% across most noise levels and component numbers, except for 20% noise and 15% noise for k=6 components and 20% noise for k=7 components. The overall accuracy in inferring tree edges is 94.8%. It is important to note, though, that we defined these error measures so that the method would not be penalized for failed marker detection in assessing component or tree edge detection nor be penalized for failed component detection in assessing tree edge detection. This decision was motivated by a desire to assess the accuracy of each step independent of the others. The reported accuracies would appear more pessimistic if we counted components correct only if all markers were detected or counted tree edges incorrect if the components they separate were not detected.

Real Tumor Data: Application of our analysis to the [38] data yielded six components corresponding to inferred cell states, in addition to a seventh normal cell type added to root the subsequent tree. The components themselves and a detailed analysis of those components and the associated mixture fractions is provided in our prior work [73] and we therefore refer the reader to that prior literature for a detailed discussion of the mixture components by themselves.

We next analyzed the components to find significantly amplified marker regions. The analysis yielded a total of 27 non-overlapping regions at which the components collectively showed significant amplification. The full set of marker

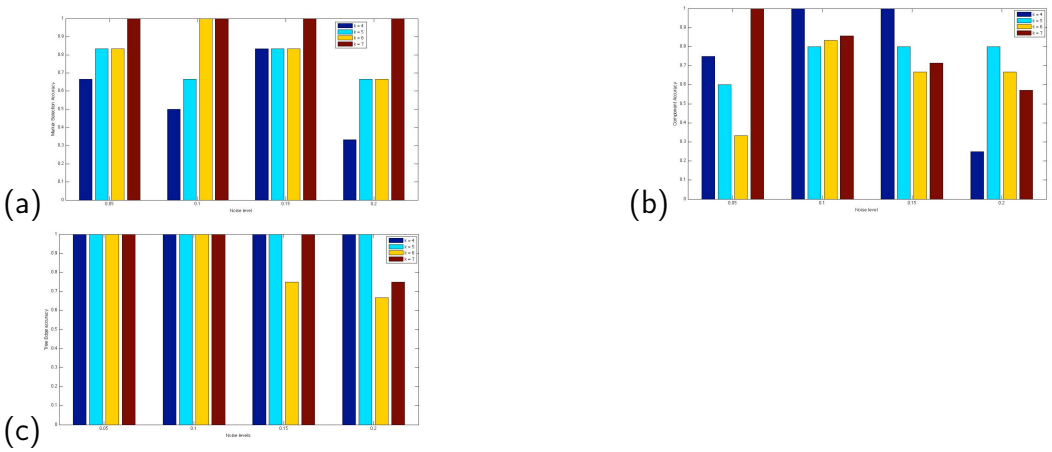


Figure 4.3: Quantification of accuracy on simulated data from $k = 4 - 7$ components and noise levels 0.05–0.20. (a) Fraction of markers correctly predicted in each experiment. (b) Fraction of components correctly identified on all identified markers in each experiment. (c) Fraction of tree edges correctly identified for the components and markers identified in each experiment.

regions is provided in Table 4.1. In addition, we provide a list of genes overlapping the regions that have some known association to cancers. Most of the regions contain at least one gene known to have some prior association with cancers, including several genes specifically associated with breast cancers (CD55, MDM4, WNT2, ERBB2, GRB7, BCAS, CCNE, CTTN, AURKA, BCL2, MYC, TNFRSF11A, ZNF217, CYP24A1). In several other cases, a region lacking known cancer associated-genes is found adjacent to one with a known association and might be presumed to be part of a common amplicon (e.g., 18q22.2-18q22.3).

These regions overlap a total of 343 genes, of which 56 (16.3%) were manually found to be associated with cancers in OMIM. It is difficult to rigorously establish a global frequency with which genes are cancer related, but we can derive an estimate by reference to the work Badjik et al. [133], who used a text-mining approach to determine that 1,943 genes as of the time of their work were annotated as cancer-related in OMIM. Comparing this number to the number of Refseq transcripts, 27,704 (NCBI genome build 35), provides an estimate that 7.01% of all genes are annotated as cancer-associated in OMIM. The comparison suggests that the marker regions identified by our study are strongly enriched for known cancer-related genes. A chi-squared statistical test shows this difference in frequencies to be highly significant (chi-square score 43.2, p-value <0.0001).

We would expect the unmixing to screen out amplifications that occur in only a small fraction of samples, leading to the discovery of fewer but more robust markers than would be found from the raw aCGH data. To test that assumption, we also ran the marker selection method on the raw aCGH data. This process yielded 47 marker regions, including 24 of the 27 found from the unmixed data.

Marker ID	Cytogenetic coordinates	Chromosome positions	Annotated cancer-related genes
1	1q32.1-1q32.2	196117366-206330147	CD55, MDM4, NR5A2,PTPN7, IL10, CNTN2,CD34
2	1q44	242649493-245131380	SMYD3
3	2p12	76777788-78642108	None
4	3q25.1-3q25.2	151037467-154216571	None
5	5p15.33-5p14.2	3485419-24119655	PAPD7, TAG, CDH18
6	5q21.1-5q21.3	100224934-106834646	None
7	5q22.3-5q23.1	115172420-118711133	TNFAIP8, ATG12, SEMA6A
8	7q31.2-7q31.31	116016939-120372452	ING3,ASZ1, WNT2, ST7
9	8q12.1	55969808-58018737	None
10	8q12.3-8q13.2	63931435-69387571	MYBL1
11	8q13.2-8q13.3	69634776-74092165	TRPA1
12	8q21.11-8q24.3	77351432-143296089	MYC
13	11q13.2-11q13.4	67830873 -70354248	CCND1, CTTN, FGF4, FGF3
14	11q14.1-11q13.4	74383378-82935709	None
15	11q23.3	115171785-116542726	None
16	12p11.22-12p11.21	28901065-33207415	ERGIC2
17	15q25.2-15q25.3	82525637-85513682	None
18	15q26.3	96434691-99661839	None
19	17q11.2	23392447-25127504	RAB34, NEK8, TRAF4, FOXN1
20	17q12-17q21.2	32705491-37628927	STAT5, ERBB2, GRB7
21	17q21.33	45403785-47282174	SPAG9, UTP18, CA10, ANKRD40, CACNA1G, PPP1R9B
22	18q21.32-18q22.2	56806538-66527883	TNFRSF11A, BCL2, SERPINB5, SERPINB13, SERPINB4, SERPINB3, CDH19
23	18q22.2-18q22.3	66607283-71314138	None
24	19q12	34017456-36812510	CCNE1
25	20q13.12	44249187 -45563781	None
26	20q13.2-20q13.32	50440150-57022263	ZNF217, CYP24A1, BCAS1, AURKA, CTCFL, ZBP1, RAB22A, GNAS, SDX16
27	20q13.33	57624055-58571221	None

Table 4.1: Marker regions determined to be significantly amplified across components for the data of Navin et al. [38]. The table provides, for each marker region, a unique identifier, cytogenetic coordinates, probe positions along the genomic axis, and gene IDs for genes identified as having some known association with cancers.

Three markers (Markers 6, 22, and 23) are found only from the unmixed data.

We next assigned states to each of the identified marker regions in each component. Table 4.2 shows the full assignment of marker states to components. We further manually examined the copy number profiles for the predicted components in each marker region. Figure 4.4 provides two illustrative examples, showing the inferred copy number data for the six components and identifying those components determined to be amplified versus non-amplified. Figure 4.4(a) shows the inferred profile for marker 1, corresponding to locus 1q32.1-1q32.2. C1, C3, C4 and C5 are determined to be amplified, which appears to provide a good correspondence to those with copy numbers significantly above one. It is worth noting, however, that there is a finer resolution of amplification apparent in the figure: C1 shows broad but low amplification across the region, C3 shows a more specific amplification of the subregion approximately from probes 5250 to 5300, and C4 shows a distinct pattern of multiple amplicons across the region. These observations suggest the marker-identification method is performing well at a coarse resolution but that there is considerable finer-scale structure that could in principle exploited by a more sophisticated marker selection strategy, particularly where contiguous regions show distinct patterns of amplification.

Figure 4.4(b) shows a second example, the inferred copy number profile for

marker 20, corresponding to an amplicon at 17q12-17q21.2. We would expect this site to be picked up as a marker and to show high amplification, since it is the site of the Her-2 locus. The region again shows a strong but selective amplification, with C5 and C6 highly amplified (although with distinct fine-scale structures), C4 slightly amplified, and others showing no amplification. The result again confirms that the method produces correct answers at a coarse resolution, although there may be finer-scale structure that could exploited by a more sophisticated method.

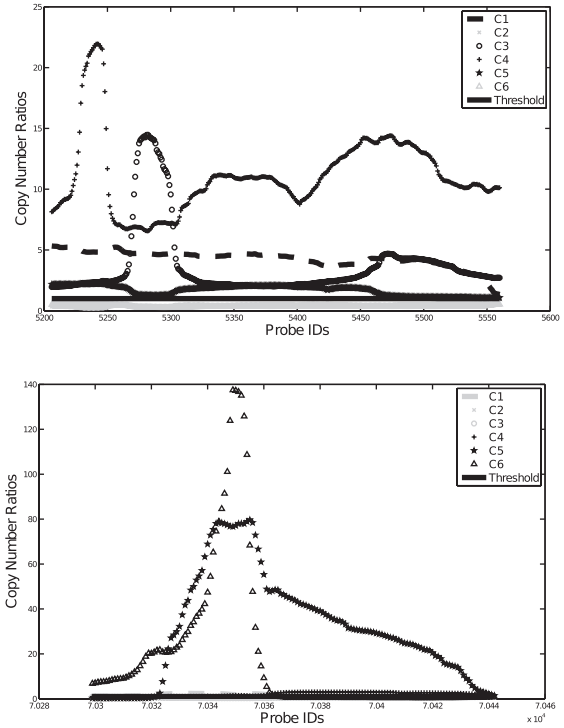


Figure 4.4: Inferred copy number profiles for mixture components in the vicinity of three markers from the data of Navin et al. [38]. The x-axis of each figure corresponds to probes within a specific marker region and the y-axis to copy number relative to the diploid control in that region for each component. The thin solid line in each plot at value 1 shows the diploid threshold. Amplified components appear in black and non-amplified in grey. (a) Marker 1, corresponding to the amplicon at 1q32.1-1q32.2. (b) Marker 20, corresponding to the amplicon at 17q12-17q21.2.

Using the resulting probes, we then performed phylogenetic inference. Figure 4.5 shows the phylogenetic tree produced from the six inferred progression components and the additional normal component manually added to the analysis. The majority of markers are gained at a unique point in the tree and never subsequently lost. Marker 9 (8q12.1) is lost in the tree in the transition to component C4. In addition, some markers are inferred to be gained more than once in

Marker ID	C0	C1	C2	C3	C4	C5	C6
1	0	1	0	1	1	1	0
2	0	1	0	1	1	1	0
3	0	0	0	1	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	1	1	1	0
6	0	0	0	0	1	0	1
7	0	0	0	0	1	0	1
8	0	0	0	1	0	0	0
9	0	1	0	1	0	1	0
10	0	1	0	1	1	1	0
11	0	1	0	1	1	1	0
12	0	1	0	1	1	1	0
13	0	0	0	0	1	1	0
14	0	0	1	0	1	0	0
15	0	0	0	0	0	0	1
16	0	1	0	0	1	0	0
17	0	0	0	0	1	0	0
18	0	0	0	0	1	0	1
19	0	0	0	0	1	1	1
20	0	0	0	0	1	1	1
21	0	0	0	0	1	1	1
22	0	0	1	0	0	0	0
23	0	0	1	0	0	0	0
24	0	1	0	0	0	0	1
25	0	0	0	1	1	1	0
26	0	0	0	1	1	1	0
27	0	0	0	1	1	1	0

Table 4.2: Phylogenetic states of all components at all identified progression markers for the data of Navin et al. [38]. Columns show the states for the six inferred components (C1–C6). The additional normal component (C0) used to root the tree is included for completeness. "1" corresponds to an amplified region and "0" to non-amplified.

the tree. Most notable of these is the collection of 17q markers, which are gained separately in the subtree leading to component 6 and that leading to Steiner node 8 and then to components 4 and 5.

Application to an Independent Data Set: Application to a second component set derived from the lower-resolution data of Pollack et al. [132] provides a secondary validation of the reproducibility of the results on distinct datasets, aCGH platforms, and unmixing methods for a common tumor type. The method identified 20 markers, shown in Table 4.3. The lower resolution of the data leads to substantially more possible genes per amplicon than were found with the Navin et al. data, making it infeasible to conduct a similar analysis of the genes identified. We therefore must compare the two results more indirectly based on markers reported by Pollack et al. in their own analysis of their data as well as known breast cancer markers found in the primary analysis of the Navin et al. data above. Pollack et al. described finding 1q, 8q, 17q and 20q as predominantly amplified regions in the data, and our method did find sizeable amplicons on each of these regions. Other amplicons appear to correspond to several important tumor markers, including the HER2, CCND1, c-myc and CCNE1 loci noted in the analysis of the Navin et al. data as well as the FGFR1 locus that is conspicuously absent from our analysis of the Navin et al. data. Of note, the CCNE1 locus is found as a significant marker when analyzing the unmixed components but is not detected by a similar marker analysis of the raw data without unmixing. All other markers

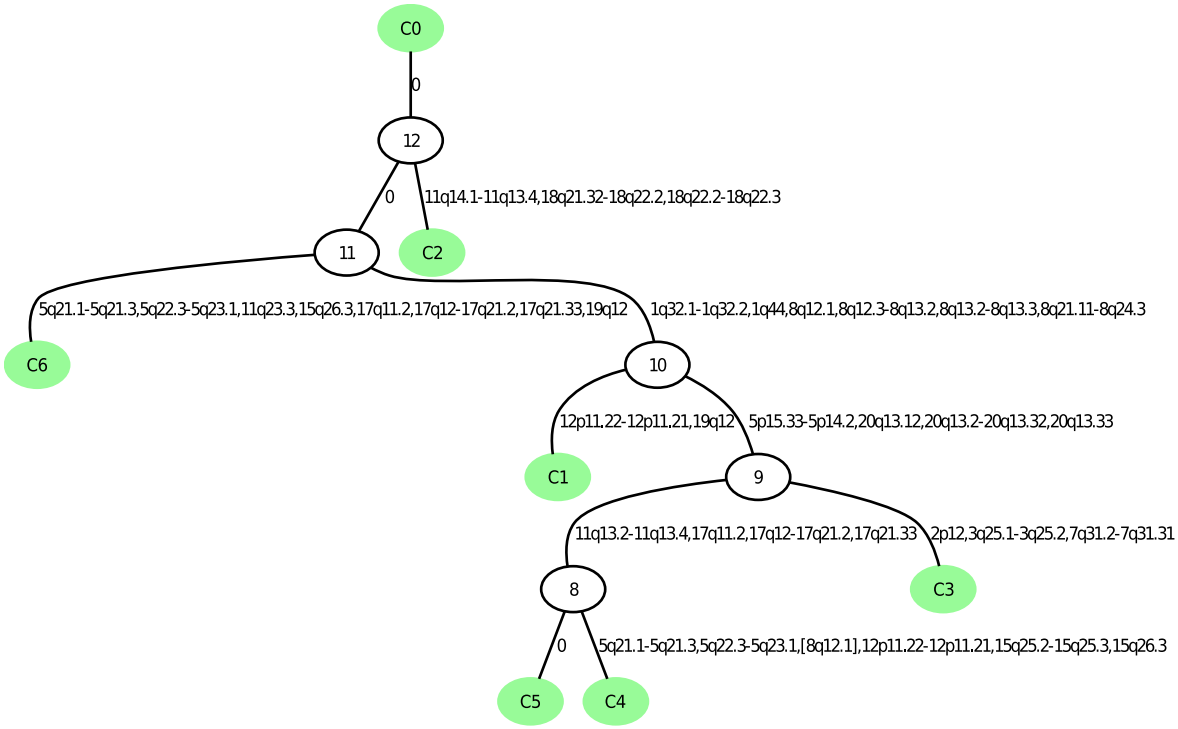


Figure 4.5: Inferred phylogenetic tree for the mixture components from the data of Navin et al. [38]. Nodes are labeled by component for the six inferred components C1-C6 and the normal component C0. Internal nodes are inferred ancestral states (Steiner nodes) and are each labeled by a unique identifier (8-12). Tree edges are labeled with the markers inferred to be amplified across each. Markers inferred to be lost along a given edge are shown in brackets and edges with no markers gained or lost are labeled "0".

found in the unmixed data are also found in the raw data, as was observed with the Navin et al. data. Figure 4.6 shows the inferred phylogenetic tree. For these data, it was not necessary to add a normal root component C0, as wsa done with the Navin et al. data, because the method directly inferred component C1 to be non-amplified at all markers and thus to serve as the expected normal root.

4.3.2 Discussion

Analysis on simulated data shows the method to have generally good accuracy at identifying amplified markers, identifying complete components with defined patterns of marker amplification, and grouping these components into phylogenies. The dependence of accuracy on various model parameters is difficult to analyze, with generally better marker-level accuracy but worse component-level and tree-edge-level accuracy as greater numbers of components are modeled. Examination of different noise levels, chosen to roughly approximate noise levels

Table 4.3: Amplified markers with probe boundaries and corresponding cytogenetic coordinates for the data of Pollack et al. [132].

Marker Id	Start Probe Id	End Probe Id	Cytogenetic Coordinate
1	1	37	1p36
2	136	272	1p34-1p22
3	330	649	1p13-1q44
4	671	790	2p24-2p13
5	1170	1210	3p25 - 3p21
6	1810	1889	5p15-5q11
7	2056	2229	5q23 - 6p21
8	2253	2331	6p21
9	2532	2865	7p22-7q36
10	2935	3106	8p12-8q24
11	3235	3264	9q22-9q31
12	3800	3926	11q12-11q14
13	4194	4255	12q12-12q14
14	4478	4522	13q22-14q11
15	4523	4566	14q11-14q12
16	4968	5367	16p13.3-17q11
17	5384	5448	17q11.2-17q21
18	5478	6056	17q21-19q13.4
19	6057	6230	20p13-20q13.33
20	6231	6312	21q11-21q22.3

observed on the real data, show no strong dependence within a range of 5%-20% noise. Overall, the results suggest that methods show good although far from perfect performance, picking out 79.2% of true markers and greater than 72.8% of true components in most scenarios and correctly identifying 94.8% of tree edges dividing the identified components. The high specificity of the marker assignment, with no false positives observed in any of the tests, suggests that there may be room to tune the methods to improve accuracy by trading off sensitivity for a somewhat higher rate of false positives. While simulated data provides some assessment of the effectiveness of the method, however, there are many features of tumor evolution that are not yet well enough understood to permit a faithful simulation of real tumor data. In assessing our methods, we must therefore rely primarily on more indirect validation on real data.

There is no closely comparable method to ours of which we are aware that we could use as a basis for comparison and we therefore validate the results on the Navin et al. data primarily by considering whether they are consistent with prior knowledge about breast tumors. One could in principle validate our results against recent work of Navin et al. [94] using single-cell analysis of the subsections of Tumor 10 analyzed here. Navin's phylogenetic approach, however, leads to progression trees dominated by changes in overall ploidy, which is not examined in our trees and precludes any direct comparison. As noted previously, a majority of the markers we find correspond to some genes with known cancer associations. These include well characterized breast cancer amplicons at 17q, 11q, and 20q [134, 135, 136]. The most notable absence among well known breast cancer markers would be the 8p locus associated with the gene FGFR1. A majority of the markers (16 of 27) include genes with some annotated relationship with cancers, although only 7 of those (markers 1, 12, 13, 20, 22, 24, and 26) are annotated in OMIM as specifically associated with breast cancers.

Of those markers lacking an annotated association with breast cancers, many

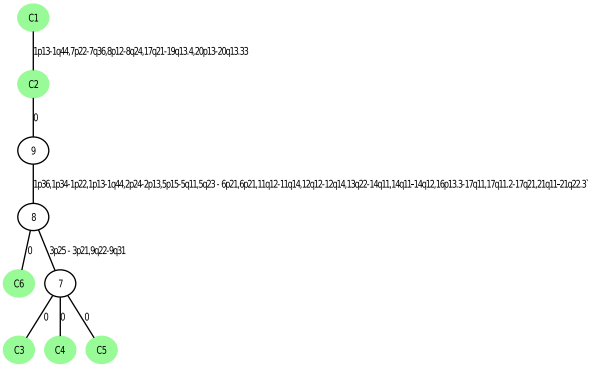


Figure 4.6: Inferred phylogenetic tree for components derived from the data of Pollack et al. [132]. Nodes are labeled by component for the six inferred components C1-C6. Internal nodes are inferred ancestral states (Steiner nodes) and are each labeled by a unique identifier (7-9). Tree edges are labeled with the markers inferred to be amplified across each. Markers inferred to be lost along a given edge are shown in brackets and edges with no markers gained or lost are labeled "0".

are in close proximity to and inherited with breast-cancer associated markers and might plausibly be assumed to contain distinct portions of common amplicons. Table 4.4 identifies those proximal markers that are co-inherited in the tree and likely reflect common amplicons. For example, 17q is interpreted as three distinct markers (markers 19–21), and although only marker 20 contains genes with an annotated breast cancer association (ERBB2/ Her-2/neu, STAT5, and GRB7), all are inherited together apparently as a common amplicon. Similar explanations can account for markers 2 on 1q, which is coinherited with marker 1 (MDM4); markers 10 and 11 on 8q, which are coinherited with marker 12 (MYC); and marker 25 and 27 on 20q, which are coinherited with marker 26 (ZNF217, CYP24A1, BCAS1, and AURKA). In other cases, however, we observe coinherited markers for which no specific explanation is available for any of the markers. It is impossible to say purely from a computational analysis whether these represent false positives, discoveries not annotated specifically in OMIM, or even novel but significant associations with breast cancer progression.

Examining the phylogeny itself allows us to further examine the possible biological significance of the data and its concordance with current knowledge about breast cancer progression. In this regard, it is helpful to interpret the tree as a set of possible progression pathways from the healthy root cell type (C0). As the tree implies, however, different progression pathways do not function in isolation but rather may share some common features in early progression.

The first internal node, Steiner node 12, is inferred to be identical to the root, but diverges at the top level into two pathways. The first such progression pathway (C0 $\rightarrow$ 12 $\rightarrow$ C2) describes a short terminal progression pathway isolated from

Co-amplified markers	Phylogeny edges
18q21.32-18q22.2,18q21.2-18q21.3	12→ C2
1q32.1-1q32.2,1q44	11 →10
5q21.1-5q21.3, 5q22.3-5q23.1	11 → C6, 8 → C4
8q12.3-8q13.2,8q13.2-8q13.3,8q21.11-8q24.3	11 →10
20q13.12,20q13.2-20q13.32,20q13.33	10 → 9
7q31.2-7q31.31	9 → C3
15q25.2-15q25.3,15q26.3	8 → C4
17q11.2,17q12-17q21.2,17q21.33	11→ C6, 9 → 8

Table 4.4: Marker regions amplified simultaneously during tumor evolution. The table provides, for each such set of marker regions, a unique identifier, cytogenetic coordinates, and corresponding specific edges or paths in the phylogenetic tree.

the rest of the tree. The progression pathway is resolved only to a single step of mutation corresponding to amplification of 11q14.1-11q13.4, 18q21.32-18q22.2, 18q22.2-18q22.3. 11q is a known breast cancer amplicon [135, 136] and harbors CCND1, which has been found to be amplified in breast cancers [137]; FGF3 and FGF4, which are known oncogenes [138]; and CTTN, which is frequently overexpressed in breast cancers [139]. The region also contains other genes, such as NPAT, with functions in cell cycle regulation that might be considered candidates for an oncogenic function. 18q21.32-18q22.2 harbors the oncogene BCL2, which is involved in the MYC pathway [140] and TNFRSF11A, which is frequently expressed in late stage breast cancers [141, 142]. The marker also harbors several SERPIN genes known to be tumor associated. 18q22.2-18q22.3 does not carry any currently known cancer-related genes but may be gained due to proximity to 18q21.32-18q22.2 as part of a common amplicon. Together, these abnormalities appear to define a distinct sub-class of breast tumor cells with early divergence from all other cell types.

Within the sub-branch rooted at Steiner node 11, one branch leads directly to a terminal node characterizing a second progression pathway (C0 $\rightarrow$ 11 $\rightarrow$ C6). This progression pathway is characterized by amplification of 5q21.1-5q21.3, 5q22.3-5q23.1, 11q23.3, 15q26.3, and 19q12 and is one of two sub-trees characterized by amplification of 17q11.2, 17q12-17q21.2, and 17q21.33. The 17q region is a well established breast cancer hotspot [134, 136], including genes ERBB2 (Her-2/neu), GRB7, and STAT5. 19q12 contains CCNE1, an important prognostic marker for breast cancer progression [143, 144]. CCNE1 amplification has been specifically associated with basal-like breast cancers [145], but has been previously identified as co-associated with particularly aggressive Her-2 positive breast tumors [146]. Our phylogeny is consistent with the notion that 17q/19q co-amplification defines a distinct sub-type of Her-2 positive tumors. Region 15q26.3 has no genes specifically noted to be breast-cancer associated in OMIM, although amplification of the locus was identified as predictive of recurrence in systematic breast cancers [147] and the region contains IGF1R, an anti-apoptotic gene broadly amplified in cancers [148]. The biological significance of the 5q amplicon is not apparent.

While 5q22.3-5q23.1 has several genes associated with cancers (e.g., ATG12, TNFAIP8, SEMA6A, which are associated with lung cancer), they are predominantly tumor suppressors. Likewise, there is no obvious relevance to the 15q amplicon, although it is close to other known 15q markers.

The next major division in the tree corresponds to the branch from Steiner nodes 11 to 10, characterized by gains in 1q32.1-1q32.2, 1q44, 8q12.1, 8q12.3-8q13.2, 8q13.2-8q13.3 and 8q21.11-8q24.3. Both 1q and 8q are rich in tumor-associated genes. 1q32.1 includes the breast cancer associated gene MDM4, a putative oncogene involved in apoptosis regulation of p53 activity [149], in addition to various genes associated with cancers more generally. 8q21.11-8q24.3 includes the MYC locus, another well known breast cancer amplicon [136]. We can suggest, then, that the 11 $\rightarrow$ 10 branch corresponds to a specific subset of progression pathways characterized by MYC amplification and suppression of apoptosis.

A third progression pathway can be identified within this branch through progression into C1 (C0 $\rightarrow$ 12 $\rightarrow$ 11 $\rightarrow$ 10 $\rightarrow$ C1). The final step on this pathway is characterized by amplifications on 12p11.22-12p11.21 and 19q12. 19q12 is the locus of CCNE1 suggesting a generic connection to cell cycle control on this pathway. 12p11.22-12p11.21 has no known cancer-related genes but carries the apoptosis-related gene DNM1L and the telomerase-related gene DDX11 [150].

Further progression pathways diverge from Steiner node 10 through Steiner node 9 with gains on 5p15.33-5p14.2, 20q13.12, 20q13.2-20q13.32, and 20q13.33. The 5p amplicon contains two genes with known cancer associations, CDH18 [151] and PAPD7 [152], although neither appears to have a known role in breast cancers specifically. 20q13.2-20q13.32 contains several genes associated with breast cancers, including ZNF217, CYP24A1, BCAS1, and AURKA [136], making it difficult to ascribe a particular mechanism to this branch.

Within the Steiner node 9 subtree, we can characterize a fourth progression pathway terminating in C3 (C0 $\rightarrow$ 11 $\rightarrow$ 10 $\rightarrow$ 9 $\rightarrow$ C3). The final step on this progression pathway corresponds to gains on 2p12, 3q25.1-3q25.2, and 7q31.31-7q31.32. The 7q31.32 marker contains the WNT2 gene associated with many cancer types, including breast cancer [153]. 7q31.31 has no known cancer related genes and is perhaps gained due to its proximity to 7q31.2. 3q25.1-3q25.2 has been previously detected as an amplicon in fraction of breast cancers [154], although we can offer no mechanistic explanation for its presence. We are not aware of any prior suggestion of an association between 2p12 and cancers.

The remaining two terminal nodes of the tree, C4 and C5, appear likely to represent two steps on a common progression pathway. Both branch from Steiner node 9 through 8 by acquisition of 17q11.2, 17q12-17q21.2, 17q21.33 (the Her-2 locus) along with 11q13.2. This subtree might thus be characterized primarily as a second Her-2 positive progression group associated with gain of CCND1, distinct from the Her-2 positive progression group terminating at C6 and associated with gain of CCNE1. C5 branches from Steiner node 8 with no changes, indicating a single progression pathway corresponding to C0 $\rightarrow$ 11 $\rightarrow$ 10 $\rightarrow$ 9 $\rightarrow$ C5 $\rightarrow$ C4.

The final step in this pathway is then characterized by a series of amplifications on 5q21.1, 5q22.3, 12p11.22, 15q25.2, 15q26.3 and loss of 8q12.1. We would not expect loss of a previously gained marker, and can suggest that this apparent loss might be better explained as a miscall of the state of that marker. Most of these loci have no annotated association with any cancers, with the only specific annotated breast cancer association being to 11q13.2, described above. This lack of associations may again represent false positive inferences specifically associated with this component. We can suggest, however, that such markers might be have been missed if they are specific only to late progression of one sub-type of Her-2 positive breast tumor. Summarizing across the tree, we can note that there is clear support in the prior literature for many of the specific markers, although there is little evidence one way or the other supporting the specific sequences of mutations suggested by our phylogeny analysis. Nonetheless, these pathways make several novel predictions that may warrant further investigation. Chief among these would be the identification of two apparently distinct pathways to Her-2/neu amplification that separate relatively early in progression and exhibit distinct sets of co-occurring amplifications.

The tree suggests several distinct patterns of co-amplification that may be useful in identifying or classifying novel sub-types, particularly with respect to Her-2 amplifying tumors. Of particular interest are the observation of two distinct Her-2 amplifying subtrees, one showing coamplification with CCND1 and c-myc and the other with CCNE1. Loden et al. have previously reported separate Cyclin-D amplified and Cyclin-E amplified subgroups of breast cancer following separate pathways of oncogenesis, with Her-2/neu overexpression and c-myc amplification accompanying both subgroups. Co-amplification of Her-2, CCND1, and c-myc is supported by additional literature, with this particular coamplification associated with later or more advanced stages of breast cancer [17, 146, 155]. Janocko et al. [17], however, does suggest that c-myc amplification should occur late in this sequence, a finding not supported by our phylogeny. Other more recent work has supported the idea of Her-2 and CCNE1 coamplification in breast cancers [156, 157] with Scaltriti et al. specifically suggesting this coamplification as a possible mechanism for Herceptin resistance in Her+ breast tumors. Other patterns of coamplication are apparent in the tree although not to our knowledge supported by prior literature or any obvious functional interpretation, e.g., the observation of coamplification of loci on 5q and 15q in both Her-2 amplifying subtrees.

Additional analysis of the Pollack et al. [132] provides little additional insight into breast tumor development, although it does provide some independent validation of our method. While the lower resolution of those data prevents an analysis of specific amplified breast tumor genes comparable to that done with the Navin et al. data, we can nonetheless observe that the method is effective at picking out those amplicons noted by the authors of that study. Furthermore, the additional markers it detects beyond those four include several of those also inferred to be important progression markers on the Navin et al. data and supported by exten-

sive prior literature, most prominently the loci of Her-2, CCND1, and CCNE1. These results show that the method can robustly find at least some prominent known tumor markers across two distinct sets of tumor samples using very different aCGH platforms and distinct unmixing methods. The tree itself provides no obvious new insights into breast tumor progression, as the method detected only four components that were actually distinct at the level of assigned markers, with three components determined to be amplified at all markers. Furthermore, all identified components were inferred to lie along a single progression pathway. It is notable that the tree implies amplification of most of the identified markers in a majority of components, perhaps because of the late clinical stages of the tumor samples and the presence of cell lines that would provide reasonably homogeneous representations of advanced states of breast tumor progression.

4.4 Conclusion

In this chapter, we have developed a computational pipeline for tumor phylogeny inference from genome-scale profiles of tumor state, specifically to test the feasibility of using computational unmixing methods to circumvent the problem of cell type heterogeneity in tumor phylogeny inference. We have developed a set of statistical tests to allow us to analyze computationally inferred mixture components — representing inferred profiles of well populated cell types from which heterogeneous tumor samples can be explained — to identify phylogenetic markers, assign them to specific inferred cell types, and use them in phylogenetic inference of tumor progression. We have demonstrated the approach with specific application to aCGH DNA copy number data, applied to a breast cancer data set [38], showing that the method is effective at locating biologically meaningful markers of tumor progression and assembling a biologically plausible model of breast tumor progression pathways. The inferred progression pathways provide several novel suggestions about possible steps in tumor evolution and key molecular abnormalities associated with progression. These inferences may provide useful guidance into the basic biology of tumor development as well as suggestions of possible targets for future diagnostics and therapeutics. Further application to a secondary lower-resolution breast tumor data set [132] and to a series of simulated aCGH data sets provides additional evidence for the effectiveness of the method at identifying markers of tumor progression, grouping them correctly into well-represented progression states, and accurately placing these states in phylogenetic trees.

Validation remains a challenge for tumor phylogeny inference, as there is no alternative method by which we can determine progression pathways with certainty for any real tumor data set. Simulated data can lend some confidence that the method works effectively relative to a model of the real data, as has been done here, but real tumor progression mechanisms are likely to be far more complex than our simulation models can capture. Comparison to single-cell approaches

like FISH [17, 63, 61] and single-cell sequencing efforts [94] can help to verify the pure cell states determined by the unmixing within a single sample and potentially validate some ancestral states predicted by the phylogenetic inferrence. FISH data provides only a few markers per cell, making it infeasible for a comprehensive validation of the results of our method, but could be used prospectively on targeted markers selected from an inferred phylogeny. Single-cell sequencing approaches could in principle eventually overcome this limitation given sufficient volumes and quality of data. Other sources of data in which more information is available about the true pathways of progression might also be useful. While we know of no such data currently available, one might in principle construct such a data set by, for example, studying discrete passages of cell lines or through the use of animal models in which one can monitor tumor development and progression over time. While gathering such a data set would be beyond the scope of the present work, it could in principle provide a basis for a more thorough future assessment of the accuracy of the pipeline implemented here or other methods for the problem of tumor phylogeny inference.

While this pilot study was intended to establish the feasibility of an unmixing approach to tumor phylogenetics, there are many ways by which the work might be advanced in the future. It will be important to further establish the reproducibility of the specific markers and phylogenetic pathways in additional breast tumor datasets. Novel markers found to be robustly predictive of particular progression pathways will ultimately need to be experimentally verified. In addition, it will be important to establish that the approach is applicable to other forms of tumors. Each of the individual steps of analysis also might benefit from improvement. The approach developed here depends on use of an unmixing method for identifying progression states, a problem which itself might benefit from improvements in the model and algorithms to more precisely fit the kind of sparse, noisy data characteristic of tumor data sets. Adapting the methods to more reliable data types, such as next generation sequencing data, may also prove valuable in that regard. The results on marker detection suggest there is room for improvement in more precisely determining the fine-scale structure of specific amplicons, especially when contiguous regions show distinct patterns of amplification across components. Likewise, there would appear to be room for improvement in better discriminating between normal and slightly elevated copy numbers. It is a weakness of the general approach that, because the unmixing models must work in linear rather than log space, they have difficulty distinguishing the relatively small linear change between normal and deleted regions. Improving sensitivity for deletions, or for subtler variations among amplification levels, may provide additional data for phylogeny construction. Finally, the phylogeny construction itself used a standard parsimony method not specifically tailored to tumor progression. This parsimony model has advantages in not requiring parameters for which there is currently no empirical basis and in allowing us to test for unexpected behavior, such as loss of previously amplified regions, that can help to validate the method.

Nonetheless, there is now sufficient data that one might in principle learn more sophisticated probabilistic models of cancer progression or of the behavior of particular amplicons and build these models into the phylogeny inference.

Chapter 5

Novel multi-sample scheme for inferring phylogenetic markers from whole genome tumor profiles

In this Chapter, we present HMM-CNA, a novel multi-sample Hidden Markov Model (HMM) capable of inferring phylogenetically informative markers of progression from whole genome tumor data. We present applications of the method to both unmixed and raw tumor data. The contents are adapted from the conference proceedings [158] and the extended journal paper [78]

Computational cancer phylogenetics seeks to enumerate the temporal sequences of aberrations in tumor evolution, thereby delineating the evolution of possible tumor progression pathways, molecular subtypes and mechanisms of action. We previously developed a pipeline for constructing phylogenies describing evolution between major recurring cell types computationally inferred from whole-genome tumor profiles. The accuracy and detail of the phylogenies, however, depends on the identification of accurate, high-resolution molecular markers of progression, i.e., reproducible regions of aberration that robustly differentiate different subtypes and stages of progression. Here we present a novel hidden Markov model (HMM) scheme for the problem of inferring such phylogenetically significant markers through joint segmentation and calling of multi-sample tumor data. Our method classifies sets of genome-wide DNA copy number measurements into a partitioning of samples into normal (diploid) or amplified at each probe. It differs from other similar HMM methods in its design specifically for the needs of tumor phylogenetics, by seeking to identify robust markers of progression conserved across a set of copy number profiles. We show an analysis of our method in comparison to other methods on both synthetic and real tumor data, which confirms its effectiveness for tumor phylogeny inference and suggests avenues for future advances.

5.1 Problem and Background

In the present work, we focus specifically on the problem of marker inference from array comparative genomic hybridization (aCGH) data providing genome-scale DNA copy number measurements. For these data, the problem corresponds to finding discrete genomic regions of DNA gain or loss that can serve as markers of tumor progression.

Existing methods for aCGH analysis include algorithms for smoothing, segmentation and combined segmentation and classification of both single- [121, 159, 160, 161, 162, 163] and multi-sample data [164, 165, 166, 167, 168, 169, 170]. Such methods can be highly effective at identifying discrete copy number variations in such data, but are poorly suited to the problem of phylogenetic inference because they do not constrain solutions to common markers across tumor samples. They thus provide no straightforward way to infer a set of robust markers with defined boundaries across patients and progression states for use in phylogenetic inference. A similar objective was considered by Picard et al. for their method, CGHSeg [171], which addresses the problem of joint segmentation and calling of multiple samples primarily as a way of improving accuracy of assignment using similarities between data. This method, though, was also not designed for the purpose of phylogenetic inference, and is inefficient for the data characteristics needed for these purposes, especially the combination of large numbers of markers with defined boundaries across a modest number of discrete samples characteristic of whole-genome datasets.

Our method is distinguished from other methodologically similar segmentation methods for CGH data primarily in that it is designed specifically to facilitate phylogenetic inference from tumor samples. We favor character-based phylogenetic methods, which allow us to intepret evolution of tumors in terms of gain or loss of specific discrete amplicons. For such inferences, we must interpret raw copy number data as sets of phylogenetic characters for which we can assign discrete states to each sample in a data set. To be useful for phylogenetic inference, such characters must describe common regions of copy number change that are shared across multiple samples. Hence, it is essential for our purposes to have a joint segmentation and calling algorithm that can output discrete phylogenetic character data. Typical segmentation algorithms, which seek only to find most plausible explanations of the raw data in terms of regions of amplification or loss, are unlikely to produce segmentations that yield common shared regions of gain or loss across samples. As detailed in Approach, our method involves a variety of innovations designed to improve its ability to find shared markers across samples useful for phylogenetic inference. In Results and Discussion, we show using simulated aCGH data that these innovations lead to an improved ability over prior methods to find markers and call them accurately in individual samples and that these improvements in marker detection translate to improved ability to reconstruct phylogenetic trees.

In the previous chapter, we had developed an approach to the problem of tumor phylogenetics based on the use of mixture models to infer discrete states of progression recurrent across tumor samples [72, 73]. We subsequently used this mixture modeling approach as the basis for a pipeline for tumor phylogeny inference [123]. For this pipeline, we developed a multi-sample segmentation method based on a simple statistical test applied to fixed-length windows of probes heuristically merged to identify amplicons from a set of inferred mixture components. The unmixing procedure in its present formulation can only reliably infer amplifications and, hence, we focus only on copy number amplifications in this work. An additional statistical test would then call presence or absence of each amplicon in each component, converting the components into discrete character arrays suitable for character-based phylogenetic inference. Validation on a set of components derived from real breast tumor data [38] showed the marker selection method to be reasonably effective at finding known breast cancer amplicons suitable for use as phylogenetic markers. The segmentation step, however, showed a poor ability to resolve fine-scale structure within amplicons, limiting the number of phylogenetic markers and the ability of the method to discriminate between subtle changes in nearby markers. In addition, separating segmentation from calling left no way to guarantee that amplicons detected in the segmentation stage would in fact be called differently in different components and thus become useful markers for phylogenetics.

The present work is aimed at developing an improved marker detection method designed to maintain the advantages of our prior work in using multi-sample segmentation from mixture components to identify a robust set of common markers usable across samples, while adapting ideas from prior single-sample methods to improve fine-scale resolution of amplicon structure. The method uses a novel HMM scheme to do joint segmentation and calling of markers simultaneously from a set of mixture components. It is thus similar in character to the method of Picard et al. [171] although with fewer assumptions about shared features of amplicons across samples. Both FLLat [170] and the HMM-mix model in [165] deal with the issue of heterogeneity inference in multi-sample aCGH data through mixture modeling. The outputs are not directly suited for phylogeny analysis of a set of input samples as they consist of representative driver aberration profiles, similar to the outputs of our mixture models, rather than phylogenetic characters derived from those aberration profiles as in the present work. Other HMM-based methods [162, 163] are either single-sample based, primarily platform-specific or focus on other issues of multi-sample analysis. Our new approach allows joint segmentation and thus detection of phylogenetically useful markers across mixture components. In contrast to our prior work, the use of the HMM scheme also allows the method to detect changes in assortments of amplicons across components within regions of amplification. We analyze the method on both simulated and real data and compare it to related methods heuristically adapted to the problem of phylogenetic calling. The results show the method to give superior perfor-

mance at both marker inference and phylogenetic reconstruction for biologically reasonable levels of experimental noise.

5.2 Approach

Our model is based on a generalization of the use of HMMs to multi-sample data for the purpose of finding a common marker set across a set of samples. It accomplishes this task by treating states of the HMM as tuples of amplification states across samples, with each copy number probe assigned one state. Any contiguous region of common state in which at least one component is called amplified can then serve as a single marker for phylogenetic inference.

5.2.1 The HMM model

5.2.1.1 Notation

Let the data **X** consist of m samples, each sample being a vector of log copy number intensity ratios at n genomic coordinates. We assume each of the m copy number profiles are ordered in genome coordinates starting from chromosome 1 to chromosome 22 and potentially X and Y. Thus **X** is a m x n data matrix where each element x_{ij} is a copy number ratio in the log domain where $i \subset \{1, 2, ...m|$ and $j \subset \{1, 2, ...n\}$. A Hidden Markov model defines the joint probability distribution of the sequence of x_{ij} in the observed matrix **X** by using another latent or hidden sequential state set. The HMM divides **X** into k distinct segments **S** where $k << n$ and each segment s_t is assigned one of the possible hidden copy number states defined below and $t \subset \{1, 2, ...k\}$. Each s_t is made up of as many members x_{ij} as its length. We denote by s_{at} an element x_{ij} that belongs to segment t of length l and is at position a in the segment where $t \subset \{1, 2, ...k\}$ and $a \subset \{1, 2, ...l\}$. An illustration of our model is shown in Fig. 5.1

We assume no linkage disequilibrium between the x_{ij}s and they are hence assumed mutually independent for all j. Further, we do not take into account whether the individuals are heterozygous or homozygous at each x_{ij}. We also note that as a preprocessing step, we smooth input data by replacing each probe value with the average over a window of five consecutive probes centered on that value.

5.2.1.2 Hidden State Space

We assume two possible copy number states for each x_{ij} : normal or aberrated (loss/gain). The normal state is indicated by 0 and aberrated by 1. The copy number states can be further assigned ploidy defintions whereby the normal state is thought of as being diploid and the aberrated state is aneuploid. Then for any

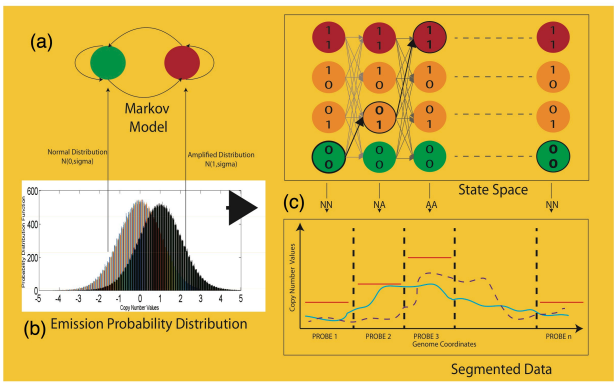


Figure 5.1: Representation of our HMM model, HMMCNA. The amplicon model (a) seeks to explain each probe in each progression state as either normal (green) or amplified (red) based on its fit to one of two copy number distributions (b). The HMM model (c) allows simultaneous maximization of the likelihood of these assignments across all probes and progression states, in the process segmenting the data and producing markers suitable for phylogenetic analysis. In the two-sample HMM example of (c), nodes labeled "1 1" (red) correspond to positions at which both samples are amplified, those labeled "0 0" (green) to positions at which neither sample is amplified, and those labeled "1 0" or "0 1" (orange) to positions at which exactly one of the two samples is amplified.

position i, the hidden state is a binary vector $\mathbf{H}_i$ of size m where each element h_i is either 0 or 1 and $i \subset \{1, 2, ...m\}$. Each $\mathbf{H}_i$ is thus one of 2^m possible state vectors in this 2-state paradigm. We, however, believe that the optimum segmentation of a dataset will normally be defined by fewer than 2^m combinations of unique state vectors. The assumption of n-tuples over $\{0, 1\}$ for n samples is particularly useful for character-based phylogenetic methods where the data must be represented as discrete states across markers.

5.2.1.3 Parameters

By definition, the sequence of states in the HMM follows a Markov model with transition probabilities defined between each pair of states. We assume the Markov model to be ergodic. Because our goal is to produce a phylogenetically useful set of amplicons rather than to infer the true amplicon structure per se, we do not learn model parameters directly from the data. Rather, we seek a model that will favor a simpler representation of the amplicon structure specifically preferring fewer and longer amplicons and preferentially finding amplicons with shared boundaries across samples. For this reason, we build into the model a prior expectation of the approximate frequency and length of amplicon expected, encoded in the HMM transition probabilities as follows:

1. Transition Probabilities (**A**)

The Markov model underlying the HMM is described in Figure 5.1.

As explained above, the basic Markov model has two possible states for each x_j : normal or 0 (N) and aberrated or 1 (A). We define four possible transitions:

(a) p_{NN}: The probability of staying in the normal state.

(b) p_{NA}: The probability of going from the normal state to an aberrant state.

$$p_{NA} = \left(\frac{p}{n * m}\right)\left(\frac{1}{2^m - 1}\right)$$

where p is a penalty set to 0.001 in the present work, effectively penalizing the model for assigning large numbers of amplicons by creating a prior expectation of 0.001 amplicons occurring by chance across the entire data set. The value of 0.001 was chosen to act comparably to a p-value of 0.001 used in statistical approaches to this problem, effectively requiring a 1000-fold excess in likelihood for amplicon versus no amplicon to identify a region as amplified.

(c) p_{AA}: The probability of going from an aberrant state to another aberrant state (or to itself; the possibilities are assumed to have the same transition rates). We set $p_{AA} = \frac{w-1}{w}$ to enforce an average amplicon width w, where we assume in the present work that $w = 20$. The other two transition probabilities are then fixed by p_{AA} and p_{NA}.

(d) p_{AN}: The probability of going from an aberrant state to normal.

$$p_{AN} = 1 - (2^m - 1) * p_{AA}$$

and

$$p_{NN} = 1 - (2^m - 1) * p_{NA}$$

which is derived by subtracting the probability of going to all other $2^m - 1$ aberrant states.

2. Emission Probabilities ($\mathbf{O}$)
Estimating Empirical Noise Levels: Before we define the emission probabilities, we introduce a measure to determine noise in copy number data that exploits the spatial dependence of the data. Empirical results on real aCGH datasets show that the data is log-Laplacian distributed [73], but we can adopt the approximation of this distribution as log-normal, modeling log copy number data as a true signal with additive Gaussian noise:

$$X_{ij} = S_{ij} + \mathcal{N}(0, \sigma^2)$$

where S is the signal. This log-normal model is commonly used for modeling aCGH data [170].

We introduce a non-standard formulation for inferring the noise in this framework that takes into consideration the spatial distribution of the probes. We developed an estimator of variance or, equivalently, standard deviation σ based on the average difference between adjacent probe values. We can pose this estimate in terms of the expectation of the difference between two normal random variables:

$$\begin{aligned}
\sum_{i,j} \frac{|X_{i,j} - X_{i,j+1}|}{m * (n-1)} &= E[|N_{i,j}(\mu, \sigma^2) - N_{i,j+1}(\mu, \sigma^2)|] \\
&= E[|\mu - \mu|] + E[|N_{i,j}(0, \sigma^2) \\
&\quad -N_{i,j+1}(0, \sigma^2)|] \\
&= \sigma E[|N_{i,j}(0,1) - N_{i,j+1}(0,1)|] \\
&= \sigma E[|N(0,2)|] \\
&= \sqrt{2}\sigma E[|N(0,1)|] \\
&= 2\sqrt{2}\sigma \int_0^\infty \frac{x}{\sqrt{2\pi}} \exp \frac{-x^2}{2} dx \\
&= \frac{2\sigma}{\sqrt{\pi}} \int_0^\infty \exp(-u) du \\
&\quad (u = x^2/2) \\
&= \frac{2\sigma}{\sqrt{\pi}} - \exp(-u)|_0^\infty \\
&= \frac{2\sigma}{\sqrt{\pi}}
\end{aligned}$$

Therefore:

$$\sigma = \sum_{i,j} \frac{\sqrt{\pi}|X_{i,j} - X_{i,j+1}|}{2m(n-1)}$$

This non-standard formula is used, rather than the conventional estimate of standard deviation, $\sqrt{E[X^2] - E[X]^2}$, in order to better separate variance due to measurement noise, which we wish to model, and true variance in the signal due to different amplicon copy numbers, which we do not want included in the noise model.

To illustrate the difference between the two measurements, we can use a model of DNA drawn from a genome with amplified segments, where we assume for illustration a fixed segment length L with alternating amplification levels of 0 and K for some K, here simplifying by assuming no true measurement noise. In the limit of an infinite number of segments, the

standard estimator would measure variance to be:

$$E[X^2] - E[X]^2 = \frac{K^2}{2} - \left(\frac{K}{2}\right)^2 = \frac{K^2}{2} - \frac{K^2}{4} = \frac{K^2}{4}$$

and thus standard deviation to be $K/2$.

Our estimator, on the other hand, would add contributions to the estimate only at boundaries between segments, giving for a single genome of infinite length an estimated standard deviation of $\frac{K\sqrt{\pi}}{2L}$ and a variance of $\frac{K^2\pi}{4L^2}$. In general, then, our estimator will suppress spurious estimates of standard deviation of the noise due to true amplification by a factor proportional to the average amplicon length. Our expectation is that this will lead to more accurate estimates of the parameter σ of our noise model for real data than will a straightforward measurement of standard deviation of the data.

We can bound variance of the noise estimator under the assumption that the input is a stream of n i.i.d. normal random variables, corresponding to consecutive probes, by noting that the estimator would then be described by a random variable of the form

$$\frac{\sqrt{\pi}}{2n}\left(\sum |Z_j - Z_{j+1}|\right)$$

where each Z_j is assumed to be an independent $N(0, \sigma^2)$ random variable. The variance in the estimator would then be given by:

$$\frac{\pi}{4n^2}\mathrm{Var}\left(\sum |Z_j - Z_{j+1}|\right)$$

This in turn is given by

$$\frac{\pi}{4n^2}((n-1)\mathrm{Var}\left(|Z_j - Z_{j+1}|\right)$$

$$-(n-2)\mathrm{Cov}\left(|Z_j - Z_{j+1}|, |Z_{j+1} - Z_{j+2}|\right))$$

for some arbitrary $1 < j < n$. That expression can be bounded as follows:

$$\frac{\pi}{4n^2}((n-1)\mathrm{Var}\left(|Z_j - Z_{j+1}|\right)$$

$$-(n-2)\mathrm{Cov}\left(|Z_{j+1} - Z_j|, |Z_{j+1} - Z_{j+2}|\right))$$

$$\leq \frac{\pi}{4n^2}(n-1)\mathrm{Var}\left(|Z_j - Z_{j+1}|\right)$$

$$
\begin{aligned}
&= \frac{\pi}{4n^2}(n-1)\mathrm{Var}\left(|N(0,\sigma^2)-N(0,\sigma^2)|\right) \\
&= \frac{\pi\sigma^2}{4n^2}(n-1)\mathrm{Var}\left(|N(0,1)-N(0,1)|\right) \\
&= \frac{2\pi\sigma^2}{4n^2}(n-1)\mathrm{Var}\left(|N(0,1)|\right) \\
&= \frac{2\pi\sigma^2}{4n^2}(n-1)(E[|N(0,1)|^2]-E[|N(0,1)|]^2) \\
&= \frac{2\pi\sigma^2}{4n^2}(n-1)\left(E[\chi_1^2]-\left(\sqrt{\frac{2}{\pi}}\right)^2\right) \\
&= \frac{\pi\sigma^2}{2n^2}(n-1)\left(1-\frac{2}{\pi}\right) \\
&\approx \frac{0.6\sigma^2}{n}
\end{aligned}
$$

The variance of our estimator can thus be bounded by a term that falls approximately linearly with the number of probes, n, which can be expected to yield accurate estimates of σ for genome-scale data. We empirically validate the performance of the estimator in the Results and Discussion below.

Defining Emission Probabilities: Once we have an estimate of the noise level, we define emission probabilities O by assuming each measured copy number x_{ij} comes from either a normal diploid distribution or an aberrant aneuploid distribution:

$$P(O_d|H) = \phi(x;\mu_d,\sigma) \text{ and } P(O_a|H) = \phi(x;\mu_a,\sigma)$$

where we assume here that diploid data has a mean $\mu_d = 0 + \mu$, where 0 corresponds to a mean ratio of one between observed data and a diploid control in the log-domain, and aneuploid data is modeled as having a mean ratio μ_a=1+μ relative to a diploid control. The additive term μ is an empirically estimated mean of the data, used to control for overall background amplification that may arise due to overall signal aneuploidy or as an artifact of the unmixing process.

3. Initial State Probabilities (π)
   The initial state probability π for all aberrated states is assumed to be $q = (p/(2^m-1)/n)$ leaving an initial probability of the normal state of $1-(2^m-1)q$.

5.2.2 Selection of Optimal States

We employ an extension of the Viterbi algorithm to determine the optimal sequence of copy number states for a given multisample copy number data set, assigning amplification or normal condition to each sample at each probe. A state here is defined, as above, as a tuple of binary normal/amplification assignments for all samples at a single probe. Our method differs from the generic Viterbi algorithm only in that our outputs are real-valued copy number measurements, rather than a discrete set of output characters, and our emission probabilities are thus drawn from log normal distributions to allow for continuous values. This extension still allows for optimal solution of the log likelihood via dynamic programming, as with Viterbi over a discrete state set. More specifically, we find a maximum likelihood solution H of hidden state assignments by optimizing for the subproblem $\hat{H}(i,j)$, defined to be the maximum likelihood assignment of amplification states to the first i probes terminating in state b_j, for some canonical ordering of amplification vectors $b_0, \ldots, b_{2^m-1}$ where b_0 is defined to be the all-diploid vector.

We solve this problem using the recurrence:

$$\hat{H}(i,j) =$$

$$\max_k \begin{cases} \hat{H}(i-1,0)p_{nn}\prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j=0 \\ \hat{H}(i-1,0)p_{an}\prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j\neq 0 \\ \hat{H}(i-1,k)p_{na}\prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j=0, k\neq 0 \\ \hat{H}(i-1,k)p_{aa}\prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j\neq 0, k\neq 0 \end{cases}$$

where x_{il} is the observed copy number of probe i in sample l and b_{jl} is the binary amplification state of sample l in state j. The optimal assignment is then derivable by identifying $\max_k \hat{H}(n,k)$ and backtracking to reconstruct the full state assignment.

The above recurrence relation admits a dynamic programming algorithm with runtime $\mathcal{O}(2^{2m}n)$. The resulting algorithm was implemented in MATLAB.

5.3 Experimental Methods

5.3.1 Synthetic Data

To assess accuracy on data of known ground truth, we simulated a series of aCGH data sets across a range of assumed experimental noise levels. We assumed a log-normal noise model $Y_{ij} = M_{ij} + \mathcal{N}(0,\sigma)$ for each sample i ($i = 1,2,\ldots,m$) and aCGH probe position j ($j = 1,2,\ldots,n$). Here, Y_{ij} is the simulated copy number ratio in the log domain, M_{ij} is the amplification model and $\mathcal{N}(0,\sigma)$ is Gaussian noise. We modeled the distribution of copy numbers in tumor data by an exponential distribution $M_{ij} = 1 + \mathbb{1}(j \subset S_i)\mathrm{Exp}(\lambda)$ where $\mathbb{1}$ is the indicator

Table 5.1: Qualitative comparison of HMMCNA with other state-of-the-art copy number segmentation methods. The table distinguishes methods based on whether they perform marker calling, whether they work on single- or multi-sample data, and whether they are generic with respect to input data or speciific to a particular data platform.

Method	Segmentation	Calling	Data	Platform specificity
CBS[121]	Yes	No	Single-sample	No
PennCNV[162]	Yes	Yes	Single-sample	SNP-Array
PICNIC[163]	Yes	Yes	Single-sample	SNP-Array
CGHSeg[171]	Yes	Yes	Multi-sample	No
GISTIC[169]	Yes	No	Multi-sample	Yes
HMMCNA	Yes	Yes	Multi-sample	No

function for the presence of site j in an amplicon S_i. We estimated the exponential rate λ from the real component data in Sec. 5.3.2 using the mean of observed probe values above 5, to minimize contamination by non-amplified probes. We then simulated a series of components to model tumor evolution over a complete binary tree of depth three. Beginning from an all-diploid root, we simulated amplicons of fixed width $w = 20$ in a hypothetical data set of 1161 probes (to match the proportion of amplifications in the real data) in 6 components, adding one new amplicon per non-root node to those present in the node's parent to model acquisition of successive amplicons over succeeding generations of progression. Amplicons were placed uniformly at random within the genome, rejecting and rerunning any placement that resulted in two amplicons within w probes of one another. We then generated observed signal values for amplified and non-amplified sites by the log-normal noise model described above. This process was repeated for 200 replicates each at noise levels $\sigma = 0$ to 1.8 in increments of 0.1.

Because our method uses an estimate of noise level derived from the data, we perform a preliminary validation of our estimates of noise level on the simulated data. Specifically, at each noise level, we apply our estimator of noise standard deviation σ to the data and evaluate its inferred value and standard deviation of that value by our estimator and a generic standard deviation computation.

For each replicate, we ran the HMM algorithm as described in Sec. 5.2. For comparison, we tested the same data on two alternatives: the single-sample method Circular Binary Segmentation (CBS) [172] using the MATLAB function *cghcbs* and the multisample *multiseg* function in the R package CGHSeg [171]. While there is no comparative method developed specifically for phylogenetics, we chose to compare with one single-sample and one multi-sample copy number segmentation algorithm. The CBS output was called at a threshold of $log_2(1.5)$ as

amplified or normal. CGHSeg returns called values for each sample. Downstream analysis was performed to extract and merge probes called amplified in at least one sample to yield recurrent markers with common boundaries, each of which serves as a character for the phylogeny inference. Our choice of the algorithms CBS and CGHSeg was based on the accessibility to code, platform non-specificity and popularity of use. We have compared our method on some major usability and functionality criteria in Table 5.1.

Phylogenetic trees were inferred by adding an all-diploid root to the set of character states and then running unweighted maximum parsimony inference using PAUP [34].

Given an accurate phylogeny reconstruction algorithm, the accuracy of the phylogenies will depend on the quality of input markers or characters. The estimated markers must first be truly representative of changes in copy number. Second, normal regions of the genome must not be assigned amplification states. Third, for each sample, the markers must only be assigned amplification states if they are indeed present in the sample and represent the correct character state assignment for that sample. Quality of the methods by these criteria was measured on three tasks. First, accuracy of amplicon detection across samples was quantified by the sensitivity, defined as fraction of genuinely amplified markers assigned to an amplicon, and specificity, defined as the fraction of markers assigned to an amplicon that were in fact amplified. Second, accuracy of marker assignment to amplicons was measured, quantified by the fraction of amplicons correctly called as amplified or non-amplified for all components. Finally, accuracy of phylogeny inference was assessed, quantified by the Branch Score Distance [173] using the *treedist* function of PHYLIP [33], a measure of agreement between the true and inferred phylogenies.

5.3.2 Experiments : Real Data

5.3.2.1 Unmixed Data

We further demonstrated our methods on real data derived from a publicly available (NCBI GEO GSE16672) primary ductal breast carcinoma aCGH dataset [38]. This data set was chosen because the cell sorting and sectioning methods underlying the tumor data extraction were developed specifically to aid phylogenetic analysis, making them well suited to our purposes, and because the data contains multiple samples per tumor, making them especially useful for studies of tumor heterogeneity and mixture analysis. The raw data comprises 87 tumor sectors obtained from 14 ductal breast cancer tumors run on a high-density ROMA platform with 83,055 probes. We confined our analysis to the twenty-two autosomal chromosomes, reducing the dataset to 78,874 probes. We converted the raw aCGH data from log to linear domain, denoised it with a total variation denoising and then subjected it to an unmixing analysis to infer 6 components, or putative tumor

progression states, as described in [73]. We next converted the data back to the log domain after recentering around a mean of 1. We then ran our method as described in Sec. 5.2 using PAUP for maximum parsimony tree building as with the simulated data.

5.4 Results and Discussion

5.4.1 Synthetic Data

Because our method relies on an estimate of noise level in its input data, we begin by verifying the accuracy of our estimator. Fig. 5.2 shows a comparison of the proposed data noise estimator with the estimated standard deviation of the data. The results show that our estimator gives a highly accurate estimate of the noise level on our simulated data sets. We note that the 1161 probes used in each simulated data set is low compared to a typical genome-scale aCGH data set and the accuracy of the estimator would therefore be expected to be greater for typical real data sets. By contrast, the standard deviation of the data provides a highly biased estimate of noise, especially at lower noise levels, because it conflates noise in the data with variance due to true amplicons.

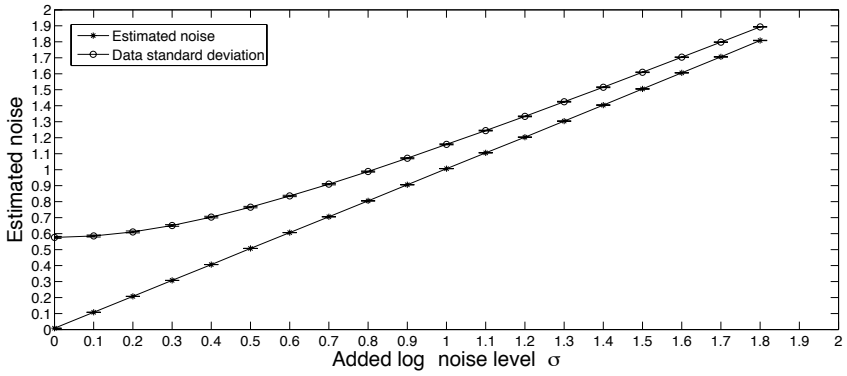


Figure 5.2: Comparison of noise estimates on simulated data derived from our method with those derived using the standard deviation of the data versus the true noise levels simulated for the data. Error bars show standard error of the estimates for each method.

We next examined the effectiveness of our HMMCNA method in comparison to the available competing methods and our own prior work on the simulated data. The results are summarized in Figure 5.3. Fig. 5.3(a,b) shows accuracy at the level of amplicon assignment. Fig. 5.3(a) shows that our method has a higher sensitivity than either of the comparative methods or our own prior method [123] at low to medium noise levels (up to about 0.6). Anecdotally, we have found that the noise inference computation described earlier yields values in the range

of 0.1–0.5 on a selection of real datasets. At higher noise levels, the sensitivity drops sharply. Fig. 5.3(b) shows that all three methods have a high specificity for amplicon calling, with no false positive calls until relatively high levels of noise. At high noise levels, CGHseg is most prone to false positive calls, CBS least prone, and our own method intermediate between the two. At lower noise levels (< 0.2), our method has the least specificity in comparison, a result expected due our method's windowing approach, which raises the likelihood of incorrectly grouping normal probes adjacent to an amplicon into the amplicon.

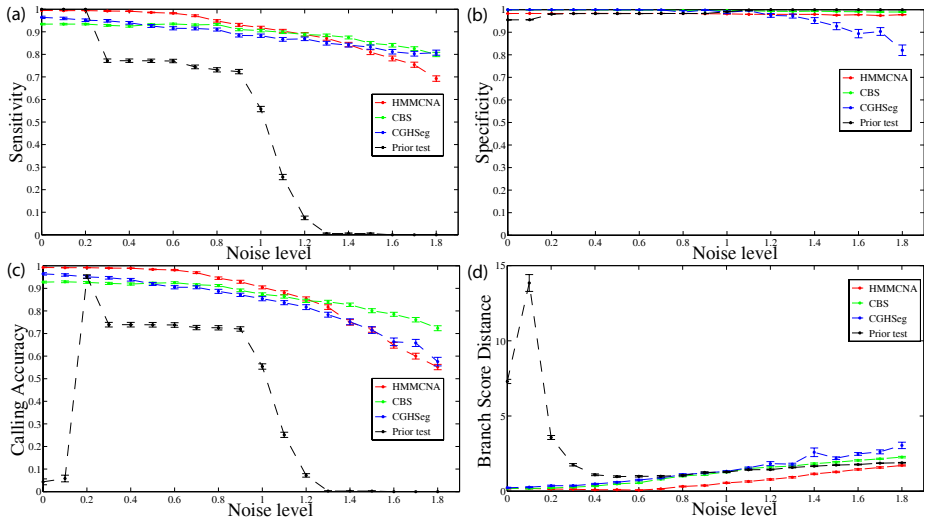


Figure 5.3: Accuracy of our method (HMMCNA), CBS, CGHseg, and our prior method on simulated data. (a, b) Accuracy in amplicon assignment, classified by the sensitivity (a) and specificity (b) of correctly assigning markers. (c) Calling accuracy, measured by the fraction of amplified markers assigned the correct amplification state. (d) Tree-building accuracy, quantified by the branch-score distance between the true and observed tree. All measures are reported as functions of the log-normal noise level σ, averaged over 200 independent runs per noise level.

Fig. 5.3(c) shows accuracy of calling amplification states within detected amplicons. All three methods closely track the sensitivity plot of Fig. 5.3(a) up to a noise level of about 1.0, suggesting that each is highly accurate in calling states given the amplicons at low to moderate noise levels. Again, our method shows a drop in calling accuracy at higher noise levels in comparison to the competitors.

Fig. 5.3(d) shows the accuracy at inferring phylogenetic trees, which is the specific goal of our method. Here, our method shows superior performance in comparison to CBS and CGHSeg across all noise levels. This result may be attributed to high calling accuracy in general combined with a specific bias of our method for finding amplicons with shared boundaries across samples, which are especially useful for phylogenetic inference. It is interesting to note that while CBS

has better calling accuracy at higher noise levels, its phylogenetic performance is not commensurate. This observation can be explained either at the marker inference step, where inconsistencies in boundary detection between samples may create problems for phylogenetic inference, or at the phylogeny-building stage itself, in that the order of phylogenetic markers can influence the topology of the resulting trees. We can thus conclude that our method does provide an advantage over the existing methods in accurate phylogeny reconstruction in the presence of moderate but biologically realistic noise levels.

5.4.2 Real Data

5.4.2.1 Results on Unmixed Data

We next applied our method to mixture components derived from the real breast cancer data set of Navin et al. [38] both for further validation and to illustrate its value in predicting progression on real tumor samples. The HMM method found 315 marker amplicons, more than a 10-fold increase compared to the 27 detected by our prior method [123]. There are, on average, 91 amplicons per component with markers spanning 74.81% of the genome. Analysis is complicated by the fact that some inferred amplicons are quite large and include many genes, which might be presumed to be predominantly passenger genes irrelevant to the progression process. It has been observed that small amplicons, in the range of a few megabases, are a distinct phenomenon from the large chromosome-scale amplifications produced by aneuploidy and translocations [174], which we believe account for the bulk of the total genome coverage. We therefore screened out inferred amplicons covering more than 148 probes (approximately 2.5 Mb) and examined enrichment of the shorter amplicons alone for known breast cancer markers. This reduced the portion of the genome found in some amplicon to 16% of the autosomal probes. We used the UCSC genome Table browser NCBI build 35 (corresponding to the aCGH array platform build) to find 3869 unique genes within the remaining small amplicons (versus 15869 for the set of all detected amplicons). We then used the Catalogue Of Somatic Mutations In Cancer (COSMIC) Database v. 57 [175] to specifically identify those associated with breast cancer, identifying 1014 breast cancer associated genes covered by short amplicons (versus 4126 in the full amplicon set) out of a total of 6973 breast cancer associated genes in COSMIC. To test whether these numbers suggest an enrichment for breast cancer-associated genes in our amplicons, we performed a chi-square test of significance of enrichment of our gene set for breast cancer markers relative to the full 23307 unique Refseq-curated human genes in NCBI build 35. The short amplicons were found to be significantly enriched for breast cancer associated genes (chi-square score 30.24, p-value < 0.0001). The set of both large and small amplicons was also strongly enriched (chi-square score 363.41, p-value < 0.0001). Anecdotally, this set of amplicons carries several important markers not identified by our earlier

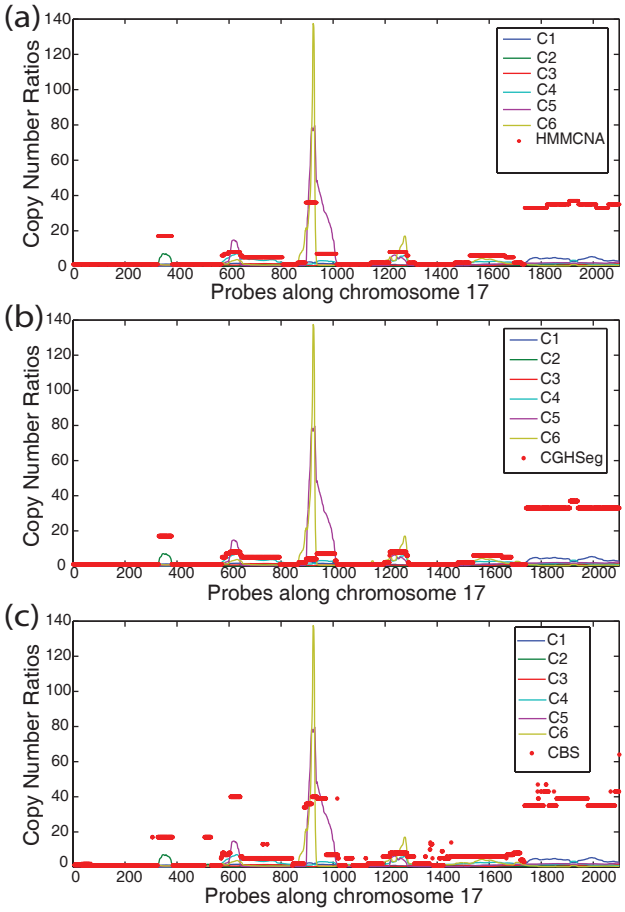


Figure 5.4: Segmentation of chromosome 17 using mixture components of Navin et al. (a) Our method, HMMCNA. (b) CGHSeg. (c) CBS.

method, notable among them being JUN, BRAF, KRAS, FGFR1, ESR1 and JAK2.

Figure 5.4 provides a visual comparison of results of our method to those of CBS and CGHSeg, using chromosome 17. Our method and CGHSeg produce similar results, although with some additional fine-scale amplicon structure identified by our method. CBS produced considerably more breakpoints than either other method. Over the entire genome, CBS produced 1425 distinct marker segments, a much higher number than our own method spanning 93.8% of the genome. We cannot definitely say to what degree these extra breakpoints reflect better sensitivity to true variations versus spurious breaks due to experimental noise. CGHSeg has substantially higher computational cost and could not complete analysis of the full genome in more than a month of processing and we therefore do not provide a full comparison to that method. It should be noted, though, that neither of these methods are designed to work with mixture components of the sort for which our method was developed, which might be expected to conform poorly to their error models.

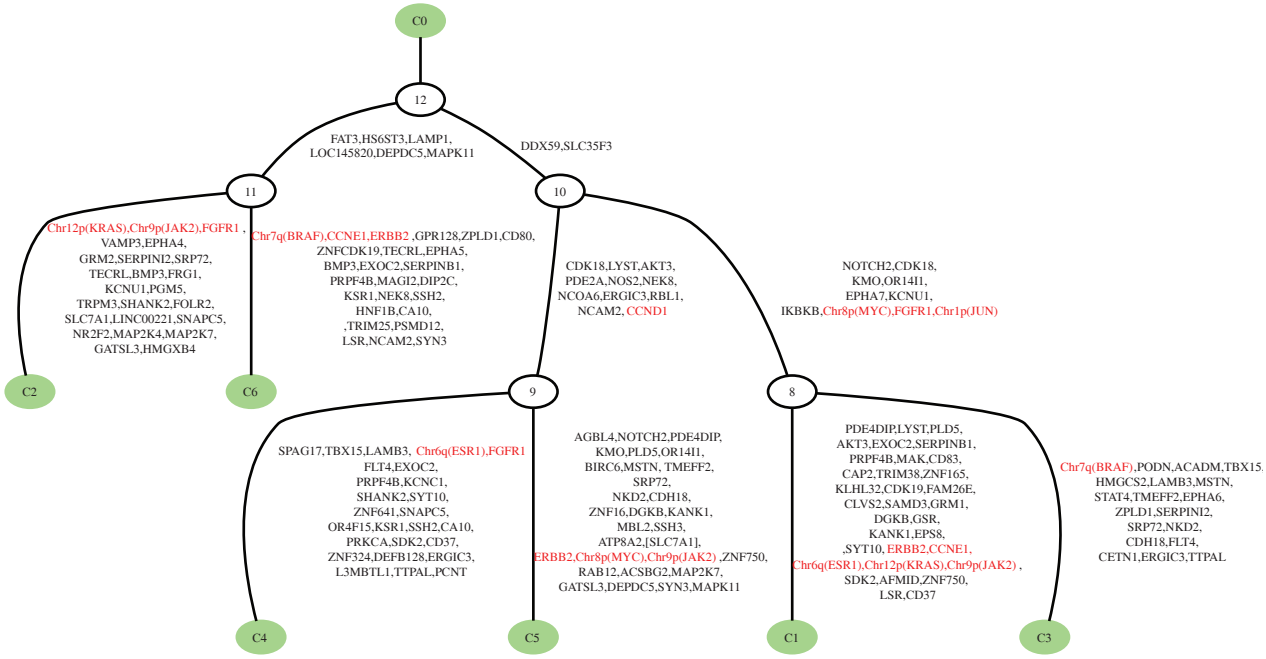


Figure 5.5: Maximum parsimony tree inferred from mixture components derived from real breast cancer data of Navin et al. [38]. Edges are labeled with putative driver genes, with those of particular note as breast cancer progression markers highlighted in red. Amplicons of 148 or fewer probes (approximately 2.5 Mb on average) are listed by gene while selected larger amplicons are listed by chromosome arm with genes of interest in parentheses. Green nodes are observed components and white are inferred ancestral states, also known as Steiner nodes.

Next, we analyzed the phylogenetic tree obtained from the markers, summarized in Figure 5.5. Nodes correspond to putative stages of progression and edges to amplicons gained during discrete steps of progression. For purposes of annotation of the phylogeny, we identified specific genes for the short amplicons, favoring those in the COSMIC breast cancer set when a short amplicon covered multiple genes and using genes cited by Navin et al. [38] in their own analysis of their data to break ties. We annotated only a subset of large amplicons manually chosen because they carry genes we expect to be particularly important to breast cancer progression.

The resulting tree is shown in Figure 5.5. The tree exhibits homosplasy (recurrent mutation) but no reversion of markers, a result we believe to improve upon that of our prior method [123], which exhibited both homoplasy and reversions. While the homoplasy might reflect genuine convergence of distinct progression pathways, it could also be explained by false positive calling errors or errors in phylogeny inference due to the maximum parsimony assumption.

Table 5.2: Computation run-time on real data for CBS, CGHSeg and our method, HMMCNA over the entire genome.

Method	Runtime
CBS[121]	1.0395h
CGHSeg[171]	41 days
HMMCNA	20.1s

Analyzing the tree in more detail reveals several features of note. The progression pathway to C5 occurs with the gain of HER2 (ERBB2) and CCND1 suggesting a distinct arm of HER2/CCND1 co-amplification. There are two other progression pathways leading to C6 and C1 that also show HER2 amplification. The pathway leading to C6 has an amplicon housing CCNE1, consistent with a notion of two distinct forms of HER2-amplifying tumors. It has been reported recently that co-occurrence of HER2 and CCNE1 leads to Herceptin therapy resistance in HER2 overexpressing breast cancer [156, 157]. The phylogeny supports this idea of distinct pathways of evolution of HER2-amplifying breast cancers, specifically including one pathway co-amplifying with CCND1 and one co-amplifying with CCNE1. We also observe late co-amplification of HER2 and a large amplicon containing MYC in both CCND1-amplifying and CCNE1-amplifying variants, as well as a CCNE1/HER2-amplifying pathway that does not co-amplify MYC.

5.4.3 Runtime Analysis

We also compared the computation run-time for all three methods. The results are shown in Table 5.2. The results show HMMCNA to be by far the most efficient method, requiring seconds per chromosome. CGH-Seg was the least time-efficient. CBS gave intermediate values. These results illustrate a secondary advantage of our method in scaling efficiently to many more probes than the alternatives, a key advantage for a method designed for working on whole-genome data.

5.5 Conclusion

We have developed a novel method for joint segmentation and calling of multi-sample genome-scale DNA copy number data, designed specifically for use in tumor phylogenetics. The method uses a novel multi-sample HMM approach to identify consistent markers across a set of samples, typically mixture components inferred from raw tumor data, for use as markers for phylogenetic inference. Comparison with a state-of-the-art multi-sample scheme and a leading single-sample scheme shows that our method has superior performance at levels of experimental noise typical of real aCGH data for the specific task of tumor phylogenetics, as well as for the more general task of tumor marker inference. Further, the method

substantially improves on our own prior work for the problem of phylogenetic inference from inferred mixture components through a novel HMM approach for multi-sample amplicon detection and improved methods for modeling noise in the data. In particular, our method outperforms the alternatives, and substantially outperforms our own prior method, in the noise range of 0.0–0.6, a region that subsumes the noise range of approximately 0.1–0.5 we have estimated for real aCGH data. These methodological improvements lead to a more than ten-fold increase in the number of markers available for phylogeny inference and detection of several important progression markers not previously found from these data.

While there is no obvious direct way to validate the results obtained from running HMMCNA on real data, we have shown indirect support for our results through comparison to established marker sets and anecdotally supported features of the inferred trees based on previously published research. The issue of assessing the true validity of our results remains a challenge since there is no known ground truth for either the quality of inferred amplicons or the reconstructed phylogeny from the amplicons.

In future work, we hope to improve on the current approach through a more realistic model of amplification distributions including handling of genomic deletions, algorithmic improvements to avoid combinatorial increase in state size with components, and improvements in the upstream unmixing and downstream phylogenetic inference steps. We further hope to explore how one might better tune the method to specifically detect markers most likely to be informative for phylogenetic inference. In addition, the method may have value for other applications of copy number data in phylogenetics and related problems.

Chapter 6

Phylogenetic analysis of tumor genome rearrangement data : A case study in the MCF-7 human breast cancer cell line

In this Chapter, we study genomic variation among MCF-7 breast tumor cell lines grown in different laboratories using phylogenetic methods. A manuscript is under preparation on this work.

MCF-7 cells are one of the most commonly used breast cancer cell lines in research. However, over rapid use and proliferation in various research settings, the cells have undergone considerable genetic and phenotypic drift over time. An understanding of these evolutionary differences is important as the field of research actively using this model system moves forward. Here, we describe a study of the evolution of a set of structural rearrangements in MCF-7 cell lines using data obtained from 7 MCF-7-sublines grown in different laboratories and 8 MCF-7- subclones derived by single cell cloning from a single subline. We apply computational phylogenetics, a set of tools to build phylogenies or evolutionary trees, on the structural rearrangement data. An analysis of the results supports conclusions from prior work and current experimental data and leads to key new observations relevant to the field.

6.1 Tumor Cell Lines

Human cancer-derived cell lines are widely used in cancer research as model systems for recapitulating tumor behavior. They often serve as the model of choice for pre-clinical studies, grown either in culture as monolayers or as xenografts in mice. The Cancer Cell Line Encyclopedia (CCLE) [176] has established a set of cell lines with data on the genome, transcriptome and epigenome profiles, which show close resemblance to primary tumors in their tissues of origin. Cancer cell line panels, which include cell lines derived from different individuals, have been used for *in vitro* tumor drug screening.
The use of in-vitro cell lines as pre-clinical models has been under debate [177] due to problems of cross contamination related to tissue or species of origin, or

the evolutionary selective effects of the lab cell culture environment which may distort the characteristics of the cell lines and as a result, the cell lines may no longer faithfully resemble the original tumor. For example, studies have shown that differences in gene expression characteristics among cell lines of different tissues are smaller than the differences between the individual cell lines and their matched primary tumor tissue.

While there has been considerable debate on the clinical relevance of cultured cell lines, proponents of cell lines cite easy propagation and reliability of results under defined experimental conditions as important factors for their use. Neve et al., [178] described a model system of 51 breast tumor cell lines, which mirrored both the patterns of heterogeneity and responses to targeted therapy observed in primary tumors. In the study, they compared transcriptional and genome copy number profiles for the cell lines with those measured for primary tumors. As new ex- vivo models of cancer continue to be developed, cancer cell lines continue to remain as one of the keys models sought in an integrative paradigm which include genetically engineered mice and xenografts among others.

In this paper, we investigate how laboratory environments and multiple passages affect genome rearrangements leading to heterogeneity in the same cancer cell line. The heterogeneity can, in some cases, provides advantages like drug resistance. Specifically, we study the MCF-7 breast cancer cell line. MCF-7 (HTB-22, ATCC) is one of the most commonly grown and studied breast cancer cell lines in the lab. It was originally isolated in 1970 at the Michigan Cancer Foundation (MCF) from a 69-year-old Caucasian woman, is positive for both estrogen- and progesterone-receptors, and has been used extensively to study hormone therapy response. However, observed phenotypic differences in MCF7 cells grown in different laboratories led many groups to name their own laboratory variants, or sublines.

Studies examining effects of different lab conditions on the cell lines have concluded major genome profile differences among these sublines. Graham et al., [179] used restriction fragment polymorphisms to establish genetic identity between MCF-7 cells from different passages and the original cell line. Osborne et al., [180] compared morphology, structural chromosome alterations and estrogen-responsiveness among 4 MCF-7 cell lines from different laboratories. Resnicoff et al., [181] studied heterogeneity among MCF-7 sub-populations demonstrating the existence of a subset of cells capable of giving rise to the different cell lines, leading to a stem- cell hypothesis in addition to the hypothesis of selective pressure under different culture conditions. Nugoli et al., [67] documented expression and cytogenetic changes in 11 MCF-7 cell lines. Overall, they observed important differences in copy number changes ranging from 28-31 breakpoints among the different profiles. Hampton et al., [182] generated a sequence level map of genomic rearrangements in MCF-7 cells. Genome re-arrangements were found to occur more frequently as LCRs (low copy repeats), are frequently repaired by the mechanism of NHEJ and can disrupt tumor suppressor genes. They also observed

that rearrangements are either clustered or dispersed and several specific gene-fusions and translocations are involved in the process that are directly associated with tumor progression.
MCF-7 is a tamoxifen responsive cell line. Coser et al., [183] studied antiestrogen-resistant subclones of MCF-7 and concluded that hormone therapy resistance arises from clonal selection of pre-existing drug resistant subpopulations as opposed to the theory that resistance is acquired on exposure to hormone therapy. Gonzalez-Malerva et al., [184] performed a series of dilution experiments where a master clone was serially diluted to obtain several other clones in an attempt to discover tamoxifen resistant clones. They assayed a signature set of tamoxifen resistant genes in 7 sub-clones derived from the master clone on serial dilution. There has been much interest in understanding the sequence of events leading to development of heterogeneity and hormone therapy resistance in MCF-7. Sharma et al., [185] reported that reversible drug tolerant states occur in tumor cell models through the acquisition of specific chromatin rearrangements associated with regulating genes like IGF-1R and KDM5-A.
Here, we document the evolutionary sequence of events underlying the Hampton et al., [182] data. We study the evolution of genome structural rearrangements among MCF7 sub-lines from different laboratories and sub-clones [184] using computational phylogenetics, which represents the progression of changes as a tree or phylogeny. In addition, we run clustering on the data to compare and identify specific advantages of tumor phylogenies. Lastly, we analyze the resulting trees to seek specific informative features of tumor biology. In Section 2, we describe the materials and methods used. In Section 3, we illustrate the results of the methods. In Section 4, we describe the analyses and key findings. We conclude with some discussion on the implications of the paper and future directions.

6.2 Materials and Methods

6.2.1 Data

We used the data generated by Hampton et al., [182], comprising two distinct datasets, which we call D1 and D2 for reference. D1 consisted of 7 MCF-7 sub-lines from different laboratories: MCF-7-B, MCF-7-ATCC, MCF-7-Neo, MCF-7-C, MCF-7-BK, MCF-7-D, MCF-7-L. D2 consisted of 9 subclones of a single master clone derived from a MCF-7 (MCF-7-LG) cell line as in [182]. The clones were labeled MCF-7-H9, MCF- 7-H7, MCF-7-G11, MCF-7-B2, MCF-7-F11, MCF-7-B3, MCF- 7-B7, MCF-7-C11. For both D1 and D2, the data comprised of discrete genome rearrangement information for a set of genome structural rearrangement breakpoints. The breakpoint locations were derived from an earlier report [182] in which a BAC library from MCF7 was sequenced and surveyed to create a genome-wide map of 157 genomic rearrangements. In this study, the specific 157

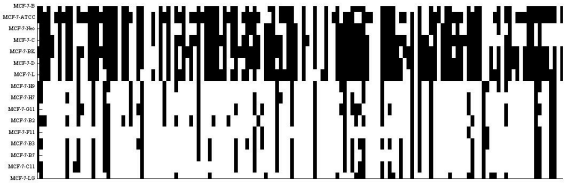


Figure 6.1: Breakpoint data for D1 and D2. Each column is a breakpoint. A black pixel represents the presence of the breakpoint.

rearrangements were examined in D1 and D2 using PCR (see table). The genomic co-ordinates and primers used for PCR have been previously reported [182] . Briefly, cell lines were lysed and DNA isolated using Qiagen DNeasy kit according to the manufacturerÕs instructions. PCR products were visualized using 2% agarose gel. If a band was detected at the correct size the rearrangenment was scored as present (1) and if not then absent (0). Not all breakpoints are found in all cell lines. After quality filtering for BAC induced breakpoints, D1 had 93 breakpoints and D2 had 133 breakpoints. The two datasets shared 80 breakpoints in common. These breakpoint regions were then targeted and sequenced for each of the samples in the two datasets. The breakpoint region sequences were subjected to PCR. Whenever a rearrangement was found, the breakpoint region for that cell-line were assigned a value 1 and if not, assigned a value 0 thus giving rise to discrete data. There is a possibility of false positives arising from multiple rearrangements. Thus, both D1 and D2 can be represented by two binary matrices where rows are specific genome breakpoint locations, and columns are the individual sub-lines or sub-clones. Each element is either a 0 or 1 representing the presence or absence of a rearrangement respectively. Figure 1 represents the combined data from D1 and D2.

6.2.2 Tumor Phylogenetics

Phylogeny reconstruction involves the inference of evolutionary events that lead to the observed breakpoints in the cell lines. With the structural rearrangement data, the phylogeny will represent the evolutionary drift among the cell lines under study in an assumed model of evolution. The cell lines are called taxa and will form the leaves and internal nodes of the tree while each edge will represent the breakpoints that occurred in transition between the nodes it connects.
Each breakpoint region serves as a character or trait for the phylogeny inference with two discrete states of 0 and 1 where 0 is absence of amplification and 1 is presence of amplification at the breakpoint region. Then each, subline or subclone or taxa in phylogeny parlance, forms the leaves of the resulting phylogeny and ancestral nodes maybe imposed during the tree building. We assumed independence of characters and unweighted unordered transitions between the two

discrete states of 0 and 1. We assumed the evolutionary model of maximum parsimony, which states that the evolution to the present taxa involved minimum number of changes when starting from ancestral states. Applying this model for phylogeny inference then results in a tree where each edge describes the specific rearrangements that occurred while transitioning between the nodes or taxa that it connects. The model favors that tree in which the sum of all the rearrangements across all the edges is minimum. We also separately apply a modification of maximum parsimony to enforce the direction of change, in other words, we consider the case that structural rearrangements once gained may not be lost to revert back to normal. We call the former version the model of unconstrained maximum parsimony and the latter the model of constrained maximum parsimony. While there is no efficient method to solve this problem in reasonable computational time, there are algorithms, that employ heuristics for solving this problem. We use the maximum parsimony function employed by the phylogeny inference software PAUP [34] for building the trees.

6.2.2.1 Unconstrained Maximum Parsimony Trees

For each dataset, we built maximum parsimony trees as described above on all breakpoints. This results in unrooted phylogenies. In order to set a reference, we impose a root node A0, composed of all zeros, i.e. no rearrangements. To establish statistical significance, we built consensus trees after bootstrapping with 50000 replicates. Tree edges with frequency of occurrence greater than 50% were used to build the consensus tree. Further, we built two phylogenies for each of the datasets on the commonly shared breakpoints. We also built one major phylogeny combining data across all two datasets on the commonly shared breakpoints.

6.2.2.2 Constrained Maximum Parsimony Trees

Constrained Maximum parsimony trees were built by imposing the constraint that amplified and re-arranged breakpoints cannot revert back to normal or move to the original place. Since no selective pressure was applied on the sub lines, this seems to be a reasonable assumption. Bootstrapped consensus trees were built as explained before. Various experiments for rooted and unrooted cases were also performed.

6.2.2.3 Clustering Analysis

To compare the results of tumor phylogenetics with standard clustering methods, we performed hierarchical clustering on the combined data with the root node imposed.

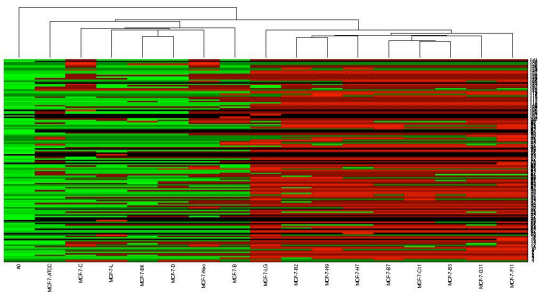


Figure 6.2: Hierarchical clustering of the breakpoint data.

6.3 Results

6.3.1 Genetic similarity amongst MCF7 sublines

We compared the overlap of genomic rearrangements between the seven MCF7 sublines. From a total of 157 rearrangements there were 31 (19.7%) that were common and in all cell lines. To determine the potential biological significance of this result, we examined whether these rearrangements were found in genic regions of DNA (within 50kb of a gene). In the total set of genomic rearrangements (n=157), 79 (50.3%) are found in genic regions. When considering the rearrangements which are shared between MCF7 sublines (n=31)24 (77%) were in genic regions, a statistically significant enrichment ($p<0.006$, fishers exact test). This suggests that these rearrangements may be under selective pressure. Importantly, there were several gene-gene fusion which were present in all MCF7 sublines. While there were 10 gene-gene fusions present in the 157 breakpoints (6.4%), the 31 common breakpoints contained 5 gene-gene fusions (16.1%) a 2.5 fold enrichement. It is possible that these gene-gene fusions are early driver events critical to MCF7 tumorigenesis.

6.3.2 Hierarchical Clustering

The sub clones and sub-lines are separated early on. MCF-7-LG appears ancestral to the rest of the sub clones. MCF-7-ATCC emerges as a separate group in relation to the rest of the sub-lines.

6.3.3 Phylogenies across sublines

We only consider the consensus trees obtained after bootstrapping as reliable trees for further analysis. All edges supported in over 50% of the boostrap replicates are indicated in red in all the figures. We first analyze the MCF-7 sub-lines in D1. Figure 3(a) shows the unconstrained maximum parsimony tree for the 7 sub-

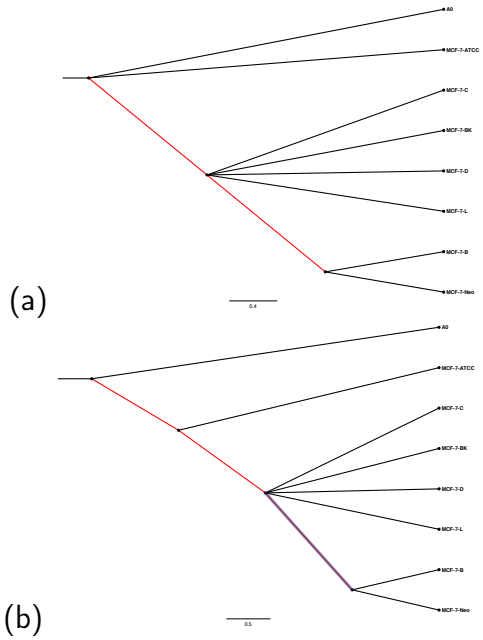


Figure 6.3: (a) Unconstrained Phylogeny across all breakpoints and A0 in D1. (b) Constrained Phylogeny across all breakpoints and A0 in D1.

lines in D1 and A0. We observe that there are three distinct clusters: (MCF-7-B, MCF-7-Neo), (MCF-7- BK, MCF-7-D, MCF-7-L, MCF-7-C) and (MCF-7-ATCC, A0). Figure 3(b) shows the constrained maximum parsimony tree for the same data. Here we see that MCF-7-ATCC is a separate distinct group of its own. MCF-7-B and MCF-7-Neo continue to group together. The phylogenies suggest an evolutionary order among the sublines.

6.3.4 Phylogenies across sub-clones

Figures 4(a) and 4(b) show unconstrained and constrained phylogenies for the 7 sub clones with the addition of A0. The tamoxifen sensitive clones MCF-7-C11 and MCF-7-B7 cluster together in unconstrained phylogenies. The constrained phylogeny does not exhibit any distinctive structure. The tamoxifen resistant sub-clones MCF-7-G11 and MCF-7-H9 occur in the same cluster.

Figure 5 represent unconstrained and constrained phylogenies for the combined dataset of D1 and D2 across all breakpoints. There are a total of 141 breakpoints. The unconstrained phylogeny shows a clustering of the tamoxifen resistant clones H9 and B2 together. The imposition of a model of evolution that only allows gain of structural rearrangements clusters the tamoxifen sensitive clones together.

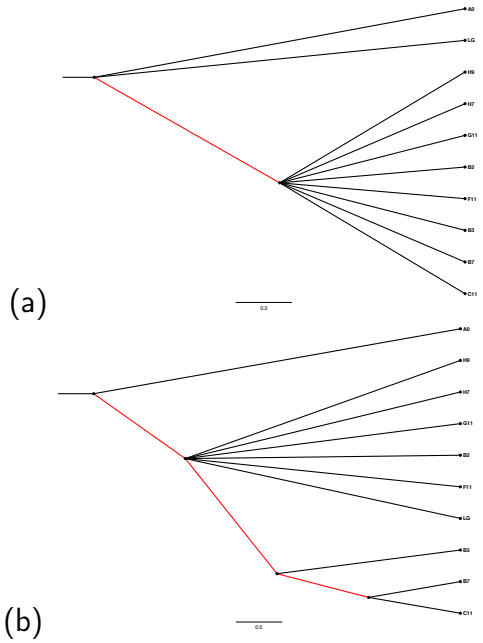


Figure 6.4: (a) Unconstrained Phylogeny across all breakpoints and A0 in D2 (b) Constrained Phylogeny across all breakpoints and A0 in D2.

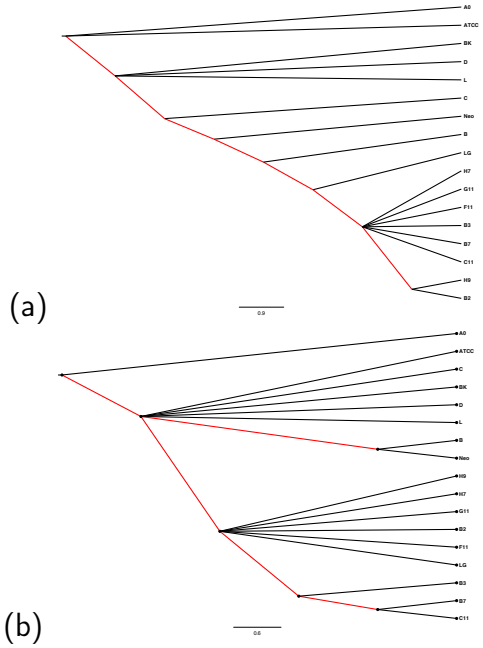


Figure 6.5: (a) Unconstrained Phylogeny across all breakpoints and A0 in D1 and D2 (b) Constrained Phylogeny across all breakpoints and A0 in D1 and D2

6.4 Discussion

While it is not clear which breakpoints maybe significant and which paths of progression are noteworthy, it is worth noting that MCF-7-ATCC branches out to its own distinct early on in the phylogeny and this is supported in prior work [180, 181, 67] . Also, B and Neo are related sublines and they appear clustered. Gonzalez-Malerva et al. [184] conclude that 2 of the 7 sub-clones G11, H9 are particularly tamoxifen-resistant and 2, B7 and C11, tamoxifen-sensitive. We note however that B2 and H9 cluster together in the consensus tree in the unconstrained setting while B7 and C11 cluster together farther away from the root in the constrained setting. MCF-7-LG is the parental cell line and this is supported in the tree.

Sharma et al., [185] have reported the existence of reversible drug tolerant states. We looked at some genome rearrangement events reported in [182]. Character 5 on the tree corresponds to the ARFGEF2/SULF2 fusion event and is an early event in all datasets. Character 35 corresponds to the PTPRG fusion event, which occurs early on in D2. Character 39 corresponds to the BCAS3-BCAS5 fusion, which is also an early event. Character 18 corresponds to the RAD51C-ATXN7 fusion event, which also occurs early on in the datasets. These maybe regions positively selected for genome instability.

6.5 Conclusions

We have demonstrated a method to determine tumor evolution in cancer cell lines using structural rearrangement data in sublines of the breast cancer cell line MCF-7. The analyses reveal that some sublines cluster together and may share evolutionary similarities. MCF-7-ATCC stands out distinctly as a separate evolutionary pathway. The presence of distinct evolutionary paths among seemingly similar cell-lines may present several caveats to conclusions garnered from their use as model systems for study. It also presents an important question for the field to consider: that of standardization and generalization of results on cell lines as a model system.

Chapter 7

A distance-based phylogenetic framework for tumor sequence data

In this Chapter, we present a strategy for reconstructing distance-based evolutionary trees from tumor whole genome sequencing data. A manuscript is in preparation on this work.

7.1 Introduction

Tumor evolution comprises the set of changes that a living tissue undergoes during its transformation from a healthy state with normal phenotype to a benign or malignant state with a cancerous phenotype with possible further progression to more advanced, aggressive and metastatic states. In his seminal paper, Weinberg described the six hallmarks of the cancerous phenotype as evading apoptosis, tissue invasion and metastasis, sustained angiogenesis,self-sufficiency in growth signals, limitless reproductive potential and insensitivity to anti-growth signals. At the molecular level, these phenotypic changes amount to changes in DNA copy number, RNA copy number, patterns of epigenetic regulation and levels of protein expression. Historically, tumor progression was studied by the pathologist using samples or biopsies of suspected tumors who described changes in phenotypes using grading systems reflecting changes in tissue morphology, and later, using immunohistochemistry to describe changes in cell surface protein expression and FISH to describe genome structural rearrangements and instability. Advances in molecular biology and genomics led to genome-wide study of RNA expression, DNA copy number and methylation patterns resulting in classification of tumor subtypes. It was thus established that while there are distinct temporal stages of tumor progression, patients showing similar phenotypes could still exhibit heterogeneity at the molecular level. Further advances in regional sectioning showed patterns of inter-tumoral heterogeneity as well projecting that subpopulations of cells can exist within the same tumor sample. Thus, the study of tumor evolution tries to capture the development and progression of both inter and intra-tumoral heterogeneity on the temporal scale.

Techniques for obtaining data for such studies have primarily either been single cell methods like FISH or genome wide methods like microarrays. Single cell methods offer the advantage of investigation at the level of individual cells while trading off on the number of molecular markers surveyed. Genome wide methods allowed higher resolution surveys of entire genomes but gave an average profile for a tissue of multiple cells. Regional sectioning accompanied with genome wide methods offered a way to combine the advantages of both single cell and genome wide methods. Unmixing approaches served to combine these advantages using computational methods. Genome wide sequencing technology offer advantages of very high resolution and also depth since we can identify changes at the single base level. Single nucleus sequencing methods offer both the advantages of genome wide sequencing and analysis at the single cell level.

The rapid innovation and technology development in genome-wide sequencing have made it possible to obtain high resolution nucleotide sequence data both faster and cheaper. The availability of massive amounts of high resolution sequence data in tumors have opened up unprecedented opportunities for molecular investigation, while also presenting unique challenges. First, the massive amounts of data present unprecedented demands in data storage, transfer and security. Second, the massiveness of the data require both the development of new algorithms for big data as well as the scaling up of current algorithms to handle large data. Third, the new technologies themselves come with new sources of error and noise which must be accounted for in the data processing model and pipeline. As such, research has made great strides in both improving next generation sequencing technology as well as development of better error correction and detection models. And this brings us to the fourth challenge of matching the right models of error [186] and data mining algorithms to the sequencing technology applied to the data under investigation and making further innovations to answer fundamental questions of scientific enquiry. In this chapter we present an attempt to understand breast tumor evolution by extending existing knowledge in computational phylogenetics to single nucleus sequencing data. In this chapter, we describe a strategy on reconstructing phylogenies from sequence data by inferring counts of k-mers derived from the sequence data. We use publicly available data [94] from a single nucleus sequencing method which employed whole genome amplification with random priming to obtain deep sequence reads with coverage of and an average read length of 35bp. The method was applied to 100 single cells combined from both primary and metastatic tumor stages in a single patient, referred to as T16P and T16M respectively, and 100 single cells from another primary ductal breast carcinoma sample; referred to as T10. Matched normal cell data was not available.

7.2 Approach

Earlier, we had established a character-based tumor phylogeny inference pipeline which consisted of three key steps. In the first step, we extract phylogenetically informative markers from the data. In the second step, we process the markers so that they are amenable to phylogeny algorithms thus yielding discrete character matrix. In the third step, we apply a phylogeny reconstruction algorithm. In previous work, we have used maximum parsimony inference. Next generation sequencing data poses some unique challenges as described in Section 1 and so we modify the pipeline to accommodate the data characteristics. First, we generalize the pipeline to include distance-based phylogenies. The first step would then consist of phylogeny marker selection, the next step as building the phylogenetic data matrix and the third step as phylogeny inference.

7.2.1 k-mer counts as a function of genome imbalance

We introduce a novel approach to understanding tumor behavior and evolution using genome sequence information. The discrete marker units we will use are k-mers or sequences of DNA of length k where k is minimum 5. For small ks, we anticipate finding all possible k-tuples of {A,T,G,C} in the data. Each sample then has a unique number of each kmer in its genome profile. We then compute the the unique count distribution of kmers found in the genome. The count distributions can then be used to build distance matrices for distance based phylogeny reconstruction. The count distribution vectors can also be manipulated for its predictive and classification potential.

7.2.2 Data noise

There are several contributions to the data noise which must be considered in the model

1. **Noise due to sequencing error** Each sequencing technology has some error rate for sequencing and this is usually a standard error rate per base. The major consequence of sequencing errors is confounding while variant calling since one cannot distinguish between variants and sequencing errors. Ideally, one would hope that sequencing errors would occur in a much smaller percentage than actual variants.

2. **Noise due to sample bias** Some samples may get a higher coverage or read depth than others. Such an error may be accounted for by correcting for the average coverage of the given sample while making comparisons among different samples.

3. **Noise due to random amplifications in the WGA** The amplification technology (E.g. Multi-displacement amplification) may pose noise as certain regions of the genome may be amplified in some samples and not in others. This maybe addressed by only looking at those regions which are amplified in all. Such an approach has the caveat that we cannot detect regions which are truly deleted or missing in the genomes of certain samples or regions which have rearranged in lengths which may not be detected by alignment. It may also lead to the analysis of a much smaller section of the genome.

4. **Noise due to alignment** The reference genome cannot account for all rearrangements and hence, detection of structural rearrangements remains a problem.

Depending on the technology used, the sequence data is likely to carry noise due to sample handling and sequencing errors. A base thresholding maybe used to rule out kmers that occur by random chance or sequencing errors. Further, each sample must be normalized in some way to account for sample-specific biases.

7.3 Methods

7.3.1 Data

Raw sequence reads from genomes of 100 primary breast tumor cells (T10) and 100 matched primary and metastatic tumor cells (T16) analyzed in Navin et al. 2011 were downloaded from NCBI SRA as fastq files.
Relevant links:
Study: http://www.ncbi.nlm.nih.gov/sra?term=SRP002535
T10 : http://www.ncbi.nlm.nih.gov/sra/SRX021401
T16P Primary breast tumor : http://www.ncbi.nlm.nih.gov/sra/SRX037035
T16M Metastatic liver: http://www.ncbi.nlm.nih.gov/sra/SRX037132

We only have primary and metastatic cells. We do not have normal cell genome counts for comparison. It would have been an advantage to have had normal cells in order to account for noise.

7.3.2 Phylogenies using kmer counts

7.3.2.1 Data Processing using Jellyfish

The fast k-mer counter Jellyfish [89] was used to count k-mers of lengths 5, 10, 15 and 20 from the individual cells . Jellyfish was downloaded from http://www.cbcb.umd.edu/software/jellyfish/. The commands "jellyfish count

-c k -o output -t 32 input.fastq " was used to create the k-mer count hash, "jellyfish dump -c output" to recover the counts of each k-mer from the hash and "count_in_file output" to merge counts from all cells to generate the count data matrix.

7.3.3 Reducing the data matrix for processing

The resulting merged kmer files increase in size with the length of k as these files are matrices of size N x M where N is the number of samples and M is the number of kmers which is on the order of 2^k. The files can be as large as 2.2T when k = 20 and this is one drawback of the method as the resolution must be then be compromised by subsampling. The subsampling must effectively retain only the informative kmers .

7.3.3.1 Accounting for data noise

First, only those kmers which were found in all samples were retained.This selection has the caveat that true deletions or mutations may be missed but reduces the space of kmers. Second, the number of reads for each config may be different due to the experimental run conditions, difference in abundance of DNA(copy number) in the sample, nature of the contig(repeats or micro satellite region) or other noise. To deal with differences in experimental run conditions, we may normalize the count of each kmer by the total number of kmers found in that sample. This leaves us with a matrix of kmer count fractions for each sample(single cell).

7.3.3.2 Distance-Based Phylogeny reconstruction

k-mers common to all cells were retrieved and a Euclidean distance matrix was built based on differences in counts or count fractions. When comparing across samples, we are comparing fractions of the genome occupied by different kmers. In other words, the distances capture the differences in genome composition across the samples. Neighbor joining trees were built using *neighbor* program in PHYLIP. 50000 bootstrap replicates were used to construct consensus neighbor joining trees.

7.3.3.3 Analyses of resulting phylogenies

To analyze the resulting trees, we defined a test statistic that would serve to capture how well the tree partitions cells or sample belong to different stages of tumor progression. We would expect that cells belonging to the same stage from the same tumor would be clustered closer together than cells from different tumors or stages. We computed a test statistic that would serve as a metric of separation as the ratio of the average distance between cells in the same class and the average

distance between cells in different classes.

We then sought to reject the null hypothesis that cells are randomly distributed in the phylogeny. We performed 10000 permutation tests to derive the distribution of the test statistic for the null hypothesis.

7.3.4 k-mer counts as features for learning tasks

To test the predictive power of the kmer counts to classify the cells as primary or metastatic, we performed classification tests on two types of data : the count distribution matrix and the binary character matrix. We applied an the MATLAB SVM function *svmtrain* and *svmclassify* with LOOCV in both cases. There are 152 primary cells across both tumors and 48 metastatic cells.

7.4 Results

7.4.1 Data substructure

The data when viewed in the principal components space can be seen in Fig 7.2. The data substructure shows that kmer counts are able to distinguish among the stages of tumor progression. Primary cells from tumors T10 and T16 mostly cluster together (green and blue) and metastatic cells cluster separately (in red).

7.4.2 Distance-based phylogenies

Figures 7.3-7.6 describe consensus distance based phylogenies obtained from both T10 and T16 from 5-, 10-, 15- and 20-mers. Figures 7.7 -7.10 describe consensus distance based phylogenies obtained from primary and metastatic cells in T16. The trees separate subsets of metastatic and primary tumors into different pathways in the tree. There is however a mixing of some sets of metastatic and primary cells and these maybe cells in transition from one state to the other. We may distinguish between 3 classes: primary cells of T10, primary cells of T16, metastatic cells in T16.The average distance between cells in the same class was computed as the sum of all pairwise distances in that class normalized by the total number of such pairs. This metric of clustering is tabulated in Table. The average distance between cells is the maximum in T10. As can be seen in the trees, there are several distinct clusters of T10 cells with some intermixing with the metastatic cells.

k-mer	T10	Primary cells in T16	Secondary cells in T16
5-mer	14.79	15.89	13.03
10-mer	19.41	15.45	12.27
15-mer	16.20	12.76	16.43
20-mer	18.86	15.32	10.9229

Table 7.1: Average Distance among cells in bootstrap consensus trees from both T10 and T16

k-mer	Test Statistic	Distribution mean, sv	p-value
5-mer	0.6484	1, 0.0048	≤ 0.0001
10-mer	0.7333	1, 0.0058	≤ 0.0001
15-mer	0.6196	0.99, 0.0058	≤ 0.0001
20-mer	0.8266	1, 0.0049	≤ 0.0001

Table 7.2: Average Distance among cells in bootstrap consensus trees from both T10 and T16

7.4.3 Comparison of trees and reliability of results

To compare the trees, both the symmetric and branch score distance were used which measure similarity among trees obtained from various kmers.

7.4.4 Phylogenetic Analysis

The test statistic obtained for 5, 10, 15 and 20 -mer trees are described in table. In all cases, the values had a p-value of lesser than 0.0001 when compared to the mean value obtained from the permutation tests.

7.4.5 Classification Tasks

For 10-mer count fractions, applying an SVM with LOOCV on the count distribution matrix gave a prediction error of 1.5% when cells from both T10 and T16 were used and a prediction error of 1% when only cells from T16 were used.

k-mer	All Data	Primary, Metastatic Data
10-mer	0.985	0.99
15-mer	0.99	0.99

Table 7.3: Prediction Accuracy of classification when using k-mer count fractions as features when k = 10 and 15

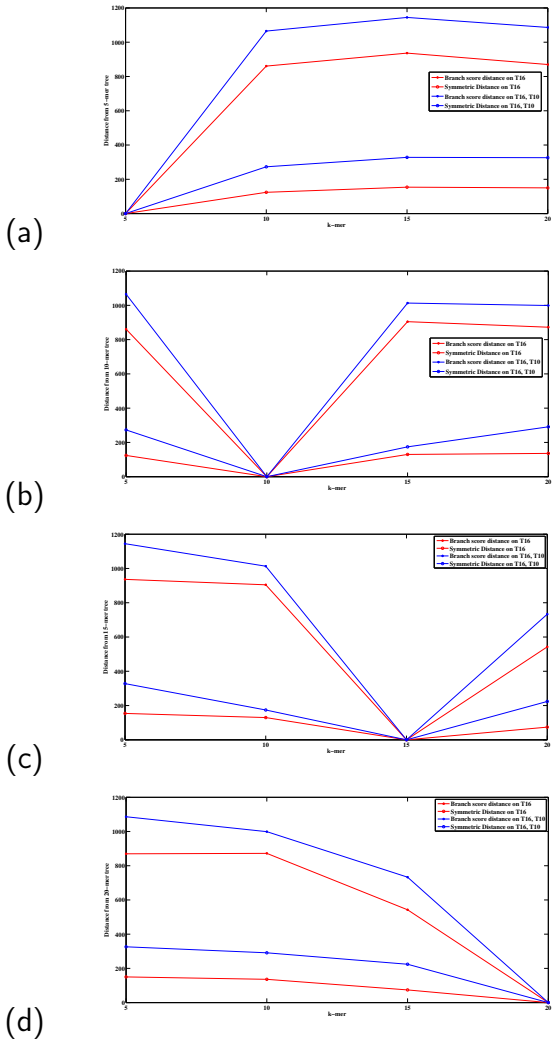
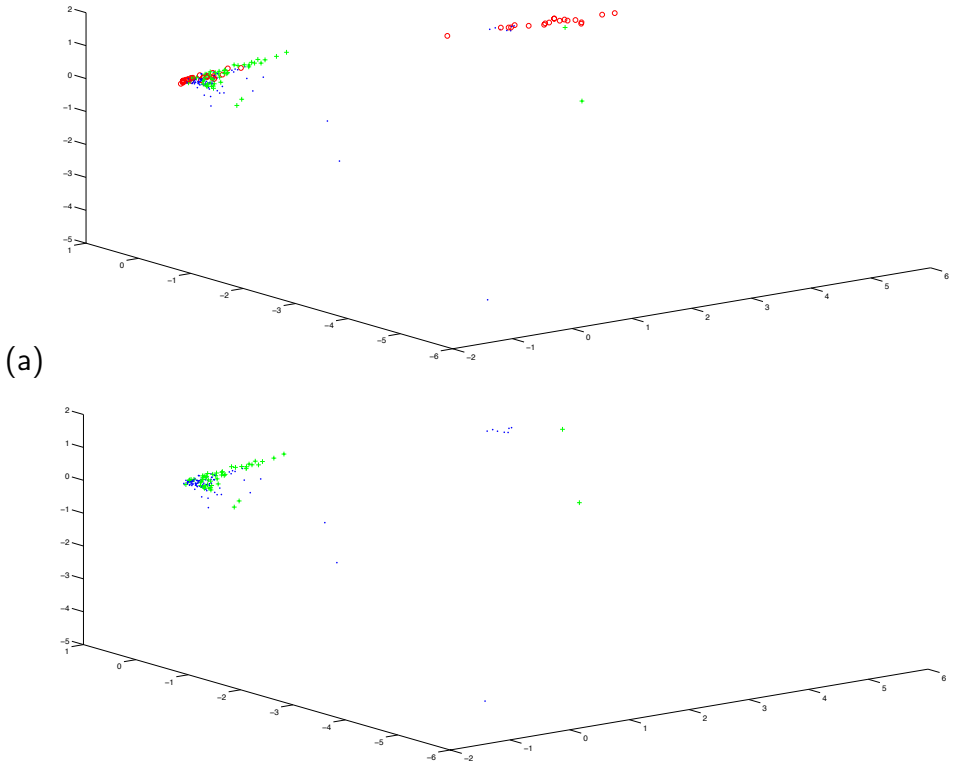


Figure 7.1: Branch score distance and symmetric distances among trees built on the dataset comprising T10 and T16 and only T16



(a) (b)

Figure 7.2: Tumor Single cells in PC space. Red is metastatic T16, Green is Primary T16, Blue is primary T10

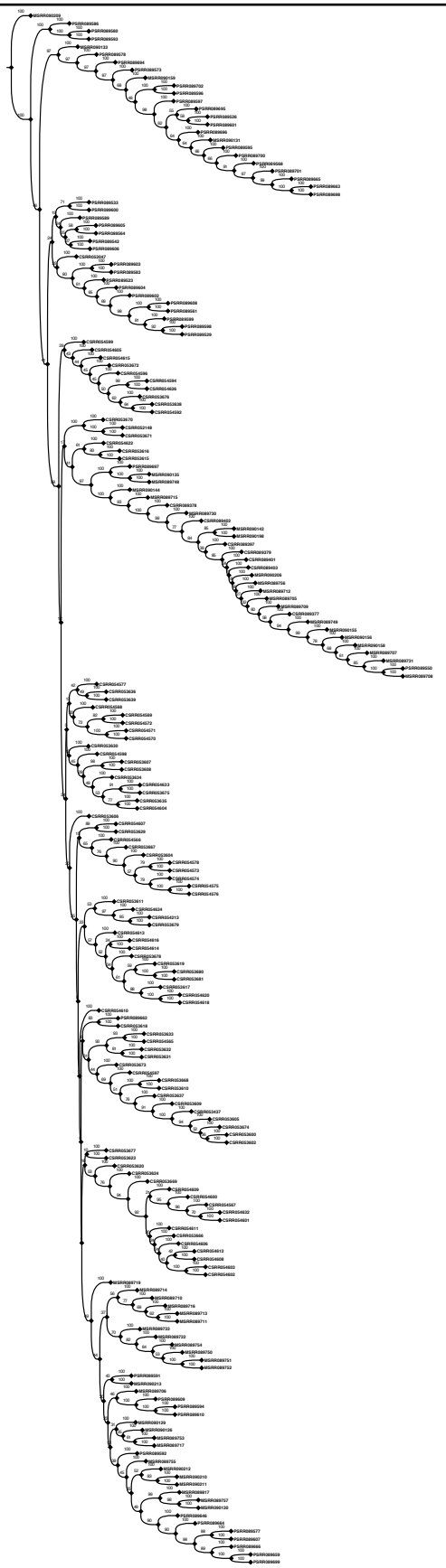


Figure 7.3: 5-mer bootstrap consensus Neighbor-joining tree built from T10 primary breast tumor cells (prefix C), T16 primary (prefix P) and metastatic data (prefix M)

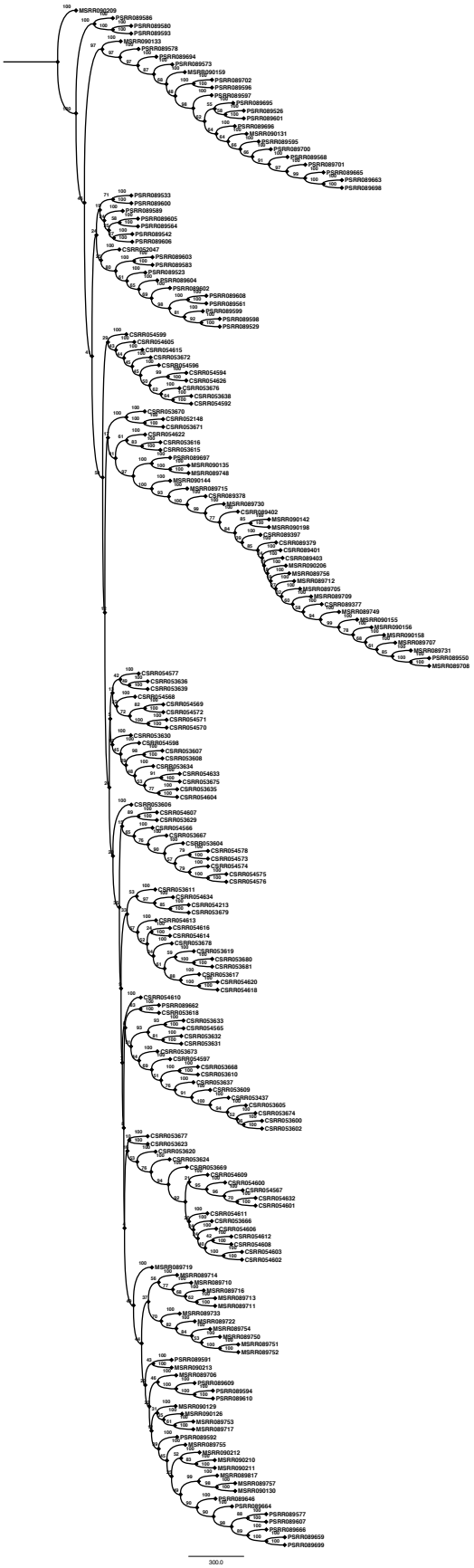


Figure 7.4: 10-mer bootstrap consensus Neighbor-joining tree built from T10 primary breast tumor cells (prefix C), T16 primary (prefix P) and metastatic data (prefix M)

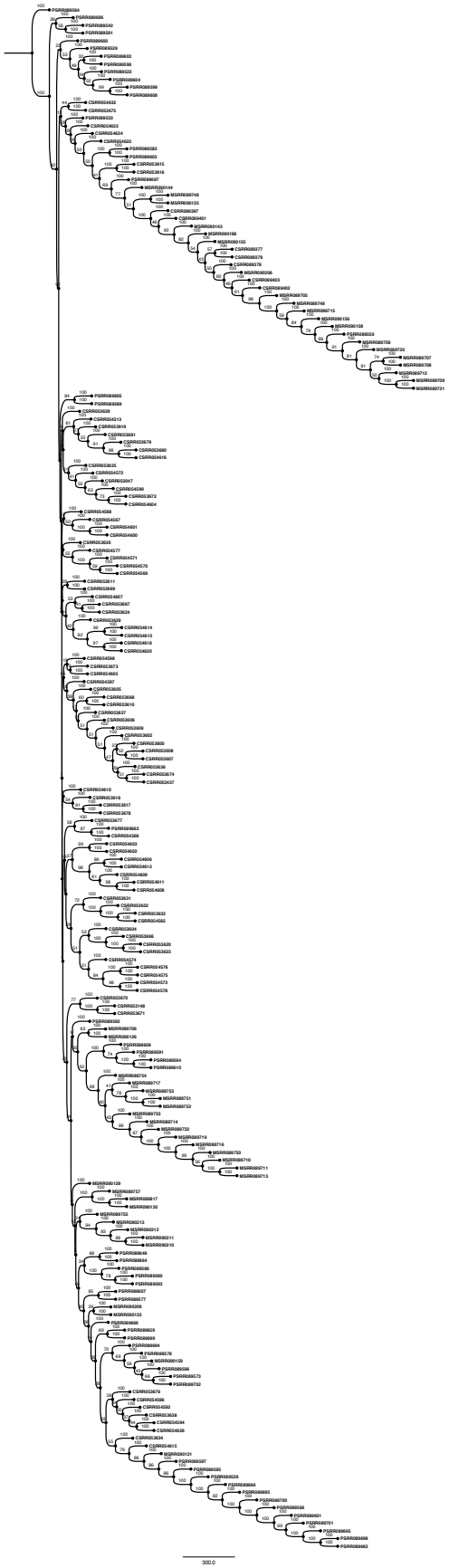


Figure 7.5: 15-mer bootstrap consensus Neighbor-joining tree built from T10 primary breast tumor cells (prefix C), T16 primary (prefix P) and metastatic data (prefix M)

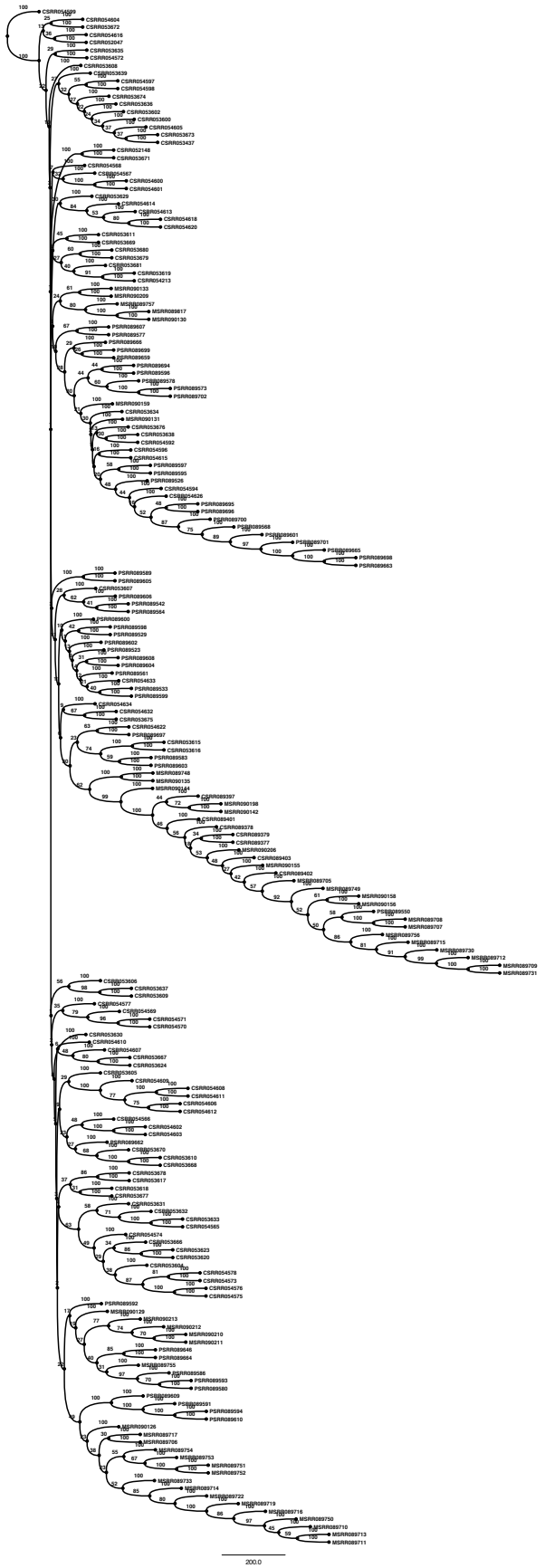


Figure 7.6: 20-mer bootstrap consensus Neighbor-joining tree built from T10 primary breast tumor cells (prefix C), T16 primary (prefix P) and metastatic data (prefix M)

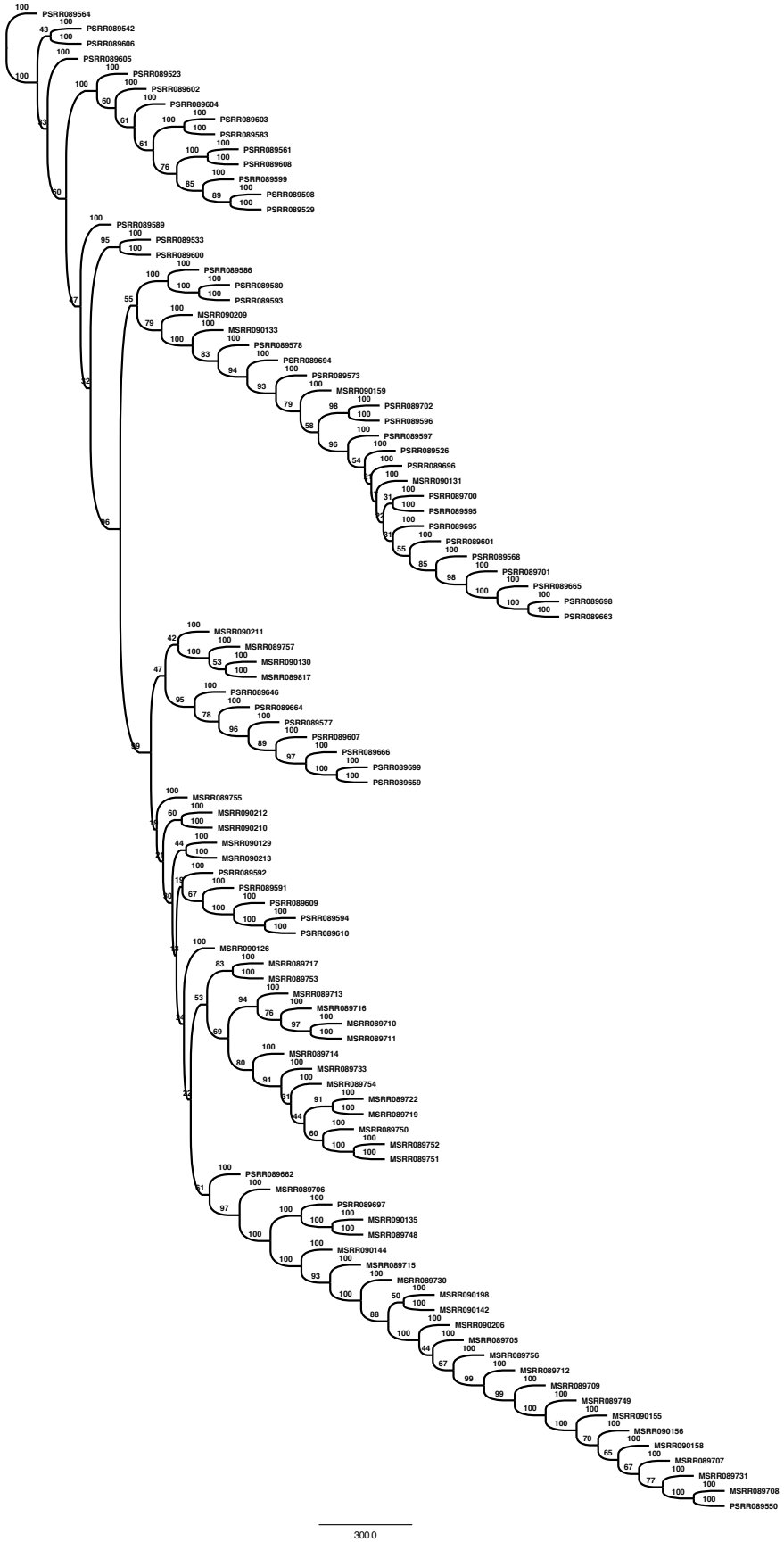


Figure 7.7: 5-mer bootstrap consensus Neighbor-joining tree built from T16 primary (prefix P) and metastatic data (prefix M)

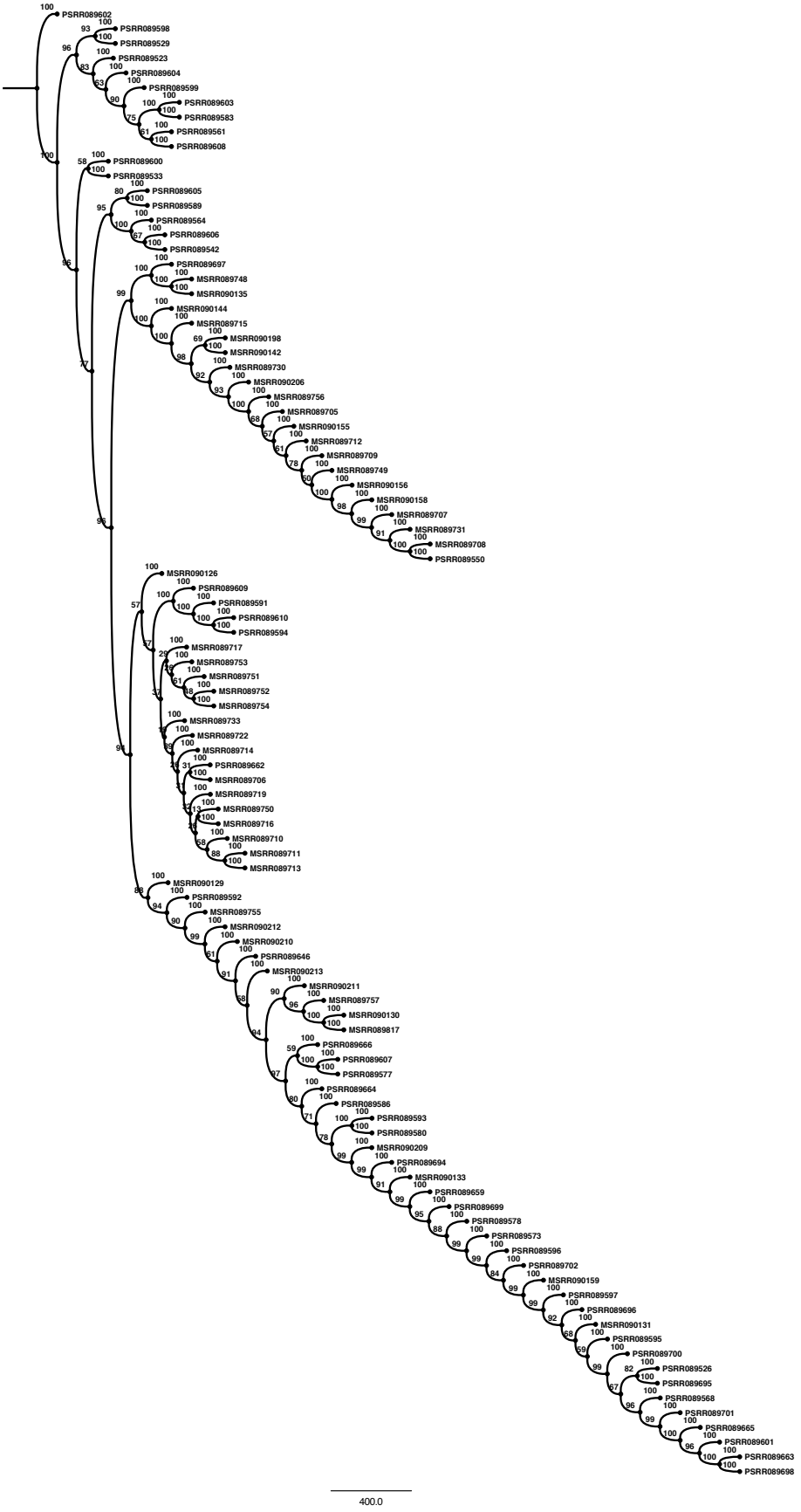


Figure 7.8: 10-mer bootstrap consensus Neighbor-joining tree built from T16 primary (prefix P) and metastatic data (prefix M)

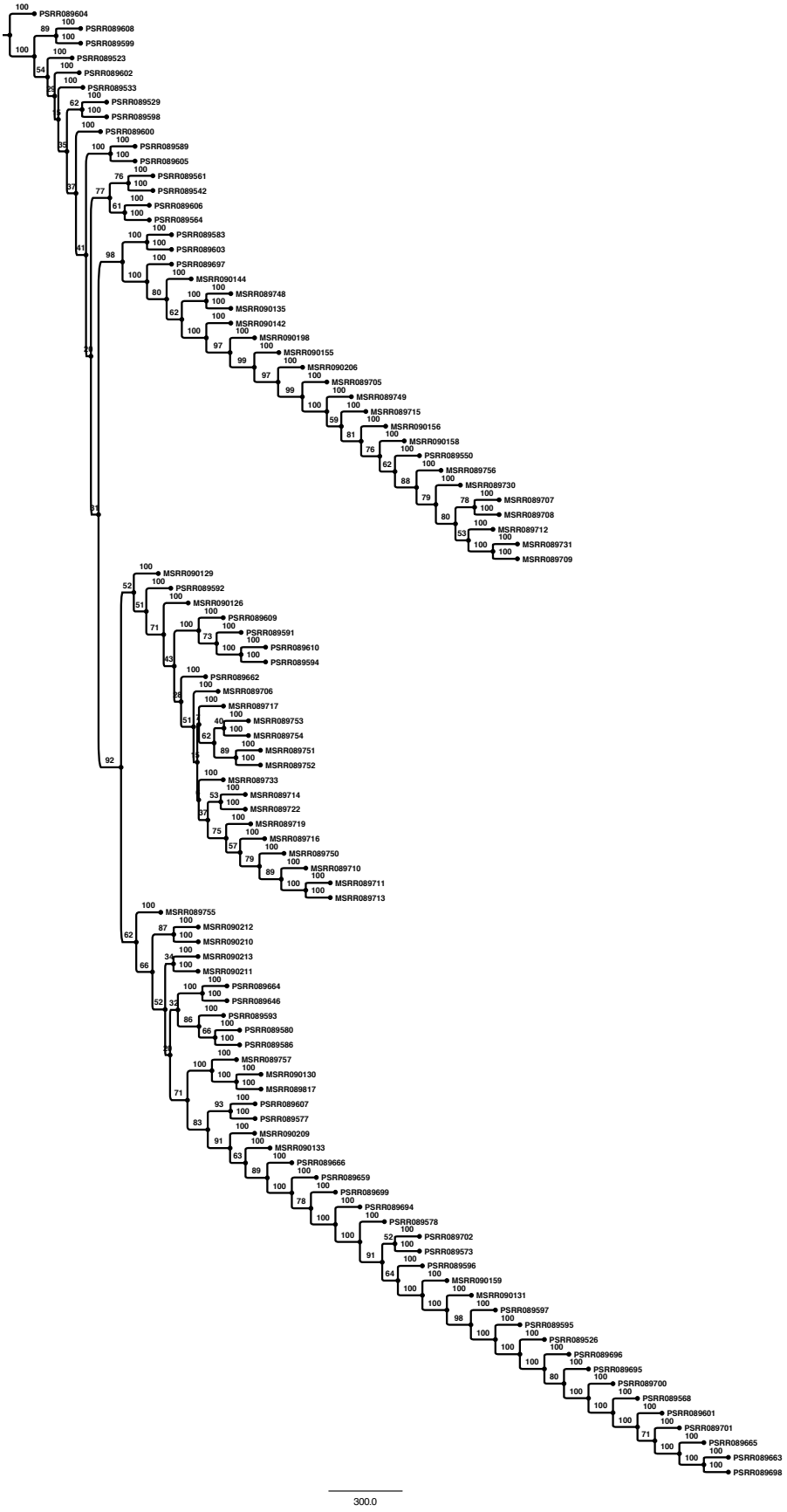


Figure 7.9: 15-mer bootstrap consensus Neighbor-joining tree built from T16 primary (prefix P) and metastatic data (prefix M)

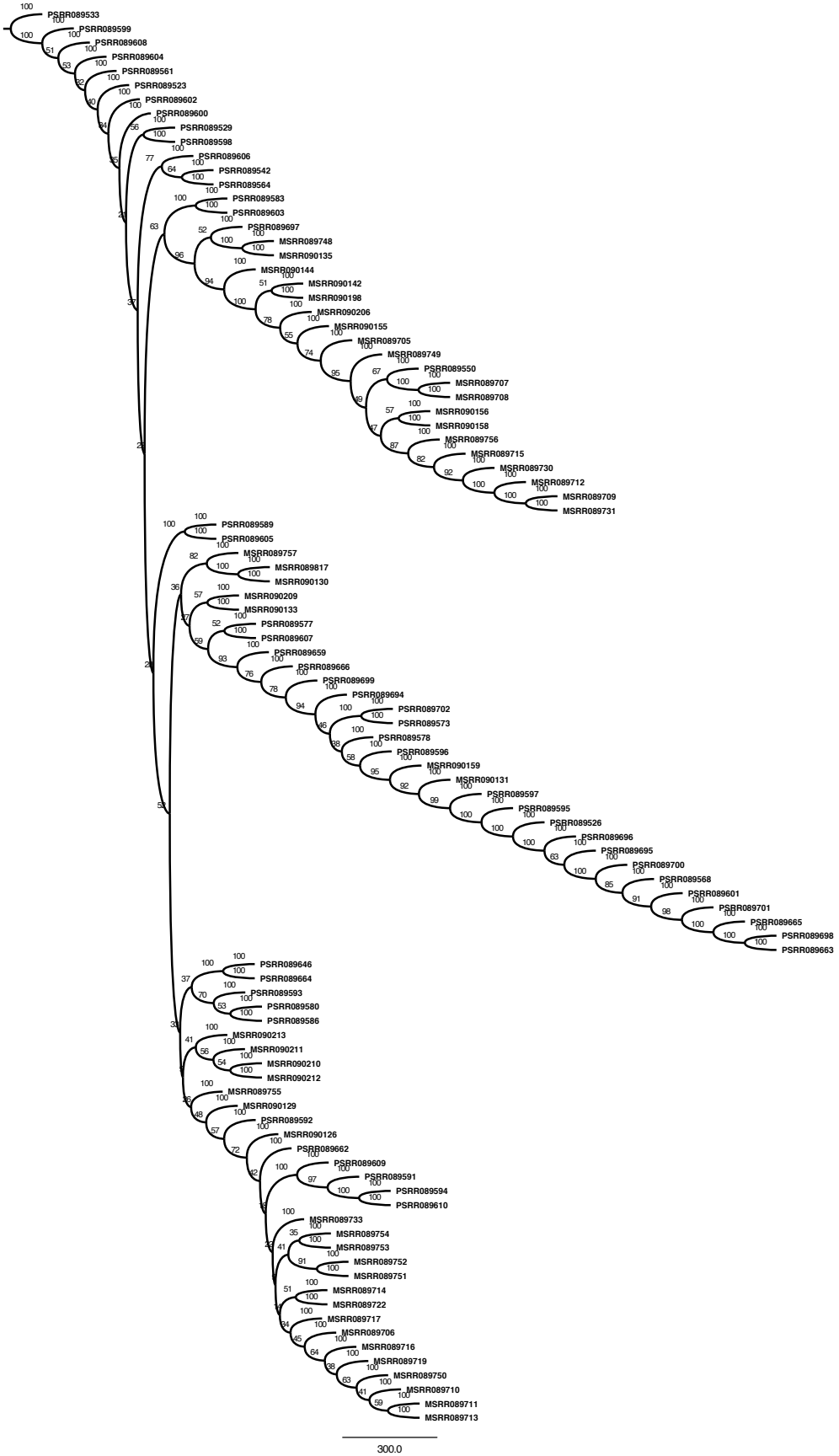


Figure 7.10: 20-mer bootstrap consensus Neighbor-joining tree built from T16 primary (prefix P) and metastatic data (prefix M)

7.5 Conclusion

We have illustrated a strategy to derive distance-based tumor phylogenies from whole genome sequencing data. The approach using kmer counts to represent genomic imbalances is much more computationally efficient that aligning sequence reads to the reference genome and then following downstream data processing steps in time, space and also algorithmic complexity. It also gets around the challenge during refrerence-genome alignment of massive genome rearrangement typical of tumor genomes. We have applied a series of noise correction measures to the resulting kmer count matrices. Finally, we demonstrate a method to analyze the resulting phylogenies as a measure of how well partition different stages of tumor progression. While the contributions are mainly methodological, future directions include applying the strategy to larger real datasets towards inferring biologically significant observations from the resulting phylogenies.

Chapter 8
Conclusions and Future Directions

In this chapter, we summarize the main contributions of this dissertation. We also discuss challenges and opportunities facing the field.

In this dissertation, we set out to answer the scientific question:

What is the sequence of events underlying the onset and progression of tumorigenesis?

We looked at various genomics data types including FISH, array copy number and whole genome sequencing in order to present a phyloegentic pipeline for reconstructing evolutionary trees capable of tracing events in tumorigenesis. The dissertation laid a specific focus on developing methods for the analysis of genomic aberrations including copy number and structural rearrangements, which can be applied to other data types like gene expression data.

I have established a phylogenetic framework for inferring tumor evolution that delineates the sequence of genome changes underlying tumor progression. Specifically, I have accomplished the following work.

1. Designed and implemented a novel phylogenetic pipeline for building tumor evolutionary trees or phylogenies from whole genome copy number variation data. This pipeline provides a step-by-step procedure for turning whole genome tumor profiles into phylogenies with biologically meaningful information. The steps included an optional computational unmixing to infer heterogeneity, a statistical method for progression-marker discovery, a statistical method for data discretization, application of character-based phylogeny reconstruction, and analyses of the resulting trees to draw biological significance.

2. Implemented an improved method HMM-CNA for discovering progression markers from cohorts of patient tumor copy number data that are especially relevant for phylogeny reconstruction via a custom multi-sample Hidden Markov Model (HMM). HMM-CNA improves upon the state-of-the-art with respect to speed of computation, accurate noise inference and the ability to analyze multiple samples at once.

3. Reconstructed evolutionary trees from single-cell sequencing data of tumor cells by designing novel features that can accurately capture the composition of the individual genome sequences. These features can be used to both build distance-based phylogenies as well as accurately classify primary and metastatic tumor data.

8.1 Challenges and Future Directions

As this dissertation has argued, oncogenesis is at its heart an evolutionary process and accurate characterization of common tumor progression pathways through phylogenetics provides a powerful method for characterizing this process and its many variations. Tumor phylogenetics is a young field but has advanced dramatically since its development due to improvements in models and algorithms for phylogenetic reconstruction of tumor data and advances in technologies for gathering the data itself. Perhaps no development has been more promising in this direction than the availability of fast, inexpensive sequencing through NGS. Phylogenetics, in turn, provides a key tool for making sense of the incredible complexity NGS tumor data is beginning to reveal. Nonetheless, we remain far from being able to truly reconstruct how even a single tumor has progressed much less, building a collective profile of the major pathways of progression across tumor populations and translating that knowledge into improved patient outcomes. In this final section, we survey some of the major challenges remaining at present and prospects for overcoming them.

8.1.1 Tumor Heterogeneity and Single-Cell Sequencing

The progress and challenges in tumor phylogenetics let us identify some features we would ideally want in a technology for profiling tumor state. Given the high heterogeneity between tumors, it should be able to produce a complete profile of potential mutation events without bias. It should further be able to accommodate heterogeneous forms of data, including copy number variations, single nucleotide changes, structural rearrangements, and epigenetic modification, as well as ancillary reporters, such as changes in gene expression. Given heterogeneity within tumors, it should be able to distinguish these variations at the level of single cells. Furthermore, it should have high enough throughput and low enough cost to be able to assay large numbers of cells in large numbers of patients. As we have seen, NGS is approaching this ideal data source, particularly with the advent of single-cell NGS, but nonetheless faces substantial technical challenges. At present, high genome coverage is achievable, leading to relatively thorough and accurate typing of point mutations [95, 96]. In principle, similar approaches could provide acceptable data with current technologies for epigenetic modifications. Accurate quantization remains a problem, however, a challenge for studying copy number

variations in tumors. Furthermore, reconstruction of rearrangements from tumor data is still in its early stages, and it remains unclear what magnitude or quality of data it will require at the single-cell level. Scales of studies still remain orders of magnitude smaller than would be ideal, with major studies either typically profiling hundreds of patients at a tumor-wide level [80] or profiling on the order of a hundred cells in individual patients [94, 95, 96] . Detailed models of major progression pathways and accurate statistics for identifying common features of major progression pathways are likely to require at least the order of hundreds of cells each for hundreds of patients, something so far only achieved with pre-genomic FISH technologies [17, 62]. Furthermore, the amount of data needed can be expected to increase in proportion to the rarity of the subtype or mutation being studied. Meeting this challenge will require the development of next-generation sequencing technologies that can handle small amounts of starting material to yield high coverage and large multiplexing capabilities to meet time, demands, and cost. Thus, we can conclude that faster and cheaper single-cell sequencing and more accurate quantization are likely to pay large dividends for tumor phylogeny approaches.

8.1.2 Computational Challenge of Tumor Phylogenetics

Another area of challenge for tumor phylogenetics comes in the computational processing of NGS data. Accurately assembling genomes remains a challenging computational problem in general, and all the more so for tumor genomes, which are likely to have undergone large rearrangements relative to available reference genomes [80, 100] . Further error comes from the assumptions in the noise models accompanying variant calling methods. While large numbers of reads can overcome inaccuracy due to random sequencing errors, more data will not overcome systematic biases or misassembly leading to incorrect inference of variations. Better models of errors in sequencing and assembly are an ongoing and active area of research [186]. Compounding this challenge is the fact that error models can differ greatly from technology to technology and will likely need to evolve along with new technologies for sequencing.

Given the limitations of the available technologies, especially for single-cell sequencing, computational models are likely to play a role for some time in controlling for biases in sampling, sequencing, and reconstruction of tumor state from NGS data. Mixture model methods developed for array technologies [72, 76] can be expected to have continued value in more accurately reconstructing profiles of tumor heterogeneity from NGS data. NGS presents substantial new challenges, though, for example, in scaling methods to handle large numbers of markers, dealing with challenging computational problems arising from mixtures of rearrangements [187] and developing methods to integrate heterogeneous forms of data (e.g., simultaneous point mutation, copy number, and epigenetic sequencing data). Advances are further needed in the phylogenetic inference step itself. Tumor phylogenetics depends on the identification of robust markers of progression

[77, 78], a problem that becomes more challenging the more data one examines. Figuring out how to separate drivers of oncogenesis from passenger mutations [125] is a challenging problem for the field of tumor biology as a whole but also central to characterizing pathways of tumor progression. Many promising approaches have been explored to attempt to reduce the scope of the problem. For copy number data, this can be approached as a kind of segmentation problem, in which one can improve statistical power by grouping nearby markers on the genome [78, 121]. For point mutation data, other groupings, such as by gene or pathway, can provide similar help in simplifying the statistical problem [94]. Nonetheless, the problem remains unsolved.

Models of evolution remain a further obstacle. Tumor phylogeny studies to date have generally modeled tumor evolution using standard generic phylogenetic algorithms, such as neighbor joining [40, 38] or maximum parsimony [77]. Some single-cell studies have used custom but still quite simplified models [50, 63].

Accurate tumor phylogenies, however, depend on having an accurate quantitative model of how tumors specifically evolve. Although tumors are indeed evolutionary systems, they are systems that behave quite differently in some ways than evolving systems of organisms, with very high mutation rates and strong selective pressures that distinguish them from typical species evolution problems. Furthermore, since damage to the cell replication machinery is a hallmark of tumor development [6], models of evolutionary events in tumor cells will be very different from those in individual organisms. We know a great deal about some of the specific mechanisms by which tumors evolve [94], and a variety of mathematical models have been developed by the field [14, 16, 188, 189, 190] that may be useful for more accurately judging the plausibility of various tumor phylogeny scenarios. Nonetheless, we do not have good models for how different tumors differ in their propensities for these various events, much less the detailed quantitative models of likelihoods of possible mutational events that would ideally be used for phylogenetic inference. Furthermore, this gap in quantitative models represents a difficult chicken-and-egg problem for the field, as we will only learn these models by studying accurate tumor phylogenies. Computational approaches can in principle allow one to solve such problems by iteratively cycling between better models from which we can learn phylogenies and better phylogenies from which we can learn models [63], but this process is likely to require a better qualitative understanding of the basic mechanisms of oncogenesis, better data, and algorithms capable of making use of such data. A final but substantial challenge is to algorithms for tumor phylogenetics. Phylogenetics is a challenging computational problem even in classic species scenarios and standard algorithms will not scale to the volumes of data NGS is making available, especially for the character-based approaches needed to reconstruct detailed events along evolutionary pathways [77]. The important role of genomic rearrangements in tumor development represents a major

challenge in itself [80, 100]. Furthermore, as improved models and new kinds of data are developed, new algorithms are likely to be needed to fit to those data. New approaches to phylogenetics will be needed to integrate heterogeneous data sources available through NGS technologies. Experimental validation of inferred phylogenies is likewise a difficult problem with no obvious solutions.

Bibliography

[1] D.P. Cahill, K.W. Kinzler, and B. Vogelstein et al. Genetic instability and darwinian selection in tumours. *Trends in Genetics*, 15(12):M57–M60, 1999. 1, 6

[2] A. Marusyk, V. Almendro, and K. Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12:323–334, May 2012. 1

[3] M. Cusnir and L. Cavalcante. Inter-tumor heterogeneity. *Hum Vaccin Immunother*, 8(8):1143–1145, 2012. 2

[4] A. Subramanian, S. Shackney, and R. Schwartz. Tumor phylogenetics in the next generation sequencing era: Strategies, challenges and future prospects. *Next Generation Sequencing in Cancer Research.*, 2013. 5

[5] N. Howlader, A.M. Noone, and M. Krapcho et al. Seer cancer statistics review, 1975-2010, national cancer institute. bethesda, md. *SEER web site*, April 2013. 5

[6] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 127(5):905–915, 2006. 5, 6, 114

[7] P.C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976. 6

[8] C. Nordling. A new theory on cancer-inducing mechanism. *Br J Cancer*, 7(1):68–72, 1953. 6

[9] A. Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA*, 68(4):820–823, 1971. 6

[10] E. R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61:759–767, 1991. 6

[11] B. S. Strauss. Hypermutability in carcinogenesis. *Genetics*, 148(4):1619–1626, 1998. 6

[12] G. Danaei, S. Vander-Hoorn, and A.D. Lopez et al. Comparative risk assessment collaborating group (cancers). causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet*, 366(9499):1784–1793, 2005. 6

[13] A. R. Anderson, A. M. Weaver, and P. T. Cummings et al. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. *Cell*, 127(5):905–915, 2006. 6

[14] Y. Iwasa and F. Michor. Evolutionary dynamics of intratumor heterogeneity. *PLoS One.*, 6(3), 2011. 6, 114

[15] F.C. Martins, S. De, and V. Almendro et al. Evolutionary pathways in brca1-associated breast tumors. *Cancer Discov.*, 2(6):503â511, 2012. 6

[16] Y.K. Cheng, R. Beroukhim, and R.L. Levine et al. A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS Comput Biol.*, 8(1), 2012. 6, 114

[17] L.E. Janocko, K.A. Brown, and C.A. Smith et al. Distinctive patterns of Her-2/neu, c-myc and cyclin D1 gene amplification by fluorescence in situ hybridization in primary human breast cancers. *Cytometry*, 46:136–149, 2001. 6, 16, 44, 62, 64, 113

[18] M. Gerlinger, A.J. Rowan, and S. Horswell et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.*, 366(10):883–892, 2012. 6, 18

[19] L.L. Campbell and K. Polyak. 7

[20] C.M. Perou, T. Sorlie, and M.B. Eisen et al. Molecular portraits of human breast tumors. *Nature*, 406:747–752, 2000. 7, 20, 43

[21] C.M. Perou. Molecular stratification of triple-negative breast cancers. *Oncologist.*, 16:61â70, 2011. 7

[22] T. Sorlie, C.M. Perou, and R. Tibshirani et al. Gene expression profiles of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*, 98:10869–10864, 2001. 7, 20, 43

[23] T.R. Golub, D.K. Slonim, and P. Tamayo et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999. 7, 20, 43

[24] C. Sotiriou, S.Y. Neo, and L.M. McShane et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA*, 100:10393–10398, 2003. 7, 20, 43

[25] M.D. Pegram, G. Konecny, and D.J. Slamon. The molecular and cellular biology of HER2/neu gene amplification/overexpression and the clinical development of herceptin (trastuzumab) therapy for breast cancer. *Cancer Treat Res*, 103:57–75, 2000. 7, 20, 44

[26] W.N Hait and T.W. Hambley. Targeted cancer therapeutics. *Cancer Res.*, 69(4):1263â1267, 2009. 7

[27] K.K. Kidd and L.A. Sgaramella-Zonta. Phylogenetic analysis: concepts and methods. *Am J Hum Genet.*, 23:235â252, 1971. 9

[28] L.L. Cavalli-Sforza and A.W.F Edwards. Phylogenetic analysis. models and estimation procedures. *Am J Hum Genet.*, 19:233â257, 1967. 9

[29] A.W.F Edwards and L.L. Cavalli-Sforza. Reconstruction of evolutionary trees. *Phenetic and phylogenetic classification.*, 6:67â76, 1964. 9

[30] J.P. Huelsenbeck, F. Ronquist, and R. Nielsen et al. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science.*, 294(5550):2310â2314, 2001. 9

[31] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.*, 4(4):406–425, 1987. 9

[32] J. Felsenstein. Inferring phylogenies. *Sinauer Associates*, 2004. 9

[33] J. Felsenstein. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989. 9, 77

[34] D. Swafford. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4., 2002. 9, 50, 77, 89

[35] A. Stamatakis, M. Ott, and T. Ludwig et al. Raxml-omp: an efficient program for phylogenetic inference on smps. *Proceedings of 8th international conference on Parallel Computing Technologies (PaCT2005)*, 3506:288â302, 2005. 9

[36] J.P. Huelsenbeck and F. Ronquist. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics.*, 17(8):754â755, 2001. 9

[37] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A.A.Schaffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comp Biol*, 6:37–51, 1999. 10, 21, 44

[38] N. Navin, A. Krasnitz, and L. Rodgers et al. Inferring tumor progression from genomic heterogeneity. *Genome Research*, 20:68–80, March 2010. viii, ix, x, xi, 1, 10, 16, 22, 28, 32, 33, 34, 44, 45, 51, 52, 54, 55, 56, 57, 63, 68, 77, 80, 82, 114

[39] F. Jiang, R. Desper, and C.H. Papadimitriou et al. Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data. *Cancer Res.*, 60(22):6503â6509, 2000. 10, 15

[40] R. Desper, F. Jiang, O.P. Kallioniemi, H.Moch, C.H. Papadimitriou, and A.A. Schaffer. Distance-based reconstruction of tree models for oncogenesis. *J Comp Biol*, 7:789–803, 2000. 10, 15, 114

[41] Z. Huang, R. Desper, and A.A. Schaffer et al. Construction of tree models for pathogenesis of nasopharyngeal carcinoma. *Genes Chromosomes Cancer.*, 40(4):307–315, 2004. 10

[42] L. Chen, C. Norlander, and A. Behboudi et al. Deriving evolutionary tree models of the oncogenesis of endometrial adenocarcinoma. *Int J Cancer.*, 120(2):292–296, 2007. 10

[43] C. Sweeney, K.M. Boucher, and W.S. Samowitz et al. Oncogenetic tree model of somatic mutations and dna methylation in colon tumors. *Genes Chromosomes Cancer.*, 48(1):1–9, 2009. 10

[44] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans Inform Theor.*, 14(3):462–467, 1968. 15

[45] A. Szabo and K. Boucher. Estimating an oncogenetic tree when false negatives and positives are present. *Math Biosci.*, 176(2):219–236, 2002. 15

[46] K. Yamamoto, S. Oba, and S. Ishii. Considering false negatives in mixture oncogenetic trees model for chromosomal alteration data. *Genome Informatics Workshop (GIW2006).*, pages 219–236, 2006. 15

[47] A. von Heydebreck, B. Gunawan, and L. Füzesi. Maximum likelihood estimation of oncogenetic tree models. *Biostatistics*, 5:545–556, 2004. 15

[48] M.D. Radmacher, R. Simon, and R. Desper et al. Graph models of oncogenesis with an application to melanoma. *J Theor Biol.*, 212(4):535â548, 2001. 15

[49] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol.*, 9(5):687â705, 2002. 15

[50] R. Desper, J. Khan, and A.A. Schaffer. Tumor classification using phylogenetic methods on expression data. *J Theor Biol*, 228:477–496, 2004. 15, 114

[51] B. Weigelt, A.M. Glas, and L.F. Wessels et al. Gene expression profiles of primary breast tumors maintained in distant metastases. *Proc Natl Acad Sci USA*, 100(26):15901â15905, 2003. 15

[52] J. Quackenbush. Microarray analysis and tumor classification. *N Engl J Med.*, 354(23):2463â2472, 2006. 15

[53] A. Perez-Diez, A. Morgun, and N. Shulzhenko. Microarrays for cancer diagnosis and classification. *Adv Exp Med Biol.*, 593:74–85, 2007. 15

[54] A. Dutt and R. Beroukhim. Single nucleotide polymorphism array analysis of cancer. *Curr Opin Oncol.*, 19(1):43â49, 2007. 15

[55] S. Zheng and Z. Zhao. Dna methylation profiling distinguishes three clusters of breast cancer cell lines. *Chem Biodivers.*, 9(5):848â856, 2012. 15

[56] R.C. OâHagan, C.W. Brennan, and A. Strahs et al. Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma. *Cancer Res.*, 63(17):5352â5356, 2003. 15

[57] S. Shackney Y. Park and R. Schwartz. Network-based inference of cancer progression from microarray data. In *International Symposium on Bioinformatics Research and Applications (ISBRA08)*, pages 268–279, 2008. 15

[58] M. Riester, C. Stephan-Otto Attolini, and R.J. Downey. A differentiation-based phylogeny of cancer subtypes. *PLoS Comput Biol.*, 6(5). 15

[59] M. Abu-Asab, M. Chaouchi, and H. Amri. Evolutionary medicine: a meaningful connection between omics, disease, and treatment. *Proteomics Clin Appl.*, 2(2):122–134, 2008. 15

[60] M. Abu-Asab, M. Chaouchi, and H. Amri. Phylogenetic modeling of heterogeneous gene-expression microarray data from cancerous specimens. *OMICS.*, 12(3):183–199, 2008. 15

[61] M. Ryott, D. Wangsa, and K. Heselmeyer-Haddad et al. Egfr protein overexpression and gene copy number increases in oral tongue squamous cell carcinoma. *Eur J Cancer.*, 45(9):1700–1708, 2009. 16, 18, 64

[62] K. Heselmeyer-Haddad, L.Y.B. Garcia, and A. Bradley et al. Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity, yet conserved genomic imbalances and gain of myc during progression. *Am J Pathol.*, 181(5):1807â1822, 2012. 16, 38, 113

[63] G. Pennington, C.A. Smith, S. Shackney, and R. Schwartz. Expectation-maximization method for the reconstruction of tumor phylogenies from single-cell data. In *Computational Systems Bioinformatics Conference (CSB)*, pages 371–380, 2006. 16, 44, 64, 114

[64] G. Pennington, C.E. Smith, S. Shackney, and R. Schwartz. Reconstructing tumor phylogenies from single-cell data. *Journal of Bioinformatics and Computational Biology*, 5:407–427, 2007. 16, 21

[65] O.R.P. Bininda-Emonds, J.L. Gittleman, and M.A. Steel. The (super) tree of life: procedures, problems, and prospects. *Annu Rev Ecol Syst.*, 33:265–289. 16

[66] A.D. Gordon. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of leaves. *J Classif.*, 3:31â39, 1986. 16

[67] M. Nugoli, P. Chuchana, and J. Vendrell et al. Genetic variability in mcf-7 sublines: evidence of rapid genomic and rna expression profile modifications. *BMC Cancer.*, pages 3–13. 16, 86, 93

[68] N. Beerenwinkel, J. Rahnenfürer, and M. Dämer et al. Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol.*, 12(6):584â598, 2005. 17, 18

[69] Niko Beerenwinkel, Jörg Rahnenführer, Rolf Kaiser, Daniel Hoffmann, Joachim Selbig, and Thomas Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, 2005. 17, 21

[70] J. Yin, N. Beerenwinkel, and J. RahnenfuÌhrer et al. Model selection for mixtures of mutagenetic trees. *Stat Appl Genet Mol Biol.*, 2006. 17

[71] R. Ketter, S. Urbschat, and W. Henn et al. Application of oncogenetic trees mixtures as a biostatistical model of the clonal cytogenetic evolution of meningiomas. *Int J Cancer.*, 12:323–334, May 2007. 17

[72] R. Schwartz and S. Shackney. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics*, 11:42, 2010. 17, 21, 22, 30, 44, 45, 47, 68, 113

[73] D. Tolliver, C. Tsourakakis, and A. Subramanian et al. Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics*, 26(12):i106–i114, 2010. 17, 20, 45, 46, 47, 51, 52, 68, 71, 77

[74] R. Etzioni, S. Hawley, and D. Billheimer et al. Analyzing patterns of staining in immunohistochemical studies: application to a study of prostate cancer recurrence. *Cancer Epidem Biomark*, 14:1040–1046, 2005. 17, 22, 45

[75] G. Quon and Q. Morris. A mixture model for the evolution of gene expression in non-homogenous datasets. *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, 25(21):2882â2889, 2008. 17

[76] G. Quon and Q. Morris. Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics.*, 2009. 17, 113

[77] A. Subramanian, S. Shackney, and R. Schwartz. Inference of tumor phylogenies from genomic assays on heterogeneous samples. *J Biomed Biotechnol.*, 2012. 17, 43, 113, 114

[78] A. Subramanian, S. Shackney, and R. Schwartz. Novel multi-sample scheme for inferring phylogenetic markers from whole genome tumor profiles. *IEEE/ACM Trans Comput Biol Bioinform.*, 2013. 17, 66, 113, 114

[79] Cancer genome atlas research network. comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.*, 455(7216):1061â1068, 2008. 17

[80] P.J. Campbell, P.J. Stephens, and E.D. Pleasance. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet.*, 40(6):722–729, 2008. 17, 113, 115

[81] T.M Kim, L.J. Luquette, and R. Xi et al. rsw-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics.*, 11(432), 2010. 18

[82] C Xie and M.T. Tammi. rsw-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics.*, 2009. 18

[83] P. Medvedev, M. Stanciu, and M. Brudno et al. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods.*, 2009. 18

[84] Z.D. Zhang, J. Du, and H. Lam et al. Identification of genomic indels and structural variations using split reads. *BMC Genomics.*, 12(375), 2011. 18

[85] J.A. Neuman, O. Isakov, and N. Shomron. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform.*, 14(1):46â55, 2012. 18

[86] A. Elsharawy, M. Forster, and N. Schracke et al. Improving mapping and snp-calling performance in multiplexed targeted next-generation sequencing. *BMC Genomics.*, 13(1):2078â2079, 2012. 18

[87] M.D. Adams, M.L. Veigl, and Z. Wang et al. Global mutational profiling of formalin-fixed human colon cancers from a pathology archive. *Mod Pathol.*, 25(12):1599â1608, 2012. 18

[88] H. Li, B. Handsaker, and A. Wysoker et al. The sequence alignment/ map format and samtools. *Bioinformatics.*, 25(16):2078â2079, 2009. 18

[89] G. Marcais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.*, 27(6):764â770, 2011. 18, 97

[90] P. Melsted and J.K. Pritchard. Efficient counting of k-mers in dna sequences using a bloom filter. *BMC Bioinformatics.*, 12(333), 2011. 18

[91] C.D. Greenman, E.D. Pleasance, and S. Newman et al. Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.*, 22(2):346â361, 2012. 18

[92] L.M. Merlo, J.W. Pepper, and B.J. Reid et al. Cancer as an evolutionary and ecological process. *Nat Rev Cancer.*, 6(12):924–935, 2006. 18

[93] T. Ried, M. Liyanage, and S. du Manoir et al. Tumor cytogenetics revisited: comparative genomic hybridization and spectral karyotyping. *J Mol Med.*, 75(11-12):801–814, 1997. 18

[94] N. Navin, J. Kendall, and J. Troge et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472, 2011. 18, 19, 58, 64, 95, 113, 114

[95] Y. Hou, L.Song, and P. Zhu et al. Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell.*, 148(5):873â885, 2012. 19, 112, 113

[96] X. Xu, Y. Hou, and X. Yin et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell.*, 148(5):886â895, 2012. 19, 112, 113

[97] F. Tang, C. Barbacioru, and Y. Wang et al. mrna-seq whole-transcriptome analysis of a single cell. *Nat Methods.*, 6(5):377â382, 2009. 19

[98] S. Yilmaz and D.J. Singh. Single cell genome sequencing. *Curr Opin Biotechnol.*, 23(3):437–443, 2012. 19

[99] F.B. Dean, S. Hosono, and L. Fang et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA.*, 99(8):5261–5266, 2002. 19

[100] P.J. Stephens, D.J. McBride, and M.L. Lin et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature.*, 462(7276):1005–1010, 2009. 19, 113, 115

[101] T. Sorlie, R. Tibshirani, and J. Parker et al. Repeated observation of breast tumor subtypes in indepednent gene expression data sets. *Proc Natl Acad Sci USA*, 100:8418–8423, 2003. 20

[102] J.H. Atkins and L.J. Gershell. From the analyst's couch: Selective anticancer drugs. *Nat Rev Cancer*, 2:645–646, 2002. 20

[103] A.H. Bild, A. Potti, and J.R. Nevins. Opinion: Linking oncogenic pathways with therapeutic opportunities. *Nat Rev Cancer*, 6:735–741, 2006. 20

[104] A. Kamb, S. Wee, and C. Lengauer. Why is cancer drug discovery so difficult? *Nat Rev Drug Discov*, 6:115–120, 2007. 20, 44

[105] S. Paik, C. Kim, and N. Wolmark. HER2 status and benefit from adjuvant trastuzumab in breast cancer. *N Engl J Med*, 358:1409–1411, 2008. 20

[106] S.E. Shackney, C.A. Smith, and A. Pollice et al. Intracellular patterns of Her-2/neu, ras, and ploidy abnormalities in primary human breast cancers predict postoperative clinical disease-free survival. *Clin Cancer Res*, 10:3042–3052, 2004. 21

[107] M HÃ¶glund, D Gisselsson, N Mandah, B Johansson, F Mertens, F Mitelman, and T SÃll. Multivariate analyses of genomic imbalances in solid tumors reveal distinct and converging pathways of karyotypic evolution. *Genes, Chromosomes and Cancer*, 2001. 21

[108] Moritz Gerstung, Michael Baudis, Holger Moch, and Niko Beerenwinkel. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, 25(21):2809–2815, 2009. 21

[109] Jun Liu, Jaaved Mohammed, James Carter, Sanjay Ranka, Tamer Kahveci, and Michael Baudis. Distance-based clustering of CGH data. *Bioinformatics*, 22(16):1971–1978, 2006. 21

[110] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901. 21, 24

[111] P.Comon. Independent component analysis. *Signal Processing*, 36:287–314, 1994. 21

[112] B. Schölkopf, A.J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998. 21

[113] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002. 21

[114] P. Lamy, C.L. Anderson, and L. Dyrskjot et al. A hidden Markov model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. *BMC Bioinformatics*, 8:434, 2007. 22

[115] R. Ehrlich and W. Full. Sorting out geology — unmixing mixtures. In *Use and Abuse of Statistical Methods in the Earth Sciences*, pages 33–46. Oxford University Press, 1987. 22, 27, 47

[116] T.H. Chan, C.Y. Chi, Y.M. Huang, and W.K. Ma. A convex analysis based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Trans. Signal Processing*, 57(11):4418–4432, 2009. viii, 22, 26, 27, 31

[117] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788â791, 1999. 26, 47

[118] A. Packer. NP-hardness of largest contained and smallest containing simplices for V- and H-polytopes. *Discrete Computational Geometry*, 28:349–377, 2002. 27

[119] Y. Zhou and S. Suri. Algorithms for minimum volume enclosing simplex in R^3. *Proceedings of the Eleventh Annual ACM/SIAM Symposium on Discrete Algorithms*, pages 500–509, 2000. 27

[120] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. 27, 33

[121] A. B. Olshen, E. S. Venkatraman, and R. Lucito et al. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, October 2004. 33, 67, 76, 83, 114

[122] S. Guha, Y. Li, and D. Neuberg. Bayesian hidden Markov modeling of array CGH data. *Harvard University, Paper 24*, 2006. 33

[123] A. Subramanian, S. Shackney, and R. Schwartz. Inference of tumor phylogenies from genomic assays on heterogeneous samples. *BCB '11 Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB) 2011.*, pages 172–181. 43, 68, 78, 80, 82

[124] D. T. Ross, U. Scherf, and M. B. Eisen et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*. 43

[125] I. Bozic, T. Antal, and H. Ohtsuki et al. Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci.*, 107(43):18545–18550, 2010. 44, 114

[126] C.A. Smith, A.A. Pollice, and L.P. Gu et al. Correlations among p53, Her-2/neu and ras overexpression and aneuoploidy by multiparameter flow cytometry in human breast cancer: evidence fro a common phenotypic evolutionary pattern in infiltrating ductal carcinomas. *Clin Cancer Res*, 6:112–126, 2000. 44

[127] N. Beerenwinkel, M. Dämer, T. Sing, J. Rahnenfürer, T. Lengauer, J. Selbig, D. Hoffman, and R. Kaiser. Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J Infect Dis*, 191:1953–1960, 2005. 45

[128] A. W. F. Edwards and C. L. L. Sforza. The reconstruction of evolution. *Heredity*, 12, 1963. 50

[129] J. Ellson, E. Gansner, and L. Koutsofios et al. Graphvizâ open source graph drawing tools. *Graph Drawing. Lecture Notes in Computer Science*, 2265:594â597, 2002. 50

[130] W. James Kent, C. W. Sugnet, and T. S. Furey et al. The human genome browser at ucsc. *Genome Research*, 12(6):996â1006, 2002. 51

[131] MD) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore and MD). National Center for Biotechnology Information, National Library of Medicine (Bethesda. Online mendelian inheritance in man, omim (tm. 2010. 51

[132] J.R. Pollack and T. Sorlie and C.M. Perou et al. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963–12968, 2002. x, 1, 51, 56, 58, 59, 62, 63

[133] C. D. Bajdik, B. Kuo, and S. Rusaw et al. Cgmim: automated text-mining of onlinemendelian inheritance in man (omim) to identify genetically-associated cancers and candidate genes. *BMC Bioinformatics*, 6, 2005. 53

[134] J. S. Ross and J. A. Fletcher. The her-2/neu oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy. *Stem Cells*. 58, 60

[135] S. Saito, K. Morita, and T. Hirano. High frequency of common dna copy number abnormalities detected by bacterial artificial chromosome array comparative genomic hybridization in 24 breast cancer cell lines. *Human Cell*, 22(1):1–10, 2009. 58, 60

[136] I. Bieche and R. Lidereau. Genome-based and transcriptomebased molecular classification of breast cancer. *Current Opinion in Oncology*, 23(1):93â99, 2011. 58, 60, 61

[137] K. Mu, L. Li, and Q. Yang et al. Detection of chk1 and ccnd1 gene copy number changes in breast cancer with dual-colour fluorescence in-situ hybridization. *Histopathology*, 58(4):601â607, 2011. 60

[138] B. M. Zaharieva, R. Simon, and P. A. Diener et al. Highthroughput tissue microarray analysis of 11qi3 gene amplification (ccnd1, fgf3, fgf4, ems1) in urinary bladder cancer. *Journal of Pathology*, 201(4), 2003. 60

[139] E. Schuuring, E. Verhoeven, W. J. Mooi, and R. J. A. M.Michalides. Identification and cloning of two overexpressed genes, u21b31/prad1 and ems1, within the amplified chromosome 11q13 region in human carcinomas. *Oncogene*, 7(2):355â361, 1992. 60

[140] S. M. Aukema, R. Siebert, and E. Schuuring et al. Double-hit bcell lymphomas. *Blood*, 117(8):2319â2331, 2011. 60

[141] H. R. Park, S. K. Min, and H. D. Cho et al. Expression of osteoprotegerin and rank ligand in breast cancer bone metastasis. *Journal of Korean Medical Science*, 18(4):541â546, 2003. 60

[142] D. H. Jones, T. Nakashima, and O. H. Sanchez et al. Regulation of cancer cell migration and bone metastasis by rankl. *Nature*, 440(7084):692â696, 2006. 60

[143] A. M. Sieuwerts, M. P. Look, and M. E. Meijer-Van Gelder et al. Which cyclin e prevails as prognostic marker for breast cancer? results from a retrospective study involving 635 lymph node-negative breast cancer patients. *Clinical Cancer Research*, 12(11):3319â3328, 2006. 60

[144] C. Sotiriou, M. Paesmans, and A.Harris et al. Cyclin e1 (ccne1) and e2 (ccne2) as prognostic and predictive markers for endocrine therapy (et) in early breast cancer. *Journal of Clinical Oncology*, 22(14). 60

[145] R. Agarwal, A. M. Gonzalez-Angulo, and S. Myhre et al. Integrative analysis of cyclin protein levels identifies cyclin b1 as a classifier and predictor of outcomes in breast cancer. *Clinical Cancer Research*, 15(11):3654â3662, 2009. 60

[146] C.B. Moelans, R.A. de Weger, and H.N. Monsuur et al. Molecular profiling of invasive breast cancer by multiplex ligation-dependent probe amplification-based copy number analysis of tumor suppressor and oncogenes. *Mod Pathol*, 2010. 60, 62

[147] K. T. Hwang, W. Han, and J. Cho et al. Genomic copy number alterations as predictive markers of systemic recurrence in breast cancer. *International Journal of Cancer*, 123(8):1807â1815, 2008. 60

[148] R. Nahta, D. Yu, and M. C. Hung et al. Mechanisms of disease: Understanding resistance to her2-targeted therapy in human breast cancer. *Nature Clinical Practice Oncology*, 3(5):269â280, 2006. 60

[149] F. Toledo and G. M. Wahl. Mdm2 and mdm4: p53 regulators as targets in anticancer therapy. *International Journal of Biochemistry and Cell Biology*, 39(7). 61

[150] P. van der Lelij, K. H. Chrzanowska, and B. C. Godthelp et al. Warsaw breakage syndrome, a cohesinopathy associated with mutations in the xpd helicase family member ddx11/chlr1. *American Journal of Human Genetics*, 86(2):262â266, 2010. 61

[151] R. Venkatachalam, E. T. P. Verwiel, and E. J. Kamping et al. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *International Journal of Cancer*, 129(7):1635â1642, 2011. 61

[152] C.Walowsky, D. J. Fitzhugh, and I. B. Castano et al. The topoisomerase-related function gene trf4 affects cellular sensitivity to the antitumor agent camptothecin. *Journal of Biological Chemistry*, 274(11):7302â7308, 1999. 61

[153] A. M. Brown. Wnt signaling in breast cancer: have we come full circle? *Breast Cancer Research*, 3(6):351â355, 2001. 61

[154] F. Forozan, E. H. Mahlamaki, and O. Monni et al. Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary dna microarray data. *Cancer Research*, 60(16):4519â4525, 2000. 61

[155] K. Al-Kuraya, P. Schraml, and J. Torhorst et al. Prognostic relevance of gene amplifications and coamplifications in breast cancer. *Cancer Research*, 64(23):8534–8540, 2004. 62

[156] E A. Mittendorf, Y. Liu, and S.L. Tucker et al. A novel interaction between HER2/neu and cyclin E in breast cancer. *Oncogene*, 29:3896–3907, July 2010. 62, 83

[157] M. Scaltriti, P.J. Eichhorn, and J. Cortes et al. Cyclin E amplification/overexpression is a mechanism of trastuzumab resistance in HER2+ breast cancer patients. *Proceedings of the National Academy of Sciences*, 2011. 62, 83

[158] A. Subramanian, S. Shackney, and R. Schwartz. Novel multi-sample scheme for inferring phylogenetic markers from whole genome tumor profiles. *Proceedings of the 8th International Symposium on Bioinformatics Research*

and Applications (ISBRA) 2012, Dallas, TX, USA, May 21-23, 2012.Lecture Notes in Computer Science., 7292:250–262, 2012. 66

[159] F. Picard, S. Robin, and M. Lavielle et al. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6, 2005. 67

[160] L. Hsu, S.G. Self, and D. Grove et al. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, 2005. 67

[161] P.H.C. Eilers and R. de Menezes. Quantile smoothing of array CGH data. *Bioinformatics*, 21(7):1146–1153, 2005. 67

[162] K. Wang, M. Li, and D. Hadley et al. Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Research*, 17(11):1665–1674, 2007. 67, 68, 76

[163] Chris D. Greenman, Graham Bignell, Adam Butler, Sarah Edkins, Jon Hinton, Dave Beare, Sajani Swamy, Thomas Santarius, Lina Chen, Sara Widaa, P. Andy Futreal, and Michael R. Stratton. Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, 11(1):164–175, 2010. 67, 68, 76

[164] R. Pique-Regi, A. Ortega, and S. Asgharzadeh. Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. *Bioinformatics*, 25(10):1223–1230, 2009. 67

[165] S.P. Shah, K.J. Cheung, and N. A. Johnson et al. Model-based clustering of array cgh data. *Bioinformatics*, 25(12):i30–i38, 2009. 67, 68

[166] V.D. Wiel, A. Mark, and R. Brosens et al. Smoothing waves in array CGH tumor profiles. *Bioinformatics*, 25(9):1099–1104, 2009. 67

[167] L.Y. Wu, H.A. Chipman, S.B. Bull, L. Briollais, and K. Wang. A Bayesian segmentation approach to ascertain copy number variations at the population level. *Bioinformatics*, 25(13):1669–1679, 2009. 67

[168] N.R. Zhang, Y. Senbabaoglu, and J.Z. Li. Joint estimation of DNA copy number from multiple platforms. *Bioinformatics*, 26(2):153–160, 2010. 67

[169] R. Beroukhim, G. Getz, and L. Nghiemphu et al. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*, 104(50):20007–20012, 2007. 67, 76

[170] G. Nowak, T. Hastie, J.R. Pollack, and R. Tibshirani. A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics*, 12(4):776–791, 2011. 67, 68, 71

[171] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin. Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, 12(3):413–428, 2011. 67, 68, 76, 83

[172] A.B. Olshen, E.S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of arrayâbased DNA copy number data. *Biostatistics*, 5(4):557–572, 2004. 76

[173] M K Kuhner and J Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468, 1994. 77

[174] C. Lengauer, K. W. Kinzler, and B. Vogelstein. Genetic instabilities in human cancers. *Nature*, 396:643–649, 1998. 80

[175] S. Bamford, E. Dawson, and S. Forbes et al. The COSMIC (catalogue of somatic mutations in cancer) database and website. *Br J Cancer.*, 2004. 80

[176] J. Barretina, G. Caponigro, and N. Stransky et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.*, 483(7391):603–607, 2012. 85

[177] S.E. Burdall, A.M. Hanby, and M.R. Lansdown et al. Breast cancer cell lines: friend or foe? *Breast Cancer Res.*, 5(2):89–95, 2003. 85

[178] R.M. Neve, K. Chin, and J. Fridlyand et al. Ta collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell.*, 10(6):515–527, 2006. 86

[179] K.A. Graham, J.M. Trent, and C.K. Osborne et al. The use of restriction fragment polymorphisms to identify the cell line mcf-7. *Breast Cancer Res Treat.*, 8(1):29–34, 1986. 86

[180] C.K. Osborne, K. Hobbs, and J.M. Trent. Biological differences among mcf-7 human breast cancer cell lines from different laboratories. *Breast Cancer Res Treat.*, 9(2):111–121, 1987. 86, 93

[181] M. Resnicoff, E.E. Medrano, and O.L. Podhajcer et al. Subpopulations of mcf7 cells separated by percoll gradient centrifugation: a model to analyze the heterogeneity of human breast cancer. *Proc Natl Acad Sci U S A.*, 84(20):7295–7299, 1987. 86, 93

[182] M. Nugoli, P. Chuchana, and J. Vendrell et al. A sequence-level map of chromosomal breakpoints in the mcf-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.*, 19(2):167–177. 86, 87, 88, 93

[183] K.R. Coser, B.S. Wittner, and N.F. Rosenthal et al. Antiestrogen-resistant subclones of mcf-7 human breast cancer cells are derived from a common monoclonal drug-resistant progenitor. *Proc Natl Acad Sci U S A.*., 106(34):14536–14541. 87

[184] L. Gonzalez-Malerva, J. Park, and L. Zou et al. High- throughput ectopic expression screen for tamoxifen resistance identifies an atypical kinase that blocks autophagy. *Proc Natl Acad Sci U S A.*., 108(5):2058–2063. 87, 93

[185] S.V. Sharma, D.Y. Lee, and B. Li et al. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell*, 141(1):69–80. 87, 93

[186] X. Yang, S.P. Chockalingam, and S. Aluru. A survey of error-correction methods for next-generation sequencing. *Brief Bioinform.*, 14(1):56–66, March 2012. 95, 113

[187] A. Mahmoody, C.L. Kahn, and B.J. Raphael. Reconstructing genome mixtures from partial adjacencies. *BMC Bioinformatics.*, 2012. 113

[188] F. Vandin, E. Upfal, and B.J. Raphael. Intratumor heterogeneity in evolutionary models of tumor progression. *Algorithms Mol Biol.*, 188(2):461–477, 2011. 114

[189] R. Durrett, J. Foo, K. Leder, J. Mayberry, and F. Michor. Finding driver pathways in cancer: models and algorithms. *Algorithms Mol Biol.*, 1(23), 2012. 114

[190] O. Podlaha, M. Riester, S. De, and F. Michor. Evolution of the cancer genome. *Trends Genet.*, 28(4):155–163, 2012. 114