Algorithms to Reconstruct Evolutionary Models of Tumor Progression

Salim Akhter Chowdhury

CMU-CB-15-101

February 18, 2015

School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15213

Thesis Committee:

Russell Schwartz, Chair
Nathan Clark
Dannie Durand
Carl Kingsford
Adrian Lee
Alejandro A. Schäffer

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy.

This research was supported in part by National Institutes of Health grant 1R01CA140214. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: Tumor Progression, Tumor Evolution, Tumor Phylogenetics, Copy Number Change, Steiner Tree, Maximum Parsimony, Fluorescence *in situ* Hybridization, Cancer Diagnosis and Prognosis

Abstract

Cancer is one of the major causes of human mortality. Extensive genetic, epigenetic and physiological variations are observed within tumor cells, which complicates the diagnosis and treatment of the disease. Despite the extensive heterogeneity within single tumors, recurring features of their evolutionary processes are observed by comparing multiple regions or cells of a tumor. Recently, phylogenetic models have begun to see widespread use in cancer research to reconstruct processes of evolution in tumor progression. Mutations that drive development and progression of solid tumors typically include changes in the number of copies of genes or genomic regions. One particularly useful source of data for studying likely progression of individual tumors is fluorescence in situ hybridization (FISH), which allows one to count copy numbers of several genes in hundreds of single cells per tumor and thus especially well suited to characterizing intratumor heterogeneity. This thesis focuses primarily on phylogenetic characterization of single tumors at the cellular level from FISH data. We first develop phylogenetic methods using single gene duplication to infer likely models of tumor progression at the cellular level from FISH copy number data and apply these to a study of FISH data from two cancer types. We next extend our single gene models to include copy number changes at the scale of entire chromosomes and the whole genome. We develop new provably optimal methods for computing an edit distance between the copy number states of two cells given evolution by copy number changes of single probes, all probes on a chromosome, or all probes in the genome. Our two proposed models for inferring phylogenies of single tumors by copy number evolution assume models of uniform rates of genomic gain and loss across different genomic sites and scales, a substantial oversimplification necessitated by a lack of algorithms and quantitative parameters for fitting to more realistic tumor evolution models. We propose a framework for inferring models of tumor progression including variable rates for different gain and loss events. Application of the phylogenies inferred by our algorithms to real cervical and breast cancer data identifies key genomic events in disease progression consistent with prior literature. Classification experiments on cervical and tongue cancer datasets lead to improved prediction accuracy for models that allow non-uniform rates over models that have uniform rates, for the metastasis of primary cervical cancers and for tongue cancer survival.

Acknowledgments

The journey to the Ph.D. is the most challenging journey in one's lifetime. I am deeply indebted to a number people for their guidance and sacrifices throughout this wonderful journey. First and foremost, I would like to thank my Ph.D. supervisor Dr. Russell Schwartz for introducing me to the field of tumor phylogenetics, and for his exceptional guidance and supervision that kept me motivated all the time. Russell has always encouraged me to challenge my limits by putting forth the intellectually challenging problems. With great patience, he has listened to all my ideas and never showed any sign of discontent when he found problems with the solutions I proposed. He has taught me how to formulate a problem and solve it in a step-by-step fashion. I have always tried to emulate his exceptional writing skill and it has helped me to improve my standards. He has always encouraged me to send my papers to the top conferences and top journals in computational biology for publication. I have also learned from Russell how to handle rejections and criticisms from the reviewers with a positive mindset.

I am also deeply grateful to Dr. Alejandro Schäffer for his exceptional mentoring during the course of my Ph.D. I have learned enormously from him when it comes to performing algorithmic and mathematical analyses. With great patience, he has gone through all the mathematical proofs I have formulated and the software codes I have written. I will miss his encouraging remarks which have kept me going during the challenging times.

I am thankful to my M.S. supervisor at Case Western Reserve University, Dr. Mehmet Koyuturk, for introducing me to the field of computational biology and for instilling the belief within me that I have the potential to make fundamental contributions to this wonderful field.

I would like to thank Dr. Nathan Clark, Dr. Dannie Durand, Dr. Carl Kingsford and Dr. Adrian Lee for serving on my thesis committee and for the helpful suggestions, ideas and comments on my thesis.

Most importantly, I would like to thank my mother, my father and my brother for supporting and encouraging me throughout my entire life. I would like to pay special thanks to my wife, Nafisa Iqbal, for her love and patience throughout the course of my Ph.D. Without the sacrifices of my parents and my wife, I could never have achieved what I have accomplished in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is one of the major causes of mortality in the world. Large research and clinical efforts have been put into improving the treatment of cancer in the last few decades. As a result of the huge efforts, our understanding of cancer biology has advanced by leaps and bounds and the mortality rate has improved for some types of cancer [185]. Still, successful early detection and reliable treatment of the disease remain elusive [185]. The main challenge remains in translating our understanding of the disease to successfully identify and control advanced disease states [185].

Cancer develops as a result of dysregulation of the cell cycle in the metazoan body. Individual cells in the metazoan body carry a full genome and are endowed with great autonomy [185]. Each cell grows and divides, and thus participates in the formation of tissues. This autonomy of cells poses a threat to the life of an organism when corruption of the genome by any of various mechanisms results in alteration of cellular growth programs or disruption of external controls on growth. This may result in large populations of cells effectively acting autonomously and no longer obeying the rules governing normal tissue organization. As a consequence of the cells' normal development going awry, cancer develops. Once the breakdown occurs, tumor cells succumb to different evolutionary pressure: making more and more copies of themselves, growing, invading, and metastasizing [185].

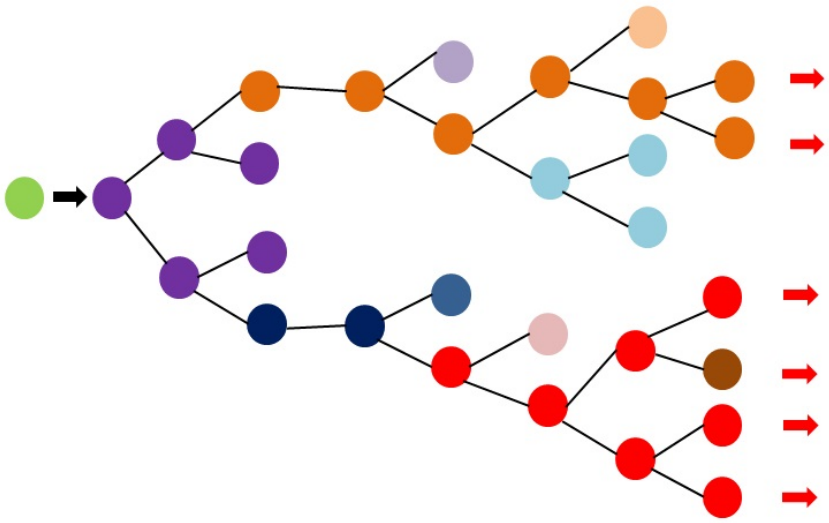

Figure 1.1: Clonal evolution model of tumor progression.

1.1 Cancer is an evolutionary process

Cancer is a multi-step process which is driven by genetic mutations and epigenetic alterations of DNA [185]. The genetic and epigenetic alterations typically affect the genes controlling growth, proliferation and survival of the cells [185]. It has long been recognized that cancer is an evolutionary process [185] which is characterized by accumulation of mutations that drive tumor initiation, progression and development of treatment resistance. This complex process requires bypassing a series of barriers between normal and cancer cells and reflects the work of evolution.

An early theory combining evolutionary biology with tumor biology was presented by Nowell [126]. The proposed model, known as "Clonal evolution model", proposes that tumors initially arise from a single mutated cell and, afterwards, biological and clinical progression ensue. During the progression of the disease, tumor cells initially undergo a brief period of heterogeneity followed by expansion of one or multiple subpopulations of cells. These subpopulations are known as subclones. Each of these expanding subclones undergoes Darwinian evolution and natural selection (Figure 1.1), meaning these clones undergo positive selection when advantageous mutations arise and negative selection when deleterious mutations arise. The advantageously positive mutations, known as "driver mutations", are the main driving force behind the expan-

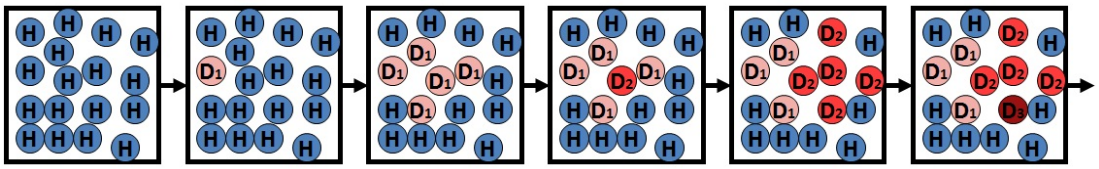

Figure 1.2: Example showing emergence and progression of cancer.

sion of the clones. Other types of mutations, known as "passenger mutations", hitchhike on the expanding clones but are assumed to be not under selective pressure.

A simpler pictorial view of initiation and progression of cancer is given in Figure 1.2. In a healthy tissue, mutations may arise in certain genes in one of the cells which result in that cell's acquiring higher proliferative potential in comparison to the healthy cells. This cell (as shown by D_1 in Figure 1.2) grows, divides faster and quickly overpopulates the healthy cells in the tissue. As tumor cells typically acquire damage to the DNA replication system, which makes them prone to further mutations [105], other mutations may arise resulting in some other cell population (as shown by D_2 cell type in the picture) with high growth potential to appear. This process of acquiring mutations, resulting in a faster growth rate for a subset of cells, goes on and ultimately we have multiple subpopulations of genomically distinct cell types forming the tumor mass. Afterwards, the tumor may become invasive, with cells migrating and colonizing to other parts of the body and potentially becoming life threatening.

Instead of a linear model of evolution proposed by Fearon and Vogelstein for colorectal cancer [50], where it is assumed that sequentially ordered mutations in a set of driver genes results in clonal sweeps of homogeneous tumor cell expansion, recent research suggests a branch-type evolutionary mechanism. Branch-type evolution was evident in findings of Anderson *et al.* [7] on fluorescence *in situ* hybridization (FISH) based analyses of acute leukemia single cells, where they observed this type of evolution in different subclones of the same patient. In-depth genome sequence analysis of 21 breast cancers also revealed branch-type evolution and subclonal variation [121]. Yachida *et al.* [190] showed that clonal tumor populations present in primary pancreatic cancers give rise to metastatic disease in a branched evolutionary pattern. From DNA

sequencing of 13 pancreatic patients, Campbell *et al.* [31] showed that genomic instability occurs early during cancer development.

One important insight into the evolutionary aspect of tumor progression has been appreciating that a tumor is a heterogeneous system [163]. Although subtypes of normal cells in humans are morphologically similar, striking dissimilarity exists within the cells in tumors. Early genomic technologies identified high tumor heterogeneity of tumors from patient to patient. In spite of this heterogeneity, some functionally organizing principles exist within the tumor cells. These organizing principles are summarized in [74].

There are experimental evidences of both intratumor and intertumor heterogeneity. Whole genome sequence studies have revealed distinct mutations in closely related tumors. Similarly, studies profiling variations among single cells have revealed large heterogeneity within single tumors. It has been shown that tumor heterogeneity temporally evolves during disease progression [163]. Through whole genome sequencing, Shah *et al.* [146] identified existence of 19 non-synonymous mutations in metastases that were not present in a primary lobular carcinoma of the breast diagnosed 9 years previously. Using DNA copy number analysis by CGH or single-cell sequencing, Navin *et al.* [117, 118] demonstrated that a single breast cancer biopsy may contain multiple intermixed karyotypic tumour populations that differ by major structural chromosomal gene amplifications. Gerlinger *et al.* [62] showed that 63-69% of all non-synonymous somatic mutations identified across multiple biopsies of two primary tumours and associated metastatic sites were not detectable in a single biopsy, suggesting that a single biopsy may underestimate the somatic mutational landscape of a tumor. Campbell *et al.* [31] demonstrated the presence of genomic heterogeneity among metastasis initiating cells and that seeding metastasis driver mutations may be beyond those required for primary tumors. Also, all of the 21 tumors in [121] harbored a dominant clone that was distinct from other subclones by thousands of mutations.

1.2 Clinical implications of intratumor heterogeneity

Cancer cells evolve by accumulating mutations and genomic rearrangements during cell division. This accumulation of mutations may result in a high level of heterogeneity not only among cancer patients but also within the cells forming a particular tumor. The intertumor and intratumor heterogeneity have been shown to have significant effect on both diagnosis and treatment of cancer [65, 163]. As cancer is a continuously evolving system, the success in treatment of patients depends not only how the disease is in current state, but also what state it will be over the course of the treatment. This "evolutionary" perspective of treatment can be expected to change significantly how the disease is treated in the future. Both intrinsic and acquired mutations complicate the treatment of chronic myeloid leukemia by the drug imatinib [60] which is a major concern in clinical practice now.

The "evolutionary" aspect of cancer progression creates various problems in the diagnosis, prognosis and treatment of the disease. The heterogeneity of tumor cells results in sampling bias, complicating the process of extracting a representative cell population [163]. The sampling bias may arise due to diversity in events in single biopsies. The dynamic nature of the subclone architecture create challenges during disease diagnosis and prognosis as the subclone that defines clinical outcome may not be detectable at diagnosis. At present, normally targeted treatment with chemotherapy is done based on the primary lesion, which may have been collected months or years in prior. Such a treatment strategy may cease to be therapeutically tractable as previously dormant or sub-dominant subclones in the primary lesion become preeminent over the course of the treatment resulting in development of therapeutic resistance and relapse of the disease.

The current procedures of biomarker discovery also get complicated with the dynamic "dominance architecture" of the tumors. Currently, only 100 of the 150,000 known biomarkers are used in clinical practice [137]. The biomarker discovery approaches that combine prediction of gene function with the help of genetic or transcriptomic analyses of tumor tissue often depend on tumor biopsies collected from the primary or metastatic site of the tumor to prioritize the

identification of candidate biomarkers for validation [163]. The highly heterogeneous nature of the tumors may result in tumor sampling bias that leads to confounding effects in the validation of biomarkers due to spurious associations of heterogeneous tumor genetic events with clinical outcome.

An important guiding principle of modern cancer research is that early detection of cancer is crucial to treatment of the disease [152]. Research in recent times has focused on detection of the disease before it becomes aggressive and invasive, although these efforts had much less impact in reducing the mortality than had been expected [18]. This observation has led to the conclusion that many earlier-detected cancers never posed a threat to the patient. Overtreatment of cancers that would never have been a threat can have significant side-effects on the patients' health. The overtreatment issue creates a dilemma because focusing solely on avoiding it may conversely result in failing to treat the tumors aggressively that have the potential to become dangerous [23].

Another aspect of disease treatment that is affected by heterogeneous characteristics of cancer is targeted therapeutics. It is assumed that initiation and progression of cancer is result of disruption in a limited number of cellular functions and pathways caused by a handful "driver" genes for any one type of cancer [73, 74]. It has been proposed that identification of the driver genes will allow us to design drug or therapeutics targeting these genes in a patient specific manner, allowing for personalized disease treatments. The goal is that this will allow us to kill more specifically the cancer cells while avoiding the disastrous side-effects of traditional chemotherapy [40, 131] which targets indiscriminately rapidly dividing normal and cancer cells. Targeted therapeutics often focus on blocking proliferation of tumor cells, whereas the goal of chemotherapy is to kill tumor cells. To date, more than 100 targeted therapeutics have been identified, facilitated by huge investments in large-scale sequencing efforts nearly identifying most of all driver genes. Targeted therapeutics has some notable successes in discovering effective drugs, such as trastuzumab for *HER-2* overexpressed breast and stomach cancer [163], vemurafenib for treating patients with metastatic melanoma that contains altered BRAF protein [32], imatinib

mesylate which targets the BCR-ABL fusion protein in leukemic cells [47]. Although targeted therapeutics has had some great success stories, however, it has fallen short of its initial promise because the normal outcome is only short-term recovery before the emergence of drug resistance that leads to relapse and mortality [53]. The emergence of drug resistance is predictable from the evolutionary point of view of cancer progression.

Drug resistance arises because application of the drugs changes cancer clone dynamics by introducing a new source of artificial selection. But similar evolutionary principles still apply. Introduction of the selective pressure for proliferation of variant cells promotes the development of resistance of therapeutics by several mechanisms [60]. Also, the surviving cancer cells may have incurred additional mutations, some of which can improve their malignant potential. Both "acquired" and "de novo" resistance have been observed in targeted therapeutics. An example of the "acquired" resistance is the emergence of a "gatekeeper" mutation in the BCR-ABL oncoprotein in relapsing chronic myelogenous leukemia (CML) cells treated with the ABL inhibitor imatinib [60]. An example of "de novo" resistance is the progression of disease in 10% of $BRAF^{V600E}$ melanomas after treatment with RAF inhibitor vemurafenib [54].

1.3 Mutations in cancer cells

In normal human cells, the rate of point mutations is 10^{-10} per base per cell division [14]. Due to the large number of mutations in cancer cells, it has been proposed that the spontaneous mutation rate in normal cells is not enough to give rise to the enormous mutation load in cancer cells [106]. The phenomenon of dramatic increase in mutation rates in cancer and ultimately the resistance to treatment can be attributed to a phenomenon known as "hypermutability", also known as a "mutator phenotype" of the tumor [83]. This phenotype is the result of mutations in genes that function in the maintenance of genomic stability and arise as a result of errors in DNA synthesis and repair of DNA damage [105]. Errors in DNA synthesis occur when the number of misincorporation of nucleotides by DNA polymerases exceeds the capacity of cells to excise

and repair the lesions. DNA damage is produced by reactive molecules generated in cells as a result of metabolic processes and also by environmental exposures. When the amount of DNA damage exceeds the cells' capacity for DNA repair, the residual nonrepaired lesions become a dominant source of mutations during DNA replication. Another source of mutations in cancer cells is the unequal partitioning of chromosomes during cell division, which results in a change in chromosome number in the daughter cells, a phenomenon known as aneuploidy.

The best known example of hypermutability is mutation in the *TP53* gene observed in a majority of human solid tumors [66], which produces characteristic patterns of Chromosome Instability (CIN) resulting in extensive gain, loss and rearrangement of whole or large fractions of chromosomes during cell division. CIN produces an imbalance in the number of chromosome in cells and an increase in the rate of loss of heterozygosity [125]. The DNA Mismatch Repair (MMR) system is used for recognizing and repairing faulty insertion, deletion and misincorporation of bases during DNA replication, recombination. Defects in MMR can produce a pattern of hyper-elevated microsatellite instability [68, 173]. Recent research of tumor mutation burdens across many tumors has revealed at least four distinct somatic point mutation hypermutability phenotypes: DNA mismatch repair defects [68, 173], two distinct signatures of DNA polymerase defects [44] and AID/APOBEC cytidine deaminase dysregulation [76, 170]. This wide array of observations supports a long standing hypothesis that hypermutability is the defining characteristic of aggressive cancers. This hypothesis also provides a simple explanation of why targeted therapeutics exhibit low success rate in cancer therapy: it may not be particularly important to future tumor progression which driver genes are active in a tumor at a particular time; rather hypermutability and plasticity allow tumors to easily shift to a new driver set once the set of current dominant drivers are targeted. The observation suggests that cancer treatment should not focus on only on the current state of tumor. Rather, it should try to identify reliable progression biomarkers to overcome therapeutic resistance and the overtreatment/undertreatment issue. The time has come to recognize cancer as a process of ongoing evolution rather than treating it as a stationary system.

Mutations in cancer genomes affect both a single nucleotide and large segments of genes, phenomena termed Single Nucleotide Variation (SNV) and Copy Number Aberrations (CNA), respectively. SNVs are variations within genomes at single base levels. CNAs result in larger rearrangements in chromosome segments and are an effective force behind the larger genomic and phynotypic heterogeneity observed in cancer. In cells, copy numbers of genes may change as a result of Single Gene Duplication (SD), Chromosome Duplication (CD) or Whole Genome Duplication (GD) events. SD events result in gain or loss of single gene copy numbers and can occur as a consequence of non-allelic homologous recombination or single-strand annealing events during DNA repair cycles [78]. Chromosome level events are gain and loss of whole chromosomes due to incomplete segregation of chromosomes during mitosis. On the other hand, genome level events result in duplication of the whole genomes due to failure to segregation of chromosomes during mitosis. In addition, cancer genes may exhibit epigenetic aberrations, such as change in DNA methylation [38].

Other types of complex patterns of aberrations are observed in cancer genomes too. Kataegis refers to the phenomenon where certain single nucleotide polymorphisms (SNP) are found in clusters in some parts of the genome. Such hypermutable regions have been observed in breast cancers [121]. In a Breakage Fusion Bridge cycle, a telomere of a chromosome breaks off and replication of that chromosome gives rise to two sister chromatides both lacking a telomere. This ultimately results in two daughter cells receiving uneven chromatids, which ultimately leads to DNA amplification [69]. Chromothripsis refers to the phenomenon of massive genomic rearrangements as a result of a single catastrophic effect during the cell's lifetime [159]. Chromothripsis is considered to be provoked by radiation exposure at a critical time point during cell cycle when chromosomes are condensed for mitosis. Such chromosome shattering events may represent the upper limit of genomic damage cells can tolerate and cells that survive these events can have selective advantage as a result of increased tumor cell growth. Chromoplexy is another source of complex genomic rearrangements where several strands of DNA are broken and then ligated together to form a new configuration. There has been evidence of chromoplexy both at

clonal and subclonal level in prostate tumors [9]. The breakpoints in chromothripsis are clustered in one chromosome, whereas chromoplexy involves multiple chromosomes.

1.4 Phylogenetic models

Phylogenetic models are widely used computational tools to study the evolutionary relationships among a set of species. Phylogenies are typically tree based models, but in complicated cases, they can take the structure of more general graphs too. Each node in a phylogenetic tree is called a taxon. In traditional phylogenies, observed taxa are placed at the leaf level and inferred internal nodes, known as "Steiner Nodes", represent extinct or unobserved taxa. Phylogeny-building algorithms can largely be divided into two categories: distance-based and character-based [51]. Distance-based phylogenetic algorithms aim to infer the evolutionary relationships among a set of taxa by using a measure of similarity or distance based on the genetic or physical characteristics of the taxa. First, a distance between each pair of taxa is computed and then a tree is built so that the phylogenetic distances between the taxa in the tree closely resemble the computed distances. Distance-based methods can further be subdivided into objective-based and non-objective-based methods. Unweighted Pair Group Method using Arithmetic Averages (UPGMA) [51] and Neighbor Joining (NJ) [141] are two of the most well-known non-objective-based methods, where trees are built in a bottom-up fashion by connecting the closest pair of taxa and then updating the distance matrix at each step until all the taxa are connected into a tree. In objective-based methods, such as minimum evolution [51], one aims to optimize an objective function, such as minimizing the total weights of the tree edges connecting the taxa. The objective-based methods are computationally expensive, but they have the advantage of providing theoretical guarantees of identifying an optimal tree characterized by a well defined objective function, often by searching through the space of all possible trees.

The second class of phylogeny building algorithms are character-based methods. Character-based methods use aligned set of characters, such as DNA or protein sequences, and build a

tree to infer the changes in characters that describe the generation of the observed characters from a common ancestral sequence [51]. Character based algorithms can be divided into three subgroups: Maximum Parsimony (MP), Maximum Likelihood (ML) and Bayesian. The goal of MP is to construct a tree so that the total number of mutations along the edges of the tree is as small as possible [51]. MP follows a variant of Occam's Razor, where the intuition is that the most plausible evolutionary history uses as small a number of mutations as possible to introduce variations. Thus, recurrent mutations can be assumed rare and optimizing for the number of mutations is a reasonable assumption. ML is a probabilistic approach where the objective is to choose a model consisting of a tree topology and other parameters from a set of all possible models so that the probability of the observing the sequence under the inferred model is as high as possible [51]. Under conditions for which the maximum parsimony assumption does not hold, the ML approach can be expected to give a better model of evolutionary history, but the ML based models can be computationally very expensive. Finally, the Bayesian approach [91] infers the posterior distribution over the set of all possible models consisting of tree topology, sequences and parameters.

Among the set of phylogenetic models, the character-based models are generally considered to give a more detailed and accurate representation of the evolutionary history among the species under consideration, but can be computationally very expensive. Normally, as the number of taxa increases, the cost of building character based models goes up and quickly becomes intractable. However, a wide variety of heuristic methods have been developed to make character based phylogenies feasible in practice if the number of taxa and markers are not too large [51].

1.5 Phylogenetic models of tumor progression

Tumor cells accumulate mutations and undergo large genomic rearrangements during the cell cycle to form distinct subpopulations of cells. Due to this evolutionary nature of the progression of cancer, phylogenetic models have been used to infer models of tumor progression from ge-

netic variation data [14]. One of the earliest use of phylogenetic models in cancer biology was developed by Desper *et al.* [42]. The proposed tree-based model, termed "oncogenetic tree", uses a wild-type node as the tree root, with each other node representing a cell type which arises as a result of a genomic change (could be a mutation or a copy number change) in the parent node. A maximum weight branching-based [48] tree construction algorithm was proposed for inferring the tumor phylogenies. Later Desper *et al.* [43] proposed a distance based tree construction method to estimate dependencies among cancer driving events. Von Heydebreck *et al.* [178] proposed a maximum likelihood based model for probabilistic Bayesian formulation of the oncogenetic tree model. In [93], both oncogenetic tree and distance-based models were used to analyze relationships among chromosomal abnormalities in clear cell renal cell carcinoma. Szabo and Boucher [164] extended the oncogenetic tree models to account for false positive and false negative observations. Beerenwinkel *et al.* [12] proposed a structural EM algorithm for identifying mixtures of oncogenetic trees, which allows a more flexible tree-fitting approach in the presence of multiple tumor subgroups. Tofigh *et al.* [174] proposed a generalization of the mixture of oncogenetic tree models by proposing a different error model and a global structural EM algorithm for learning mixtures of trees.

Advances in next generation sequencing have made it possible to infer SNVs in a population of cancer cells. Due to the large number of SNVs in cancer cells, it is computationally difficult to infer a complete model of tumor progression explaining the observed data. To infer phylogenies from SNV frequencies, Nik-Zainal *et al.* [121] used an infinite site assumption (no mutations occurs twice in the progression of cancer) and impossibility of back mutations (no mutation is lost) assumption to solve the problem in practice. Strino *et al.* [160] proposed a linear algebra-based parsimony approach to implement these two assumptions. Their proposed method works for up to 25 abnormalities. A recent method was developed to generalize the approach proposed in [191] for multiple samples per patients. Subclone identification is required for inferring phylogenies based on SNV frequencies data. Miller *et al.* [113] proposed a Variational Bayes mixture model for identification of subclones and Fischer *et al.* [52] proposed a Hidden Markov Model

making use of information present in CNAs and SNVs.

An alternative to sequencing distinct tumor cells or regions is to computationally reconstruct intratumor heterogeneity via mixture modeling. For inference of tumor phylogeny from gene expression data, Schwartz *et al.* [143] proposed an unmixing of tumor samples and then used the inferred mixtures of individual tumor states to identify possible evolutionary relationships among tumor cells. Application of the proposed method to a lung cancer dataset [143] showed that the method suggests possible evolutionary relationships that show good correspondence with current understanding of lung tumor development. One of the limitation of the proposed approach was that it handled noisy and outlier data poorly. To overcome this issue, they proposed a more robust unmixing method [175] that reduces the sensitivity to noise and outliers. Application of the method to CGH data yielded fast and accurate separation of tumor states. In [161], Subramanian *et al.* developed a pipeline for inferring tumor phylogenies from the deconvoluted heterogeneous tumor samples. Later, in [162], they proposed a novel Hidden Markov Model (HMM) for inferring phylogenetically significant high-resolution markers of progression through joint segmentation and calling of multisample tumor data.

Other methods are designed to infer phylogenetic history of single tumors profiled at regional level. Navin *et al.* [119] used a neighbor joining method on CGH breast cancer data to show that the breast cancers they studied could be subdivided into one homogeneous group and one genetically heterogeneous group. Letouze *et al.* [100] presented a method named TuMult, that works on DNA breakpoints instead of full genomic profiles. They used a breakpoint distance to compute the number of genomic events between any two copy number profiles and then used this distance for tree reconstruction. Greenman *et al.* [67] proposed a graph theoretical approach to identify the most likely assignment of gene copy number to either allele using external linkage information, which ultimately finds the most likely clonal ordering over the whole set of genomes of a tumor. Using a minimum evolution criterion, Schwarz *et al.* [145] proposed a method called MEDICC to jointly solve the problems of phasing (process of finding the correct assignment of major and minor copy-number to the two parental alleles) and tree reconstruction. MEDICC

infers the shortest number of amplification and deletion events required to transform one genomic profile into another using finite state transducers [144].

A number of methods have also been developed for inferring clonal progression of single tumors by using cell to cell variation in single tumors [132, 133, 134]. Two types of data are used for extracting genomic information at single cell level: (i) for a limited number of loci, fluorescence *in situ* hybridization (FISH) is used to identify copy number information of the targeted genes across hundreds to thousands of cells, (ii) on the whole genome scale, recently single-cell sequencing has been used effectively [148] as a source of data for single tumor phylogeny inference. In the earliest work employing FISH data, Pennington *et al.* [132, 133, 134] developed a model inference framework using an Expectation Maximization (EM) [41] approach for concurrently inferring mutation rates of gene copy number evolution and iteratively optimizing the tree fit relative to the model and vice versa. Martins *et al.* [109] used FISH data from 55 *BRCA1* breast cancers and developed computational methods to predict the termporal order of somatic events of the genes *BRCA1, PTEN*, and *p53*. The other emerging type of single-cell data is single-cell sequencing data at a genome scale. Navin *et al.* [118] used low-coverage single nucleus sequencing to infer evolutionary histories of cancer lineages using the neighbor joining algorithm, employing a Euclidean distance metric on the discretized integer copy number profiles for tree construction. Xu *et al.* [189] performed performed single-cell exome sequencing on a clear cell renal cell carcinoma tumor and its adjacent kidney tissue and found small number of mutant genes present in a large fraction of individual cells and a significantly greater number of genes mutant present in only one or a few cells. Hou *et al.* [87] performed whole-exome single-cell sequencing of 90 cells from a *JAK2*-negative myeloproliferative neoplasm patient and found that this neoplasm represented a monoclonal evolution. Since myeloid tumors are hematopoietic tumors and hence subject to fewer evolutionary pressures. Thus, the finding of monoclonality in a myeloid tumor is not a direct contradiction to intratumor heterogeneity observed in solid tumors.

1.6 Our contributions

Although there have been widespread efforts put in to developing methods for inferring tumor phylogenies, some characteristic features of tumor evolution have mostly been ignored in the models that have previously been proposed. The evolution of cancer is different from evolution of species in a number of important ways. Progression and evolution of cancer is characterized by rapid mutations, widespread and frequent copy number and other structural rearrangements, and complex selective pressures [78, 81, 99]. These characteristic features are problematic for traditional phylogenetic models and algorithms that were developed for inferring ancestral relationship among species. The chromosome instability (CIN) phenomenon, which is a major driver of cancer progression [125], results in abnormal patterns of gains and losses of genomic materials at large scales and at a rapid pace, which is very different from the mutation patterns assumed in standard models of species or population-level evolution. There is no standard algorithm that takes into account our rich knowledge of these characteristic patterns of tumor evolution. It is imperative that tumor phylogeny building algorithms accurately model the rapidly evolving tumor genomes through gain and loss of genes and chromosomes if they are to characterize accurately the progression of the disease at a cellular level.

Another characteristic feature of tumor evolution that separates it from species evolution is high genetic and epigenetic heterogeneity within the cell population. Both FISH and single-cell sequencing study have shown that single tumors may contain hundreds of genetically distinct clones, suggesting one may need to sample thousands of cells per tumor to capture intra-tumor heterogeneity reliably [7, 146]. The phylogeny building algorithms needed to process these data must be able to handle datasets with at least hundreds of cells per tumor to have hope of modeling the role of rare cell populations in tumor progression. A farther implication of high inter-tumor heterogeneity is that tumor phylogenetics must be performed on patient populations to identify common characteristics of the progression mechanism predictive of tumor progression.

Previous methods for cell-level tumor phylogenetics could handle only a small number of

copy number probes [132, 133, 134]. Other methods [109] used greatly simplified models of CIN evolution. More advanced phylogenetic models have been developed for datasets of regional tumor variations consisting of 10-20 biopsies per tumor [157] employing sophisticated Bayesian algorithms [11], but such approaches scale poorly with large datasets. There is a need for tumor phylogenetic algorithms that can capture the evolution of tumors by CIN and that are also capable of handling hundreds to thousands of cells across tens of patients to infer more realistic models of tumor progression adequate to capture observed levels of intratumor and intertumor heterogeneity. Traditional phylogeny building algorithms, such as neighbor joining [141], which have been used in analyzing single-cell sequencing [117, 189] are capable of handling large datasets, but they are not designed to model copy number evolution.

The work in this thesis is intended to address the need for scalable algorithms for inferring tumor phylogenies at the cellular scale by copy number variation. The first key contribution of the thesis is the development of novel methods that can model copy number evolution on the scales needed to capture intratumor and intertumor heterogeneity. The thesis describes a novel algorithm for inferring tumor phylogenies from single-cell gene copy number data using a simple model of tumor progression assuming that gene copy number change events happen by gain or loss of single gene probe. The algorithm extends the theory for Rectilinear Steiner Tree Problem (RSMT) [58] which provides a basis for inferring maximum parsimony models describing evolution by single gene copy number changes in order to make it possible to apply them to large data sets. Two approaches are developed, an optimal but exponential-time method and a fast heuristic method based on the median joining algorithm [10]. Novel theory is also developed to reduce the search space of the exponential-time algorithm to handle larger datasets. The proposed algorithms are used to infer phylogenies on FISH dataset. FISH makes it possible to count copies of small numbers of gene probes across hundreds to thousands of cells, which provides a way to characterize relative frequencies of CIN at gene, chromosome or whole genome scale and to characterize intratumor and intertumor heterogeneity. Although other technologies such as single-cell sequencing allow one to gather whole-genome variation data, it is currently imprac-

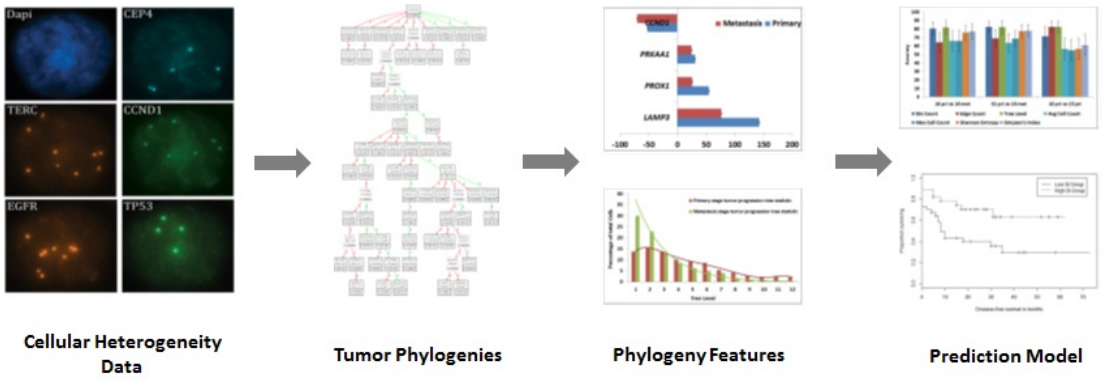

Figure 1.3: The algorithms in this thesis infers tumor progression models using single-cell gene copy number data. The phylogeny models are used to extract meaningful quantitative measures (features) of tree topologies that characterize the mutation process of a single tumor. These features are then used for clinically relevant tasks: to predict future tumor progression and for prediction of overall and disease free survival times.

tically expensive to sequence hundreds of cells across tens of patients, as is needed to identify predictive common features of tumor phylogenies conserved across patient populations.

The next contribution of the thesis is development of novel algorithms for extension of the single gene probe level evolutionary model to add chromosome-level events (gain or loss of whole chromosomes) and genome-level events (duplication of the whole genome). All these three types of events are common in tumor evolution [78] and failure to account for all three of them would distort inferred tree topologies. Allowing all these three events makes the tree inference problem complex, as one cannot rely on the RSMT theory that first made these inferences tractable for large datasets. A provable algorithm with theoretical guarantees is developed to compute the maximum parsimonious single gene duplication/loss, chromosome duplication/loss and genome duplication events required to transform one copy number profile into another. This algorithm provides a major component of the phylogeny building algorithm: inference of evolutionary distances among pairs of taxa. The novel algorithms result in a comprehensive model of tumor progression considering all three types of gene copy number change events.

Another key contribution of the thesis is the development of algorithms for tree inference and parameter estimation to learn unknown quantitative parameters of tumor evolution models. The

comprehensive model of tumor evolution developed in the thesis initially assumes that gene level, chromosome level and whole genome level duplication events happen with same frequency. But specific gene probes and chromosomes are under distinct selective pressure in cancers and thus their copy number change events are associated with own distinct frequencies. Accurate tumor phylogeny inference depends on accurate models of the corresponding rate parameters. Such models were unavailable in the prior literature and no methods were available either to learn these parameters and to apply them in phylogeny inference. The thesis develops novel algorithms for computing maximum weighted parsimony combinations of gene level, chromosome level and genome level duplication events when each of these events occur with different frequencies. An Expectation Maximization (EM) type algorithm is also developed for the joint inference of event frequencies and tree topology.

The final key contribution of the thesis is the use of tumor phylogenies in developing composite biomarkers that are associated with traditional measures of tumor progression such as metastasis or clinical stage. Tumor phylogenetics has largely been used for discovery in basic cancer research, primarily as a way of distinguishing driver from passenger mutations [29], and to show how similar types of tumors are phylogenetically related to each other [43]. There has been little effort to show clinical significance of tumor phylogenies in guiding patient treatment, however. Work in thesis develops the idea that one can identify characteristic features of inferred tumor phylogenetic models that statistical and machine learning algorithms can employ to draw robust inferences about the progression mechanism. The algorithms developed in this thesis make it possible to build tumor phylogenies from hundreds to thousands of cells across tens of patients, which makes it possible to produce phylogenies complex and numerous enough to mine them for quantitative prediction. The work in this thesis demonstrates that the tree topologies capture important characteristics of tumor progression mechanisms and that numerical features quantifying the basic tree topologies can scale as crucial predictors of future progression. The numerical features are applied to two prediction tasks: supervised prediction of future tumor progression, and unsupervised prediction of overall and disease free survival time. These prediction tasks are

validated with four tumor types — breast, prostate, cervical and tongue — that are each major human health concerns. Figure 1.3 gives a summary of the key contributions of the thesis as well as a snapshot of the overall pipeline of inferring the tumor phylogneny from single-cell gene copy number data to prediction of tumor progression and patient survival time.

1.7 Datasets used in this thesis

In this thesis, we applied our algorithms to FISH datasets on four kinds of human cancers: cervical cancer (CC), breast cancer (BC), oral (tongue) cancer (TC) and prostate cancer (PC). The datasets used are as follows:

- A set of CC [183] FISH data consisting of 47 samples organized into 16 primary samples of metastatic patients, 16 paired metastasis samples from the same patients, and 15 primary samples from patients who did not progress to metastasis. Each sample consisted of $223 - 250$ cells profiled on four FISH probes: *LAMP3* [95], *PROX1* [187], *PRKAA1* [89] and *CCND1* [56]. All of these four genes are oncogenes, which typically show copy number gains in tumor cells. Each of the genes belongs to a distinct chromosome.
- A set of BC [83] FISH data consisting of 13 paired (from the same patient) ductal carcinoma in situ (DCIS) and invasive ductal breast carcinoma (IDC) samples with $76 - 220$ cells per sample profiled on eight FISH probes: *COX-2* [88], *MYC* [188], *CCND1* [56], *HER-2* [168], *ZNF217* [122], *DBC2* [72], *CDH1* [17] and *TP53* [179]. The first five genes in this list are oncogenes and the last three genes are tumor suppressors. In tumor cells, tumor suppressors are typically associated with loss in copy numbers.
- Dataset TC [Wangsa et al., submitted] consists of 65 single samples collected from tongue cancer patients probed for four genes located on distinct chromosomes (*TERC, EGFR, CCND1, TP53*), with tumor stages (ranging from 1 (least advanced) up to 4 (most advanced)) available on all patients and tobacco usage (a known risk factor), survival, and

disease-free survival out to 73 months available on most patients.

- An additional cervical cancer dataset, CC2, which consists of following set of cervical samples: (a) one early pre-cancerous lesion (denoted by CIN1), (b) ten late pre-cancerous lesions (denoted by CIN3) and (c) ten cancerous lesions (denoted by CA). Each sample was probed on eight genes: *COX, ING5, FHIT, TERC, TERT, MYC, CHEK1, ZNF217*. *ING5, FHIT, CHEK1* are tumor suppressors and all others are oncogenes. *FHIT* and *TERC* lie on chromosome 3, while the others all reside on distinct chromosomes.

In cervical cancer, normally prognosis is good with early detection, but drops sharply with lymph node metastasis [180]. The CC dataset is extensively used in this thesis for prediction of separating samples from different stages of cancer based on features extracted from the phylogenies. CC remains the gold standard for prediction of progression of cancer to metastasis in this thesis. This problem is clinically very important, as it allows us to predict which primary tumors are going to become aggressive and thus give proper treatment to the patients with early detected cancer. Improved prediction will help address the undertreatment/overtreatment issue, which is a major problem in cancer treatment.

Breast cancer is the most prevalent type of cancer in females and is one of the most well-studied cancers. Breast cancer has been used extensively to study intratumor heterogeneity [83]. Breast cancer also has historical importance because it was the first model for the use of targeted therapeutics (Herceptin) in treating cancer. BC dataset is used in this study to identify the subset of phylogenetic features that helps improving the separation of DCIS samples from the IDC ones. In addition, statistical tests are performed on this dataset to illustrate the differential selective pressures working on DCIS and IDC stages of breast cancer.

Tongue cancer patients used in this study were followed up to 73 months from the time of initial diagnosis, and overall and disease free survival times were reported for each patient. As a result of the presence of this metadata, the TC dataset gives us an opportunity to use it as a baseline for prediction of survival time, because no prediction biomarker has been validated to

have predictive power beyond that already available from cancer staging information. In this thesis, the tongue cancer dataset is used to validate the survival time prediction performance of the tumor phylogeny statistic based subgrouping of patients.

Prostate cancer has been a well-studied system to pursue questions related to overdiagnosis and overtreatment. As prostate cancer is a slow-growing tumor of primarily elderly people, it is now normally treated without the goal of eliminating the tumor. However, the untreated cancers can ultimately prove to be dangerous to the patients. In this thesis, phylogenetic features are used to predict the recurrence of prostate cancer, which can be useful in identifying which tumors need aggressive early treatment.

1.8 Thesis organization

Chapter 2 gives a detailed description of an exact and a heuristic approach for inference of tumor phylogenetic models using single-cell gene copy number data under single gene duplication/loss model of tumor evolution. Machine learning algorithms are used for predicting future tumor stage using numerical features characterizing tumor progression. Chapter 3 extends the single gene duplication based model developed in Chapter 2 to add chromosome level duplication/loss and whole genome duplication events for inference of tumor phylogenies. Simulation experiments are performed to evaluate the reconstruction accuracy of the proposed comprehensive algorithm for inferring progression trees with known topology. As an extension of the algorithms in Chapter 3, Chapter 4 describes a framework for inferring the rates of different gene and chromosome copy number change events and to infer the maximum parsimonious trees under the new weighted-event model. Chapters 5 and 6 provide applications of tumor phylogenetic models in understanding tumor progression and predicting partient survival time across two different types of cancers. Chapter 7 proposes a novel distance measure for computing difference between phylogenetic trees built on arbitrary sets of taxa. Finally, Chapter 8 summarizes the findings of the studies and their conclusions and outlines possible future directions on the topic of tumor

phylogenetics.

Chapter 2

Phylogenetic Analysis of Multiprobe Fluorescence in situ Hybridization Data from Tumor Cell Populations[1]

Recent studies of genetic variation in solid tumors have revealed massive intratumor heterogeneity in the spectrum of genomic changes within single tumors [62, 83, 117, 118]. These observations suggest the importance of understanding cell-to-cell variability, but profiling large numbers of single cells and building coherent models of their evolution remain challenging problems. Fluoresence *in situ* hybridization (FISH) is a technique that can be used to count the copy number of DNA probes for specific genes or chromosomal regions that has proven useful in studying cancer. Gene gains and losses are common in solid tumors and FISH provides a practical and reliable method for monitoring such changes in large numbers of individual cells from single tumors [84, 92]. FISH is even more useful when one uses multiple colors to monitor multiple genes simultaneously [83, 109, 183]. In this chapter, we develop new methods to model and analyze the progression of copy number changes, as measured by multicolor FISH. Our methods include

[1]This chapter was developed from material published in "Chowdhury *et al.*, Phylogenetic analysis of multiprobe fluorescence *in situ* hybridization data from tumor cell populations, *Bioinformatics*, 29.13 (2013): i189-i198" [34].

analysis of multiple samples from the same patient, typically from different cancer stages.

We apply the new methods to published data on cervical cancer with four gene probes [183] and breast cancer with eight gene probes [83]. The data for each sample are presented as a matrix, where each column is one of the probes and each row is a "cell count pattern" of four probe counts, such as 2,3,4,1 (or eight counts for breast cancer) and the number of cells matching that pattern. A normal cell has the count pattern of all 2s. Both data sets include paired samples from an earlier stage and a later stage in the same patient.

It is of interest to study cervical and breast cancer, as we do, because the number of cases of cervical cancer and breast cancer diagnosed early has increased due to Pap smears and mammograms respectively. Early diagnosis is important because lymph node metastasis is one of the best predictors of poor outcome [28, 49]. Paradoxically, it has been shown statistically that early diagnosis of breast cancer has not led to a substantial decrease in deaths because most of the cancers diagnosed early would not progress to be life-threatening if left untreated [18]. These are large-scale studies that do not address the benefits of early detection in individual cases. Therefore, a consistent explanation of these findings is that earlier detection could be of clinical value if there were a better understanding of tumor evolution to help identify the early-stage cancers likely to progress to dangerous forms.

The problem of modeling tumor progression has been studied by a variety of techniques and using different kinds of tumor data [12, 33, 42, 63, 67, 109, 134, 150, 161, 177, 189]. Several of these methods used techniques from the area of phylogenetics (reviewed by [8]) because of the insight that tumor genomes evolve [29, 126]. Most prior studies have used either comparative genome hybridization or sequencing of cell populations, which have the advantage that one can do genome-wide analysis, but the disadvantage that the input data are explicitly or implicitly averaged over many cells of the same tumor. Other data types can offer distinct advantages, such as the microsatellite data used by Shlush *et al.* ([150]), which can allow some inference of useful population genetic parameters generally difficult to assess with tumor data.

As discussed in Chapter 1, FISH is the only currently available reliable technique that allows

measurements on enough individual cells to model the evolution of substantial intratumor heterogeneity. The disadvantage of FISH is that it uses only a small number of preselected markers. Only two of the previous studies analyzed FISH data ([109, 134]). These studies were limited either to two probes ([134]) or to three probes and coarsely distinguishing only loss (copy number < 2), neutral (copy number 2), and gain (copy number > 2) ([109]).

We address a need for new methods capable of handling the larger numbers of cells and probes in recent FISH data sets. More specifically, we aim to develop a theoretical foundation for efficient handling of large copy number data sets. Towards this goal, we develop theory and algorithms for a model of tumor progression driven by gains and losses associated with FISH probes. Our methods handle in principle any number of probes and any range of copy numbers 0 through UB (default 9). The use of UB limits the combinatorial search and hence running time of our methods on inputs where the measured copy numbers exceed this limit. The work is intended to establish a framework capable of giving useful tree inferences on state-of-the-art FISH data, which might be extended in future work to handle even harder problem instances and more realistic models of tumor evolution.

Our contributions in this chapter include:

1. Reducing a model of the problem of modeling progression of FISH probe cell count patterns to the Rectilinear Steiner Minimum Tree (RSMT) problem and thus bringing prior theory on the RSMT problem to bear on the FISH phylogeny problem.
2. Design and implementation of an exponential-time exact method and a polynomial-time heuristic method to construct trees modeling the progression of cell count patterns.
3. Mathematical proof and software implementation of a new inequality that speeds up the RSMT-based computation.
4. Definition and evaluation of new test statistics based on the trees computed by our methods. These test statistics give novel insight into the selective pressures in tumor progression, compared to test statistics derived from the cell count patterns alone.

5. Definition and evaluation of "features" based on the tree structures that can be used with machine learning to classify the tumors. For example, we show improved effectiveness at distinguishing the cervical tumors that metastasize from those that do not.

2.1 Methods

In this section, we describe a set of algorithms to identify a most parsimonious tree of copy number changes consistent with a data set on cell-level tumor copy-number heterogeneity. We first describe an exponential-time exact algorithm. We next propose a set of valid inequalities to reduce the running time of the algorithm. We then propose a heuristic approach that returns an approximate solution in polynomial time. Both the exact and heuristic methods are implemented in the C++ software package FISHtrees (ftp://ftp.ncbi.nlm.nih.gov/pub/FISHtrees).

Data sets

We give a description of the datasets used in this chapter for evaluating our tree inference method. The datasets were described in section 1.7. For clarity, we reiterate here the sample composition and the genes on which each dataset was profiled.

The cervical cancer (CC) dataset contains genomic copy numbers of the four oncogenes *LAMP3* [95], *PROX1* [187], *PRKAA1* [89] and *CCND1* [56] on samples from 16 lymph node positive and 15 lymph node negative patients [183]. For the lymph node positive patients, this dataset contains a sample from the primary tumor and another from the metastasis, making the total number of samples 47. The number of cells per sample ranges from 223 to 250, after filtering to remove cells that likely had cut nuclei and those in the process of division, as described previously [83]. The breast cancer (BC) dataset contains copy numbers of 5 oncogenes, typically but not always gained *COX-2* [88], *MYC* [188], *CCND1* [56], *HER-2* [168] and *ZNF217* [122]- and 3 tumor suppressor genes, typically lost *DBC2* [72], *CDH1* [17] and *p53* [179] from 26

paired samples, one from the ductal carcinoma in situ (DCIS) and one from an invasive ductal carcinoma (IDC), from 13 patients. The number of interphase cells ranges from 76 to 220. The FISH protocol filters out cells that are in the process of DNA replication using the fact that these cells have recognizable FISH probe doublets [183].

2.1.1 Rectilinear Steiner Minimum Tree (RSMT) problem

For each patient sample, each cell assayed will have some non-negative integer number of copies of each probe. If we consider measurements on d probes in c cell count patterns for a given patient, then that patient's information can be represented by a two-dimensional array D with c rows and $d+1$ columns where entry $D(i,j)$, for $j = 1, \ldots, d$, represents the copy number of gene j in sample pattern i, and column $d+1$ has the number of cells with this count pattern. All counts above UB are reduced to that value. Each row of D can be treated as a point in R^d. Our goal is to explain the observed data via a phylogenetic tree of single gene duplication and loss events. We use the L_1, or rectilinear, distance metric for inferring the Steiner nodes in R^d. If we are given a set S of points in R^d, and we build a Steiner tree T spanning S, then for any particular edge e, joining points $x = (x_1, x_2, \ldots, x_d)$ and $y = (y_1, y_2, \ldots, y_d)$, the rectilinear distance $w(e)$ is defined by $w(e) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_d - y_d|$. The problem of identifying a minimum weight tree including all the observed points and, as needed, unobserved Steiner nodes with the rectilinear metric is known as the Rectilinear Steiner Minimum Tree (RSMT) problem [75, 154].

The RSMT problem is NP-complete [58] and thus does not have an efficient exact algorithm. One potential advantage of reducing to Steiner trees is that there are high-quality implementations of sophisticated branch-and-cut methods that solve large instances to optimality [96, 136]. To keep our implementation free and self-contained, we implemented a simpler domain-specific method for our instances of RSMT. We developed an inefficient exact algorithm and a heuristic algorithm based on the median-joining algorithm for maximum parsimony phylogenetics [10] adapted for RSMT using theoretical results from Hanan [75] and Snyder [154].

2.1.2 Exact algorithm for the RSMT problem

An exact algorithm for the RSMT problem in two dimensions was first proposed by Hanan ([75]). Hanan's theorem implies that if we draw lines parallel to the two axes through each of the points in S, then there exists an RSMT of S whose Steiner nodes will be located at the intersections of those lines, a two-dimensional grid known as the Hanan grid that identifies a finite set of positions where the Steiner nodes might be found. Snyder [154] generalized Hanan's theorem to any dimension d. To formally present Snyders theorem, assume S is a set of points in a d dimensional space R^d and that $x_1, x_2, \ldots, x_d$ are the coordinate axes of R^d. Let $P = (p_1, p_2, \ldots, p_d)$ is a point belonging to S. There are d hyperplanes orthogonal to the coordinate axes that contain P. Suppose $N(P, i)$ is one of those hyperplanes which is orthogonal to the axis x_i. Hanans grid in d dimensional space is formed by taking the union of all $N(P, i)$ for all points P and all dimensions i. Formally, the Hanan grid $H(S)$ in d dimensions is defined as, $H(S) = \cup N(P, i), \forall P \in S$ and $1 \leq i \leq d$. For each subset of the form $N(P_1, 1), N(P_2, 2), , N(P_d, d)$, there is a point at which the d hyperplanes intersect. If the set of all of these intersection points are denoted as $IH(S)$, then the generalized Hanan's theorem, proposed and proved by Snyder is the following:

Theorem 1. *[154] For a given set of points $S \subset R^d$, there exists an RSMT T of S, such that if Q is a Steiner point of T, then $Q \in IH(S)$.*

Algorithm 1 Exact algorithm for generating RSMT

Require: A point set S
Ensure: Steiner tree including the set of inferred Steiner nodes and weight of the Steiner tree

```
Infer Steiner node set Q using generalized Hanan's theorem
Identify MST on S and let min_weight = weight(MST(S))
for k ← 1, |Q| do
    Enumerate all size-k subsets T of Q
    for each T_k ∈ T do
        Identify MST on {S ∪ T_k} and let current_mst_weight = weight(MST(S ∪ T_k))
        if current_mst_weight < min_weight then
            min_weight = current_mst_weight, steiner_tree = MST(S ∪ T_k)
    return steiner_tree and min_weight
```

According to Theorem 1, the possible Steiner nodes are the intersection points of the Hanan

grid $H(S)$. For each possible number of Steiner nodes k, Algorithm 1 enumerates all subsets of k potential Steiner nodes from those allowed by the generalized Hanan's theorem. For each such subset, the algorithm constructs a minimum spanning tree (MST) using the observed data points and those k specific Steiner nodes. The minimum cost tree over all such subsets and all possible values of k is returned as the optimal tree. This method is guaranteed to find an optimal solution to the RSMT problem if k is large enough, but in practice we limit to $k \leq 3$. More efficient approaches, such as that proposed in [46], cannot be used in our case, since they assume that all the terminal nodes must be leaf nodes in the Steiner tree, while in tumor progression trees, a terminal node can be a parent node of other terminal nodes. Below, we show that the Algorithm 1 run time is at worst exponential in the number of potential Steiner nodes and the size of the probe set.

Theorem 2. *The time complexity of Algorithm 1 is exponential in the number of potential Steiner nodes.*

Proof. If $UB + 1 = m$, then, by Theorem 1, the total number of possible Steiner nodes to be considered is $s = m^d$. To find the exact solution, we consider each possible subset of the inferred Steiner nodes and build a minimum spanning tree on the set of terminals and subset of Steiner nodes under consideration. The total number of subsets of a set with cardinality n is 2^n. We implemented Prim's algorithm for MST and its complexity is $O(nlogn)$. So, the total running time of Algorithm 1 is $O(2^s nlogn)$. □

2.1.3 Pruning Steiner node subsets

Since the time complexity of Algorithm 1 depends on the number of calls made to the MST routine, we can reduce its runtime by checking before hand if a call to that procedure cannot lead to a solution of lower cost than $current\_mst\_weight$. We propose a lower bound on the weight of the MST and we add checks in every for loop in Algorithm 1 to test if the lower bound is higher than the minimum weight MST generated so far. If so, then we do not generate the MST.

Theorem 3. *Suppose we would like to build an MST on a graph that has n nodes, of which nodes $1, \ldots, r$ might have degree 1 in an MST and hence be eligible to be its root, while nodes $r+1, \ldots, n$ are required to have degree > 1 in the MST and hence are not eligible to be its root. By construction, a Steiner node in the graph must have degree > 1 in the MST because otherwise, its inclusion cannot reduce the weight of the MST.*

Assume, the weight matrix of the graph is

$$w_{11}, w_{12}, w_{13}, \ldots, w_{1n}$$

$$w_{21}, w_{22}, w_{23}, \ldots, w_{2n}$$

$$\ldots$$

$$w_{n1}, w_{n2}, w_{n3}, \ldots, w_{nn}$$

Then,

$$W(MST(n)) \geq (w_1 + w_2 + \ldots + w_n) - \sup(w_1, w_2, \ldots, w_r) \tag{2.1}$$

where $W(MST(n))$ is the total weight of the MST with n nodes and

$$w_i = \inf(w_{i1}, w_{i2}, \ldots, w_{i(i-1)}, w_{i(i+1)}, \ldots, w_{in})$$

.

Here, for a list L, $\sup(L)$ and $\inf(L)$ denote the lowest upper bound and greatest lower bound of L, respectively.

Proof. We define the difference on the right hand side of the inequality (claim) as $Q(n)$. For each non-root node v, define $p[v]$ to be the weight of the edge connecting v to its parent in the MST, and define $p[root] = 0$. We can readily see that $p[v] \geq w_v$. $p[v]$ cannot be smaller than w_v as

edges do not get split in the MST building process. For a graph with n nodes, $W(MST(n)) = p[1] + p[2] + \ldots + p[n]$. If we assume node 1 is the root node, then $W(MST(n)) = p[2] + p[3] + \ldots + p[n]$. We divide into two cases depending on the value of $\sup(w_1, w_2, w_3, \ldots, w_r)$. If $\sup(w_1, w_2, w_3, \ldots, w_r) = w_1$, then $Q(n) = w_1 + w_2 + w_3 + \ldots + w_n - w_1 = w_2 + w_3 + \ldots + w_n$. Since $w_v \leq p[v]$ for any node v, we have, $Q(n) = w_2 + w_3 + \ldots + w_n \leq p[2] + p[3] + \ldots + p[n] = W(MST(n))$. On the other hand, if $\overline{w} = \sup(w_1, w_2, w_3, \ldots, w_r) > w_1$, then, $w_1 + w_2 + w_3 + \ldots + w_n - \overline{w} < w_2 + w_3 + \ldots + w_n \leq p[2] + p[3] + \ldots + p[n] = W(MST(n))$ So, $W(MST(n)) \geq (w_1 + w_2 + \ldots + w_n) - \sup(w_1, w_2, \ldots, w_r)$. □

Distinguishing between the potential Steiner nodes and the non-Steiner nodes in (2.1) makes the claim more complicated, but leads to a direct simplification of the algorithm. Fewer calls to the MST procedure are made because the lower bound exceeds the current best MST weight more often.

2.1.4 Heuristic algorithm for the RSMT problem

We also propose a heuristic method that can find a potentially suboptimal solution in polynomial time. Our proposed heuristic method uses the median joining principle of iteratively identifying Steiner nodes (known as median nodes) that allow one to more parsimoniously link some triplet of nodes, using the generalized Hanan theorem to enumerate possible medians. The method, described in Algorithm 2, begins by constructing a minimum spanning network, corresponding to the union of edges in all minimum spanning trees. It then enumerates triplets of nodes (u, v, w) such that at least two nodes of each triplet are connected in the network, followed by enumerating possible median nodes, consisting of combinations of coordinate values of u, v, and w. It then tests whether introducing the given possible median as a Steiner node reduces the cost of the minimum spanning tree. If so, then the median node is added. The process is continued until no additional cost-saving median node can be added.

Theorem 4. *The time complexity of Algorithm 2 is polynomial in the cardinality of the terminal*

Algorithm 2 Heuristic algorithm for generating RSMT

Require: A point set S

Ensure: Steiner tree including the set of inferred Steiner nodes and weight of the Steiner tree

```
Calculate Minimum Spanning Network (MSN) on S using the approach described in [10]
Identify MST on S and let min_weight = weight(MST(S))
Identify all 3 node subsets of MSN, T, where at least two nodes out of the 3 nodes are connected
for each T_i ∈ T do
    Identify candidate Steiner node set L by taking combination of the values of coordinate axes of the points in T_i
    for each L_i ∈ L do
        Identify MST on {S ∪ L_i} and let current_mst_weight = weight(MST(S ∪ L_i))
        if current_mst_weight < min_weight then
            min_weight = current_mst_weight
            S = S ∪ L_i
            steiner_tree = MST(S)
    return steiner_tree and min_weight
```

set.

Proof. The running time of the heuristic algorithm is dominated by the number of triples that are considered during the Steiner node inference process. The maximum number of triples considered for inferring the Steiner node is $\binom{n}{3}$. If we are considering d probes, then the maximum num-ber of Steiner nodes is $\binom{n}{3}3^d$. So, the total running time of the heuristic approach is $O(3^d n^4 logn)$. □

2.1.5 Experimental procedure

We began statistical analysis with a basic test of imbalance in tree topologies to determine whether differential evolutionary pressures in primary/DCIS versus metastatic/IDC environments might be reflected in the trees. To obtain sufficient counts and detect statistically significant trends, we grouped cells into bins by subtrees based on the child of the root from which each cell traces its ancestry. The root node represents a cell type with a copy number count of 2 for each gene probe (i.e., a healthy diploid cell). Direct children of the root are those nodes distinguished by an increase or decrease of one copy in a single probe. For example, for four gene probes in

the CC case, the copy number profiles of the eight children of the root are (1,2,2,2), (2,1,2,2), (2,2,1,2), (2,2,2,1), (3,2,2,2), (2,3,2,2), (2,2,3,2) and (2,2,2,3). We refer to all descendants of one second level node as a "bin". We counted the total number of cells in that bin for each of the eight sub-trees separately for the 16 pairs of primary and metastasis samples. The resulting eight-dimensional vectors for each primary-metastasis pair were compared by a chi-square test to test the null hypothesis of independence between bin counts and primary vs. metastasis labels.

To illustrate the difference between the dynamic views of relationships among cell types offered by the trees relative to the static snapshot offered by raw probe counts, we next examined two different measures of the net mutational bias in the CC and BC trees: one based on imbalance of copy numbers in cell counts and one on imbalance in tree edges. These statistics provide two different views of the net evolutionary process of mutation and selection. For the cell count data in CC/BC, we aggregated all 16/13 patients primary/DCIS and metastasis/IDC information separately, computing average difference in copy number of individual cells relative to diploid, excluding the contribution of all-diploid cells. For tree-based calculation of gene gain/loss, we measured the net gain or loss of each gene by the number of tree edges showing gain minus those showing loss over all trees generated by FISHtrees.

We also performed a series of experiments on the use of progression tree statistics for classification tasks related to tumor progression and prognosis. In each case, we examined the use of tree statistics as features for prediction methods in comparison to prediction from features derivable solely from raw cell counts. As feature sets, we used:

1. Fractions of cells in the 8/16 subtrees rooted at children of the diploid root: We defined tree-based features consisting of 8/16 features corresponding to the fraction of cells in each of the subtrees corresponding to immediate children of the diploid root.
2. Fractions of edges exhibiting gain or loss of each gene: We used 8/16 features corresponding to the fraction of total tree edges showing gain or loss of each gene.
3. Fractions of cells at each level from one to ten in the trees: We used ten features corre-

sponding to the fraction of cells at each level in the tree from one to ten. The root (the node representing normal cells) of the tree is assumed to be located at level one.

Fractional rather than absolute counts are used for each measure, so that the sum of the values is normalized to be 1 and the test statistics are not distorted by variability sample-to-sample in the number of cells. These features were compared to four non-tree-based features:

1. Mean gain or loss in each gene individually.
2. Maximum copy number of each gene individually.
3. Shannon index [130], an information theoretic measure. For each gene G, each distinct combination of gene copy numbers and cellular ploidy represents a species. If p_i denotes the frequency of species i among all tumors, then the Shannon index H for G is given by $H = -\sum p_i log_2(p_i)$.
4. Simpson index [130], $D = \sum p_i^2$.

We further performed simulation tests to evaluate the correctness of the phylogenetic trees inferred by our algorithm in terms of the underlying tumor progression mechanism. Trees were simulated to approximate true FISH progression trees by expanding from an initial diploid root node by selecting a Poisson number of children of each node (possibly with repetition) and expanding each node selected recursively until the process terminates. To produce trees comparable to the real data, we reject those with fewer than 50 or greater than 120 distinct cell types. Individual cells are then chosen uniformly from the nodes in this topology until 250 cells are sampled.

More specifically, we simulated progression trees by the following protocol, which takes a random number $r \in [0, 1]$, a number gene probes Np to be simulated and the number of cells K to be sampled:

1. Fix the root node to be $R = 2^{Np}$.
2. Initialize the tree $T = R$ and depth of the tree $d = 1$.

3. Extract all nodes N generated at depth d.

4. For each node $n \in N$, repeat the following:

   (a) Generate a uniform random number $U \in [0, 1]$.

   (b) if $U > r$, then do the following:

      i. Select a gene probe q from the copy number profile of n randomly.

      ii. Select a direction gl of unit copy number change (gain or loss) from the copy number profile of n randomly.

      iii. Generate a child C of n with copy number profile based on q and gl.

      iv. If the copy number profile of C has not been generated previously then put C into T.

      v. Go to step 4(a)

5. If no new child node is generated at this iteration then stop, otherwise increment d and go to step (3)

6. For each $k \in K$, generate a uniform integer random number I_R between 1 and $|T|$ and assign k to the copy number profile of the node indexed by I_R.

Simulations were conducted for the present work with parameters $r = 0.5, Np = 4$ and $K = 250$. In order to compare the similarity between the simulated tree T_S and the inferred tree T_I, we used a weighted matching based metric [102]. We first identified the set of non-trivial bipartitions in both simulated and inferred tree, represented by π_{T_S} and π_{T_I} respectively. Then we defined a complete weighted bipartite graph $G(A, B, E)$, where A and B consist of π_{T_S} and π_{T_I} respectively. Since T_S and T_I can have non disjoint node sets, we calculated the weight between two vertices of G by dividing the cardinality of shared nodes set between the bipartitions represented by those two vertices with the total number of nodes in the simulated tree. Then the distance between T_S and T_I is calculated by identifying the maximum weight matching in $G(A, B, E)$. After dividing the weight of the matching by the total number of nodes

in T_S, we calculated the fractional similarity in bipartitions between the two trees. We calculated the matching accuracy by multiplying the fractional similarity with 100.

2.2 Results

In this section, we present the results of experiments to evaluate the utility of tumor phylogeny inference for understanding the developmental processes of these tumor types. We also explore the prognostic value of tumor phylogenies by using them to derive features for classification experiments and comparing to features that do not rely on tree inference. For these experiments, we built tumor progression trees of the CC and BC data using the ploidyless heuristic approach to phylogenetic inference described in Methods.

Figure 2.1 shows representative examples of tumor progression trees from the CC data set. Figure 2.1(A) shows a tree inferred from the primary tumor of patient 1 of the CC dataset. Figure 2.1(B) shows the tree for the paired metastatic sample from patient 1. The primary stage tree has more nodes and is more balanced and broader in shape compared to the metastatic stage tree. The distinct topologies of the trees may indicate the fact that cells residing in primary and metastatic sites of the tumor face different selective pressures.

2.2.1 Comparison of exact and heuristic algorithms

To evaluate the quality of the solution generated by the heuristic algorithm, we generated tumor progression trees using both the exact and the heuristic algorithms. We report a comparative study of the two algorithms in Figure 2.2. For each example run, we report the number of probes considered, the total number of terminal nodes in the given dataset and a comparison of the weights of the RSMTs generated by the exact and heuristic approach. The heuristic approach returns an optimal solution about 80% of the time. For the cases where the heuristic solution is not optimal, the excess weight is very small. From the runtime comparison of the two approaches,

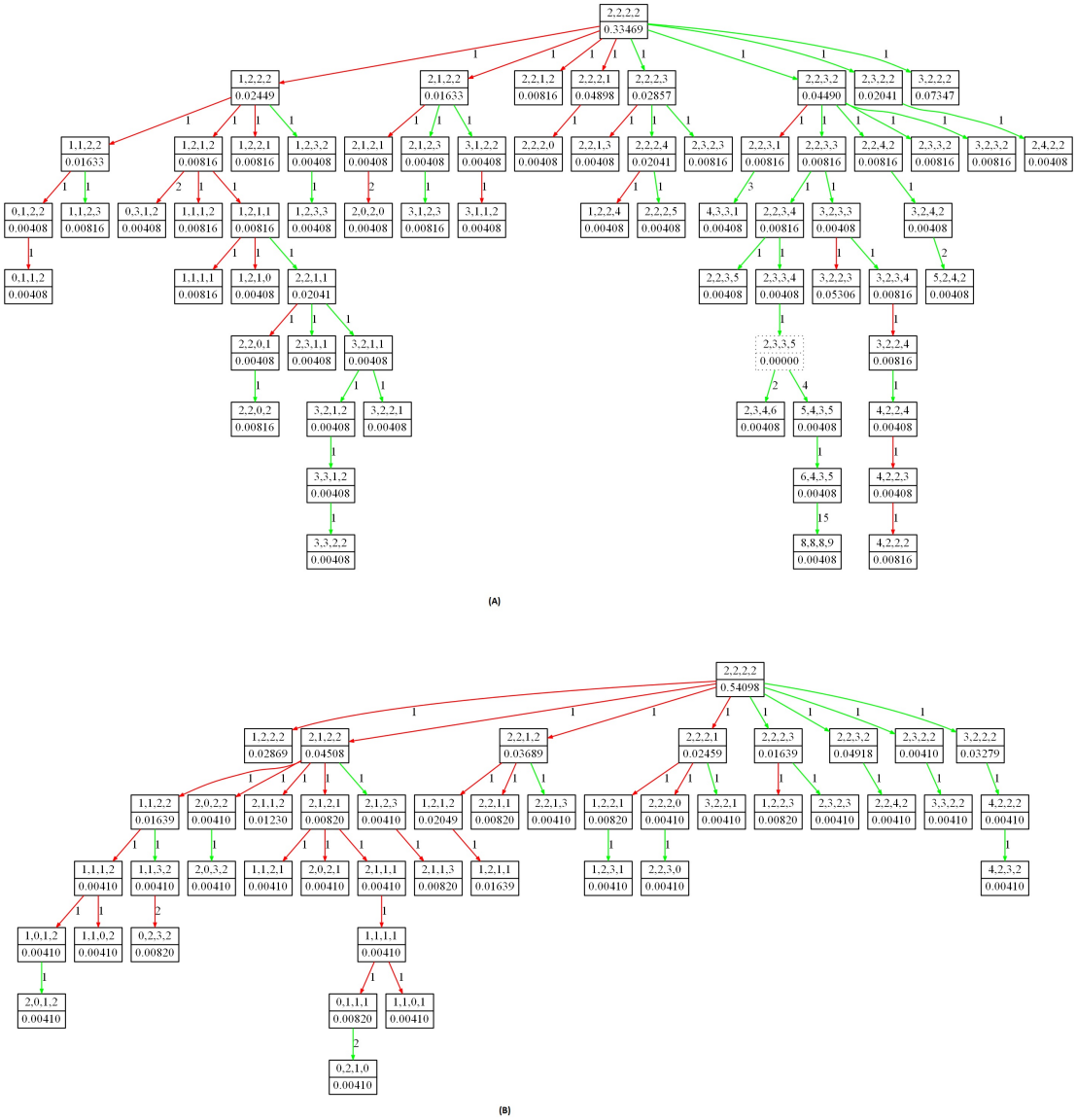

Figure 2.1: Phylogenetic trees showing progression of (A) primary and (B) metastasis stage cervical cancer in patient 1. The trees are built from single cell-copy number data using the ploidyless heuristic approach implemented in FISHtrees. Each node in the trees represents a copy number profile of the four gene probes *LAMP3, PROX1, PRKAA1* and *CCND1* respectively. Nodes with solid borders represent cells present in the collected sample, while nodes with dotted borders represent inferred Steiner nodes. Green and red edges model gene gain and gene loss respectively. The weight value on each edge connecting two nodes x and y is the rectilinear distance between the states of x and y. The weight on each node describes the fraction of cells in the sample with the particular copy number profile modeled by that node; Steiner nodes are assigned weight 0.

			Exact Approach		Heuristic Approach		
Case	Total Probes	Number of Terminal Nodes	RMST Weight	Total Runtime in Seconds	RMST Weight	Total Runtime in Seconds	Percentage of times Call to MST generation routine is avoided in the exact approach
1	3	96	126	910	126	1	80.30
2	3	40	51	10	51	1	85.77
3	3	74	94	318	94	1	67.04
4	3	32	35	1	35	1	99.42
5	3	78	93	150	93	1	63.20
6	3	40	42	1	42	1	96.11
7	3	78	91	47	91	1	94.12
8	3	42	43	1	43	1	99.99
9	3	70	85	16	85	1	91.04
10	3	37	42	1	42	1	99.47
11	3	62	70	7	70	1	89.06
12	3	33	35	1	35	1	99.94
13	3	68	79	4	81	1	95.60
14	3	31	33	1	33	1	93.53
15	3	73	89	165	89	1	61.68
16	3	107	123	100	124	1	97.75
17	3	39	52	87	52	1	75.62
18	3	10	14	1	14	1	36.02
19	3	72	91	473	91	1	67.93
20	3	65	81	154	83	1	83.17
21	4	87	90	10	90	1	99.88
22	4	57	60	4	60	1	99.19
23	4	69	73	57	73	1	93.25
24	4	61	70	236	71	1	99.15
25	4	65	79	813	79	1	88.34
26	4	59	65	11	66	1	99.55
27	4	63	73	87	74	1	98.75
28	4	29	33	11	33	1	94.32
29	4	31	38	16	38	1	90.82
30	4	63	76	250	76	1	97.10
31	4	33	42	35	42	1	90.64
32	4	58	68	129	68	1	99.98
33	4	21	32	63	32	1	28.57
34	4	62	72	28	72	1	98.81
35	4	50	55	8	55	1	98.79
36	4	118	128	157	128	1	99.98
37	4	48	55	18	56	1	97.17
38	4	80	83	56	83	1	99.98
39	4	70	73	17	73	1	99.98
40	4	76	83	30	83	1	99.21
41	4	63	76	140	76	1	99.74
42	4	32	43	144	43	1	94.79
43	4	39	47	82	47	1	98.66
44	5	39	50	35	52	1	96.12
45	5	29	31	12	32	1	99.98
46	5	25	26	125	26	1	99.98
47	5	25	37	1966	37	1	35.75
48	5	15	18	167	18	1	99.82
49	5	39	46	193	46	1	99.50

Figure 2.2: Example runs of the exact and heuristic approach showing the total number of probes considered, total terminal nodes in the particular case, RSMT weight, total runtimes of the exact and heuristics approach respectively and percentage of times the MST building routine in Algorithm 1 is not executed resulting in reduced running time of the exact approach.

we see that the heuristic approach returns a solution within 1 second every time. The runtime of the exact approach varies from 1 second to 1966 seconds. When the number of probes is higher

than 5, the total running time of the exact approach becomes impractical. The heuristic algorithm can return a solution in less than one minute even when using all eight probes in the BC data set (data not shown).

In Figure 2.2, we also report the percentage of total calls to the MST generation routine in Algorithm 1 that are avoided as a result of the inequality we proposed in Theorem 3. For 75% of the examples, the lower bound in Inequality 2.1 exceeds the current best MST weight more than 90% of the time. As most of the entries in Inequality 2.1 are computed just once and used throughout, this results in a huge decrease in the runtime of the exact approach.

2.2.2 Statistical analyses of tumor phylogenies

Cervical Cancer Primary vs. Metastatic Samples

Table 1

Patient ID	Chi-Square Test p-value
1	1.91E-15
2	4.17E-20
3	1.14E-11
4	4.39E-15
5	2.52E-12
6	3.69E-05
7	9.78E-06
8	1.47E-70
9	2.99E-37
10	1.99E-69
11	3.90E-14
12	4.83E-50
13	2.67E-50
14	4.32E-38
15	1.22E-11
16	1.76E-40

(A)

Table 2

Patient ID	Chi Square test p-value
1	1.15E-49
2	1.44E-20
3	1.96E-07
4	1.75E-42
5	5.85E-11
6	8.92E-64
7	7.72E-26
8	6.23E-56
9	3.00E-39
10	2.03E-34
11	9.15E-64
12	2.33E-35
13	7.29E-15

(B)

Table 3

Mutation	Total Number
LAMP3 Gain	7
PRKAA1 Gain	7
PROX1 Gain	5
CCND1 Gain	4

(C)

Table 4

Mutation	Total Number
DBC2 Loss	8
CDH1 Loss	6
COX-2 Gain	4
CCND1 Gain	4
HER-2 Gain	3
ZNF217 Gain	1
p53 Loss	1
MYC Gain	1

(D)

Figure 2.3: P-values from chi square tests comparing the number of descendants in the (A) eight children of the root in the primary tumor tree vs. the metastasis tree in the same CC patient, (B) sixteen children of the root in the DCIS tree vs. the IDC tree in the same BC patient. The total number of (C) CC and (D) BC patients for which each bin for gain of oncogenes or loss of tumor suppressor genes shows significance in individual 2×2 chi square tests.

We first examined cervical cancers, looking at paired primary tumor and metastasis samples.

Table 1 in Figure 2.3(A) reports p-values for chi square tests on all 16 pairs of patients. For each patient, the chi square tests compare two 8-element vectors, one for the primary tumor and one for the metastasis, in which element i is the number of descendants of the i-th child of the root. All p-values in this and other tests are corrected for multiple testing. In Figure 2.3(A), all 16 p-values are statistically significant. The same is true for an analogous chi square test of DCIS vs. IDC in 13 BC patients in Figure 2.3(B). That these comparisons are significant indicates significant imbalance between tree geometries of the two tumor stages. It may suggest that distinct evolutionary pressures act on growth in the primary tumor versus the metastasis.

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10	Patient 11	Patient 12	Patient 13	Patient 14	Patient 15	Patient 16
Gain of CCND1	17	9	0	0	0	3	4	0	7	186	1	0	4	0	3	10
Gain of PRKAA1	54	4	1	103	1	83	8	1	7	2	1	1	4	161	3	6
Gain of PROX1	6	7	0	7	7	0	1	0	10	3	2	11	21	0	44	0
Gain of LAMP3	18	2	115	4	17	4	11	0	3	1	82	188	6	0	36	12

(A)

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10	Patient 11	Patient 12	Patient 13	Patient 14	Patient 15	Patient 16
Gain of CCND1	7	6	4	3	0	8	1	1	3	0	4	3	0	11	17	3
Gain of PRKAA1	13	36	4	41	2	28	12	3	9	1	3	6	0	8	4	0
Gain of PROX1	2	2	3	37	18	3	2	8	1	2	1	3	20	3	4	0
Gain of LAMP3	10	3	133	0	9	3	3	71	3	81	14	15	4	5	14	114

(B)

Figure 2.4: Number of cells in the subtrees rooted at the nodes (termed a "bin") directly connected to the root node in the cervical cancer dataset. Data is shown for (A) primary and (B) metastatic stage tumor for bins representing gene gains, with the most populated bin highlighted in each case.

There is, however, high variability from patient-to-patient in the nature of the imbalance. Figure 2.4 shows cell counts for the bins associated with gain of the four genes, with the largest bin of each tree highlighted. The bin accounting for Gain of *LAMP3* is the most frequent dominant bin in both primary and metastatic samples. This bin is also the only dominant bin across multiple pairs of primary and metastatic samples. This finding is consistent with the ubiquitous gain of *LAMP3* in CC reported in [183]. We also performed chi square tests on individual bins in each pair of primary and metastasis samples using 2×2 contingency tables. Table 3 in Figure 2.3(C) reports the total number of patients for which each bin representing gene gains or losses shows significant association with tumor stage. The results suggest a net trend towards *LAMP3* and *PRKAA1* gains, with again a significant difference between primary and metastasis. We infer from these results that *LAMP3* has a dominant role both in initiation and development

of different stages of CC.

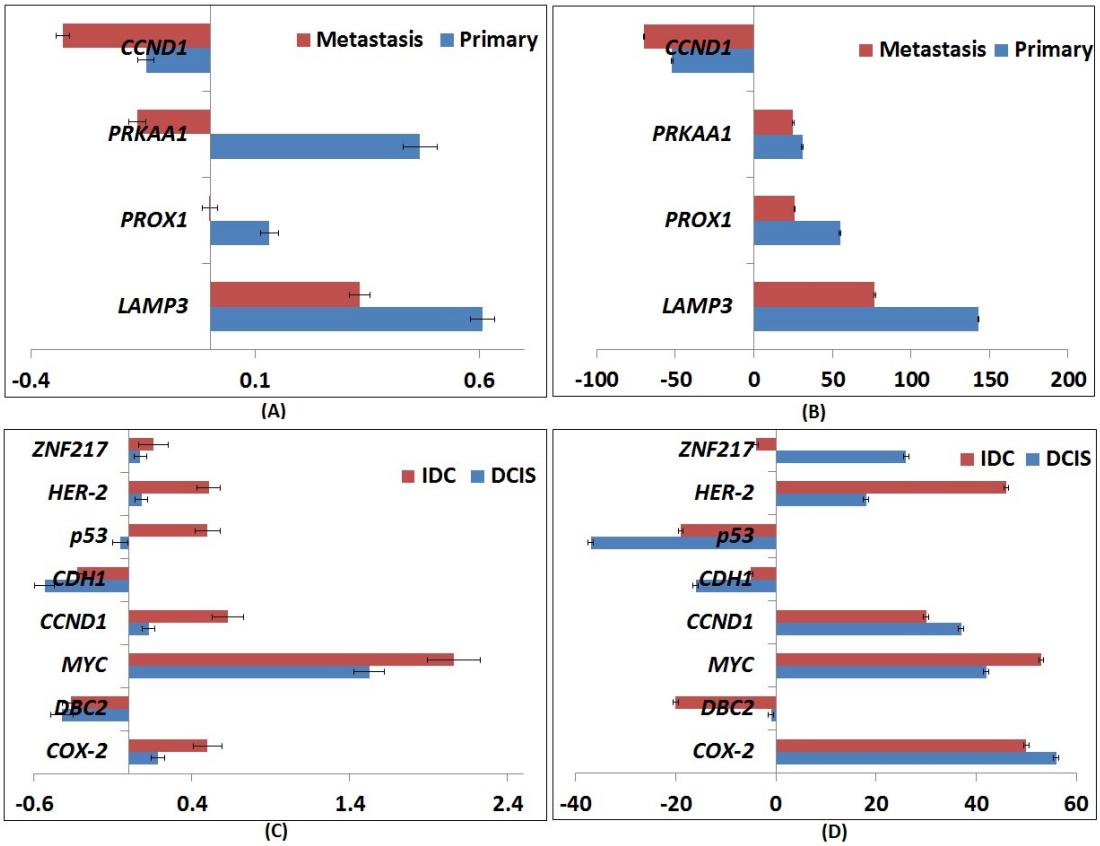

Figure 2.5: Increase and decrease in copy number count of *LAMP3, PROX1, PRKAA1* and *CCND1* (A,B) across 16 CC patients and *COX-2, DBC2, MYC, CCND1, CDH1, p53, HER-2* and *ZNF217* (C,D) genes across 13 BC patients. Copy number count is calculated using (A,C) average of cell count data and (B,D) net tree edge changes. The units on the x-axis differ in the two adjacent subfigures due to the different types of data used.

Figure 2.5 shows the results of cell-count and tree-edge-based analysis of gene gain/loss statistics. Cell counts (Figure 2.5(A)) and edge counts (Figure 2.5(B)) show similar trends in the gain and loss of the marker genes except for two cases. In the first case, cell counts show no net gain or loss of *PROX1* in metastasis, while edge counts show a gain. The latter result is supported by the literature [183] associating gain of *PROX1* with metastasis. Likewise, the two measures suggest opposite trends with respect to *PRKAA1* in metastasis, with cell counts suggesting net loss but edge counts net gain. Again, net gain has been previously associated with progression to metastasis [183]. These results suggest that quantifying progression via evolutionary events,

as enabled by the trees, provides a clearer view of the selective pressure than does quantification by cell counts.

Analysis of Breast Cancer DCIS vs. IDC Samples

We performed a comparable statistical analysis on the paired BC DCIS and IDC samples to understand how the evolutionary process varies between early stages and late stages of tumor development. The BC dataset includes copy number counts for eight gene probes, yielding 16 potential children of the diploid root node representing single copy number gain and loss of individual gene probes. We again treated the subtrees rooted at each of these sixteen children as bins and counted the total number of cells in each bin for each DCIS and IDC tree. We then performed a chi square test using the 16×2 contingency table defined by each DCIS/IDC pair. The results of the chi square tests are presented in Table 2 in Figure 2.3(B). As with the CC data, the table consistently shows significant p-values, which again may indicate differences in the evolutionary processes at different stages of tumor development.

	DCIS												
	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10	Patient 11	Patient 12	Patient 13
Loss of DBC2	14	22	48	103	4	0	2	0	37	0	0	17	12
Loss of CDH1	0	11	3	0	6	128	0	160	4	0	0	105	3
Loss of p53	6	8	1	5	8	0	0	5	1	0	1	2	5
Gain of COX-2	0	0	6	6	0	91	0	8	0	0	2	5	2
Gain of MYC	0	1	0	8	55	0	0	0	1	159	0	0	0
Gain of CCND1	0	1	4	1	0	0	61	0	3	0	167	1	0
Gain of HER-2	10	12	3	0	0	0	0	0	0	1	0	2	25
Gain of ZNF217	129	0	5	2	3	0	0	0	1	0	0	0	0

(A)

	IDC												
	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10	Patient 11	Patient 12	Patient 13
Loss of DBC2	145	160	5	8	2	4	67	0	0	1	159	8	8
Loss of CDH1	0	8	2	111	48	9	3	0	89	0	6	21	8
Loss of p53	14	0	9	2	3	6	0	0	0	0	3	21	5
Gain of COX-2	5	0	6	0	0	7	0	93	0	0	0	48	70
Gain of MYC	0	0	0	0	15	0	0	0	0	0	0	2	0
Gain of CCND1	2	0	1	0	0	98	0	0	0	0	12	14	0
Gain of HER-2	6	4	3	2	0	15	0	0	0	0	7	69	2
Gain of ZNF217	3	0	5	3	2	6	0	0	5	0	0	0	4

(B)

Figure 2.6: Number of cells in the subtrees rooted at the nodes (termed a "bin") directly connected to the root node in the 13 Breast cancer patients. Data is shown for (A) DCIS and (B) IDC stage tumor for bins representing gain of oncogenes and loss of tumor suppressor genes, with the most populated bin highlighted in each case.

We report bin counts for gain of oncogenes and loss of tumor suppressor genes in Figure 2.6.

Examination of individual bin counts shows that the precise biases tend to differ from patient to patient, with the most frequent dominant bins being loss of the two tumor suppressor genes *DBC2* and *CDH1*.

Table 4 in Figure 2.3(D) shows the number of times each of the bins representing gain of oncogenes or loss of tumor suppressor genes shows statistical significance for individual chi square tests on each pair of DCIS and IDC trees. This table again shows bias towards loss of the two tumor suppressor genes *DBC2* and *CDH1*. Loss of *DBC2* and *CDH1* is part of a dominant imbalance clone reported in [83] where it is inferred that cells with this imbalance clone have a growth advantage in DCIS and IDC. Our analysis supports this argument.

We next calculated gain/loss statistics based on raw cell count data and based on tree edges, as we did with the CC data. We present the results in Figures 2.5(C) and 2.5(D) respectively. The trends are qualitatively generally consistent between cell count and tree-based statistics. With two exceptions, oncogenes are amplified and tumor suppressor genes lost in DCIS and IDC by both measures. One exception is the tumor suppressor gene *p53*, which shows amplification rather than the expected loss when analyzed by cell count statistics (Figure 2.5(C)) but not with tree statistics (Figure 2.5(D)). The difference may reflect an occasional amplification of *p53* concurrent with the rest of chromosome 17 due to aneuploidy. The other exception occurs with respect to the oncogene *ZNF217*, which shows net gains by both statistics for DCIS, but loss rather than gain in IDC for tree statistic. The discrepancy appears to be due to one case, in which 90% of cells in IDC show *ZNF217* deletion. This unusual case might be due to the loss of chromosome 20 (on which *ZNF217* resides) in the IDC stage of the tumor for that patient.

2.2.3 Use of tree statistics for classification

A key question in studying mathematical models of tumor progression is whether an understanding of tumor evolutionary pathways will lead to improved prognostic or diagnostic capabilities. We performed classification experiments on the CC dataset to understand how features derived

from progression trees can help differentiate samples from different cancer stages. We used support vector machines (SVM), as implemented in MATLAB, with leave-one-out cross-validation (LOOCV). The performance of each classifier was assessed by percentage of samples correctly classified (Accuracy). We performed 500 rounds of bootstrapping and assessed mean Accuracy as well as their standard deviations.

We performed experiments exploring the predictive power added by the three different types of tree-derived features. The three classification experiments on the CC dataset are:

1. Distinguishing primary from metastatic samples using 16 paired primary and metastatic samples,
2. Distinguishing primary from metastatic samples using 16 metastasizing and 15 non-metastasizing primary tumor samples versus 16 metastatic samples, and
3. Distinguishing 16 primary samples that later metastasized from 15 primary samples that did not metastasize.

Classification of Cervical Cancer Samples

Figure 2.7 reports performance of the two feature sets for the SVM classifier in terms of mean accuracy, along with confidence interval of one standard deviation. Tree-based statistics lead to improved classification accuracy in all experiments. The best accuracy on the three tasks is achieved using the tree-based level-count features, at 81.91% accuracy for distinguishing primary tumors from their paired-metastasis samples, 82.26% for distinguishing all primary tumors(metastasizing and non-metastasizing) from the metastasis samples, and 82.58% for distinguishing metastasizing versus non-metastasizing primary tumors. This result suggests that the qualitative observation that primary trees appear broader and deeper than metastatic trees (Figure 2.1) captures a robust quantitative property of progression trees distinguishing primary from metastatic samples. Among the non-tree based features, Simpson index shows the best classification performance, yielding average accuracies of 76.94%, 78.12% and 61.08% on the same tasks.

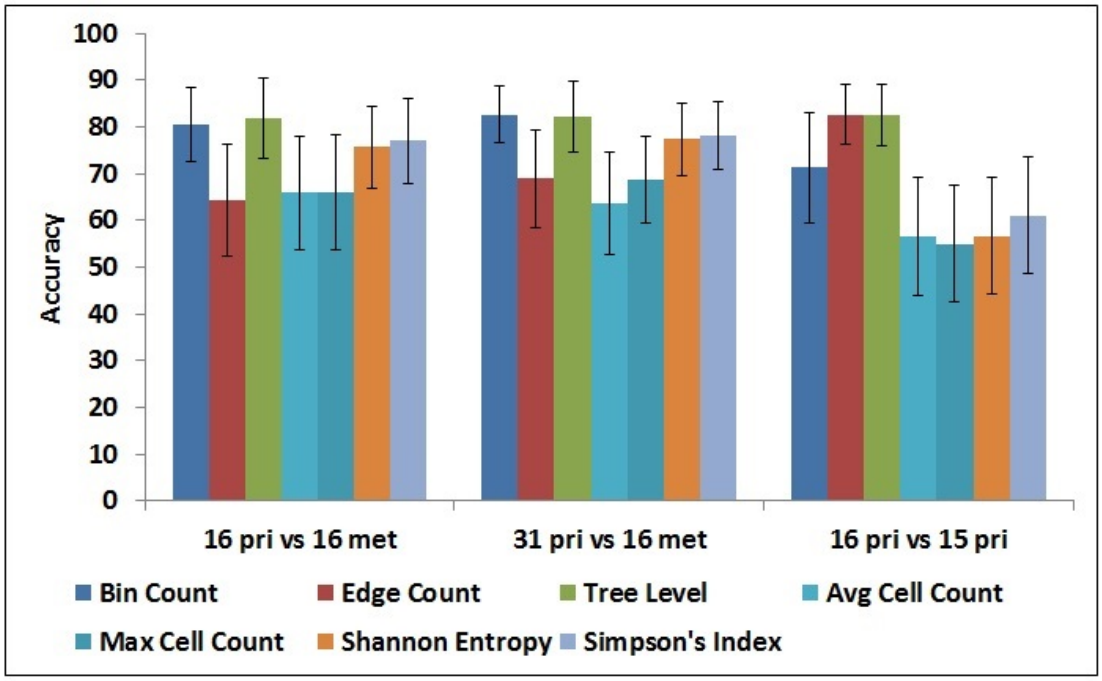

Figure 2.7: Accuracy of tree-based versus cell-based features in classification tasks using an SVM classifier. Each chart shows accuracy of three tree-based and four cell-based feature set on the three defined prediction tasks.

Average and maximum copy number counts show worse performance in all three classification tasks.

To follow up on the observation that tree topology seems to be the most informative feature type, we examined how this feature varies between primary and metastatic trees. Figure 2.8 shows the distribution of the aggregated cell counts across different levels of 31 primary stage and 16 metastatic stage tumor progression trees. In the primary stage tumors, around 70% of the cells are distributed in the first 6 tree levels and the cell count decreases gradually when level of the tree is increased. In contrast, for the progression trees of metastatic stage tumors, the total cell count shows an exponential decrease with more than half (53%) of the cells located in the first two tree levels. This topological difference could reflect the fact that in the primary stage tumors, the clones have more time to continue diversifying. An alternative hypothesis is that the difference reflects stronger purifying selection for clones that must evolve to be able to migrate to and survive outside their native microenvironment. The BC study was designed to

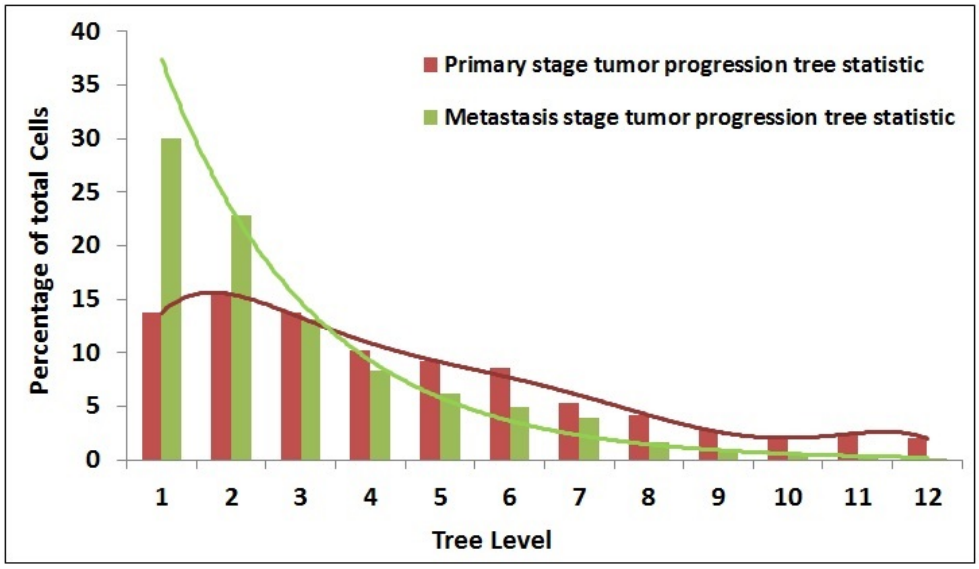

Figure 2.8: Distribution of cells across different levels of tumor progression trees, counted for primary and metastatic trees separately.

include only patients in whom the diagnosis of IDC and DCIS was concurrent [83], which has the effect of making the time of evolution for each sample in a pair comparable, consistent with the hypothesis of increased purifying selection in IDC. The data here, however, are insufficient to reject either hypothesis.

Informative feature selection

		Bin Count			Edge Count			Tree Level Cell Count		
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Cervical Cancer	16 pri vs 16 met	78.13	0.82	0.75	68.75	0.63	0.75	78.13	0.81	0.75
	31 pri vs 16 met	78.72	0.56	0.84	70.00	0.31	0.90	80.85	0.75	0.84
	16 pri vs 15 pri	72.00	0.69	0.80	87.10	0.88	0.87	77.40	0.75	0.80
Breast Cancer	13 DCIS vs 13IDC	80.77	0.85	0.77	76.90	0.85	0.69	80.76	0.77	0.85

(A)

		Bin Count	Edge Count
Cervical Cancer	16 pri vs 16 met	*LAMP3, PRKAA1, PROX1*	*LAMP3, CCND1*
	31 pri vs 16 met	*LAMP3, PRKAA1,PROX1,CCND1*	*PRKAA1*
	16 pri vs 15 pri	*LAMP3, PRKAA1*	*PRKAA1, PROX1, CCND1*
Breast Cancer	13 DCIS vs 13 IDC	*DBC2, MYC, CCND1, HER-2, ZNF217*	*MYC, CCND1, p53, CDH1*

(B)

Figure 2.9: (A) Classification performance for particular subsets of features that show best prediction accuracy among all possible subsets on CC and BC datasets. (B) Sets of gene probes that show best classification accuracy.

We applied feature selection to identify the most informative features. We exhaustively enumerated subsets of features and tested the cross-validated predictive accuracy of each. Figure 2.9 shows, for each classification experiment and feature type, the optimal SVM prediction accuracy over all subsets. For the most challenging task, distinguishing metastasizing from non-metastasizing samples based on the primary sample, accuracy peaks at 72% for Bin Counts, 87.1% for Edge Counts and 77.4% for Tree Level topological features. Interestingly, the best performance over all tests at identifying metastasizing tumors (87.1% accuracy) comes from the Edge Count features, despite poorer performance of Edge Count in most tasks. Among the Bin Count features, *LAMP3* was an informative feature in all three tasks reinforcing our statistical result that *LAMP3* is an important gene in CC progression.

In previous work on the same classification task, Wangsa *et al.* [183] reported sensitivity and specificity of 0.75 and 0.87 respectively, with composite FISH markers using percentages of cells with amplified signals for each individual marker, on the CC dataset, but this was done without LOOCV [183]. The optimal feature set identified here improved substantially on the robustness and sensitivity of that result while keeping equal specificity.

We performed similar classification experiments for informative feature selection on the BC data to distinguish DCIS from IDC samples. When we used all the features for classifying the DCIS samples from IDC, Bin Count and Edge Count measures showed 50% accuracy and Tree Level topological features showed 57% accuracy. Figure 2.9(A) shows that feature selection improved accuracy to 80.7% for both Bin Count and Tree Level feature subsets. The poor performance while using all features might be due to the high intra- or inter-tumor heterogeneity [83].

When we selected the most informative subsets, the feature sets for BC DCIS versus IDC samples classification (Figure 2.9(B)) differed depending on whether the Bin Count or the Edge Count measure was considered, although both agreed on the selection of *MYC*. *MYC* was reported in [83] to be a prognostic marker in the progression of DCIS to IDC. Deletion of *CDH1* was also reported in [83], and was selected here in the Edge Count case.

2.2.4 Simulation Results

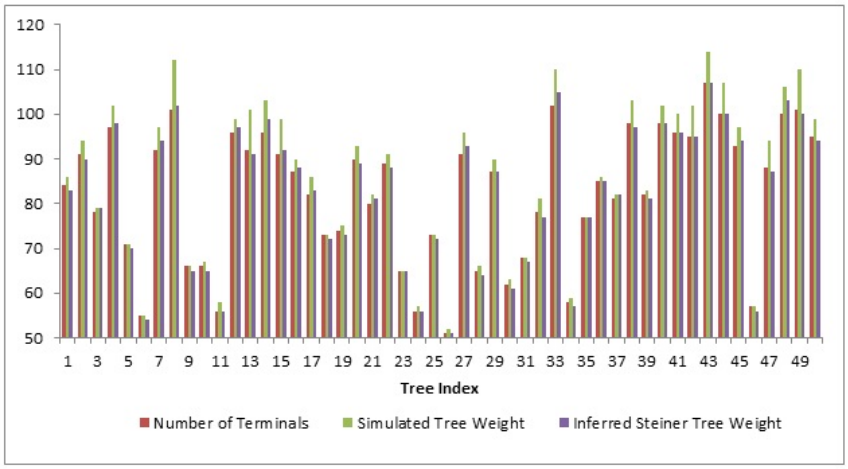

Figure 2.10: Comparison of simulated and inferred Steiner tree weights for fifty simulated trees. For each tree, the total number of terminals, real tree weight and inferred Steiner tree weight are shown.

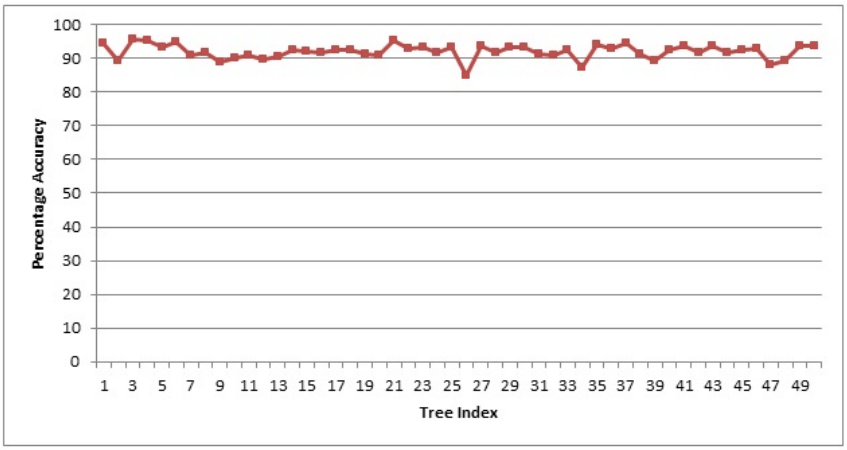

Figure 2.11: Percentage accuracy of the set of bipartitions for each inferred Steiner tree with respect to the bipartitions in the corresponding simulated tree.

Because we cannot know the ground truth for real data with certainty, we used simulated trees to test accuracy of tree inferences. A comparison of the 50 trees is shown Figure 2.10. For each case, Figure 2.10 shows the total number of terminals and the weights of the simulated and inferred tree. We can see from the comparison that the weight of the inferred tree is generally

close to that of the simulated tree but never exceeds it. This observation would suggest that the heuristic is effective in finding parsimonious trees but that the parsimony model leads to a small bias towards underestimating true tree cost.

We show the percentage matching accuracy for each of the 50 test in Figure 2.11. The high accuracy for all the cases indicates that the simulated and inferred trees are topologically similar. The mean and standard deviation of the percentage matching accuracies across all fifty examples are 92% and 2.13% respectively. In order to statistically quantify the accuracies, for each simulation case, we generated 100 random trees by randomly picking edges among the nodes in the terminal set and repeated the analysis procedure above to calculate the percentage matching accuracy. We then calculated the p values for the inferred tree matching accuracy by scoring it on the distribution generated by the random trees. For 80% of the cases, the p value is statistically significant. Some of the discrepancies might result from the fact that our algorithm does not generate any Steiner node with degree 2, as these Steiner nodes do not decrease the weights of the maximum parsimony progression trees.

2.3 Conclusions

We developed exact and heuristic algorithms for building tumor progression trees using copy number information and applied our methods to two different types of cancer. To reduce the complexity of the exact algorithm, we further developed an inequality that can prune up to 99% of the solution space, resulting in substantial reduction in the runtime of the algorithm. The heuristic approach returns potentially sub-optimal solutions in reasonable time for datasets with large numbers of probes. These algorithms have been implemented in a publicly available C++ software package, FISHtrees. Copy number changes can evolve using additional basic operations such as changing the entire ploidy by 1 or doubling and FISHtrees includes an implementation of another method that allows these "operations" [134]. Analyses of statistics developed using different features of the tumor progression trees identify some important recurring markers of

tumor progression and highlight the different selective pressures at work on different stages of the tumor. Use of tree statistics as features for classification further illustrates the importance of models of the evolutionary process to predicting future progression, a problem of importance to cancer treatment and diagnosis. Further improvements in tree algorithms, analysis of even larger and more complex data sets, and investigation of the resulting trees can be expected to yield further insight into both recurring features of tumor evolution and the ways in which these features vary patient-by-patient.

Chapter 3

Algorithms to Model Single Gene, Single Chromosome, and Whole Genome Copy Number Changes Jointly in Tumor Phylogenetics[1]

In this chapter, we develop new methods to advance the theory of phylogenetic inference for reconstructing evolutionary histories of cell populations in solid tumors. The work is specifically designed for use in tracking tumor evolution by gain and loss of genomic regions as assessed by multicolor fluorescence *in situ* hybridization (FISH), which measures the copy numbers of targeted genes and chromosomes in potentially hundreds of individual cells of a tumor. This technology was the basis of the earliest methods for phylogenetic reconstruction of single tumors [132, 134]. FISH remains uniquely valuable for such studies because the large number of cells that FISH can profile makes it possible to collect data on enough tumors in enough de-

[1]This chapter was developed from material published in "Chowdhury *et al.*, Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics, *PLoS Computational Biology*, 10.7 (2014): e1003740" [35].

tail to build cell-by-cell phylogenies for populations of tumors and begin to study the common features of these phylogenies. In the present chapter, we specifically extend our previously developed inference algorithms in Chapter 2 to encompass a more complicated but realistic model of evolution of FISH probe counts, accounting for gain and loss of genetic material at the level of single gene probes, multiple probes on a single chromosome, or a probe set distributed across the whole genome. We demonstrate the value of these algorithmic improvements to more accurate phylogenetic inference and improved effectiveness of the resulting phylogenies in downstream prediction tasks.

The present chapter adds to the growing list of phylogenetic methods in cancer modeling, which were reviewed through 2008 in [8]. These include methods for analyzing comparative genomic hybridization (CGH) or other genetic gain/loss data in a single tumor type [12, 13, 21, 22, 42, 43, 111, 164], for defining the cell type lineage of single tumors [55, 132, 134, 150], for organizing a taxonomy of tumor types [103], for reconstructing a partial order of genetic changes in multiple samples from one patient [100], and for reconstructing progression from cell types inferred from bulk genomic assays [161]. Recent high-throughput sequencing studies have also used ad hoc phylogenetic methods to infer putative tumor progression scenarios, e.g., [30, 87, 124, 169]. Like many of these methods, the present chapter is aimed at building tree models that provide a proposed partial order on the observed cell states, a strategy motivated originally by the work of Fearon and Vogelstein, proposing a linear order for four types of events in colorectal cancer and associating each event with a tumor stage [50]. Other ordering methods have been proposed, mostly for CGH or breakpoint data [16, 63, 70, 86, 100, 120, 127, 147] and, more recently, sequencing data [67, 138].

The present chapter specifically advances the reconstruction of phylogenetic histories of single tumors from intratumor cellular heterogeneity data. The use of phylogenetic methods to reconstruct histories of single tumors was first developed in [132, 134] by taking advantage of the ability of FISH to profile genetic changes in large numbers of single cells, allowing one to survey hundreds of cells per tumor in populations of tens of tumors [92]. This early work showed

that even small numbers of markers could reveal numerous genetically distinct cell populations in single tumors, which could be resolved by phylogenetic inference to reveal multiple distinct pathways of progression between tumors and even within single tumors. Numerous studies since then, using multicolor FISH [83, 84, 92, 109, 134, 153, 165] and, more recently, single-cell sequencing [62, 117, 169, 189] have greatly increased our ability to identify distinct cell populations and, in the process, revealed far more extensive intratumor heterogeneity than had been suspected prior to 2010 (reviewed in [110]). The repeated observation of intratumor heterogeneity has necessitated a reconsideration of Nowell's [126] theory that tumors evolve clonally, showing that a tumor may contain many subpopulations relevant to the clinical prognosis of the patient [45] and that rare subpopulations may be more relevant to prognosis than the most common ones [176]. Furthermore, a simulation study has suggested that methods based on average copy number data perform poorly when there is substantial intratumor heterogeneity [158]. Such findings suggest a need for improved methods for organizing the dozens or hundreds of observed cell states in single tumors to infer the evolutionary processes that produced them.

Despite extensive work on tumor phylogenetics, however, the study of algorithms for reconstructing tumor evolution from large numbers of single cells has lagged far behind advances in data generation. The standard in practice for single-cell tumor phylogenetics remains the use of simple generic phylogeny algorithms (e.g., neighbor-joining [141]) that are not designed to model the patterns of copy number changes one would expect from evolution by chromosome abnormalities that largely drive tumor evolution. Until recently, algorithms designed specifically for inferring phylogenies of single tumors from FISH data have been limited to just a few probes per cell and lacked robust, publicly available software implementations [109, 132, 134]. In Chapter 2, we developed algorithms to find copy-number phylogenies for in principle arbitrary numbers of probes and cells. That work, however, was itself limited to a simple model in which tumor cells evolve by events of gain or loss of a single copy number of a single probe at each mutation step. In real tumors, gene copy numbers can change due to a variety of mechanisms, including:

1. Single gene duplication/loss events (SD), in which one copy of a genetic region covered by a single probe is gained or lost.
2. Chromosome duplication/loss events (CD), in which entire chromosomes are unequally distributed among daughter cells during mitosis along with potentially several probes.
3. Whole genome duplication events (GD), in which a cell fails to divide during mitosis leading to doubling of all genetic material and all probe counts.

These events are illustrated schematically in Figure 3.1. While more complex probabilistic models of tumor evolution have been developed for inference of small phylogenies, with approximately ten taxa per tumor corresponding to distinct biopsies (e.g., [157]), the class of inference algorithms such models require would not be expected to scale to phylogenies of hundreds of single cells per tumor such as those examined in the present work.

The work presented in this chapter seeks to fill this need for scalable phylogenetic algorithms capable of fitting more realistic models of tumor-like evolution to data sets of hundreds of single cells per tumor. We improve on our prior work for inferring tumor evolutionary models considering only SD events [34] to now include CD and GD events, which are also frequently observed in tumor progression. We specifically focus on the problem of accurately inferring evolutionary distances between distinct cells in terms of maximum parsimony combinations of SD, CD, and GD events. The major contributions of this chapter are:

1. algorithms to compute minimum evolutionary distances D between pairs of cell states in terms of SD and CD events and in terms of SD, CD, and GD events;
2. a heuristic Steiner tree method based on the median-joining method [10] and our prior work on SD-only inference [34];
3. software implementation of the new methods to compute D and use of those methods to construct tumor progression trees;
4. evaluation of the new methods on simulated data, which shows that they do better than the SD-only approach at recovering simulated tree topologies;

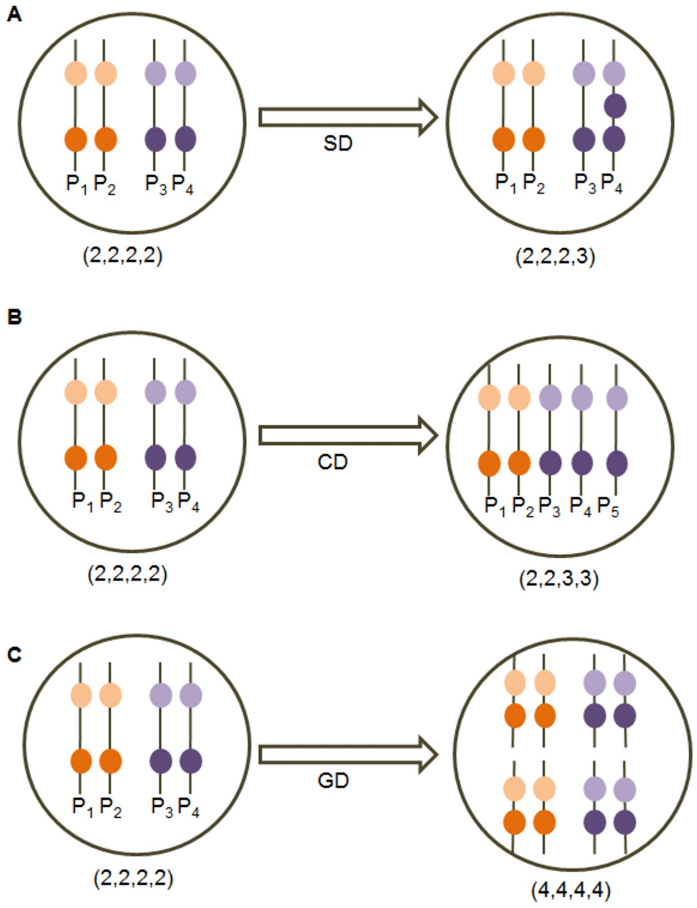

Figure 3.1: Example showing the three mechanisms of copy number changes in a hypothetical cell. A copy number profile of four genes is shown as an ordered set for homologous chromosome pairs P_1, P_2 and P_3, P_4 respectively, where the gene located on the top position in the chromosome precedes the gene located on the bottom position in the ordering. After the (A) Single gene duplication event, the copy number of a gene located on P_4 gets increased by 1. After the (B) Single chromosome duplication event, the chromosome P_4 gets duplicated and the cell has one extra copy of that chromosome as chromosome P_5. After the (C) Whole genome duplication event, all the chromosomes are duplicated and the total number of chromosomes in the daughter cell is twice the number of chromosomes in the mother cell.

5. application of the methods to published data on cervical cancer (CC, [183]) and breast cancer (BC, [83]);

6. demonstration of improved ability to classify tumor types from phylogenetic features using a strategy in the spirit of the genomic progression scores (GPS) of Rahnenführer et al. [139].

The work addresses a critical need in modern cancer research for algorithms capable of inferring evolutionary trajectories of hundreds of single cells per tumor under plausible models of evolution including both gene-specific and chromosome abnormalities that are central drivers of true tumor evolution.

3.1 Methods

Our main theoretical result is a method for inferring minimum distances between two states within a copy number phylogeny when duplication/loss of single genes (SD), duplication/loss of all genes on a common chromosome (CD), and duplication of all genes in the full genome (GD) events are possible. We first establish some mathematical results and then develop an algorithm for accurate distance computation. This algorithm then becomes a subroutine in a heuristic Steiner tree algorithm for inferring copy number phylogenies in the presence of SD, CD, and GD events.

We introduce some notation required for specifying and proving the theoretical results:

1. $C(g_1, g_2, \ldots, g_d)$: A set of copy numbers of one or more genes $g_1, g_2 \ldots, g_d$, which we call a "configuration". When $g_1, g_2 \ldots, g_d$ are clear from the context, we use C as shorthand.
2. $L_1(C^i, C^j)$: L_1 or rectilinear distance between two configurations C^i and C^j.
3. $D^{s,ch}(C^i, C^j)$, $D^{s,g}(C^i, C^j)$, $D^{s,ch,g}(C^i, C^j)$: Distance between two configurations C^i and C^j when considering SD+CD (s,ch), SD+GD (s,g), or SD+CD+GD (s,ch,g) events, respectively.
4. $O^c_g(C^i)$, $O^c_l(C^i)$, $O^c(C^i)$: Operations corresponding to single chromosome (CD) events corresponding to either gain (g), loss (l), or either (no subscript) of all genes belonging to the same chromosome c from starting configuration C^i, while keeping the copy numbers of genes on other chromosomes unchanged.
5. $O^g(C^i)$, $H(C^i)$: Operations corresponding to doubling (O^g) or halving (H) counts of all

genes in configuration C^i. In the case of halving, it is assumed that all genes in C^i have even counts.

6. *even*, *odd* configuration: A configuration (copy number profile) $C(g_1, g_2, \ldots, g_d)$ is denoted an *even* configuration if $\forall g_i \mod (g_i, 2) = 0$. Otherwise, it is denoted an *odd* configuration.

7. $G^E(C(g_1, g_2, \ldots, g_d))$: The set of "nearest even" values for each g_i in C, i.e., if $C(g_1, g_2, \ldots, g_d) = (x_1, \ldots, x_d)$ then $G^E(C(g_1, g_2, \ldots, g_d)) = \{(y_1, \ldots, y_d) | ((y_i \mod 2) = 0) \wedge ((y_i = x_i) \vee (y_i = x_i \pm 1) \vee (y_i = x_i \pm 2))\}$. For example, $G^E((7, 2)) = \{(6, 2), (8, 2), (6, 0), (8, 0), (6, 4), (8, 4)\}$.

8. An operation F is *valid* on a configuration $C(g_1, g_2, \ldots, g_d)$ if $(x_1, x_2, \ldots, x_d) = F(C(g_1, g_2, \ldots, g_d))$ satisfies $LB \leq x_i \leq UB$ for all $i = 1, \ldots, d$ given predefined lower-bound LB and upper-bound UB. Otherwise, F is *invalid* on C. LB=0 and UB=9 is used in the software, but the theory only requires that UB $>$ LB.

9. A sequence of operations $F_1, \ldots, F_k$ is *boundary-sensitive* on configuration C if $(x_{j1}, x_{j2}, \ldots, x_{jd}) = F_j(F_{j-1}(\ldots F_1(C(g_1, g_2, \ldots, g_d))))$ satisfies $LB \leq x_{ji} \leq UB$ for all $i = 1, \ldots, d$ and $j = 1, \ldots, k$. We use *boundary-insensitive* to refer to a sequence on which this condition has not been checked.

3.1.1 Progression model considering SD and CD events

We develop the theory for inference of the Steiner (unsampled or extinct cell configurations) nodes in the paths formed by the sequence of gene copy number gains and losses from an initial configuration $C^s(g_1, g_2, \ldots, g_d)$ to a final configuration $C^t(g_1, g_2, \ldots, g_d)$. We first extend the prior theory to account for SD and CD events. Our model assumes that on division of a tumor cell, the configuration can change either by gain or loss of one copy of a single gene (SD event) or by gain or loss of one copy of each gene on a single chromosome (CD event). For example, a configuration of four genes $(2, 2, 2, 2)$ with the first two genes on the same chromosome might evolve in a single mutational event to $(3, 2, 2, 2)$ by an SD event or to $(3, 3, 2, 2)$

by a CD event. We propose Algorithm 3, to calculate the minimum number of steps required to transform $C^s(g_i, g_{i+1}, \ldots, g_j)$ into $C^t(g_i, g_{i+1}, \ldots, g_j)$ considering SD and CD events, where, without loss of generality, we assume that the genes on a common chromosome have consecutive indices $(g_i, g_{i+1}, \ldots, g_j)$ in C. Algorithm 3 also identifies a minimum-length sequence of events, although this sequence is not necessarily unique. For example, if there are four genes on one chromosome and we want to get from configuration $(1, 1, 1, 1)$ to configuration $(2, 4, 3, 2)$, then a shortest sequence of SD and CD events would be CD to $(2, 2, 2, 2)$, SD to $(2, 3, 2, 2)$, SD to $(2, 4, 2, 2)$, and SD to $(2, 4, 3, 2)$. Other orders of the same four events are also possible.

The above example focuses on a single chromosome because as explained below, the problem of finding the shortest SD+CD path can be solved one chromosome at a time. We begin by establishing the following lemmas:

Lemma 5. *A minimum-length boundary-insensitive sequence of CD and SD events cannot have both a gain of chromosome* c_i *and a loss of the same chromosome* c_i*.*

Proof. By contradiction. Suppose S is a sequence of events that has both a gain and a loss of the same chromosome. Then removing one gain and one loss produces a new sequence that is 2 shorter and has the same final state. □

Lemma 6. *For any gene* g_i*, a minimum-length boundary-insensitive sequence of events cannot have both a gain of* g_i *and a loss of* g_i*.*

Proof. By contradiction. Suppose S is a sequence of events that has both a gain of G and a loss of G. Then removing one gain and one loss produces a new sequence that is 2 shorter and has the same final state. □

Lemma 7. *The following sequence of events describes a minimum-length boundary-insensitive sequence of SD and CD events for transforming* $C^s(g_i, g_{i+1}, \ldots, g_j)$ *into* $C^t(g_i, g_{i+1}, \ldots, g_j)$*:*

1. *Perform CD events in arbitrary order starting from* C^s *so that each successive event decreases the* L_1 *distance between the intermediate configurations* $C^{int}(g_i, g_{i+1}, \ldots, g_j)$ *and*

$C^t(g_i, g_{i+1}, \ldots, g_j)$ *until any further CD event will increase the* L_1 *distance. We define the final configuration reached after this step to be* $C^f(g_i, g_{i+1}, \ldots, g_j)$.

2. *Perform SD events in abitrary order starting at* $C^f(g_i, g_{i+1}, \ldots, g_j)$ *so that the* L_1 *distance between* $C^{int}(g_i, g_{i+1}, \ldots, g_j)$ *and* $C^t(g_i, g_{i+1}, \ldots, g_j)$ *decreases on each step until the distance becomes zero. The total number of events required will be* $L_1(C^f, C^t)$.

Proof. Since the sequence of events is boundary-insensitive and addition is commutative, we can change the order of events without changing the endpoints or the cost. Therefore, we assume that all CD events precede all SD events. The construction of the above sequence of the events ensures that it uses a maximum number of possible CD events. If we denote the number of genes on the common chromosome by k and the number of CD events by c, then the total number of events required is $L_1(C^s, C^t) - (k-1)c$. If there exists a shorter sequence of events to transform C^s to C^t, then that sequence must have a larger number c of CD events, which is contradicted by the construction. Thus, the number of events is minimized. □

The above lemmas show how to construct a minimum-length boundary-insensitive sequence of events. We now establish that this sequence can be used to derive a minimum-length boundary-sensitive sequence of events:

Lemma 8. *For any boundary-insensitive minimum-length sequence of SD and CD events* S *transforming* C^s *to* C^t*, there exists a boundary-sensitive sequence of SD and CD events* S' *such that* S *and* S' *have equal length.*

Proof. We analyze one chromosome at a time because in this section the events on different chromosomes are independent. By Lemma 5, on any specific chromosome all the CD events are gains or all the CD events are losses. We analyze in detail the case in which all CD events are losses; the case of all gains is symmetric.

The proof is constructive. Specifically, we will show that the upper part of Algorithm 3 will transform a boundary-insensitive S to a boundary-sensitive S' of equal cost solely by reordering

events. Without loss of generality, suppose the only CD events in S are chromosome losses. There is a symmetric algorithm, shown as the lower part of Algorithm 3, for the case where all the chromosome events are gains. We add the following definition:

A gene G is defined as unidirectional with respect to S if there are no gains of G in S. A gene G is defined as bidirectional with respect to S if S includes gains of G. For unidirectional genes, the order of chromosome losses and gene losses can never cause a boundary to be crossed because the copy numbers are monotonically decreasing. The situations we need to avoid are:

1. A bidirectional gene G has copy number UB and the next operation affecting G is a gain of G.
2. A bidirectional gene G has copy number LB and the next operation affecting G is a chromosome loss.

Chromosome gains are excluded by Lemma 5 and our assumption without loss of generality that all CD events are losses. Gene losses for bidirectional genes are exluded by Lemma 6.

To prove correctness of the algorithm, we note that S' can never cross LB for the unidirectional genes because their net loss equals their total loss. S' can never cross LB for the bidirectional genes, because when their copy number is at LB, a gene gain must still be pending and the gene gains alternate in the first while loop until no chromosome losses or gene gains are remaining. S' can never cross UB for the unidirectional genes because they have only losses. S' can never cross UB for the bidirectional genes because of the test $N^{g_i} < UB$ (line 8) before any gene gain is done. Further, all the chromosome losses will be used because one chromosome loss happens on each pass through the first while loop, if any chromosome losses remain. All gene gains in S will be used in the first while loop because the net change for any gene must keep its copy number below UB. All the gene losses for the unidirectional genes are used in the second while loop. The unordered set of events and total change in each gene is thus preserved between S' and S, while S' guarantees that the sequence is boundary-sensitive. □

We use the preceding result to derive the main theorem of this section, which estabishes a

Algorithm 3 Converts a set of boundary-insensitive events to boundary-sensitive events; lines 3-17 are used for chromosomes on which all CD events are losses and lines 18-32 are used for chromosomes on which all CD events are gains

Require: $S \leftarrow$ Boundary-insensitive list of events treated here as a multi-set and processed one chromosome at a time.

Ensure: $S' \leftarrow$ Boundary-sensitive list of events that when viewed as a multi-set is identical to S.

1: $Gain(g_i) \leftarrow$ Single gene gain event on gene g_i.
2: $Loss(g_i) \leftarrow$ Single gene loss event on gene g_i.
3: $C^L \leftarrow$ Number of chromosome loss events in S not yet done. ▷ beginning of the part assuming all CD events are losses
4: $CLoss \leftarrow$ Chromosome loss event.
5: $N^{g_i} \leftarrow$ Copy number of gene g_i.
6: $\forall$ bidirectional genes g_i, $G^{g_i} \leftarrow$ Number of gene gains of g_i in S not yet done.
7: **while** $((C^L > 0)||(\exists$ bidirectional gene $g_i : G^{g_i} > 0))$ **do**
8: **for** $(g_i : G^{g_i} > 0$ & $N^{g_i} < UB)$ **do**
9: $S' \leftarrow S' +\!\!+ Gain(g_i)$ ▷ $+\!\!+$ denotes the concatenation operator
10: $G^{g_i} \leftarrow G^{g_i} - 1$
11: **if** $C^L > 1$ **then**
12: $S' \leftarrow S' +\!\!+ CLoss$
13: $C^L \leftarrow C^L - 1$
14: $\forall$ unidirectional genes g_i, $L^{g_i} \leftarrow$ Number of gene losses of g_i remaining.
15: **while** ($\exists$ unidirectional genes $g_i : L^{g_i} > 0$) **do**
16: $S' \leftarrow S' +\!\!+ Loss(g_i)$
17: $L^{g_i} \leftarrow L^{g_i} - 1$ ▷ end of the part assuming all CD events are losses
18: $C^G \leftarrow$ Number of chromosome gain events in S not yet done. ▷ beginning of the part assuming all CD events are gains
19: $CGain \leftarrow$ Chromosome gain event.
20: $N^{g_i} \leftarrow$ Copy number of gene g_i.
21: $\forall$ bidirectional genes g_i, $L^{g_i} \leftarrow$ Number of gene losses of g_i in S not yet done.
22: **while** $((C^G > 0)||(\exists$ bidirectional gene $g_i : L^{g_i} > 0))$ **do**
23: **for** $(g_i : L^{g_i} > 0$ & $N^{g_i} > LB)$ **do**
24: $S' \leftarrow S' +\!\!+ Loss(g_i)$
25: $L^{g_i} \leftarrow L^{g_i} - 1$
26: **if** $C^G > 1$ **then**
27: $S' \leftarrow S' +\!\!+ CGain$
28: $C^G \leftarrow C^G - 1$
29: $\forall$ unidirectional genes g_i, $G^{g_i} \leftarrow$ Number of gene gains of g_i remaining.
30: **while** ($\exists$ unidirectional genes $g_i : G^{g_i} > 0$) **do**
31: $S' \leftarrow S' +\!\!+ Gain(g_i)$
32: $G^{g_i} \leftarrow G^{g_i} - 1$ ▷ end of the part assuming all CD events are gains

method to find a minimum-length sequence of SD and CD events transforming C^s to C^t. As in the proof of Lemma 8, we can consider each chromosome separately since each SD and CD event affects only one chromosome.

Theorem 9. *Assume we partition the gene list by chromosomes such that each chromosome* $c_i \in \{c_1, \ldots, c_q\}$ *corresponds to a consecutive subset of genes* $g_{i,1}, \ldots, g_{i,d_i}$. *Further define* $C^s(g_1, g_2, \ldots, g_d) = (s_1, \ldots, s_d)$ *and* $C^t(g_1, g_2, \ldots, g_d) = (t_1, \ldots, t_d)$. *Then we can construct a minimum-length boundary-sensitive sequence of events transforming* $C^s(g_1, g_2, \ldots, g_d)$ *to* $C^t(g_1, g_2, \ldots, g_d)$ *by constructing a minimum-length boundary-sensitive sequence of events* S_i *transforming* $(s_1, \ldots, s_{i,1}, \ldots, s_{i,d_i}, \ldots, s_d)$ *to* $(s_1, \ldots, t_{i,1}, \ldots, t_{i,d_i}, \ldots, s_d)$ *for each chromosome* c_i *and interleaving each* S_i *in arbitrary order.*

Proof. The distance function can be decomposed into individual parts for genes belonging to distinct chromosomes as follows:

$$D^{s,ch}(C^s, C^t) = \sum_{i=1}^{q} D^{s,ch}(C^s(s_{i,1}, \ldots, s_{i,d_i}), C^t(s_{i,1}, \ldots, s_{i,d_i}))$$

Because the distance cost can be decomposed in this way and each CD or SD event contributes to only a single term of the outer sum, we can minimize the cost of events for each chromosome independently and combine the events from distinct chromosomes in arbitrary order without changing the value of the objective function. Likewise, since these each chromosome affects a disjoint subset of genes, boundary-sensitive sequences for each chromosome will yield a boundary-sensitive sequence across all genes. □

3.1.2 Progression model combining SD, CD and GD events

We now extend the theory from the prior section to include SD, CD, and GD events. We assume in the proofs and discussion below that $C^s \prec C^t$, where $\prec$ denotes lexicographical ordering. This assumption reduces the number of cases in several proofs. If instead, $C^t \prec C^s$, the proofs

are identical or symmetric except that GD events may be used in the wrong direction (halving instead of doubling). The use of halving events is corrected heuristically by a procedure of subtree pruning and regrafting at line 24 of the pseudocode of Algorithm 6, described below, and in FISHtrees. We will produce the complete proof by deriving a series of lemmas for three cases that together will cover all possible C^s and C^t:

Lemma 10. *For an an even configuration C^t, if there exists an optimal sequence of copy number change events from C^s to C^t composed of one or more SD and CD events and a single GD event, then the following sequence of events is of minimum length:*

1. *SD and CD events to transform C^s into $H(C^t)$, constructed as described in the first named subsection of Methods*
2. *A single GD event to transform $H(C^t)$ into C^t.*

Proof. We prove the statement by considering the three different ways that can be used to transform C^s to C^t using single GD and multiple SD and CD events. The statement of the lemma presents one case and the remaining two possibilities are as follows:

1. A single GD event to transform C^s into $O^g(C^s)$ and then multiple SD and CD events to transform $O^g(C^s)$ into C^t.
2. Multiple SD and CD events to transform C^s to an intermediate configuration C^i, a single GD event to transform C^i into C^j, and multiple SD and CD events to transform C^j into C^t.

We show that for either of these alternative cases, we can produce a sequence satisfying the conditions of the lemma with equal or smaller length. For the first case, we have to show that

$$D^{s,ch}(C^s, H(C^t)) < D^{s,ch}(C^t, O^g(C^s))$$

It can be seen that

$$L_1(C^s, H(C^t)) = \tfrac{1}{2} L_1(C^t, O^g(C^s))$$

If all genes are located on distinct chromosomes, then,

$$D^{s,ch}(C^s, H(C^t)) = \tfrac{1}{2} D^{s,ch}(C^t, O^g(C^s))$$

and the claim follows directly.

Now, assume the genes are partitioned into sets of chromosomes such that each chromosome $c_i \in \{c_1, \ldots, c_q\}$ corresponds to a consecutive subset of genes $g_{i,1}, \ldots, g_{i,d_i}$. We focus on a specific chromosome c_i and consider the problem of updating just genes of that chromosome from their values in $O^g(C^s)$ to their values in C^t. Either zero or a positive even number of CD events must be performed to convert these genes from $O^g(C^s))$ to C^t and along with zero or a positive even number of SD operations on each gene. If an odd number of CD operations are performed on $O^g(C^s))$, then we get an odd configuration and at least one or an odd number of SD operations must be performed on each gene of this odd configuration to convert it to the even configuration C^t. But a combination of single SD operations acting on each of the individual genes in $g_{i,1}, \ldots, g_{i,d_i}$ has the same effect as a single CD operation on chromosome c_i and this combination therefore cannot be minimal. Therefore, the number of CD operations is even. If a total of m CD operations and n SD operations are needed to convert C^i to C^j, then a total of $\frac{1}{2}m$ CD operations and $\frac{1}{2}n$ SD operations are needed to convert $\frac{1}{2}C^i$ to $\frac{1}{2}C^j$. So,

$$D^{s,ch}(C^s, H(C^t)) < D^{s,ch}(C^t, O^g(C^s))$$

For alternative 2, we can write the distance function as:

$$D_1^{s,ch,g}(C^s, C^t) = D^{s,ch}(C^s, C^i) + 1 + D^{s,ch}(O^g(C^i), C^t)$$

The distance function for our proposed optimal sequence can be written as:

$$D_2^{s,ch,g}(C^s, C^t) = D^{s,ch}(C^s, C^i) + D^{s,ch}(C^i, H(C^t)) + 1$$

As shown for alternative 1, we can write:

$$D^{s,ch}(O^g(C^i), C^t) > D^{s,ch}(C^i, H(C^t))$$

which implies $D_1^{s,ch,g}(C^s, C^t) > D_2^{s,ch,g}(C^s, C^t)$. □

Lemma 11. *For an odd configuration C^t, if the optimal sequence of copy number change events from C^s to C^t is composed of one or more SD and CD events, followed by a single GD event, followed by one or more SD and CD events, then the configuration from which the final set of SD and CD events take place is a member of $G^E(C^t)$.*

Proof. We denote the intermediate configuration following the GD event to be C^{int}. We will show by contradiction that if there exists any optimal sequence of events for which $C^{int} \notin G^E(C^t)$ then there must exist an alternative, shorter sequence of events. Define the full sequence of events from C^s to C^t to be $\vec{p}$, subdivided into the subsequences $\vec{p}_1, \{GD\}, \vec{p}_2$. First, we note that if there is any duplicated event in $\vec{p}_2$ then we can construct a more parsimonious solution by replacing the duplicate in $\vec{p}_2$ with a single copy of the event in $\vec{p}_1$. Therefore, no event appears more than once in $\vec{p}_2$. There are exactly two SD and CD events that can increase the count of any given probe (SD of that probe or CD of its chromosome) and similarly exactly two events that can decrease the count of any probe. Thus, no probe's value changes by more than ± 2 in the transition from C^{int} to C^t in $\vec{p}_2$. Finally, we note that since C^{int} immediately follows a GD event, it must be an even configuration. Together, these assertions establish that $C^{int} \in G^E(C^t)$ for any optimal path $\vec{p}$. □

Lemma 12. *For an odd configuration C^t, if the optimal sequence of copy number change events from C^s to C^t is composed of one or more SD and CD events and a single GD event, then the optimum sequence of events follows the following path:*

1. *Generate $C^{int} = G^E(C^t)$.*
2. *SD and CD events to transform C^s into $H(C^{int})$.*
3. *A single GD event to transform $H(C^{int})$ into C^{int}.*
4. *SD and CD events to transform C^{int} into C^t.*

The optimal sequence is an element of the set of sequences generated using this procedure.

Proof. The proof follows from application of Lemma 10 and Lemma 11. As C^t is an odd configuration, the final step cannot be a GD event. So, the last steps have to be a combination of SD and/or CD events; in that case, Lemma 11 shows that the configuration reached as a result of GD must be a member of $G^E(C^t)$, which we denote by C^{int}. Lemma 10 shows that to reach any member of $G^E(C^t)$, which are even configurations, the optimal sequence of events is to generate SD and CD events to transform C^s into $H(C^{int})$ first and then to perform a GD event to transform $H(C^{int})$ into C^{int}. This sequence of events matches the sequence proposed in the lemma. □

The above lemmas allow us to derive Algorithm 4 to transform C^s to C^t using a minimum-length combination of SD, CD and GD events. To illustrate the algorithm, suppose $C^s = (3, 1)$ and $C^t = (7, 5)$, where we will assume we have two probes on a single chromosome. Since C^t is an odd configuration, we first generate its nearest even neighbors $G^E(C^t) = ((6, 4), (6, 6), (8.4), (8, 6))$ and calculate $H(G^E(C^t)) = ((3, 2), (3, 3), (4, 2), (4, 3))$. The algorithm tests for two stopping conditions by which a solution can be constructed (lines 22 and 24 in Algorithm 4), neither of which applies to any of the solutions at this point. $((3, 2), (3, 3), (4, 2), (4, 3))$ are therefore considered for the next iteration. $(3, 2)$, $(3, 3)$, and $(4, 3)$ are odd configurations, so we generate their neighbor sets $G^E((3, 2)) = \{(2, 2), (4, 2), (2, 0), (4, 0), (2, 4), (4, 4)\}$, $G^E((3, 3)) = \{(2, 2), (4, 2), (2, 4), (4, 4)\}$, and $G^E((4, 3)) = \{(2, 2), (2, 4), (4, 2), (4, 4), (6, 2), (6, 4)\}$. One stopping condition is satisfied for each of the elements of these neighbor sets, so $(3, 2)$,$(3, 3)$, and $(4, 3)$ are each considered in turn as the next candidate neighbor. $(4, 2)$ is an even configuration, so we only need to consider one possible stopping condition (line 11), which it satisfies, so it is also considered as a possible next candidate neighbor. Among the four possibilities, we will conclude that using $(3, 2)$ as the immediate neighbor will lead to the smallest possible number of steps when accumulating SD+CD events from C^s to the candidate, a single GD event from the candidate to its double, and SD+CD events from that double to C^t. Following some postprocessing updates (procedure CHECKSRCNEIGHBOR), the algorithm computes a minimum-length

solution of $(3,1) => (3,2) => (6,4) => (7,5)$ and returns the corresponding length 3.

Algorithm 4 satisfies the following theorem, which constitutes the major result of this section:

Theorem 13. *Algorithm 4 returns the minimum distance between two configurations C^s and C^t, where $C^s \prec C^t$.*

Proof. We use induction on the minimum number of steps to get from C^s to C^t, which we denote by $M(C^s, C^t)$.

Base Case: For the base case, we have $M(C^s, C^t) = 1$. We must consider two sub-cases: (i) $C^t = 2C^s$ and (ii) $M(C^s, C^t) = 1$. For case (i), C^t is an even configuration. The condition at line 11 in Algorithm 4 fails and $\frac{1}{2}C^t = C^s$ is considered for the next iteration. In the next iteration, if C^s is an even configuration then the condition at line 11 is now satisfied and $M(C^s, C^t)$ is assigned the value 1 in CHECKSRCNEIGHBOR procedure called at line 12 in the main procedure. If C^s is an odd configuration, then the condition at line 22 is satisfied for each of the even neighbors of C^s and $M(C^s, C^t)$ is assigned the value 1 in the CHECKSRCNEIGHBOR procedure called at line 23. For case (ii), one of the conditions at line 11 or line 22 is satisified in the first iteration of the algorithm depending on whether C^t is an even or odd configuration and $M(C^s, C^t)$ is assigned the value $D^{s,ch}(C^s; C^t) = 1$ at line 12 or 23.

Induction Step: For the induction hypothesis, we assume that the the algorithm uses the minimum number of steps for all cases where $M(C^s, C^t) \leq m$. Then, suppose that an adversary selects an example that has complexity $M(C^s, C^t) = m + 1$. Let us assume that the penultimate configuration in the optimal solution is C^{int}. If C^t is an even configuration, then it can be reached from C^{int} by using (i) a GD event, (ii) an SD event, or (iii) a CD event. According to the induction hypothesis, for each of these cases, Algorithm 4 uses the minimum number of m steps to generate C^{int} from C^s. If there is at least one GD event in the optimal solution, then Algorithm 4 first calculates $C^{int} = \frac{1}{2}C^t$. The induction hypothesis ensures that $M(C^s, C^t) \leq m$ and thus, Algorithm 4 returns a solution with a maximum length of $m + 1$. If there is no GD event in the optimal solution from C^s to C^t, then Algorithm 4 uses the procedure described in the

Algorithm 4 Algorithm for finding the shortest directed distance between two configurations using SD, CD, and GD events.

procedure MINIMUMDISTANCE(C^s, C^t) ▷ Minimum distance between configurations C^s and C^t
Require: C^s, C^t
Ensure: $D^{s,ch,g}(C^s, C^t)$, a minimum-distance sequence of events, is stored via the parent function
    $prev \leftarrow \{C^t\}$ ▷ $prev$ stores the configurations to be considered in the next iteration
    $dist(C^s) \leftarrow \infty$ ▷ $dist(C^i)$ stores the optimal distance between C^i and C^t calculated so far
    $D^{s,ch}(C^s, C^t) \leftarrow$ Length of the optimal path between C^s and C^t constructed using the procedure defined by Theorem 9.
    while $true$ **do**
        $nextStates \leftarrow \emptyset$
        for $i \leftarrow 1, |prev|$ **do**
            $prevConf \leftarrow prev(i)$
            if $prevConf$ is an even configuration **then**
                $prevHalf \leftarrow \frac{1}{2}(prevConf)$
                if $D^{s,ch}(C^s, prevConf) \leq (D^{s,ch}(C^s, prevHalf) + 1)$ **then**
                    CHECKSRCNEIGHBOR($C^s, prevConf, srcNeighbor, dist$)
                else
                    $nextStates \leftarrow nextStates \cup prevHalf$
                    $dist(prevHalf) \leftarrow dist(prevConf) + 1$
                    $parent(prevHalf) \leftarrow prevConf$
            else
                $prevNeighborSet \leftarrow G^E(prevConf)$
                for $j \leftarrow 1, |prevNeighborSet|$ **do**
                    $prevNeighbor \leftarrow prevNeighborSet(j)$
                    $prevNeighborHalf \leftarrow \frac{1}{2}(prevNeighbor)$
                    if $D^{s,ch}(C^s, prevConf) \leq (D^{s,ch}(prevConf, prevNeighbor) + D^{s,ch}(C^s, prevNeighborHalf) + 1)$ **then**
                        CHECKSRCNEIGHBOR($C^s, prevConf, srcNeighbor, dist$)
                    else if $D^{s,ch}(C^s, prevNeighbor) \leq (D^{s,ch}(C^s, prevNeighborHalf) + 1)$ **then**
                        $dist(prevNeighbor) \leftarrow dist(prevConf) + D^{s,ch}(prevConf, prevNeighbor))$
                        $parent(prevNeighbor) \leftarrow prevConf$
                        CHECKSRCNEIGHBOR($C^s, prevNeighbor, srcNeighbor, dist$)
                    else
                        $nextStates \leftarrow nextStates \cup prevNeighbor$
                        $parent(prevNeighbor) \leftarrow prevConf$
                        $dist(prevNeighbor) \leftarrow dist(prevConf) + D^{s,ch}(prevNeighbor, prevConf)$
                        $nextStates \leftarrow nextStates \cup prevNeighborHalf$
                        $parent(prevNeighborHalf) \leftarrow prevNeighbor$
                        $dist(prevNeighborHalf) \leftarrow dist(prevNeighbor) + 1$
        if $nextStates == \emptyset$ **then**
            $break$
        $prev \leftarrow nextStates$
    $return D^{s,ch,g}(C^s, C^t) \leftarrow dist(C^s)$
end procedure

Algorithm Algorithm for finding the shortest directed distance between two configurations using SD, CD, and GD events (continued)

procedure CHECKSRCNEIGHBOR($C^s, prevConf, srcNeighbor, dist$) ▷ Checks if $prevConf$ can become the new candidate neighbor of C^s
    $testDistance \leftarrow dist(prevConf) + D^{s,ch}(C^s, prevConf)$
    if $dist(C^s) > testDistance$ **then**
        $srcNeighbor \leftarrow prevConf$
        $dist(C^s) \leftarrow testDistance$
end procedure

first named subsection of Methods to calculate the optimal path from C^{int} to C^t and combining it with the optimal solution from C^s to C^{int}, it returns the optimal path between C^s and C^t. Now, if C^t is an odd configuration, then going from the penultimate configuration C^{int} to C^t can only be achieved using either an SD or a CD event. For odd C^t, Algorithm 4 first generates its even neighbors C^N which are steps ≥ 1 from C^t. If $C^{int} \in C^N$, the proof follows directly from the inductive hypothesis. If $C^{int} \notin C^N$, then there is a $C^n \in C^N$ such that C^{int} is located on the optimal path between C^n and C^t formed using SD and CD events only. If k is the total number of genes with odd copy number values in C^t, then $D^{s,ch}(C^n, C^t) = k$ and $D^{s,ch}(C^n, C^{int}) = k - 1$. Using the induction hypothesis, we can write,

$$M(C^s, C^n) \leq m - k + 1$$

As Algorithm 4 uses the procedure described in the first named subsection of Methods to construct the optimal path between C^n and C^t, we can see that it returns a path with $M(C^s, C^t) \leq m + 1$. □

3.1.3 Runtime analysis of Algorithm 4

We provide an upper bound on the runtime of Algorithm 4 as a function of the number of genes d and their copy numbers. Considering all three events, where $C^s \prec C^t$, the maximum number of doublings required is $\left\lceil \log_2 \left(\frac{C^t(g_i)}{C^s(g_i)}\right) \right\rceil$, where g_i denotes the copy number of the first gene where $C^s(g_i) < C^t(g_i)$ and $C^s(g_i) > 0$. At each stage of the algorithm, the maximum number of

nodes generated as a result of a G^E operation is 3^d. d SD and CD events are used to create each of those 3^d nodes in the case of an odd configuration. So, the maximum number of required L_1 operations is $\left\lceil \log_2 \left(\frac{C^t(g_i)}{C^s(g_i)}\right) \right\rceil d3^d$. Therefore, the number of operations performed during the execution of Algorithm 4 is $\mathcal{O}\left(\left\lceil \log_2 \left(\frac{C^t(g_i)}{C^s(g_i)}\right) \right\rceil d3^d\right)$.

3.1.4 Generating tumor phylogenies

We implemented Algorithm 4 and integrated it with our approximate median-joining-based algorithm from our prior SD-only FISHtrees [34] code. The key steps of this algorithm are summarized in Algorithm 6, which we describe at a high level here. The phylogeny algorithm first relies on Algorithm 4 to derive a matrix of pairwise distances between observed cell configurations, which are treated as states on a truncated integer lattice of dimension d with a maximum value (UB) set to 9 in the current code. It then repeatedly samples triplets of nodes, identifying as potential Steiner nodes those that agree in each dimension with at least one of the triplet. Those Steiner nodes that lead to reduced minimum spanning tree cost are added to the node set, with the process is repeated until there is no further improvement. Finally a series of post-processing steps are performed to prune Steiner nodes that are not needed for the final tree and to apply subtree regrafting to correct for a potential source of suboptimality arising from the fact that the core phylogeny algorithm assumes symmetric distances but GD operations are asymmetric.

3.1.5 Inferring tumor Phylogenies using Neighbor Joining (NJ) and Maximum Parsimony (MP) methods

Neighbor Joining (NJ) and Maximum Parsimony (MP) methods have been commonly used for building single-tumor phylogenies [118, 161] and we therefore compared their accuracy to that of our own methods in inferring copy number phylogenies. We applied these two traditional phylogenetic tree building methods to build tumor progression trees using the individual copy number profiles as taxa and compared them with the trees built using our algorithms. We used

Algorithm 6 Main steps in the algorithm to generate tumor progression trees; generate_distance_matrix uses Algorithm 4 on each distinct pair of nodes in the set of nodes it is passed. To compute Minimum Spanning Tree (mst), we implemented Prim's algorithm.

```
observed_nodes are parsed from the input
steiner_nodes ← ∅
matrix_1 ← generate_distance_matrix(observed_nodes)
min_weight = mst(observed_nodes, matrix_1)
while (min_weight is improved) do
    initialize the median Steiner network msn on node set {observed_nodes ∪ steiner_nodes} to have all singleton nodes and no edges
    for all possible edges e in increasing order of distance do
        if adding e would connect two distinct components in msn then
            Add e to msn
            Update components of msn
    for all triplets of nodes u, v, w in msn do
        Find the truncated integer lattice spanned by vertices u, v, w in d-dimensional space
        for all lattice points s in the hyper-rectangle do
            if s is not an observed state then
                matrix_2 = generate_distance_matrix(observed_nodes ∪ steiner_nodes ∪ {s})
                new_weight = mst(observed_nodes ∪ steiner_nodes ∪ {s}, matrix_2)
                if (new_weight < min_weight) then
                    min_weight ← new_weight
                    Store the new tree as the current best tree
                    steiner_nodes ← steiner_nodes ∪ {s}
                    Record that the min_weight has improved
Prune unnecessary Steiner nodes (having degree ≤ 2 in the final tree)
Root the minimum spanning tree at (2, 2, ..., 2)
Perform subtree pruning and regrafting step to remove edges along which GD events are inferred from the tail node configuration to the head node configuration
Display the tree
```

implementations of both approaches in MEGA version 6 [167]. For NJ, we used both Euclidean and Rectilinear distances between cell copy number profiles to build the pairwise distance matrix. For MP, we treated copy number profiles of the genes in individual cells as sequences of arbitrary phylogenetic characters. We used the "Close-Neighbor-Interchange on Random Trees" search method. For the parameters "Number of Initial Trees" and "MP search level", we used values of 10 and 1 respectively.

3.2 Results

As in Chapter 2, we used data collected from cervical cancer (CC) [183] and breast cancer (BC) [83] patients to evaluate our methods. Figure 3.2(A) shows a tumor progression tree inferred from one of the cervical cancer samples. For comparison, Figure 3.2(B) shows a progression tree inferred on the same sample using our prior SD model [34]. Visual inspection shows that large regions of the two trees are identical but that allowing CD and GD events leads to some rearrangement and a reduction in tree depth and overall size. Next we evaluate the changes induced by adding SD, CD and GD events, using simulated data to show effectiveness of the methods in finding more parsimonious solutions to the broader model and using the real CC and BC data to show the biological relevance of the improvements. We further show that our algorithms infer trees with higher accuracy than the prevailing alternative algorithms for single-tumor phylogenetic inference. Finally, we perform statistical experiments to evaluate the effects of tumor sample size on the performance of our tree building algorithm.

3.2.1 Simulation experiments

To measure accuracy of the methods for FISH datasets with a known ground truth, we generated a dataset of 100 trees with six probes, two of which were treated as being on the same chromosome. Each tree was generated by starting from a diploid root node and executing a branching process

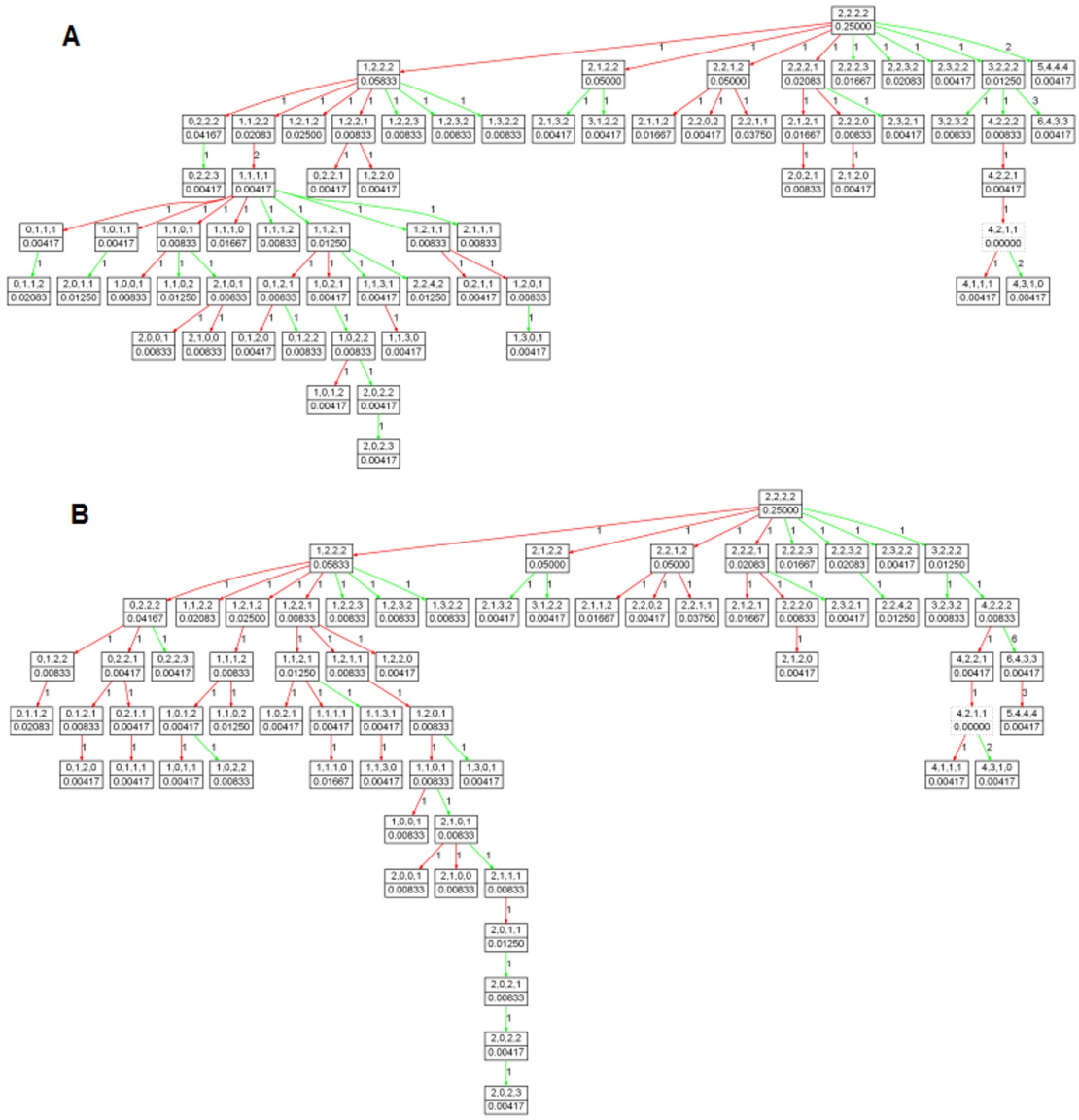

Figure 3.2: Phylogenetic trees showing tumor progression in a cervical cancer patient. Trees are built considering (A) all of SD, CD and GD and (B) only SD model of tumor evolution. Each node represents a configuration of the four gene probes *LAMP3*, *PROX1*, *PRKAA1* and *CCND1*. Nodes with solid and dotted borders represent cells present in the collected sample and inferred Steiner nodes respectively. Green and red edges model gene gain and gene loss, respectively. The weight value on each edge connecting two nodes x and y is the distance between the states of x and y, computed using the particular model of tumor progression under consideration. The weight on each node describes the fraction of cells in the sample with the particular copy number profile modeled by that node; Steiner nodes are assigned weight 0.

in which each node was recursively assigned a number of children drawn from a geometrically distributed random variable with mean 0.50. Each child was distinguished from its parent by

selecting an SD, CD, or GD event with probability 0.1167 for each of the six possible SD events, 0.18 of a CD event, and 0.12 of a GD event. This process terminated when all leaf nodes had been assigned zero children by the sampling. We then generated simulated FISH data for each tree by uniformly sampling 300 cells from the nodes in this topology. The simulated data corresponds to counts of probes for each sampled cell in the tree. We applied Algorithm 6 (see Methods) to find a minimum-cost tree for each of four event models: (i) SD only, (ii) SD and CD, (iii) SD and GD, and (iv) SD, CD and GD.

We quantified the accuracy of tree inference by comparing each simulated true tree to its corresponding inferred tree derived from the sampled cells. This assessment was performed at the level of accuracy of tree edges by the following procedure:

1. We pruned the real tree so as to remove any subtree for which no cell in the tree was sampled. This step was intended to avoid penalizing for "impossible" inferences of subtrees unsupported by any data.

2. We computed a maximum matching of edges between the real subtree and the inferred tree, with each pair of edges weighted by the maximum number of nodes in agreement between the corresponding parts of the bipartitions that the two edges define [34, 102]. We used the Hungarian algorithm [97] for computing the maximum matching (applying the function"Hungarian" by Alexander Melin from the Matlab Central File exchange).

3. We calculated a reconstruction error R of the inferred tree using the following formula which is discussed in more detail in Chapter 6:

$$R = \left(1 - \frac{W}{|T| \times (|P_r| + |P_i|) - W}\right) \times 100$$

where W is the weight of the maximum matching, T is set of taxa in common between the real and inferred trees, and P_r and P_i represent the sets of nontrivial bipartitions in the real and inferred trees, respectively.

Intuitively, this formula measures the fractional agreement between bipartitions of the trees relative to the total number of bipartitions. We use a matching-based formula, rather than the more

familiar Robinson-Foulds metric [140], both because of its greater sensitivity to small changes in trees and because the Robinson-Foulds measure is not defined for trees with different node sets. We also note that we use a different normalization factor than in our prior work [34], normalizing essentially by the total number of edges between the two trees, to control properly for the fact that different inference methods may infer different numbers of tree edges. The reconstruction error R ranges in value from 0, if the real and inferred trees are isomorphic, to an upper bound of 100 in the limit of complete disagreement.

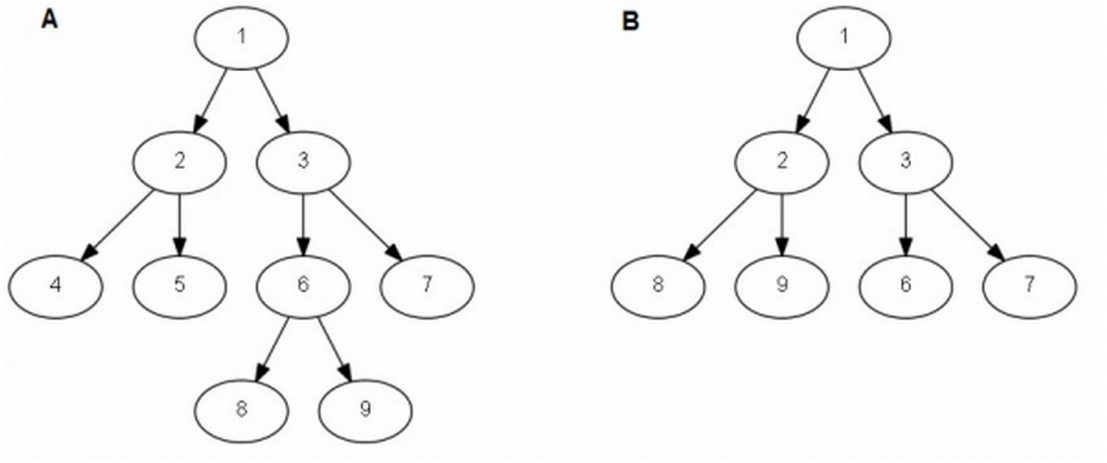

Figure 3.3: Example simulated and inferred trees illustrating key terms in the formula for calculating the reconstruction error. (A) A hypothetical simulated ground truth tree on the set of taxa $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. (B) Example inferred tree built on the sampled set of taxa $\{1, 2, 3, 6, 7, 8, 9\}$ on the dataset resulting from the ground truth tree.

To illustrate the meanings of the terms of the equation for R, we present a simple example using a hypothetical ground truth and an inferred tree presented in Figure 3.3(A) and Figure 3.3(B), respectively. The set of nontrivial bipartitions in the ground truth are

$$\{\{\{1, 3, 6, 7, 8, 9\}, \{2, 4, 5\}\}, \{\{3, 6, 7, 8, 9\}, \{1, 2, 4, 5\}\}, \{\{6, 8, 9\}, \{1, 2, 3, 4, 5\}\}\}$$

and the nontrivial bipartitions in the inferred tree are

$$\{\{\{1, 3, 6, 7\}, \{2, 8, 9\}\}, \{\{1, 2, 8, 9\}, \{3, 6, 7\}\}\}.$$

If we apply the matching algorithm on these two sets of bipartitions, the first and second bipartitions in the ground truth tree are matched with the first and second bipartitions in the

inferred tree, respectively. The weight W of the matching is 10. The number of common taxa between these two datasets is $|T| = 7$. The total number of nontrivial bipartitions in the real and inferred trees are $|P_r| = 3$ and $|P_i| = 2$. Plugging these values into the equation for R, we calculate $R = 60\%$.

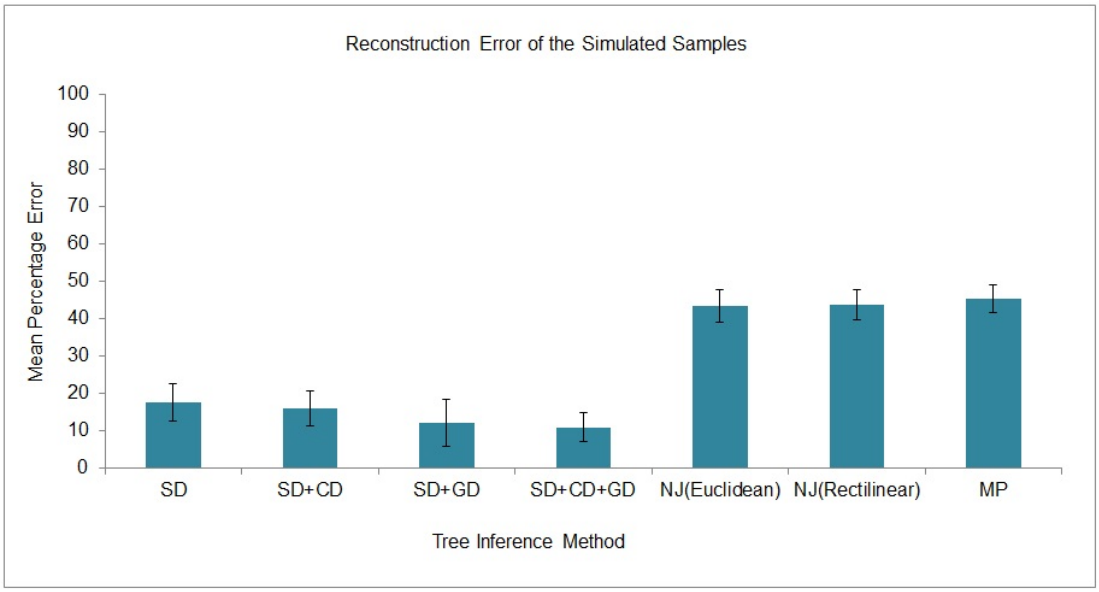

Figure 3.4: Accuracy of phylogenetic inference on simulated copy number data for varying algorithms. Variants of our phylogenetic algorithms and two competing methods from the literature were applied to simulated FISH datasets describing evolution by combinations of single-gene (SD), chromosome (CD), and whole-genome (GD) duplication and loss events. Results are reported for inference by our methods from 100 simulated trees, allowing for SD events alone, SD+CD events, SD+GD events, and SD+CD+GD events. We compared these results to inference by neighbor-joining (NJ) and pure maximum parsimony (MP) as implemented in MEGA, version 6. Accuracy is assessed by mean reconstruction error of bipartitions between true and inferred trees. Error bars show plus or minus one standard deviation across the samples for each method.

A comparison of the four models is presented in Figure 3.4. The SD model showed 17.43% reconstruction error with standard deviation (s.d.) of 5.1% across the 100 trees. The SD+CD model yielded 15.91% error with s.d. 4.59%. SD+GD yielded 12.01% error with s.d. 6.4%. The full SD+CD+GD model yielded 10.84% error with s.d. 3.88%. Collectively, the results suggest that one can reconstruct reasonably accurate trees even from the SD-only model, despite the fact that the trees were generated from a model of all three event types, although accuracy improves

with each event type added. Accounting for GD events made a larger difference in accuracy than accounting for CD events, presumably because a missed GD event might require many SD or CD events to explain it, while a missed CD event could be explained with just two SD events. The reconstruction error for the full model is reduced by more than 1.7-fold relative to the SD-only model considered in our prior work.

We further compared these results to those derived using generic phylogenetic methods that have been used in much of the single tumor phylogenetics work to date [118, 161]. We tested the accuracy of reconstruction of the 100 simulated trees described above using generic neighbor joining (NJ) with Euclidean and Rectilinear distances, and pure maximum parsimony (MP) treating copy numbers as arbitrary characters, approaches chosen because they have been the primary alternatives to our specialized algorithms in the single-tumor phylogeny literature. We omit here comparison to more complicated Bayesian phylogenetic models (e.g., [157]) because such approaches are not scalable to the numbers of cells we examine. We then used the weighted matching based similarity method, described above, to calculate the mean percentage reconstruction error R between the inferred and the ground truth trees. The mean reconstruction errors for NJ with Euclidean and Rectilinear distance metrics were 43.23% (s.d. 4.24%) and 43.57% (s.d. 4.00%), respectively. The reconstruction error for MP was 45.21% (s.d. 3.86%). When compared to the error of 10.84% (s.d. 3.88%) for the SD+CD+GD algorithm proposed in this chapter, the test demonstrates that when the underlying evolutionary process includes cancer-like chromosome abnormalities, errors are substantially reduced by using an algorithm designed for that model relative to standard off-the-shelf algorithms still widely used for single-tumor phylogenetics work.

We performed additional experiments to evaluate the effects of different evolutionary parameters on the accuracy of inference of tumor progression trees by FISHtrees. For this experiment, we selected five different combinations of probabilities of SD, CD and GD events for generating the ground truth trees and then used SD, SD+CD, SD+GD and SD+CD+GD models to infer the tumor phylogenies. These data sets again each used six probes with two of the six

Table 3.1: Comparison of mean percentage reconstruction error (with standard deviation) of different phylogeny models on simulated data for different combinations of SD, CD and GD event probabilities.

Probabilities of (SD,CD,GD) Events	SD	SD+CD	SD+GD	SD+CD+GD
(0.125,0.05,0.2)	17.97(4.49)	16.89(4.32)	9.85(3.51)	9.25(4.18)
(0.1,0.2,0.2)	25.58(4.50)	21.82(3.98)	13.81(3.62)	10.96(3.99)
(0.15,0.07,0.03)	16.02(4.15)	14.96(4.16)	11.92(4.29)	11.71(4.77)
(0.1,0.3,0.1)	23.13(4.37)	20.02(4.50)	15.43(4.60)	13.42(4.64)
(0.1166,0.18,0.12)	17.43(5.10)	15.91(4.59)	12.01(6.40)	10.84(3.88)

Mean percentage reconstruction error on 100 simulated samples are shown for four tree-building models considering (i) SD, (ii) SD+CD, (iii) SD+GD and (iv) SD+CD+GD across five different combinations of SD, CD, and GD probabilities.

on a common chromosome. The selected five combinations of (SD,CD,GD) event probabilities are: $(0.125, 0.05, 0.2)$, $(0.1, 0.2, 0.2)$, $(0.15, 0.07, 0.03)$, $(0.1, 0.3, 0.1)$ and $(0.1166, 0.18, 0.12)$. These combinations of event probabilities were chosen to yield trees of comparable complexity to the real data while producing test sets enriched in distinct combinations of the three event types. They thus allow us to consider how robust our algorithms are to contributions from each of the three event types, singly or in combination. We report the reconstruction error for 100 trees for each of these combinations of event probabilities in Table 3.1. These results again show that accuracy improves with each event type added. When the probability of SD events is high (as in combination 3), the SD model results in highly accurate trees (mean reconstruction error of 16.02% with s.d. 4.15%). Accounting for GD events in combination with SD events always result in larger improvement in the reconstruction error in comparison to the SD+CD models, even when the CD events are very frequent (as in combinations 2 and 4). Finally, accounting for GD events in combination with SD and CD events results in the largest improvements when the probability ratio of GD events to SD+CD events is highest, as can be seen from comparison of parameter sets 1 and 2.

Next, we performed simulation tests to evaluate the effects of non-uniform distributions of cells across different levels of the trees on the performance of our tree inference method. In our initial simulation experiments described above, we assumed that observed cells were sampled

Table 3.2: Comparison of mean percentage reconstruction error (with standard deviation) of different phylogeny models on simulated data for different sampling distributions of the cells.

Distribution	SD	SD+CD	SD+GD	SD+CD+GD	NJ	MP
Uniform	17.43(5.10)	15.91(4.59)	12.01(6.40)	10.84(3.88)	43.23(4.24)	45.21(3.86)
Skewed ($\gamma = 1.1$)	22.74(4.49)	19.09(4.47)	14.75(4.64)	11.92(4.64)	47(3.76)	47.38(3.72)
Skewed ($\gamma = 1.3$)	29.93(7.37)	26.35(6.56)	18.89(7.24)	15.36(6.78)	50.63(5.89)	50.32(5.74)

Mean percentage reconstruction error on 100 simulated samples are shown for six tree-building models considering (i) SD, (ii) SD+CD, (iii) SD+GD, (iv) SD+CD+GD (v) NJ and (vi) MP when the sampling distribution of cells is varied.

uniformly across clones. In real tumors, the distribution of cells would not typically be uniform due to differences in age and fitness of clones. In order to test robustness of our method to non-uniformity of clone frequencies, we sampled the cells following a non-uniform model in which the sampling frequency of a clone varies geometrically with its depth in the tree with a parameter γ. We used values of 1.1 and 1.3 for γ in our experiments. When $\gamma = 1.1$, 25% of the total cells are located in the first three levels of the trees, while for $\gamma = 1.3$, this fraction is 55%. We generated 100 trees in each case with probabilities of SD, CD and GD events fixed at $0.1167, 0.18$ and 0.12. We again used SD, SD+CD, SD+GD and SD+CD+GD models to infer the tumor progression trees. We present the results from this experiment in Table 3.2, where we also show the results from the uniform sampling of the cells. Additionally, we report the results on the trees inferred using NJ and MP for these three different cell distributions. From the table, we can see that the reconstruction error increases with increasing γ for all methods. The SD+CD+GD model, however, shows the best performance among all the models for all three values of γ and the least loss of performance with increasing γ.

Table 3.3: Comparison of mean percentage reconstruction error (with standard deviation) of different phylogeny models on simulated data for two different probe settings.

Number of Chromosomes with 2 Genes	SD	SD+CD	SD+GD	SD+CD+GD
1	17.43(5.10)	15.91(4.59)	12.01(6.40)	10.84(3.88)
2	19.01(5.61)	15.65(5.26)	11.49(4.18)	8.94(3.46)

Mean percentage reconstruction error on 100 simulated samples are shown for four tree-building models considering (i) SD, (ii) SD+CD, (iii) SD+GD and (iv) SD+CD+GD for two different cases when the number of chromosomes harboring two genes is 1 or 2.

Finally, we performed simulation experiments to understand the effects of varying the numbers of chromosomes with multiple probes. We created a simulated dataset of 100 trees with eight probes where two pairs of probes each reside on two different chromosomes and the remaining four probes reside on four separate chromosomes. The probabilities of each of the SD, CD and GD events were fixed at $0.1167, 0.09$, and 0.12, respectively. We report the results from this experiment in Table 3.3, which compares the results from this experiment with our earlier result using only a single chromosome with two probes and four other probes located on separate chromosomes. The table shows that inclusion of the extra possible CD event results in higher accuracy for all the models except for the SD only model. The performance drop in the SD model is expected, as it would require more SD events to explain a greater number of missed CD events. The highest gain in performance is observed for SD+CD+GD model. These results show that our algorithm will tend to yield comparatively more advantage over the earlier work with more complicated scenarios of sharing probes across chromosomes, suggesting its utility will increase as improvements in technology allow for larger probe sets.

3.2.2 Application to real cervical and breast cancer data

We applied the algorithm to two sets of real data which were described in section 1.7. For clarity, here we reiterate the sample composition and the genes on which each dataset was profiled:

- A set of CC [183] FISH data consisting of 47 samples organized into 16 primary samples of metastatic patients, 16 paired metastasis samples from the same patients, and 15 primary samples from patients who did not progress to metastasis. Each sample consisted of $223 - 250$ cells profiled on four FISH probes: *LAMP3* (Entrez Gene Id 27074) [95], *PROX1* (5629) [187], *PRKAA1* (5562) [89] and *CCND1* (595) [56]. All of these four genes are oncogenes, which typically show copy number gains in tumor cells. Each of the genes belongs to a distinct chromosome.
- A set of BC [83] FISH data consisting of 13 paired (from the same patient) ductal carci-

noma in situ (DCIS) and invasive ductal breast carcinoma (IDC) samples with $76 - 220$ cells per sample profiled on eight FISH probes: *COX-2* (5743) [88], *MYC* (4609) [188], *CCND1* [56], *HER-2* (2064) [168], *ZNF217* (7764) [122], *DBC2* (23221) [72], *CDH1* (999) [17] and *TP53* (7157) [179]. The first five genes in this list are oncogenes and the last three genes are tumor suppressors. In tumor cells, tumor suppressors are typically associated with loss in copy numbers.

Among the eight genes in the BC dataset, *DBC2* and *MYC* reside on chromosome 8 and *HER-2* and *TP53* reside on chromosome 17. The other four genes belong to distinct chromosomes. The oncogene Cyclin D1 (*CCND1*), which plays a role in many solid tumor types, is in both the BC and CC datasets. However, in some other tumor types, such as oral cancer, *CCND1* is part of a larger region with recurrent copy number gains on chromosome 11 and other nearby genes have also been suggested to play a role in oncogenesis [90].

We evaluated the SD+CD+GD method by its effectiveness in reducing the parsimony score (total number of mutation events) of the resulting trees relative to the prior SD-only model. With the primary CC samples, the SD+CD+GD method found a lower-cost tree in 21 of 31 cases, a tree of equal weight in 4 cases, and a higher-cost tree in 6 cases. In each case of increased weight, the increase was by 1 and appears to result from the subtree regrafting heuristic used in handling GD events (see Methods). These results suggest that the heuristic tree search may more often yield a suboptimal result for the SD+CD+GD model than it does for the SD-only model. The benefit of the more realistic model, however, outweighs the cost of this suboptimality in a large majority of instances. For trees derived from metastatic samples, 12 of 16 trees had lower weight for the full SD+CD+GD model and the remainder all had equal weight for the two models. Metastatic data sets tend to have fewer distinct cell types than do primary trees and thus may represent an easier optimization challenge. For the BC samples, 13 of 13 DCIS (samples 1-13) and 12 of 13 IDC (samples 14-26) had lower weight for the full model, with the remaining one sample having equal weight. Parsimony scores by tree are provided in Figures 3.5 and 3.6.

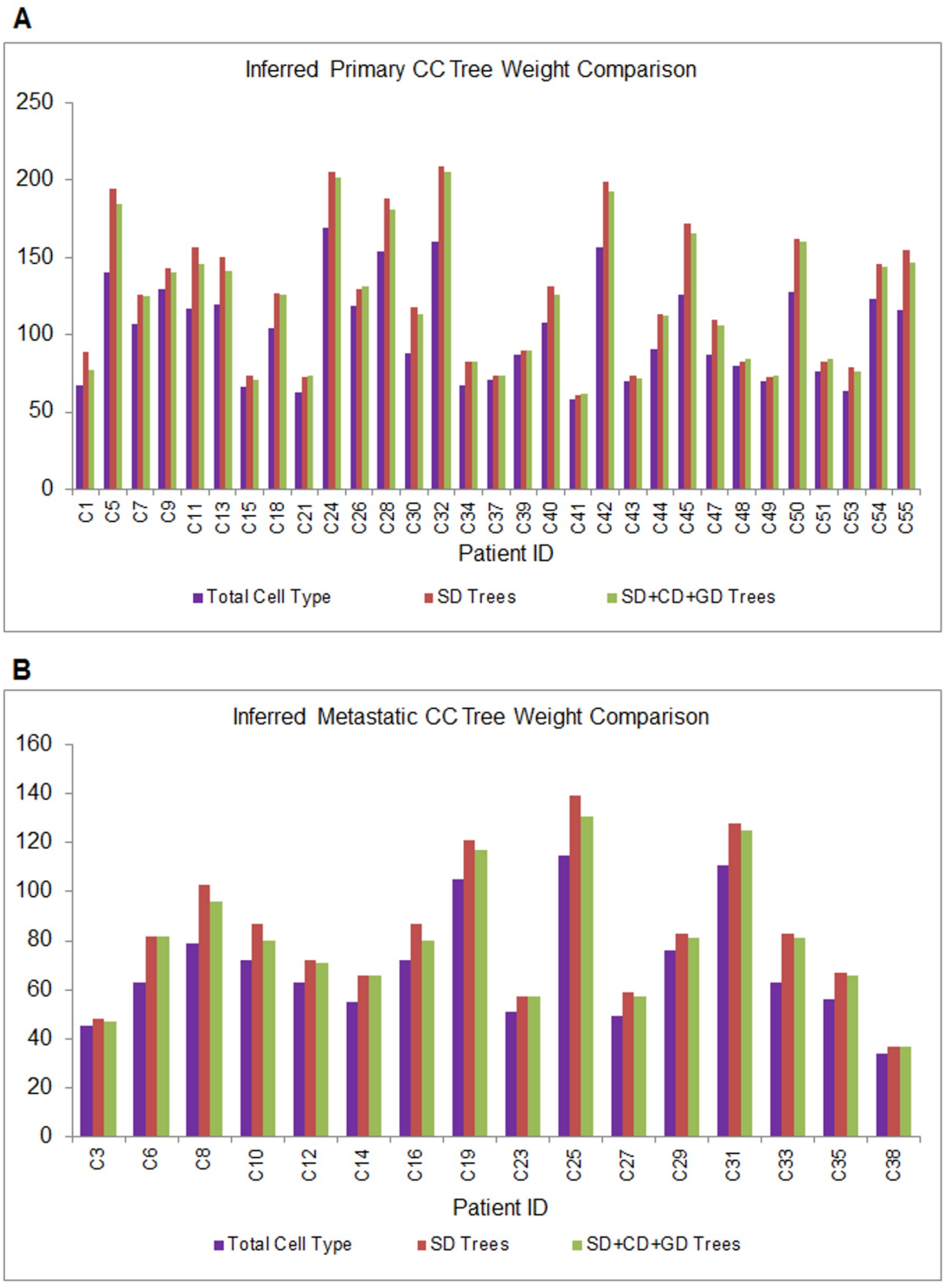

Figure 3.5: Parsimony score comparison on the CC samples. Comparison of (A) Primary and (B) Metastatic CC tumor progression tree weights built considering only SD and combined SD, CD and GD models. "Total Cell Type" refers to the total number of unique probe copy number configurations in the dataset, providing a lower bound on the minimum possible parsimony score for a given data set.

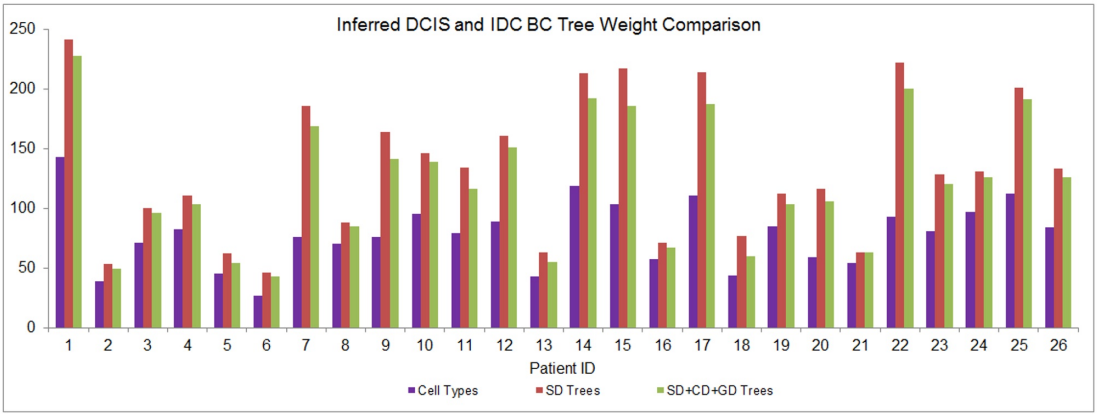

Figure 3.6: Parsimony score comparison on the BC samples. Comparison of DCIS (id 1-13) and IDC (id 14-26) BC tumor progression tree weights built considering only SD and combined SD,CD and GD models. "Cell Types" refers to the total number of unique probe copy number configurations in the dataset, providing a lower bound on the minimum possible parsimony score for a given data set.

We next evaluated effects of the improved model on overall tree topology, based on results of our prior work [34] that tree topology can significantly distinguish trees drawn from distinct progression stages of a given tumor type, with possible implications for the varying balance of diversification and selection acting on different stages of tumor progression. Figure 3.7 quantifies the topology for each sample set based on fractions of cells inferred at each tree depth from 1 to 12. The figure shows similar qualitative trends for both SD and SD+CD+GD methods, although with small quantitative differences. For example, both SD and SD+CD+GD trees recapitulate a tendency for CC primary trees to show relatively broad topology (Figure 3.7(A)) while CC metastatic trees prune rapidly beyond the first few tree levels (Figure 3.7(B)). There is, however, an overall shift to lower depth in the SD+CD+GD trees. For CC primary trees, 92.6% of cells are located in the first 12 tree levels for SD versus 97.09% for SD+CD+GD. For CC metastatic, 99.2% of cells are located in the first 12 tree levels for SD versus 99.6% for SD+CD+GD. For BC, the comparable numbers of cells in depths $1-12$ are 86.5% for SD versus 93.9% for SD+CD+GD in DCIS and 82.67% for SD versus 92.6% for SD+CD+GD. These results suggest that the overall tree topology is not greatly sensitive to the combination of event types, although there is a noticeable shift towards lower depth in the full model.

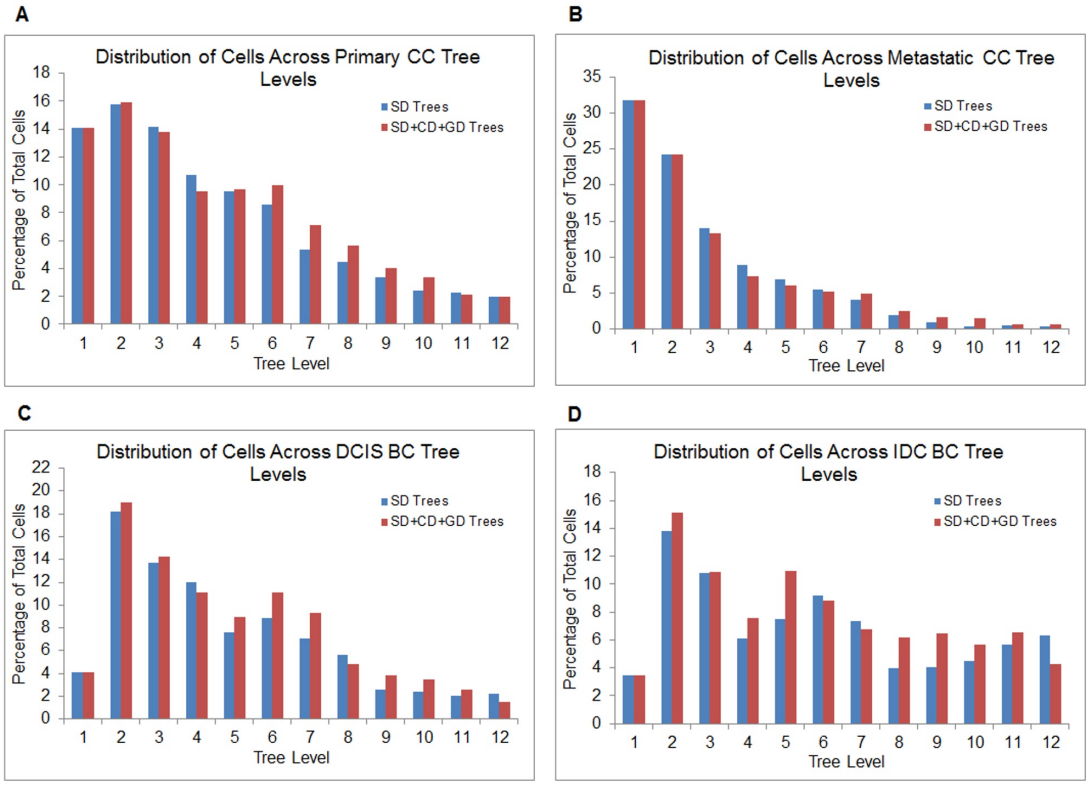

Figure 3.7: Distribution of cells across different levels of tumor phylogenies. Distribution of cells across different levels are shown for (A) Primary and (B) Metastatic CC, and (C) DCIS and (D) IDC BC tumor progression trees.

An additional evaluation was possible for the BC trees, because for the BC data, a probabilistic model and expert annotation based on two additional centromere probes made it possible to estimate the cell ploidy [83], which we define as the mode among the number of copies of the twenty-two autosomal chromosomes in a cell. Each cell in that dataset is thus annotated with an expert-curated overall ploidy estimate. We used these ploidy estimates to validate our inference of GD events based on whether edges assigned to GD events in our trees correspond to doubling of annotated ploidy. The percentage agreement by edge between GD events and annotated doubling in ploidy is 65% across DCIS trees and 64.44% across IDC trees. In 31.6% of all inferred GD events, at least one endpoint of the corresponding edge is a Steiner node, and the uncertainty among whether a GD event occurred prior to or after the emergence of the Steiner node may explain why the per-edge agreement is not higher. Nonetheless, the data support the conclusion that inferred GD events are correct in a majority of cases.

As a final step, we repeated an approach developed in our prior work [34] to both validate the biological relevance of the trees and develop a practical application of them by treating the trees as sources of features for classification tasks applied to the CC data. For this purpose, we developed several sets of quantitative features based on inferred trees as well as comparative features derived from raw FISH probe counts. We used the following set of tree-based features:

1. Edge count: 8 features corresponding to fraction of progression tree edges showing gains and losses of each gene.
2. Tree level cell percentage: 10 features corresponding to the fraction of cells at each of the first 10 levels for the progression trees.

We omitted a third feature set, bin count, used in our prior work because it is not easily comparable between SD and SD+CD+GD trees. We compared these features to four features derived directly from FISH probe counts without reference to the trees:

1. Mean gain and loss of individual genes.
2. Maximum copy number of individual genes.

3. An information theoretic measure, Shannon index [130]. For each gene, each combination of gene copy number and cellular ploidy represents a species. If we denote the frequency of species i among all tumors by p_i, then Shannon index is given by the formula $H = -\sum p_i \log_2(p_i)$.
4. Simpson's index [130], which is defined as $\sum p_i^2$.

We used each feature set as input to the Matlab support vector machine (SVM) classifier with a quadratic kernel using 500 rounds of bootstrap replicates per test with leave-one-out cross-validation to compute mean and standard deviation of accuracy. We used Matlab functions "svmtrain" and "svmclassify" for training and testing of the SVM classifier.

We then applied these methods for three classification tasks: (i) distinguishing primary samples that progressed to metastasis from their paired metastatic samples, (ii) distinguishing all primary samples from all metastatic samples, and (iii) distinguishing primary samples that metastasized from primary samples that did not metastasize. The first two tasks are relevant to identifying features that help us understand the differences in evolutionary mechanisms of primary and metastatic samples. The third is intended to model an important practical problem in cancer treatment: determining whether a given primary tumor will metastasize.

Figure 3.8 shows results on each task. For task (i), allowing SD+CD+GD events increased accuracy relative to SD trees from 64.31% to 80.77% for edge counts and from 81.91% to 84.63% for tree level cell count. The SD+CD+GD tree level cell count was the most effective of all features, tree-based or not. For task (ii), we similarly saw a substantial improvement in prediction accuracy for SD+CD+GD trees relative to SD trees. Classification accuracy improved from 68.87% to 84.06% for edge count features and from 82.26% to 87.79% for tree level features. In this case, both SD+CD+GD tree feature sets outperformed all other features sets, tree-based or otherwise. These results provide an indirect validation that using a more general tree model gets closer to the biological ground truth. For task (iii), we saw no improvement, with identical results for SD and SD+CD+GD trees for either feature set. All tree-based feature sets significantly

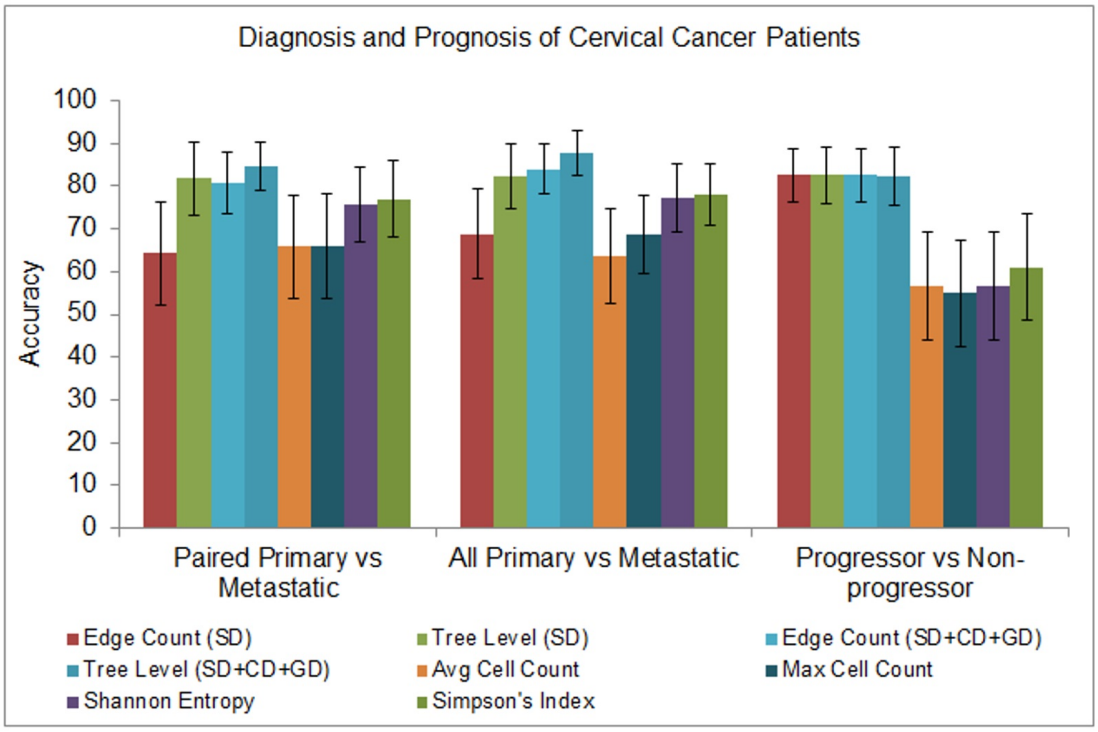

Figure 3.8: Classification results on the CC dataset. Prediction accuracy on three different classification tasks of CC samples of an SVM classifier using tree-based and cell-based features. Each of the two tree-based features, edge count and tree level cell percentage, is derived from phylogenetic trees built using two different models of tumor progression, namely SD and combination of SD, CD and GD. Two cell-based features, average gain/loss and maximum copy number of each gene, and two information theoretic measures of cell heterogeneity, Shannon entropy and Simpson's index, are used.

outperformed all non-tree-based feature sets for this task. We conclude that the more realistic evolutionary models appear not to reveal any more information to the classifiers for predicting which primary samples will go on to metastasize than the SD trees, which were already quite effective for that task.

3.2.3 Dependence on data size

A key advantage of FISH for profiling tumor heterogeneity is that it makes it cost-effective to profile much larger numbers of cells than alternatives such as single-cell sequencing. To assess the practical importance of this advantage, we asked two related questions: (1) how many cells

do we need per tumor to accurately reconstruct single-cell phylogenies and (2) how many tumors do we need to examine to identify reproducible, statistically significant features across trees.

We first assessed the number of cells needed per tumor by using our first simulated dataset of 100 trees described above with subsamples of varying numbers of cells per tumor, measuring reconstruction error of our SD+CD+GD algorithm with the weighted matching algorithm. The mean reconstruction errors calculated across 100 cases for subsamples of 20, 50, 100, 150 and 200 cells were 33.66% (s.d. 14.40%), 20.43% (7.97%), 15.28% (6.38%), 11.79% (4.03%), and 11.70% (4.4%) respectively. We can thus conclude that accuracy improves noticeably with increasing numbers of cells to at least 100 cells per tumor before plateauing at approximately 10% error.

We next assessed numbers of tumors needed to identify meaningful statistically significant properties of tumor classes by analysis of the 32 CC paired and primary samples. We randomly subsampled from among the 32 pairs and, for each subsample, calculated the following three tree statistics on progression trees inferred from our SD+CD+GD algorithm:

1. Shannon index based on distribution of cells across different tree levels.
2. Weighted mean depth of the trees.
3. Sum of differences of fractional gain and loss of each gene across the tree edges.

We then compared distributions of each statistic on primary vs. metastatic trees by a Wilcoxon signed rank test. As the samples were selected randomly, no ordering among the samples was considered. Figure 3.9 shows the 1-sided p-values of the three statistical tests when the number of randomly selected samples are increased from 5 to 32. The figure shows that ability to distinguish the two tumor subsets improves with increasing number of tumors. While the threshold for significance varies by statistic, each reaches weak significance ($p < 0.05$) between 10 and 24 tumors. We can thus conclude that finding reproducible features distinguishing the tree types requires on the order of tens of tumors, at least for the candidate probe sets examined here.

Taken together, these two results demonstrate that building accurate trees on a large enough

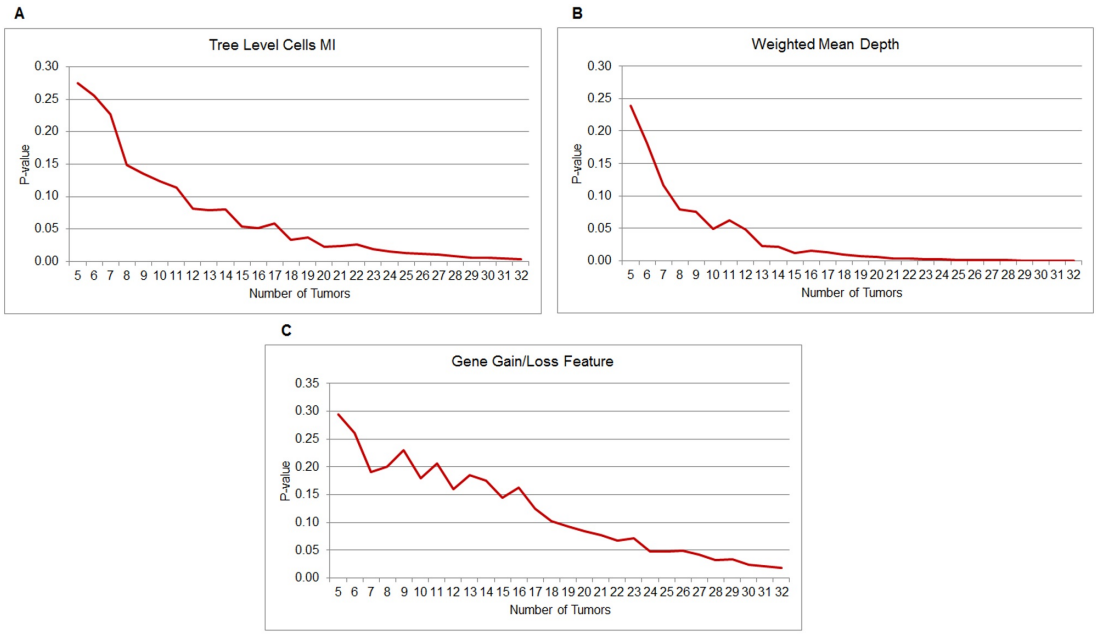

Figure 3.9: Wilcoxon signed rank test results for separating primary CC samples from the metastases. Wilcoxon signed rank test 1-sided p-values for separating the primary CC samples from the metastases across subsets of increasing numbers of randomly selected tumor samples. For each set of i tumors, i samples were randomly selected from 32 paired CC primary and metastatic tumors with atleast one of each type and then Wilcoxon signed rank test was used to calculate the p-values for separating the primary from metastases based on three different statistics: (A) Shannon index calculated using the distribution of cells across different tree levels, (B) weighted mean depth of the trees and (C) sum of differences of fractional gain and loss of each gene across the tree edges.

scale to distinguish meaningfully primary from metastatic trees requires data sets with roughly the order of thousands of single cells (hundreds of cells per tumor for tens of tumors), a scale of data that has so far been achieved only by FISH studies of tumor heterogeneity. We note, however, that one would expect these numbers to vary depending on the degree of tumor heterogeneity, the classes of trees one wishes to distinguish, and the specific markers examined.

3.3 Discussion

This chapter has presented novel theory and algorithms for reconstructing evolutionary trajectories of gene copy numbers in solid tumors in terms of a model of tumor evolution incorporating

changes at the scale of single gene probes, full chromosomes, or all probes in the genome. We have derived algorithms to reconstruct maximum parsimony sequences of events, and thus estimates of evolutionary distance, between pairs of cells assayed by FISH probes. We have further incorporated these inferences into a method for building phylogenies of hundreds of cells in single tumors. These methods have been added to FISHtrees [34], our software for inferring tumor phylogenies from single-cell copy number data. Experimental results on simulated data confirm the ability of the new methods to improve phylogenetic inference accuracy relative to simpler models by adding CD and GD events that model chromosome-scale and whole-genome copy number changes that are frequently observed in tumor evolution. Application to observed human tumor data shows that these extended evolutionary models are able to yield more parsimonious tree reconstructions and that the resulting trees lead to improved accuracy in prediction tasks related to diagnosis and prognosis.

In future work, we hope to extend the theory developed here to handle even more realistic models and more challenging data types. One important direction will be advancing the theory developed here to improve upon the heuristic approximations used in the Steiner tree inference to better approach the goal of finding globally optimal trees for the most computationally challenging FISH data sets. The evolutionary models, likewise, might be further extended to go beyond the three mutational event types considered here to better approximate the numerous distinct mutational mechanisms by which copy number profiles of tumor cells might evolve. The data sets studied here do not include geographical information about locations of individual cells in the tumor, but other data sets for analyzing tumor heterogeneity do include such geographical information [4, 62]. We expect it would be interesting to construct phylogenies with distance functions that combine spatial distance in three dimensions with combinatorial distance measures between the cell count patterns, as we have studied here. Further, while FISH for the moment retains a unique advantage in the large number of cells it can profile, one can reasonably anticipate that single-cell sequencing will eventually become practical for comparable cross-tumor studies. There would thus be value in extending the theory developed here to single-cell sequencing data,

a goal that would pose substantial algorithmic challenges due to the much larger number and variety of markers it can reveal as well as the more complicated error models it would entail. Finally, we hope to make more use of these single-tumor phylogenetic models in clinically relevant prediction tasks and further explore the biological insights one can gain from more accurate tumor phylogenies.

Chapter 4

Inferring Models of Multiscale Copy Number Evolution for Single-Tumor Phylogenetics[1]

Tumor development and progression are evolutionary processes [126] and it has become ever more apparent that evolution is fundamental to public health problems in cancer treatment, such as the failure of therapy due to drug resistance [53]. The evolutionary nature of cancers prompted the observation that one might reconstruct cancer progression processes using methods from phylogenetics, i.e., evolutionary tree-building [42]. Cancer phylogenetics was initially applied at the level of populations of cancers by modeling individual tumors or tumor types as species [14, 42]. Later, variants were developed to study evolution of single tumors at the regional [158] or cellular [109, 134] levels. Phylogenetic models have proven valuable for distinguishing driver genes from passengers in tumor genomic data, explaining intratumor heterogeneity [110], and predicting future tumor progression [176]. See [14] for a recent review.

Although the idea of adapting methods for reconstructing species evolution to the study of

[1]This chapter was developed from material submitted in "Chowdhury *et al.*, Inferring models of multiscale copy number evolution for single-tumor phylogenetics, Submitted for Publication".

tumors has proven powerful, the analogy has limits because single cells in a tumor evolve differently from organisms within a population. For example, cancers typically exhibit hypermutability, which can take the form of any of a number of known "mutator phenotypes", each with a distinct pattern of elevated mutation rates [104]. The most recognized of these is a pattern of chromosome instability (CIN) arising from dysfunction of *TP53* [66]. Other known sources of hypermutability include microsatellite instability (MSI) resulting from defects in DNA mismatch repair [173] and elevated point mutation rates resulting from DNA polymerase defects [44] or AID/APOBEC1 cytidine deaminase dysregulation [76]. These mutator phenotypes result in mechanisms of genomic diversification different from those generally assumed in species tree inference. For example, CIN hypermutability results in evolution primarily via copy number variations, requiring mathematical models and algorithms different from those generally used to study species evolution.

Although much is known about the specialized molecular mechanisms behind tumor evolution, work in tumor phylogenetics has largely relied on conventional phylogeny algorithms designed for inferring species evolution [14]. In the previous chapters, we sought to address this gap by developing phylogenetic algorithms specifically to infer evolution by cancer-like CIN mechanisms of copy number variation. Even appropriate algorithms for tumor-like mechanisms of evolution are not enough to generate reliable trees, though, because phylogenetics relies on accurate estimates of relative frequencies of different evolutionary events to decide between distinct possible explanations of extant genomes. Given the heterogeneity of mutator phenotypes and the many ways they might interact in single tumors, rates of different types of aberrations can be expected to vary widely between tumor types, between individual tumors, or even between clonal lineages of single tumors.

There has been limited work to date to estimate evolutionary parameters of tumors, none to our knowledge scalable to single-cell data sets. Approaches using rate parameters for different events have been applied to comparative genomic hybridization data outside the context of phylogenetic algorithms (e.g., [86, 120]) and several groups proposed estimating rates via maximum

likelihood from bulk sequencing data from different sections of a tumor [67, 138], but not for data on multiple single cells. Maley and colleagues [158] have inferred tumor evolution parameters at a regional level by using Bayesian phylogeny models, which are effective for small numbers of regions but scale poorly with numbers of taxa. Even with very efficient approximate Bayesian computation (ABC) algorithms [11], such approaches have been used only for small numbers of sections (10-20) per tumor. Fluorescence *in situ* hybridization (FISH) allows one to probe copy numbers of small numbers of genomic markers in thousands of single cells per study, and such studies have shown that single tumors can have hundreds of genetically distinct cell types [83, 153, 165]. Large-scale single-cell sequencing studies, which offer a much more complete picture of the genome than FISH but for many fewer cells, have supported this view of extensive intercellular heterogeneity at the cellular level [182], suggesting that tumor phylogeny approaches and their underlying models will need to scale to hundreds or thousands of taxa per tumor to produce reliable models of the evolution of cellular heterogeneity in single tumors. To date, phylogenetic model inference with event rate estimation on comparable numbers of single cells has, to our knowledge, been achieved only for specialized data sets involving just two probes per cell [134].

In the present chapter, we address the need for algorithms for evolutionary model inference for tumor phylogenetics capable of handling large single-cell data sets, with specific application to FISH copy number data. We build on prior work in Chapters 2 and 3 on maximum parsimony inference using a multiscale model of genomic copy number variation by replacing an unweighted formulation of the problem to a weighted version for which we can then infer rate parameters. Our major theoretical results include algorithms, substantially different from prior methods [35], to construct weighted parsimonious sequences of single gene gains/losses, whole chromosome gains/losses, and whole genome duplications to infer the distance between configurations of gene copy numbers between any pair of cells. These new methods allow us to infer trees from models of distinct evolutionary rates of gain or loss for different genes and at different scales within a genome. We use these tree inferences with an expectation-

maximization (EM)-like [41] model inference method to combine estimation of the gain/loss rates jointly with inference of tumor progression models. We apply this collection of novel algorithms to cervical, breast and tongue cancer data sets of hundreds of single cells per tumor, although they can be expected to scale to orders of magnitude larger data sets as they become available. We show that the resulting models lead to improved power to predict tumor progression and patient survival relative to prior methods. Our new methods are implemented in our software, FISHtrees, for which C++ source code and two data sets are available at `ftp://ftp.ncbi.nlm.nih.gov/pub/FISHtrees`.

4.1 Methods

FISH data obtained from tumor cells consist of integer counts of a set of d copy-number probes per cell g_i for $i = 1, \ldots, d$. Typically, each probe is used to count the copy number of a particular gene, so we refer to g_i as *genes*. We refer to a collection of copy-number counts observable within a cell as a *configuration*. In the actual data, we restrict counts to be between $LB = 0$ and $UB = 9$. Between any two configurations, there are one or more mutational paths. We assume that mutations may result in gain/loss of single genes (SD), gain/loss of one copy of each gene on a common chromosome (CD), and duplication of all genes in the full genome (GD).

SD gain/loss events for each gene, CD gain/loss events for each chromosome, and GD events are each assigned a distinct nonnegative cost, or *weight*. The weight for a particular event is derived from the probability p of observing that event by the rule $w = -\log p$. Thus, forming a minimum weight tree maximizes the log likelihood of the events in the tree. Forming a phylogenetic tree based on FISH data involves three tasks: estimating probabilities of each type of event; using the estimated probabilities to efficiently estimate the minimum-weight (maximum likelihood) path between pairs of configurations; and finding an approximate minimum-weight phylogenetic tree, possibly containing Steiner nodes that represent unobserved or extinct configurations.

Algorithm 7 Infers the rates of each of SD, CD and GD event iteratively using statistics from tumor phylogenies. GENERATESTEINERTREE() uses Algorithm 10 to infer Steiner trees based on the set of cell states and parameter values.

1: **function** ESTIMATEPARAMETERS($\mathcal{N}_0$, p_0, ϵ, max_iter)
2: $k \leftarrow 1$
3: **while** true **do**
4: $\mathcal{T}_k \leftarrow$ GENERATESTEINERTREE($\mathcal{N}_{k-1}, p_{k-1}$)
5: c $\leftarrow$ EDGETYPECOUNTS($\mathcal{T}_k$)
6: $p_k^i \leftarrow (1 + c_i)/\sum_j (1 + c_j)$ **for** $i = 1, \text{length}(c)$
7: **if** $k = \text{max\_iter}$ or $\sum_i |p_k^i - p_{k-1}^i| \leq \epsilon$) **then**
8: **return** $p_k, \mathcal{T}_k$
9: $\mathcal{N}_k \leftarrow \text{nodes}(\mathcal{T}_k)$
10: $k \leftarrow k + 1$
11: **end function**

4.1.1 Algorithms

Estimating rate parameters

We apply an Expectation Maximization (EM)-like algorithm, presented as Algorithm 7, to identify the rate (probability) of each possible SD, CD and GD event. We initialize the method with uniform probability estimates, effectively leading to unweighted parsimony. Then, at each iteration of the algorithm, we infer a minimum-weight directed Steiner tree (applying Algorithm 10) using the parameter values inferred at the previous iteration. We treat this as the E-step of the algorithm. This step is simplified relative to strict EM in that it uses a single optimal model fit, rather than an expectation over the solution space in the E-step as in our prior work [134], but should yield comparable results to true EM in the limit of large numbers of tree edges. In the M-step, we then update the parameter values for each event based on the fraction of times that event is inferred across the tree edges, with the addition of a pseudocount of 1 (line 6) to account for events with inferred counts of 0.

Algorithm 8 CALCULATEMINCOST computes the minimum cost of converting a copy number profile $C^s(g_1, \ldots, g_d)$ to another copy number profile $C^t(g_1, \ldots, g_d)$ using combinations of SD, CD and GD events. $D_w^{s,ch}$ provides the minimum cost of an SD+CD path, as computed by Algorithm 9. B is a table providing duplication points of minimum-weight SD+GD paths, whose construction is discussed in more detail in section 4.1.2.

function CALCULATEMINCOST(C^s, C^t, B)
    return $\min\{\text{SINGLEPATHCOST}(C^s, C^t, B, k) \mid k \leftarrow 0, m\}$
end function
function SINGLEPATHCOST(C^s, C^t, B, k)
    $\text{cost} \leftarrow 0$
    for each set of genes $(g_q, \ldots, g_r)$ on the same chromosome **do**
        $\text{path} \leftarrow \text{DUPLICATIONPATH}(B, k, C^s, C^t, g_q, \ldots, g_r)$
        $\text{cost} \leftarrow \text{cost} + \text{sum}\left\{D_w^{s,ch}(\text{path}(i,\cdot), \text{path}(i+1,\cdot)) \mid i \leftarrow 1, k+1\right\}$
    return cost
end function
function DUPLICATIONPATH(B, k, C^s, C^t, g_q, …, g_r)
    for $p \leftarrow 1, r-q+1$ **do**
        $i \leftarrow \text{path}(1,p) \leftarrow C^s(g_{p+q-1})$
        $j \leftarrow \text{path}(k+2,p) \leftarrow C^t(g_{p+q-1})$
        for $\ell \leftarrow k+1$ downto 2 **do**
            $j \leftarrow \text{path}(\ell,p) \leftarrow B(i,j,\ell-1)$
end function

Cost estimation

We present here novel algorithms for estimating the cost of the minimum weight SD+CD+GD path between two configurations. Proofs of correctness of all of the claims in this subsection are provided in section 4.1.2 along with a special case for handling probes with zero copy number.

Our algorithms are motivated by the observation that for two configurations, and for a fixed number k of genome duplication events, a shortest-length SD+GD path may be quickly generated. In short, the condition that the SD+GD path be of minimal length requires that GD events be taken as late as possible to minimize the total number of SD+GD events. For a fixed k, only a limited number of duplication points need to be considered. Moreover, because duplication increases copy number exponentially, it suffices to consider paths with $0, 1, 2, \ldots, m$ duplication events, where $m = \lceil \log_2(UB) \rceil$. In our code, $UB = 9$, so $m = 4$.

Using an algorithm to compute a shortest length SD+GD path, and an algorithm for finding

an optimal SD+CD path between any two configurations, one may obtain a minimum-weight SD+CD+GD path. This process is presented as Algorithm 8, which works in three steps. In the first step, we calculate the shortest-length SD+GD path between configurations C^s and C^t for paths having $k = 0, \ldots, m$ genome duplication events. Each of these paths defines a set of zero or more genome duplication points, configurations at which a genome duplication occurred. In the second step, we connect the endpoint of each genome duplication point with the start point of the next, or with C^t for the last genome duplication, using minimum-weight SD+CD paths. In the third and final step, we choose the lowest-weight of these five SD+CD+GD paths; ties are irrelevant because we only need the path's weight.

It remains to define an algorithm to identify a minimum-cost SD+CD path between two configurations, which we denote s and t. Because SD and CD steps on separate chromosomes may be reordered, it suffices to consider the case in which all genes are on one chromosome. The algorithm for computing the SD+CD distance is centered around the concept of a *zigzag subpath*, which is so named because its construction focuses on alternations between consecutive gain and loss events. An optimal SD+CD path may start with a series of zigzag (subpath) steps to an intermediate state r and end with an SD path from r to t. Because the weight of an SD path is trivial to compute, we need only define a subroutine to determine the cost of an initial, possibly zero-length zigzag path and its endpoint. The pseudocode is presented as Algorithm 9, which defines an input variable σ (for *sense*) that takes the value -1 to indicate a zigzag loss (a zigzag subpath having only CD losses) or 1 to indicate a zigzag gain. At most one of $\sigma = 1$ or $\sigma = -1$ may result in a beneficial series of zigzag steps and a nonzero cost (section 4.1.2). Therefore, the two values for σ may be tried in successive uses of Algorithm 9.

Constructing a phylogenetic tree

In Algorithm 7, the E step involves generating phylogenetic trees, using Algorithm 8 as the key subroutine of a heuristic median-joining-based algorithm for inference of Steiner nodes in the tu-

Algorithm 9 Compute an optimal zigzag path of sense σ from s on the way to t.

Require: Start point s, end point t, sense $\sigma \in \{-1, 1\}$, the cost γ of a CD step of sense σ, a vector a representing the weight of SD steps of sense σ, and a vector b representing the cost of SD steps of the opposite sense.

Ensure: On exit, r is the endpoint of the zigzag path and 'fullcost' is the cost of the path.

$r \leftarrow s$
fullcost $\leftarrow 0$
while true **do**
    cost $\leftarrow \gamma$
    benefit $\leftarrow 0$
    for $k \leftarrow 1, \text{size}(t)$ **do**
        if $r_k \neq 0$ **and** $\sigma(r_k - d_k) \geq 0$ **then**
            cost $\leftarrow$ cost $+ b_k$
        else if $r_k \neq 0$ **then**
            benefit $\leftarrow$ benefit $+ a_k$
    if cost $\geq$ benefit **then return**
    for $k \leftarrow 1, \text{length}(t)$ **do**
        if $r_k \neq 0$ **and** $\sigma(r_k - d_k) < 0$ **then**
            $r_k \leftarrow r_k + \sigma$
    fullcost $\leftarrow$ fullcost $+$ cost

Algorithm 10 Main steps in the algorithm to generate tumor progression trees with particular rates for each of the SD, CD and GD events.

function GENERATESTEINERTREE($\mathcal{N}$, p)
    $\mathcal{T} \leftarrow$ DIRECTEDMINIMUMSPANNINGTREE($\mathcal{N}, p$)
    for all $(u, v, w) \leftarrow$ TRIPLETS($\mathcal{T}$) **do**
        for all $s \leftarrow$ LATTICEPOINTS(u, v, w) **do**
            if $s \notin \mathcal{T}$ **then**
                if TREEWEIGHT($\mathcal{T}, p$) $>$ TREEWEIGHT(MEDIANJOIN($\mathcal{T}, s$), p) **then**
                $\mathcal{T} \leftarrow$ MEDIANJOIN($\mathcal{T}, s$)
    return $\mathcal{T}$
end function

mor phylogenies. The key steps of this tree-building algorithm are summarized in Algorithm 10. The code builds a directed minimum spanning tree $\mathcal{T}$ based on the observed cell types. It then iterates over each node triplet in $\mathcal{T}$ for which one node is the parent of the other two nodes. We define the lattice points of a triplet to be the set of configurations that agree in each dimension with at least one of the triplet. Each lattice point is considered in arbitrary order as a possible Steiner node. If a lattice point is not already in $\mathcal{T}$, a new tree, called the median tree, is created

by adding that lattice point as a node and by connecting it via an edge to all three points in the triplet. If the resulting weight, caculated as explained in the previous subsection, is less than the weight of the previous best tree, the lattice point is added to the tree. The best tree found by this procedure is returned.

4.1.2 Theoretical analyses

In this section, we provide proofs of correctness for the algorithms presented in the previous section. Our main theoretical result is a method for inferring minimum distances between two states within a copy number phylogeny when duplication/loss of single genes (SD), duplication/loss of all genes on a common chromosome (CD), and duplication of all genes in the full genome (GD) are possible and each event type is associated with a weight parameter. We first establish some mathematical results and then develop an algorithm for accurate distance computation. This algorithm then becomes a subroutine in a heuristic Steiner tree algorithm for inferring copy number phylogenies in the presence of weighted SD, CD, and GD events. Finally, we develop an iterative algorithm to infer the rate of different event types from the observed data using the weighted Steiner tree inference algorithm as a subroutine.

In what follows, we will consider sequences of SD, SD+CD, SD+GD and SD+CD+GD events, which we call *paths*. A *boundary-insensitive* path is one in which the copy numbers of intermediate configurations can take on any integer values and for which zero copy number is not treated specially. Zero is special, however, because once the copy number of a gene is reduced to zero, it is generally assumed that the gene cannot be gained back to get to copy number one. We also define *boundary-sensitive* paths for which intermediate copy-numbers must lie between positive bounds, denoted by LB and UB. When we present pseudocode for computing paths, we will discuss how zero copy number is handled.

We introduce some notation required for specifying and proving the theoretical results:

1. The observed data consists of copy-number counts of probes g_i for $i = 1, \ldots, d$. Typically,

each probe allows one to count the copy number of a particular gene, so we refer to the g_i as *genes*.

2. We define a *configuration* $C(g_1, g_2, \ldots, g_d)$ to be a vector of length d of integers representing the copy numbers of each gene. A configuration is the state that might be observed for a single cell. When the collection $g_1, \ldots, g_d$ is clear from context, we will just write C as a shorthand.
3. $w_{g_i}^{\{g,l\}}, w_{c_i}^{\{g,l\}}, w_d$: Cost/weight of gain ($w_{g_i}^{g}$) or loss ($w_{g_i}^{l}$) associated with individual gene g_i, individual chromosome c_i ($w_{c_i}^{\{g,l\}}$) or cost/weight of whole genome duplication event (w_d). The weight for a particular event is derived from the probability p of observing that event by the rule $w = -\log p$.
4. We denote the length of the shortest-length boundary-insensitive SD path between configurations C^i and C^j by $L_1(C^i, C^j)$. As we discuss below, the length is precisely the rectilinear, or L_1, distance between the configurations, justifying the notation.
5. The weight of the minimum-cost SD path between C^i and C^j is denoted $R^w(C^i, C^j)$.
6. We let $D_w^{s,ch}(C^i, C^j)$ denote the weight of the minimum-cost boundary-insensitive SD+CD path between C^i and C^j.

A *feasible* configuration is one in which all counts are between LB and UB. A feasible path consists entirely of feasible configurations. An *infeasible* path has at least one infeasible configuration. Every infeasible path is boundary-insensitive, but a boundary-insensitive path may be either feasible or infeasible.

SD and SD+CD events have the desirable property that the order of SD or SD+CD events can be rearranged arbitrarily [35]; such a property does not hold for paths with GD events. In our previous work, we established the following two lemmas for the unweighted SD and CD cases [35]:

Lemma 14. *A shortest unweighted boundary-insensitive sequence of CD and SD events cannot have both a gain of chromosome c_i and a loss of the same chromosome c_i.*

Lemma 15. *For any gene g_i, a shortest unweighted boundary-insensitive sequence of events cannot have both a gain of g_i and a loss of g_i.*

We now prove that the natural generalizations of the two lemmas hold in the weighted case too:

Lemma 16. *A minimum-weight boundary-insensitive sequence of weighted CD and SD events cannot have both a gain of chromosome c_i and a loss of the same chromosome c_i.*

Proof. By contradiction. Suppose S is a sequence of events that has both a gain and a loss of the same chromosome. Then removing one gain and one loss produces a new sequence that weighs $(w^g_{c_i} + w^l_{c_i})$ less and has the same final state. □

Lemma 17. *For any gene g_i, a minimum-weight boundary-insensitive sequence of weighted events cannot have both a gain of g_i and a loss of g_i.*

Proof. By contradiction. Suppose S is a sequence of events that has both a gain of g_i and a loss of g_i. Then removing one gain and one loss produces a new sequence that weighs $(w^g_{g_i} + w^l_{g_i})$ less and has the same final state. □

From these lemmas, it follows that the length of the shortest-length SD path between configurations C^i and C^j is precisely the rectilinear distance, justifying our use of the notation $L_1(C^i, C^j)$. In contrast, $R^w(C^i, C^j)$ is not a distance because it is not symmetric, but it can be expressed as a simple sum

$$R^w(C^i, C^j) = \sum_{C^i(g_k)<C^j(g_k)} (C^j(g_k) - C^i(g_k))w^g_{g_k} + \sum_{C^i(g_k)>C^j(g_k)} (C^i(g_k) - C^j(g_k))w^l_{g_k}$$

For other types of path, the length of the shortest-length paths and the cost of the minimum-weight path are not so easily expressed. Developing algorithms to compute weights of minimum-weight paths is the topic of the rest of this section.

Progression model considering SD and CD events

One of our main theoretical contributions consists of novel theory for inference of minimum-weight paths of single-gene (GD) and single chromosome (CD) events from a starting configuration $C^s(g_1, g_2, \ldots, g_d)$ to a terminal configuration $C^t(g_1, g_2, \ldots, g_d)$. Our model assumes that on division of a tumor cell, the configuration can change either by gain or loss of one copy of a single gene (SD event) or by gain or loss of one copy of each gene on a single chromosome (CD event). For example, a configuration of four genes $(2, 2, 2, 2)$ with the first two genes on the same chromosome might evolve to $(3, 2, 2, 2)$ by a single SD event or to $(3, 3, 2, 2)$ by a single CD event.

First, we establish the correctness of Algorithm 9, which calculates the minimum weight sequence of SD and CD events to transform $C^s(g_i, g_{i+1}, \ldots, g_j)$ into $C^t(g_i, g_{i+1}, \ldots, g_j)$. We focus on a single chromosome because, as explained below, the problem of finding the minimum-weight SD+CD path can be solved one chromosome at a time. In the case in which there is only data for single gene probe on a chromosome, one cannot distinguish CD events from SD events. We treat such cases mathematically by setting the weight of the corresponding CD events to infinity. In practice, one may simply calculate an SD path for those probes.

Thus far, we have discussed boundary-insensitive paths, which might contain problematic intermediate cases with zero copy-number or absurd cases with negative copy number. However, if one finds an optimal boundary-insensitive path, one may construct an optimal boundary-sensitive path of the same weight.

Theorem 18. *If there is a minimal-weight boundary-insensitive sequence of SD and CD events between two feasible configurations C^s and C^t, where the smallest feasible copy number is at least one, then there is a boundary-sensitive sequence of SD and CD events with the same weight.*

Proof. Assume without loss of generality that all genes are on the same chromosome. If the boundary-insensitive path contains no CD event then it is an SD path and all gene counts change monotonically, so the theorem holds.

We will suppose that all the CD events are chromosome losses; the argument for CD gains is symmetric. If placing a chromosome loss next on the path would result in an infeasible configuration, then the copy number of some genes must be at LB. If these genes did not have any SD gains in the path, then the ultimate configuration would be infeasible, which by assumption is not. Therefore, we place a single SD gain for each of the genes with copy number at the LB next on the path in arbitrary order. Adding these SD gains cannot produce an intermediate configuration that is infeasible because the genes are at copy number LB, and these gains will just increase the count for each gene by one.

Next, we place the chromosome loss on the path. The resulting configuration is feasible. We repeat the process until there are no more chromosome losses. By construction, once the final chromosome loss has been placed, we attain a feasible intermediate configuration via a feasible path, and all further events are SD gains. Thus, thereafter, we only have monotonic gene gain/losses between two feasible configurations. Thus, the entire path is feasible. □

In the Results section of this chapter, we analyze data with minimum copy number (LB) zero and maximum copy number (UB) nine, but in this subsection we assume positive copy number. Zero copy number is handled as a special case in a later subsection.

Now, we prove the following results, which exclude the possibility of gains and losses of the same chromosome or the same gene on the boundary-sensitive paths.

Corollary 19. *In a feasible optimal path, where the smallest feasible copy number is at least one, there cannot be both a gain and loss of the same chromosome or a gain and loss of the same gene.*

Proof. The proof is by contradiction. Take an optimal feasible path, which is by definition comprised of a boundary-sensitive sequence of operations. It is also a valid boundary-insensitive path. If it contains paired gain/losses, we can rearrange and cancel to produce a lower weight path. By the previous Theorem, there is also a boundary-sensitive path with this lesser weight. Hence, the original boundary-sensitive path could not have been optimal. □

Algorithm for computing the SD+CD distance

The algorithm for computing the SD+CD distance is centered around the concept of a *zigzag subpath* from C^s to C^t, which is so named because its construction focuses on alternations between consecutive gain and loss events. For any *zigzag* path, we first choose a predominant *sense* or *direction* as either *gain* or *loss*, where the sense of the zigzag path is the sense of the CD events in that path. When the *zigzag* sense is loss, then we determine the set of genes that are on the affected chromosome and for which the copy number at C^s is less than or equal to the copy number at C^t. For each such gene, we insert an SD gain in the path in arbitrary order. Then we insert a single CD loss. We define the path symmetrically when the sense of the zigzag path is gain. We use the shorthand "zigzag gain" (respectively, "zigzag loss") to refer to a zigzag path of sense gain (loss).

Thus, a zigzag path is a series of zero or more SD changes followed by a single CD change with opposite sense (gain/loss). The sense of a zigzag step is the sense of the final CD event.

Lemma 20. *There is a CD step on an optimal SD+CD path from C^s to C^t if and only if there is a (possibly different) optimal path from C^s on the way to C^t that starts with a zigzag subpath of the same sense, affecting the same chromosome.*

Proof. We consider a CD loss and argue symmetrically for a CD gain. If there is a CD loss on the optimal path, then by Lemma 19, the only CD events on the optimal path are CD losses. We choose one such CD loss and let that choice determine the affected chromosome.

If the copy number of a gene at C^s is less than or equal to the copy number of that gene at C^t, then the path must contain at least one SD gain for that gene. Thus, one may rearrange the optimal path to create an equal-weight path that starts with a zigzag loss.

The converse is true because a zigzag subpath contains a CD step, by definition. □

Lemma 21. *Consider a specific zigzag path. Let ℓ be a vector such that $\ell_k = 1$, if the copy number of gene g_k changes after the zigzag path and zero otherwise. Let m be another vector for*

which $m_k = 1$ if gene g_k is on chromosome affected by the CD step, but the copy number of gene g_k does not change after the entire zigzag path. Then,

$$\ell_k + m_k = \begin{cases} 1 & \textit{if gene } g_k \textit{ is on the chromosome lost or gained} \\ 0 & \textit{otherwise.} \end{cases}$$

Moreover, after a zigzag loss, the copy number of every gene is the same or lower, and after a zigzag gain, the copy number of every gene is the same or higher.

Proof. This Lemma is a consequence of the definition of zigzag paths and their senses. If a gene is on the chromosome affected by a the CD step, but is not matched by a corresponding SD step of the opposite sense, the copy number of the gene must change. The sense of any change in gene copy number is the same as the sense of the CD step. □

Lemma 22. *If C^{int} is an intermediate configuration created by taking a zigzag step of either sense from C^s on the way to C^t,*

$$L_1(C^s, C^t) = L_1(C^{int}, C^t) + \sum_k \ell_k,$$

where ℓ is defined as in Lemma 21.

Proof. We consider only zigzag losses; one can argue symmetrically for zigzag gains. By definition of a zigzag loss, exactly those genes with copy number greater at C^s than C^t have a copy number change after the zigzag loss, and the copy number of each of these genes decreases by one. But those genes are precisely the genes with indices k such that $\ell_k = 1$. □

Lemma 23. *If C^{int} is an intermediate configuration reached by taking a zigzag path from C^s on the way to C^t, then*

$$R^w(C^s, C^t) = R^w(C^{int}, C^t) + a^T \ell,$$

where ℓ is defined as in Lemma 21, and a is a vector for which a_k represents the cost of an SD step for gene g_k of the same sense as the zigzag path.

Proof. By definition of a zigzag loss, exactly those genes with copy number greater at C^s than C^t have a copy number change after the zigzag loss, and the copy number of each of these genes decreases by one. By Lemma 22, those genes are precisely the genes for which $\ell_k = 1$. For a zigzag loss, then a_k represents the weight of the loss of gene g_k. Thus, the weight of an optimal SD subpath from C^{int} to C^t differs from the weight of an optimal SD subpath from C^s to C^t by exactly $a^T\ell$.

We can argue symmetrically for zigzag gains. □

Next, we develop the rule indicating when a CD and a zigzag step is possible.

Theorem 24. *When there is no SD+CD path between C^s and C^t of strictly lower weight than an optimal SD path between C^s and C^t, a CD step from C^s will result in an intermediate configuration C^{int} for which*

$$R^w(C^{int}, C^t) + w_c \geq R^w(C^s, C^t),$$

where w_c is the cost of a CD step with the same sense as the zigzag path.

Proof. The weight of a CD event is precisely w_c. One may then take an SD path from C^{int} to C^t for total path weight $R^w(C^{int}, C^t) + w_c$. If this is strictly less than the weight of an SD path from C^s to C^t, then this constitutes an SD+CD path that has lower weight than the optimal SD path. □

A similar result holds for zigzag paths. By Lemma 20, if there is a CD step on the *optimal* path, the path may be rearranged so that there is a zigzag path.

Lemma 25. *When there is no SD+CD path between C^s and C^t of strictly lower weight than an optimal SD path between C^s and C^t, a zigzag step of the same sense from C^s will result in an*

intermediate configuration C^{int} for which

$$R^w(C^{int}, C^t) + w_c + b^T m \geq R^w(C^s, C^t), \tag{4.1}$$

where b is a vector such that b_k is the cost of an SD step affecting gene g_k in the sense opposite to the sense of the zigzag path. For instance, if the path is a zigzag loss, b_k is the cost of an SD gain of gene g_k.

Proof. The sum of the weights of the SD steps and the CD step that make up the zigzag step is precisely $w_c + b^T m$. One may then take an SD path from C^{int} to C^t for total path weight $R^w(C^{int}, C^t) + w_c + b^T m$. If this is strictly less than the weight of an SD path from C^s to C^t, then this constitutes an SD+CD path that has lower weight than the optimal SD path. □

Theorem 26. *When there is an SD+CD path between C^s and C^t that has lower weight than an optimal SD path, then for any CD step on the SD+CD path, taking a zigzag step from C^s on the way to C^t of the same sense and affecting the same chromosome results in an intermediate configuration C^{int} for which*

$$R^w(C^{int}, C^t) + w_c + b^T m < R^w(C^s, C^t), \tag{4.2}$$

where m is defined as in Lemma 21 and b is defined as in Lemma 25.

Proof. The proof is by induction on $L_1(C^s, C^t)$. If $L_1(C^s, C^t) = 0$, then C^s and C^t are the same configuration and $R^w(C^s, C^t) = 0$. Any nonempty SD+CD path has nonnegative weight, and so cannot have lower weight than the optimal SD path, and the claim holds.

Take as induction hypothesis that the claim holds whenever $L_1(C^s, C^t) < n$. Now let $L_1(C^s, C^t) = n$. Let C^{int} be an intermediate point such that $L_1(C^{int}, C^t) < n$, so that the induction hypothesis applies to paths between C^{int} and C^t.

To simplify the exposition, assume without loss of generality that all genes are on the same

chromosome. Furthermore, let us consider the case in which an SD+CD path between C^s and C^t containing a CD loss has lower weight than an optimal SD path; the argument for paths contain a CD gain is symmetric.

Let C^{int} be an intermediate configuration generated by a zigzag loss from C^s on the way to C^t. Since we assume inequality (4.2), taking the zigzag path must result in a change in the copy number of at least one gene. Thus, by Lemma 22, it must be that $L_1(C^{int}, C^t) < n$. Thus, one may apply the induction hypothesis on paths from C^{int} to C^t.

Suppose there is no SD+CD path from C^{int} to C^t that has lower weight than the optimal SD path. In such a case, the SD path is itself optimal as an SD+CD path from C^{int} to C^t. Thus, any path constrained to start with a zigzag path from C^s to C^{int}, and continuing on to C^t, has weight at least $R^w(C^{int}, C^t) + w_c + b^T m$. But by the assumptions of the theorem, there is an SD+CD path containing a CD loss and having weight less than the optimal SD path, and by Lemma 20 such a path may be constrained to start with a zigzag loss. Thus, inequality (4.2) holds.

Suppose, therefore, that there is an SD+CD path from C^{int} to C^t that has weight less than that of the optimal SD path and that contains a CD loss. Then by Lemma 23,

$$R^w(C^s, C^t) = R^w(C^{int}, C^t) + a^T \ell \tag{4.3}$$

Let C^u be another intermediate configuration, generated by taking another zigzag path from C^{int} of the same sense. It holds that

$$R^w(C^{int}, C^t) = R^w(C^u, C^t) + a^T \widehat{\ell}, \tag{4.4}$$

where $\widehat{\ell}$ is a vector for which $\ell_k = 1$ for each gene whose copy number decreases between C^t and C^u. By the induction hypothesis

$$R^w(C^u, C^t) + w_c + b^T \widehat{m} < R^w(C^{int}, C^t), \tag{4.5}$$

where $\widehat{m}$ is a vector representing the SD gains in the zigzag step.

It must be that

$$a^T\widehat{\ell} \leq a^T\ell \text{ and } b^T m \leq b^T\widehat{m}, \tag{4.6}$$

because the nonzero entries of ℓ denote the set of genes with copy number strictly greater at C^s than C^t, and the nonzero entries of $\widehat{\ell}$ denote the set of genes with copy number strictly greater at C^{int} than C^t. But copy numbers can only decrease after a zigzag loss. Furthermore, by Lemma 21, $\ell_k + m_k = \widehat{\ell}_k + \widehat{m}_k = 1$, for all genes g_k on the relevant chromosome.

Combining (4.3)–(4.6), we find

$$\begin{aligned}
R^w(C^{int}, C^t) + w_c + b^T m &= R^w(C^u, C^t) + w_c + b^T m + a^T\widehat{\ell} && \text{by (4.4)} \\
&\leq R^w(C^u, C^t) + w_c + b^T\widehat{m} + a^T\ell && \text{by (4.6)} \\
&< R^w(C^{int}, C^t) + a^T\ell && \text{by (4.5)} \\
&= R^w(C^s, C^t) && \text{by (4.3)}
\end{aligned}$$

completing the induction step and the proof. □

Now, we develop the rule to identify which direction of zigzag step, if any, needs to be applied on the optimal boundary-sensitive path to convert C^s to C^t. We use the following shorthand notation for four possible partial paths:

1. ZZL: the zigzag loss path.
2. ZZG: the zigzag gain path.
3. SDL: SD losses of those genes that have a lower copy number in C^t than in C^s.
4. SDG: SD gains of those genes that have a higher copy number in C^t than in C^s.

Lemma 27. *Taking either ZZL or SDL leads to the same intermediate state. Taking either ZZG or SDG leads to the same intermediate state.*

Proof. The definition of ZZL is that it has one CD loss followed by SD gains of the genes for

which the copy number at C^{int} is less than or equal C^t. Those genes have compensating losses and gains. In ZZL, the remaining genes, which have a higher copy number in C^s than in C^t have SD losses. That is the definition of SDL. The proof for ZZG and SDG is symmetric. □

The weight of ZZL is the sum of the cost of a CD loss and the sum of costs of SD gains of genes that have lower or equal copy number in C^s than in C^t. The weight of the SDL subpath is the sum of costs of single SD loss events for those genes which have lower copy number in C^t than in C^s. The weights of the ZZG and SDG partial paths are defined analogously. In the following theorem, we propose two alternative tests to identify the sense of the zigzag partial path to use, if any, as the value of the input paranter σ in Algorithm 9.

Theorem 28. *At most one of the following tests can be successful:*

1. *If the cost of ZZL is lower that that of SDL, take ZZL by setting* $\sigma = -1$.
2. *If the cost of ZZG is lower that that of SDG, take ZZG by setting* $\sigma = 1$.

Proof. The proof is by contradiction. Sort the costs of the four partial paths in increasing order, breaking ties arbitrarily. If both tests 1 and 2 succeed, then the most costly of the four paths must be SDL or SDG. Without loss of generality, assume the most expensive path is SDG, so that both $cost(ZZG) < cost(SDG)$ and $cost(ZZL) < cost(SDG)$. The path SDG has a proper subset of the single-step events in the path ZZL. Therefore, it is not possible that $cost(ZZL) < cost(SDG)$, a contradiction. □

We use the proof of correctness of Algorithm 9 to derive the main theorem of this subsection, which establishes a method to find a minimum-cost sequence of weighted SD and CD events for transforming C^s to C^t. Again, we can consider each chromosome separately since each CD and GD event affects only one chromosome.

Theorem 29. *Assume we partition the gene list by chromosomes such that each chromosome* $c_i \in \{c_1, \ldots, c_q\}$ *corresponds to a consecutive subset of genes* $g_{i,1}, \ldots, g_{i,d_i}$. *Further define* $C^s(g_1, g_2, \ldots, g_d) = (s_1, \ldots, s_d)$ *and* $C^t(g_1, g_2, \ldots, g_d) = (t_1, \ldots, t_d)$. *Then we can con-*

struct a minimum-cost boundary-sensitive sequence of events transforming $C^s(g_1, g_2, \ldots, g_d)$ *to* $C^t(g_1, g_2, \ldots, g_d)$ *by constructing a minimum-cost boundary-sensitive sequence of events* S_i *transforming* $(s_1, \ldots, s_{i,1}, \ldots, s_{i,d_i}, \ldots, s_d)$ *to* $(s_1, \ldots, t_{i,1}, \ldots, t_{i,d_i}, \ldots, s_d)$ *for each chromosome* c_i *and interleaving each* S_i *in arbitrary order.*

Proof. The distance function can be decomposed into individual parts for genes belonging to distinct chromosomes as follows:

$$D_w^{s,ch}(C^s, C^t) = \sum_{i=1}^{q} D_w^{s,ch}(C^s(s_{i,1}, \ldots, s_{i,d_i}), C^t(s_{i,1}, \ldots, s_{i,d_i}))$$

Because the distance can be decomposed in this way and each CD or SD event contributes to only a single term of the outer sum, we can minimize the cost for each chromosome independently and combine the events from distinct chromosomes in arbitrary order without changing the value of the objective function. Likewise, since these each chromosome affects a disjoint subset of genes, boundary-sensitive sequences for each chromosome will yield a boundary-sensitive sequence across all genes. □

Progression model considering SD, CD and GD events

In this subsection, we provide additional methodological details and a proof of correctness for a method for finding minimum-cost SD+GD paths, which we call Algorithm 11. This algorithm is called as a subroutine in the full SD+CD+GD method.

For any fixed number of genome duplication events, the shortest-length SD+GD path between two configurations with the specified number of GD events may be computed one component at a time. For each component, the shortest length path may be found by adding SD losses or gains preceeding and following genome duplication events to choose the most favorable duplication events given the starting and ending copy number. But for a single component, shortest-length SD+GD paths have a well-defined structure.

Theorem 30. *For a fixed number $g \geq 1$ of GD steps, the minimum-weight path between $i \geq 1$ and $j \geq 1$ must end with either a GD event, if j is even, or a GD event followed by a single SD event, if j is odd. The result holds whether or not the SD events must have the same sense, or may be mixed.*

Proof. Consider the final GD event in the minimum-weight path and the SD events that follow it, necessarily of all the same sense (otherwise one would cancel a gain and a loss to get a shorter path). Suppose there are two or more SD events following the GD event. If the SD events after the GD event are gains, or the SD events after the GD events are losses but the start point of the GD event is greater than 1, then one could create a minimum-weight path by inserting an SD event of the same sense before the GD event, using that new point as the start point of a GD event and eliminating two of the SD events that follow it. Because we replaced two SD events with an SD event of the same sense, it is irrelevant whether SD events are restricted to a single sense.

The remaining case is when the GD event is from 1 to 2, but is followed by more than one SD loss. Since counts cannot go below zero, this scenario entails that $j = 0$, contrary to the assumptions of this theorem. Therefore, a minimum-weight path must end with a GD event followed by at most one SD event.

Since the endpoint of a GD event must be an even number, if j is an even number, it would not be possible to arrive at j by following a GD event with a single SD event. Thus, if j is even, it must be the endpoint of the GD event. Similarly, if j is odd, a single SD event must follow the final GD event. □

The preceding theorem suggests an algorithm for finding the duplication points in the shortest SD+GD path between copy numbers i and j for a fixed number k genome duplications. One need only start at j, consider the one or two possible duplication points for the last GD event in a shortest-length SD+GD path terminating at j, and then choose the better of the duplication points by finding the shortest-length SD+GD path of length $k-1$ between C^s and the duplication point. An immediate corollary of the theorem is that for a maximum copy number of nine, which

Algorithm 11 Fill tables representing a shortest length SD+GD path for a single gene.

Require: A number n representing the maximum copy number of a gene and m representing the maximum number of GD events to be considered.

Ensure: On exit, $B(i,j,k)$ contains the k^{th} duplication point and $L(i,j,k)$ the number of SD gains and losses for the shortest length SD+GD path from i to j, with the constraint that k duplications occur.

```
1: for k ← 0, m do
2:     for i ← 1, n do
3:         for j ← 1, n do
4:             if k = 0 then
5:                 L(i, j, 0) ← max(i − j, j − i)
6:             else if j is even then
7:                 B(i, j, k) ← j/2
8:                 L(i, j, k) ← L(i, j/2, k − 1)
9:             else
10:                lower ← L(i, (j − 1)/2, k − 1) if j > 1 otherwise ∞
11:                upper ← L(i, (j + 1)/2, k − 1) if j < n otherwise ∞
12:                if lower < upper or (lower = upper and i ≤ j) then
13:                    B(i, j, k) ← (j − 1)/2
14:                    L(i, j, k) ← L(i, (j − 1)/2, k − 1) + 1
15:                else
16:                    B(i, j, k) ← (j + 1)/2
17:                    L(i, j, k) ← L(i, (j + 1)/2, k − 1) + 1
```

we use in our experiments, it suffices to consider at most four genome duplication events.

Corollary 31. *If $1 \leq j \leq 2^m$, there can be at most m GD steps in a minimum-weight SD+GD path between $i \geq 0$ and j.*

Proof. By induction on j. If $j = 1$ then there are no GD steps in a minimum-weight SD+GD path from i to j and the result holds. Otherwise, if j is even, then a minimum-weight path from i to j must end with a GD event starting at $k = j/2$. But then $k < 2^{m-1}$ and the induction hypothesis may be applied. Similarly, if j is odd, then the path must end with a GD step to $j-1$ or $j+1$, followed by an SD step. In either case, the GD step must start from $k \leq (j+1)/2$. Thus $k < 2^{m-1}$, and again the induction hypothesis applies. □

Because we restrict copy numbers to be at most nine, four genome duplication events is sufficient. For a fixed $k > m$, there is still a minimum length SD+GD path with k genome

duplications, but such a path must have a cycle where a copy number of 1 is duplicated by a GD event to become 2 which is followed by an SD loss to return to one. Though for a singe gene probe, a path containing a cycle would not be optimal, when a shortest-length SD+GD path is computed for several genes, some of the genes may exhibit such a cycle.

For $k = 0, \ldots, m$ (for our code, $UB = 9$ so $m = 4$), Algorithm 11 creates a table of duplication points for the shortest-length SD+GD path between copy numbers of an individual probe for given a fixed number of genome duplication events.

In Algorithm 11, cases in which the copy number of some gene is zero in C^s and C^t are special. When a copy number starts at zero, there is no simple biological mechanism for daughter cells to regain that gene. Thus, we do not attempt to calculate paths for which a copy number changes from zero to nonzero, but rather just assign all such paths an infinite weight. Cases in which the copy number end at zero are also special. In such cases, the optimal SD+GD path is to lose all copies of the gene and then cycle from 0 to 0 on all genome duplication events. For brevity, we exclude such cases from the pseudocode, though code to handle these cases is implemented in the FISHtrees software.

4.2 Results

We applied our parameter inference algorithms to FISH datasets on three kinds of human cancers: cervical cancer, breast cancer and oral (tongue) cancer. The datasets were described in section 1.7. Here, we reiterate the sample composition and the genes on which each dataset was profiled for clarity:

- Dataset CC [183] consists of paired primary and metastatic cervical cancer samples collected from 16 patients and primary samples collected from 15 patients whose tumors did not metastasize probed on four oncogenes residing on distinct chromosomes (*LAMP3, PROX1, PRKAA1* and *CCND1*).

- Dataset TC [Wangsa et al., submitted] consists of 65 single samples collected from tongue cancer patients probed for four genes located on distinct chromosomes (*TERC, EGFR, CCND1, TP53*), with tumor stages (ranging from 1 (least advanced) up to 4 (most advanced)) available on all patients and tobacco usage (a known risk factor), survival, and disease-free survival out to 73 months available on most patients.
- Dataset BC [83] consists of paired DCIS and IDC samples collected from 13 breast cancer patients. Each sample was probed on eight genes, out of which, *COX2, MYC, HER2, CCND1* and *ZNF217* are oncogenes and *DBC2, CDH1* and *P53* are tumor suppressors. Out of the eight genes, *DBC2* and *MYC* reside on chromosome 8 and *HER2* and *P53* reside on chromosome 17. Other genes reside on distinct chromosomes.
- An additional cervical cancer dataset, CC2, which consists of following set of cervical samples: (a) one early pre-cancerous lesion (denoted by CIN1), (b) ten late pre-cancerous lesions (denoted by CIN3) and (c) ten cancerous lesions (denoted by CA). Each sample was probed on eight genes: *COX, ING5, FHIT, TERC, TERT, MYC, CHEK1, ZNF217*. *ING5, FHIT, CHEK1* are tumor suppressors and all others are oncogenes. *FHIT* and *TERC* lie on chromosome 3, while the others all reside on distinct chromosomes.

BC and CC2 datasets allow us to test in practice the utility of modeling chromosome gain/loss (CD) events. In each case, the pair of collocated genes are one oncogene and one tumor suppressor, so our weighted models need to balance between CD events in which both genes have the copy number changing in the same direction and GD events, which would usually be gains for the oncogenes and usually be losses for the tumor suppressors. Each of the four datasets has a different study design from a clinical point of view. Yet, we can derive meaningful qualitative inferences, suitable to the particular study design, from our tumor progression models for all four data sets.

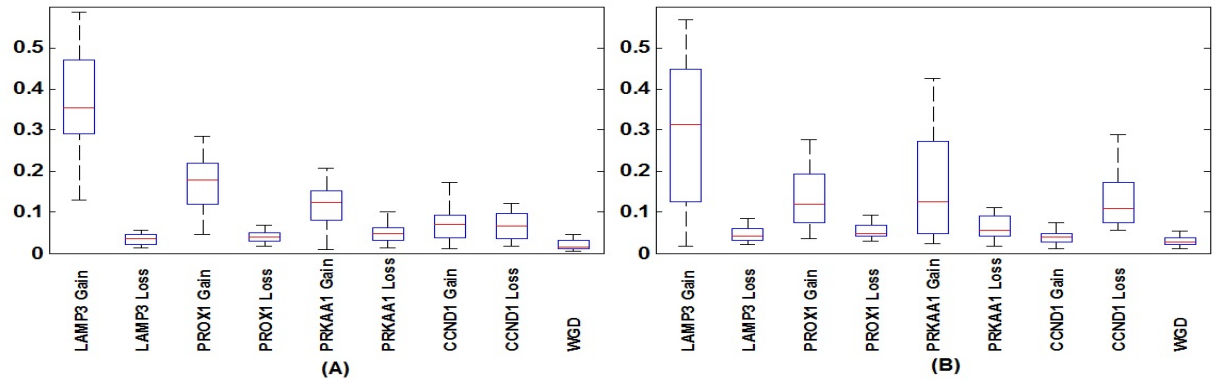

Figure 4.1: Inferred parameter values for primary (A) and metastatic (B) cervical samples. WGD refers to the rate of whole-genome duplications.

4.2.1 Identifying progression markers in cervical cancer (CC) data

We applied our algorithm on each of the samples in each of the datasets separately and inferred probabilities of each SD and GD event. We show the boxplots for the inferred parameter values in the CC dataset in Figure 4.1, across 31 primary (Figure 4.1(A)) and 16 metastatic (Figure 4.1(B)) samples. "Gain of *LAMP3*" is the most frequent event in both primary and metastatic samples, similar to the findings reported in our past work [34].

Next, for each pair of 16 primary and 16 metastatic samples, we performed a statistical test based on the "tree edge count" statistic, which quantifies, for each tumor phylogenetic tree, the total number of edges across which gain and loss of each gene is inferred. For each pair of samples and each gene, we built a 2×2 contingency table of computed gain events versus the total number of remaining events in primary versus metastatic trees and performed a chi-square test of independence on this table (or Fisher's exact test if any of the entries of the table was less than or equal to 10). We then calculated the total number of times each gene showed a statistically significant (with P-value < 0.05) difference in proportions between the primary and metastatic samples out of the 16 pairs. *LAMP3* most often produced significant results (9 times), followed by *PROX1* (5) and *PRKAA1* (2).

We examined whether the statistically significant results are consistently due to higher proportion of gain in primary or in metastatic samples. Out of nine statistically significant compar-

isons of *LAMP3*, seven were due to higher proportion of gain in metastasis and two were due to higher proportion of gain in primary samples. For *PROX1*, four significant comparisons were observed due to higher proportion of gain in primary samples and one was due to higher proportion of gain in metastatic samples. This result suggests that "Gain of *LAMP3*" is associated with the metastatic stage of cervical cancer, while "Gain of *PROX1*" is associated with the primary tumor.

4.2.2 Identifying progression markers in BC and CC2 data

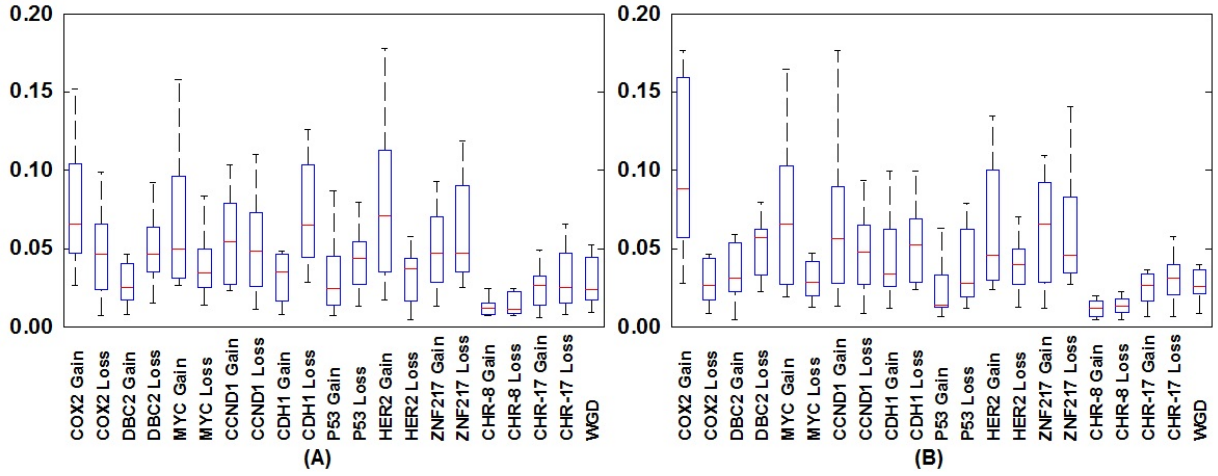

Figure 4.2: Inferred parameter values across ductal carcinoma in situ (DCIS) (A) and invasive ductal carcinoma (IDC) (B) samples.

We applied our tree building and rate estimation algorithm on the BC dataset. Boxplots for the estimated rates of each event are shown in Figure 4.2 for DCIS and IDC samples separately. As a first statistic, we ranked the different events across all of the 13 DCIS and 13 IDC samples separately based on their median parameter values. The most frequent events (with median parameter values ≥ 0.05) for both of the DCIS and IDC cases were "Gain of *COX2*", "Gain of *MYC*" and "Gain of *CCND1*". "Loss of *DBC2*" and "Gain of *ZNF217*" appeared as the most frequent events in the IDC samples only.

Similarly to the CC dataset, whose analysis is shown in the main document, we next per-

formed an edge count-based statistical test of separation of the DCIS and IDC samples based on the gain/loss values of individual genes. "Gain of *MYC*" (total 5 out of 13 pairs of samples), "Gain of *COX2*"(4) and "Gain of *CCND1*"(2) showed statistically significant separation of DCIS and IDC samples more times than any other event. This is interesting because "Gain of *MYC*" did not appear to have a significant effect in the progression dynamics of DCIS to IDC in our previous analysis using an unweighted single gene duplication model (Chapter 2), but was shown by other methods to have important effects during progression from DCIS to IDC [83]. Our current analysis, based on the more realistic model of tumor progression developed in the present work, supports the conclusion from [83].

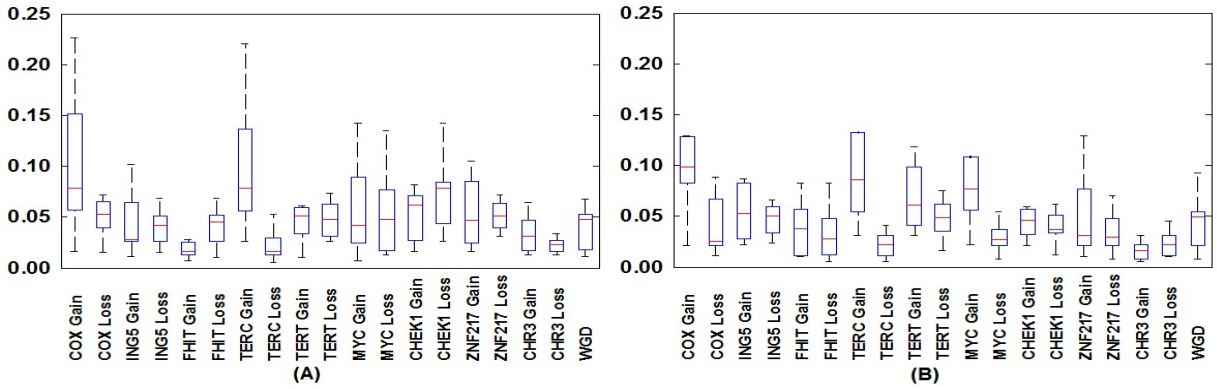

Figure 4.3: Inferred parameter values across pre-cancerous (A) and cancerous (B) cervical tumor samples.

We applied our algorithm on each of the 21 samples in the CC2 dataset. Boxplots for inferred parameter values across the pre-cancerous and cancerous samples separately are shown in Figure 4.3. Similarly to the BC dataset, we again ranked the events based on the median parameter values across all of the pre-cancerous and cancerous samples separately. The most frequent events (with median parameter values ≥ 0.05) across both of the pre-cancerous and cancerous samples were "Gain of *COX*" and "Gain of *TERC*". "Loss of *CHEK1*" appeared more frequently in the cancerous samples only.

4.2.3 Classification of cervical samples

We performed classification experiments using tree-based features to separate samples from different stages of cervical cancer in CC. We used these experiments to validate our models and demonstrate their value, based on our past observation that tree progression models allow one to distinguish between trees drawn from distinct current or future progression states (Chapter 2, Chapter 3). We used the following set of tree-based features: (1) Edge count: eight features corresponding to fraction of progression tree edges showing gains and losses of each gene; (2) Tree level cell: Features corresponding to the fraction of cells at each depth in the progression trees; (3) Parameter values: nine features corresponding to inferred gain and loss probability of each gene (SD), and the probability of a whole genome duplication event (GD).

Similarly to our approaches in Chapter 2 and Chapter 3, we applied these methods for three classification tasks: (i) distinguishing primary samples that progressed to metastasis from their paired metastatic samples, (ii) distinguishing all primary samples from all metastatic samples, and (iii) distinguishing primary samples that metastasized from primary samples that did not metastasize. We compared the classification performance of the features from our current model with the SD-only (pure rectilinear) model [34] and unweighted SD+GD [35] model of tumor progression. Since each gene resides on a distinct chromosome in CC, CD events are irrelevant. We performed 500 rounds of bootstrapping and computed mean accuracy and standard deviations of accuracy.

The results are presented in Figure 4.4. The parameter value-based features (i.e., the inferred phylogenetic rate models themselves) are the most accurate predictors for the first and second tasks of separating primary samples from the metastatic ones. For the clinically important problem of determining whether a given primary tumor will metastasize, tree level features show improved prediction accuracy by at least 3.5% over all other feature sets considered, including comparable feature sets from the earlier unweighted models.

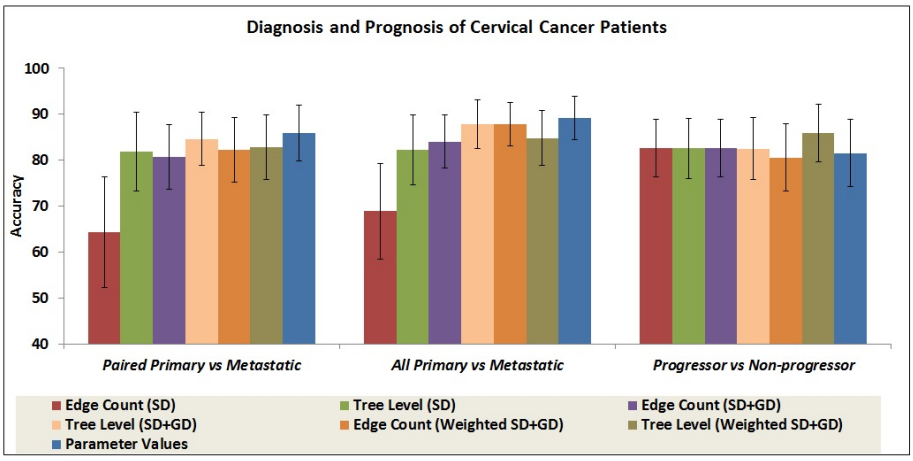

Figure 4.4: Classification results on the CC dataset.

4.2.4 Survival analysis in the TC dataset

For the TC dataset, we focused our analyses of the tree-progression models on trying to identify predictors of survival. Based on earlier work suggesting that the distribution of node depth was a useful predictor of progression in CC [34], we investigated whether the tree level cell distribution is also a predictor of overall and disease-free survival time in TC. Similarly to the cervical cancer samples, we considered distribution of cells across all the levels of the tumor phylogenetic trees inferred on the tongue cancer samples. Using the cell distribution vector as features, we performed K-means clustering to partition the samples into two subgroups. We used "Euclidean" as the distance measure for the clustering experiment and performed 10 replicates of clustering using new initial cluster centroid position.

We performed Kaplan-Meier (KM) analysis (survdiff function in R) to compare either the survival time or the disease-free survival time between the two groups obtained from the two subgroups of samples (Figure 4.5). The subgrouping of patients yielded a significant difference in overall (p-value = 0.0443) and disease-free (p-value = 0.0371) survival between the two patient groups. We repeated the same clustering procedure and KM analyses using trees derived from our previous unweighted SD+GD algorithm [35] but did not observe statistically significant differences in overall (p-value = 0.0784) or disease-free (p-value = 0.14) survival between the

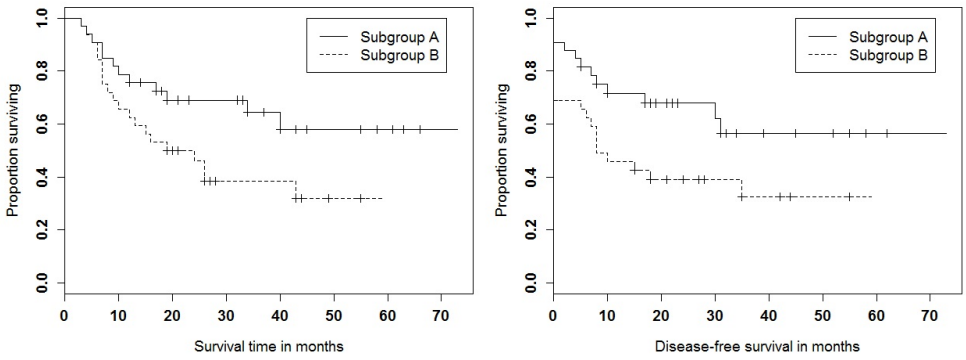

Figure 4.5: KM curves for the test of association between overall (A) and disease-free (B) survival time and tree level cell count statistics based subgrouping of patients.

Multivariate Survival Analysis with Cox Proportional Hazards Model								
	Overall survival				Disease-free survival			
	Global P value	HR	95% CI	P value	Global P value	HR	95% CI	P value
Tree level cell Subgrouping	1.03E-05	2.198	1.054-4.580	3.56E-02	3.76E-05	2.424	1.159-5.067	1.86E-02
Tumor stage		2.153	1.534-3.021	9.20E-06		2.01	1.432-2.822	5.52E-05

Figure 4.6: COXPH analysis to test the correlation between tree level cell count statistic-based subgrouping of patients and tumor stage with disease-free and overall survival time.

two patient groups with the older methods.

We then performed multivariate survival analysis using the Cox proportional hazards (COXPH) model (survfit function in R) to test whether the new test statistics can predict survival or disease-free survival independent of tumor stage. The results are presented in Figure 4.6. The combined p-value is statistically significant, showing that the two covariates are independently associated with overall (disease-free) survival time. The hazard ratio is higher to a significant degree for tree statistic-based subgrouping compared to tumor stage, meaning there is a higher risk of death (disease) if a patient is assigned to the higher risk category by the tree statistic subgrouping, independent of tumor stage.

4.2.5 Distribution of cells across primary and metastatic CC trees:

We previously showed that the differential selective pressures working on different stages of cancer is reflected in the distribution of cells located across different levels of the trees built on samples collected from these sites of tumors. We tested how this cell distribution across tree

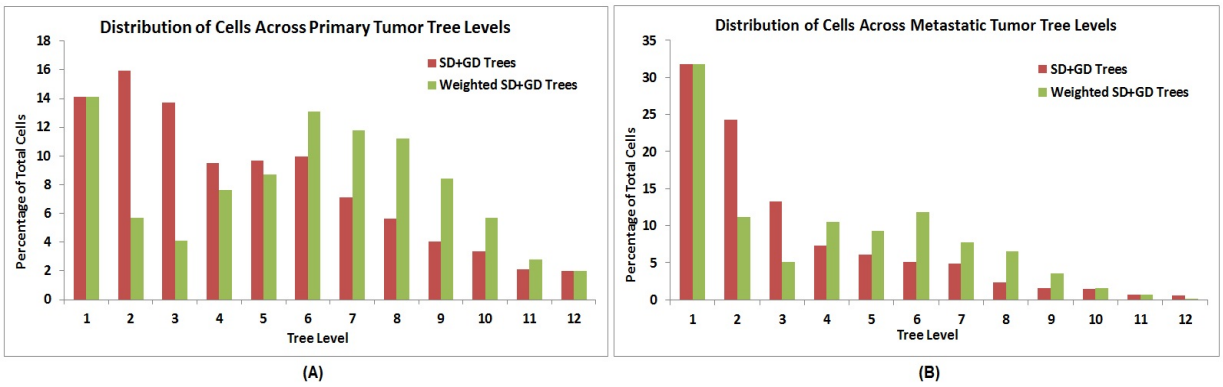

Figure 4.7: Distribution of cells across different tree levels of primary (A) and metastatic (B) tumor phylogenies.

levels is changed in trees built using our new method. In Figure 4.7, we show the distribution of cells across different phylogenetic tree levels of the primary and metastatic samples separately. Comparison between the weighted and unweighted versions of the SD+GD tree cell distributions reveals different dynamics of tumor progression under these two models. In both primary and metastatic cases, the unweighted event trees have more cells located in the first few levels in comparison to the weighted SD+GD trees, where we observe a bimodal distributions of cells with one mode located at deeper levels in the trees. In the primary case, 72.88% of the total cells are located in the first 6 levels of the unweighted SD+GD trees, while for the weighted SD+GD trees, this value is 53.29%. Similar bias in cell distribution is observed for metastatic trees too, where 87.96% and 79.65% of the total cells are located in the first 6 levels of the unweighted and weighted SD+GD trees, respectively. Our newly proposed model leads to results more consistent with the Nowell model of tumor progression, which predicts that initially tumor cells undergo a brief period of heterogeneity followed by expansion of one or more clones [126].

4.3 Discussion

We have developed algorithms for the problem of inferring tumor-specific mutation parameters and applying these to improve single-tumor phylogenetic tree inference at the cellular level, with

specific application to inferring multiscale copy number evolution from single-cell FISH data. This work involved developing efficient algorithms for a weighted parsimony model of copy number evolution, which required substantially different methods than the unweighted model in [35]. We then used an EM-like inference method to learn weight parameters jointly with tree building, providing for the first time scalable algorithms capable of learning tree models for hundreds to thousands of single cells isolated from individual tumors. This work addresses a key need for learning tumor-specific evolutionary models capable of dealing with realistic levels of cellular heterogeneity in single tumors. We showed that the resulting models provide insight into tumor-specific variation and lead to improved prediction of future tumor progression in multiple tumor types.

This work makes an important step towards scalable algorithms for inferring cancer-specific evolution in single tumors, although much remains to be done to realize the full potential of the approach. One important limitation is the focus on FISH data. FISH is currently the only technology for which it is practical to profile genomic variation in hundreds of single cells per patient for moderate-sized patient populations, an essential feature for tumor progression prediction. The present work thus focused on models of copy number evolution specifically, as it is both the most common form of hypermutability in tumor evolution [81] and the mechanism most easily profiled by FISH. FISH, however, offers a far more limited portrait of variation of each cell than does single-cell sequencing [117]. While single-cell sequencing is not yet practical for the numbers of cells needed to study variation in evolutionary mechanisms across patient populations, one can reasonably anticipate that it will eventually overcome that limit, motivating new algorithmic problems to deal simultaneously with hundreds to thousands of cells, potentially millions of markers of variation, and with the more classes of genomic variation that sequencing data can reveal.

Chapter 5

Application of Tumor Phylogenetics in Understanding Prostate Cancer Progression Mechanism[1]

Tumor phylogenies are used to infer the evolutionary history of tumors. Despite the growing recognition in the field of the evolutionary aspect of tumor progression, there has been limited application of tumor phylogenies in the literature to questions of direct clinical value. The use of tumor phylogenies has mostly been limited to distinguishing driver from passenger mutations [6, 29] as well as understanding how drivers interact in development of single tumors. Phylogenies have also been used to show how similar subtypes of tumors are phylogenetically related to each other and identifying which events occur early in tumor progression [43]. But no methods have yet been introduced by which evolutionary history of a single patient's tumor could help guide treatment decisions for that patient.

One of the limiting factors in the use of tumor phylogenies for clinical decision making has

[1]This chapter was developed from material published in "Heselmeyer-Haddad *et al.*, Single-cell genetic analysis reveals insights into clonal development of prostate cancers and indicates loss of *PTEN* as a marker of poor prognosis, *American Journal of Pathology*, 184, 2671-2686 (2014)" [82].

been scarcity of data. One has to sample large number of cells from a single tumor to build a robust model of tumor progression. Furthermore, tumor data has to be collected from a large cohort of patients to build statistically robust predictors of future tumor progression. Currently, fluorescence *in situ* hybridization (FISH) is the only technology that can be used reliably to collect genomic data across hundreds to thousands of cells from each site of tumors. As our algorithms take advantage of FISH data in building tumor progression models, they allow us to develop a framework for understanding the full potentials of tumor phylogenies in different clinical decision making procedures. The algorithms proposed in this thesis make it possible to build phylogenies from hundreds to thousands of cells across large number of patients, which makes it feasible to identify robust and statistically sound predictors from properties of the resulting tumor phylogenies.

In Chapters 2, 3 and 4, our main focus was on designing novel algorithms for inferring phylogenetic models of tumor progression from single cell gene copy number data. In these chapters, we also developed statistical and machine learning framework for identifying characteristic features of the tumor progression models, which were used to draw robust inferences about the tumor progression mechanism. These features were also shown to be highly informative about future tumor progression. Our proposed framework is the first instance of use of tumor phylogenetic models in clinically relevant tumor progression prediction task. In this chapter and the next one, we make further advancement towards the goal of using tumor phylogenetics for clinical decision making by building a mathematical model of cellular-level progression for each tumor sample, and deriving test statistics and numerical features useful for distinguishing patients subgroups, and predicting future tumor progression. We have applied these methods particularly to samples collected from prostate cancer (this chapter) and tongue cancer (next chapter). A central aim of the work in this and the next chapter is to evaluate whether the tumor phylogeny-based markers can be used to predict progression in prostate cancer and prognosis in oral cancer independent of tumor stage. A novel feature of our approach is the use of phylogenetic analysis of tumor progression to infer models of cellular evolution that serve as the basis for dividing

patients into groups with putatively differential prognoses.

In the prostate cancer study, we applied two layers of analysis to a panel of seven FISH probes. First, we pursued the conventional strategy of profiling the probe set in subsets of tumor samples from radical prostatectomy (RP) patients with different clinical outcomes (here, non-progressors and progressors) to identify combinations of probes that distinguish patient groups. Then, using the single-gene duplication based model developed in Chapter 2, we pursued a novel, unconventional goal of building a mathematical model of cellular-level progression for each tumor sample and derived test statistics that can collectively distinguish the two groups of patients. The contribution of this thesis to that study is application of the single-gene duplication based method on the prostate cancer dataset to understand the progression of the disease, which we discuss in this chapter.

5.1 Analysis of tumor heterogeneity in prostate cancer

Prostate cancer is the most commonly diagnosed non-cutaneous neoplasm among American males (238,590 estimated cases in 2013) and is the second leading cause of cancer-related death (29,720 estimated deaths) [151]. Disease incidence exceeds mortality by a factor of eight; this suggests that a large proportion of prostate cancers do not result in disease-associated death. This observation is attributable to the fact that many prostate cancers do not progress to metastatic disease. Patients with more indolent tumors would benefit from an "Active Surveillance" approach. Men with aggressive disease, however, need immediate and often adjuvant therapy after radical prostatectomy (RP) to improve survival. While serum-level screening for prostate specific antigen (PSA) has increased detection of prostate cancer at earlier stages [25], sensitive and specific tests to distinguish men with indolent disease from men with aggressive prostate cancers are still lacking, which creates a dilemma in how to adapt risk-associated treatments [107].

In this chapter, we explore intra-tumor heterogeneity of prostate tumors using a special break-apart probe for the *TMPRSS2-ERG* fusion [135] and six single-gene probes selected based on a

prior array CGH study [129]. In that study, Paris *et al.* [129] screened prostate cancers treated with radical prostatectomy from patients with similar high recurrence risk, but different clinical outcomes, for chromosomal aberrations with array CGH [129]. Comparison with an independent set of metastases revealed approximately 40 candidate markers associated with metastatic potential. For the current study, we chose six of these markers (listed here in chromosome order) –*TBL1XR1* (3q26.23), *CTTNBP2* (7q31.2), *MYC* (8q24.21), *PTEN* (10q23.1), *MEN1* (11q13), and *PDGFB* (22q13.1)–to be tested for their potential use as indicators of progressive disease. An *ERG* break-apart probe [135] was also chosen for determining whether the fusion status of *TMPRSS2-ERG* could serve as an additional progression marker. The markers and two centromeric control probes (CEP8 and CEP10) were applied as FISH probes to single cell suspensions prepared from archived formalin-fixed paraffin-embedded (FFPE) material for a subset of cases from the original study [129], i.e., seven prostate cancers from patients with recurrence and six tumors from patients without recurrence following RP. Our novel approach of multiplexing FISH probes [83] allowed signal enumeration in the same cells.

5.2 Signal count based tumor progression analysis

We refer to the ordered list of count values for each probe as a "signal count pattern". To explore the possibility that tumors with progression are more or less heterogeneous than tumors without progression, we compared three measures of diversity in the distribution of FISH signal count patterns: (i) instability index, (ii) Shannon index, and (iii) Simpson index [130]. The indices were computed for each sample and were then compared between the progressor and non-progressor distributions either by comparing the mean or by a Wilcoxon signed-rank test. The instability index is defined as 100 times the number of cell count patterns divided by the number of cells. To define the other two indices, let p_i be the probability of the i-th cell count pattern. Then, the Shannon index, which is commonly used in information theory and also known as entropy, is $-\sum p_i \log_2(p_i)$. The Simpson index, which is commonly used in population genetics, is $\sum p_i^2$.

5.3 Modeling tumor progression and analysis of node depth

For each of the 13 tumors, we modeled the progression of copy number changes using the single-gene duplication method, as developed in Chapter 2 and implemented in our software FISHtrees [34], which infers phylogenetic trees describing progression among the observed cell types as distinguished by their probe copy numbers. A tree is inferred from the data for each tumor so as to heuristically seek to minimize the total number of copy number changes across the tree. In this analysis, we used the six gene probes and we used the break-apart probe to assess the copy number of *ERG*. We did not use the fusion status or the centromere probes. For each tumor, FISHtrees generated a tree model in which the normal state (2,2,2,2,2,2,2) is at the root of the inferred tree and each edge moving away from the root to a new node corresponds to a change in copy number of one gene. For each node, we also stored the number of cells observed to match the 7-component signal count pattern for that node.

The number of steps away from the root is usually called the "depth" of a node. Our prior work on cervical cancer progression trees [34] showed that the distribution of cells by depths provides a measure of tree topology that is predictive of progression potential. To test whether this characterization of tree topology is similarly predictive of prostate cancer progression, we performed an analogous test. We computed the percentages of cells represented by nodes at each depth in the tree, as described previously [34]. We used percentages of cells at each depth rather than total cells to normalize for differing numbers of cells analyzed for different tumors. We visualized the distribution of the depths of cells in progressors vs. non-progressors by a bar graph. We then tested for significance of the difference between average depths for non-progressors vs. progressors by a Wilcoxon signed-rank test. Since we hypothesized that the trees derived from progressor samples would have greater average depth, the Wilcoxon P values were one-sided.

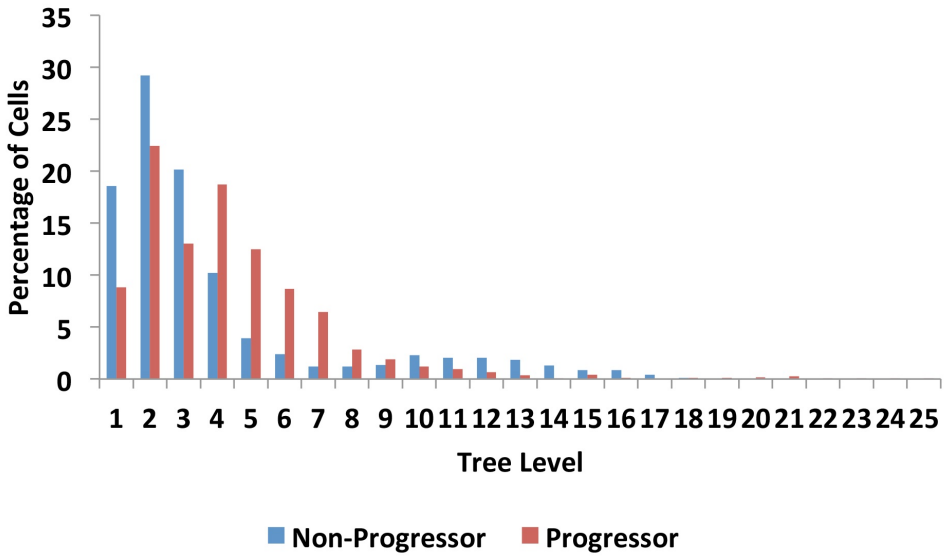

Figure 5.1: Distribution of cells across different levels of tumor progression trees for non-progressor and progressor cases.

5.4 Tree models of tumor progression show a different pattern of changes in non-progressors vs. progressors

The instability index of the prostate carcinomas with or without progression after radical prostatectomy (RP) is very similar. Besides, wilcoxon tests of the instability index, Shannon index, and Simpson index comparing non-progressers with progressers showed no statistically significant differences in. But the frequency of genomic imbalances is substantially higher in the progressor group, with 1.9 gains or losses per case versus 0.5 gains or losses in the non-progressors. To visualize the pattern of progression in each case, we constructed tree models of progression using the software FISHtrees, which infers phylogenetic trees describing likely evolution of the set of observed signal counts within each tumor from an initially diploid root cell so to heuristically minimize total copy number changes across the tree. To evaluate whether there were statistically significant differences between the inferred phylogenetic trees of non-progressors vs. progressors, the cell distribution across tree levels was calculated (Figure 5.1). The analysis showed that

Table 5.1: Significance of Wilcoxon signed-rank test of weighted average node depth in the trees of non-progressors vs progressors.

Number of levels used	1-sided P value of test on weighted-average level
5	0.067
6	0.018
7	0.011
8	0.011
9	0.018
10	0.018
11	0.037
12	0.037

in the non-progressors, 70% of all cells were distributed within the first three tree levels, which was true for only 44% of the cells of the lesions that progressed. This observation indicates that cells of lesions that have a higher propensity to progress to advanced disease on average deviate more from the normal diploid status compared to cells from non-progressing lesions. This observation can be formalized statistically by computing weighted average depth of the nodes up to some level L for the 6 non-progressors and the 7 progressors. The weighted average depths for L=5, . . . ,12 for the two sets (non-progressors vs. progressors) were compared by a Wilcoxon signed-rank test, which shows that the weighted average depth in the progressors is statistically significantly greater (Table 5.1). For example for L=10, the P value of the test is 0.018. The node depth is the distance away from the normal signal count pattern (2,. . . , 2) expressed in terms of the count of copy-number changes. Thus, the cells in the progressor samples have in general a trend towards more total chromosomal changes. This trend is not captured by previously proposed measures of diversity (Shannon index, Simpson index) [130].

5.5 Discussion

Men with slowly progressing prostate cancers could be treated with Active Surveillance approaches instead of immediate, more aggressive treatment, including surgery, which can have considerable side effects. This subset of patients will become larger as populations age and

more tumors are detected early by screening efforts. Distinguishing patients with aggressive or indolent prostate carcinomas would help in designing risk-adapted neoadjuvant and adjuvant treatments. In this chapter, we built evolutionary models of tumor progression for each patient using the gene probes for *TBL1XR1, CTTNBP2, MYC, PTEN, MEN1, PDGFB* and *ERG* in prostate cancer. Using the tumor progression trees, we derived test statistics that are significantly associated with tumor progression, and thus can help identifying patients in need of aggressive treatment. Further investigation of the tree-based models may continue to throw light onto the genetic basis of prostate cancer progression mechanisms.

Chapter 6

Application of Phylogenetic Analysis to Oral Tongue Squamous Cell Carcinoma[1]

Oral tongue squamous cell carcinoma (OTSCC) is associated with poor prognosis, with increasing incidence seen among young adults [149]. Known environmental risk factors for OTSCC include tobacco usage, either via smoking [101] and chewing [20], and alcohol consumption [108]. Patients diagnosed at earlier stages (I or II) have a significantly better prognosis. Numerous studies using single-markers attempted to improve disease prognostication could not be validated [5].

In this chapter, we describe work using enumerated copy numbers of four genes and one centromere probe in 65 cases of OTSCC with detailed patient data and follow-up including disease-free and overall survival. Our approach of enumerating all probes within the same cells allowed us to analyze intratumor heterogeneity and co-occurrence of copy number changes. We analyzed the oncogenes *TERC* on 3q, *EGFR* on 7p, *CCND1* on 11q, and the tumor suppressor *TP53* on 17p. These genes were selected because they have been frequently implicated in the progression of oral cancer, specifically, the first three have been suggested to be among the primary targets of

[1]This chapter was developed from material submitted in "Wangsa *et al.*, Phylogenetic analysis of multiple FISH markers in oral tongue squamous cell carcinoma suggests that a diverse distribution of copy number changes is associated with poor prognosis, In Revision".

copy number gains on chromosome 3q (*TERC*,[155]), 7p (*EGFR*,[59]), and 11q (*CCND1*,[59]), respectively.

A central aim of this study was to evaluate whether the combination of multiple FISH markers can be used to predict prognosis in OTSCC independent of tumor stage. A novel feature of our approach, and the specific contribution of the work in this thesis, is the use of phylogenetic analysis of tumor progression to infer models of cellular evolution. These models serve as the basis for clustering patients into groups with putatively differential prognoses.

Our approach utilizes multiple probe single-cell FISH data to build tree models of tumor progression using the single-gene, single-chromosome and whole-genome duplication based method developed in Chapter 3. We then derive summary statistics based on all four genes from the tree models generated and we test for associations between those summary statistics and survival, while taking into account tumor stage and smoking history. We show that this evolutionary approach has more predictive power than using the static gene copy number counts for survival analysis.

6.1 Multivariate survival analyses

We used statistics calculated from FISH copy number data and either clustering or phylogenetic trees inferred on these FISH samples to test for association with disease-free survival and overall survival time using KM analysis. We then performed multivariate survival analysis using the Cox proportional hazards (COXPH) model as implemented in R [166, 171]. The variables used in the multivariate analysis include: tumor stage, smoking status, and various test statistics derived from the phylogenetic analysis, as explained below. The objective of the multivariate analysis is to test whether the new test statistics can predict survival or disease-free survival independent of tumor stage and smoking status.

For cluster-based subgrouping, we followed convention and assigned the cluster associated with longer (shorter) survival of patients as the lower (higher) risk group. For smoking behavior,

we assigned smokers (non-smokers) to the higher (lower) risk category based on previous epidemiological results [26, 37, 101, 123]. Because of these assignments of categories, we expected the hazard ratios (HRs) to be > 1. Nevertheless, P-values are two-sided.

6.2 Clustering of samples by gain/loss patterns

For each cell count pattern in each sample, we estimated whether each gene is amplified or lost by comparing the copy number counts of the centromere probe with the copy number counts of that gene to identify one of three possible gain/loss conditions for each gene: No Change (N), Gain (G) and Loss (L) [183]. Considering all four genes yields $3^4 = 81$ possible combined gain/loss conditions, each of which we call a "comparator", for each cell. Although losses of *EGFR* and gains of *TP53* have been found in some cases of oral cancer [79], we considered only those comparators that include gain, no change for oncogenes and loss, no change for the tumor suppressor for a total of $2^4 = 16$ comparator strings for each cell. The comparators represent the genes *TERC, CCND1, EGFR, TP53* in this order from left to right. We calculated the total number of cells that match each comparator for each sample. Because there is little variation in the number of cells per sample, we counted total cells rather than fractions of cells. Each sample is then represented by an ordered list of 16 integers that are considered "features" (a term of art in machine learning meaning a quantitative measure by which one can distinguish meaningfully different groups of objects). The transformed dataset becomes a 65×16 dimensional matrix with integer-valued count entries.

We used K-means clustering (with K=2) as implemented in Matlab to partition the 65 samples into two non-overlapping subgroups. We then performed KM analysis via the survfit function in R [171] to compare either the survival time or the disease-free survival time between the two groups obtained from the unsupervised clustering of samples and derive associated two-sided P-values. We derived KM curves derived using (1) two simplified representations of the dataset ("Binarization" and "Thresholding") and (2) three different distance measures in the K-means

clustering algorithm (L1/Rectilinear, Euclidean, Cosine).

We now give an example of the thresholding and binarization procedure. Consider a smaller example where we have a dataset consisting of five samples S1-S5 and two genes for each of which we consider {gain,no change} events. The example dataset is shown in Figure 6.1(A). Under the "thresholding" operation, all the entries in the dataset less than a particular threshold are replaced with 0 leaving all the other entries unchanged. So, in this case, if we decide that the threshold is 4, then the thresholded dataset looks like Figure 6.1(B). The "binarization" operation replaces all the entries below a certain threshold with 0 and all the other entries greater than or equal to that threshold with 1. In the above example, the binarized dataset would be as in Figure 6.1(C).

Comparator String	S1	S2	S3	S4	S5
NN	5	6	4	4	3
NG	4	3	8	4	2
GN	1	3	2	4	2
GG	5	5	2	4	3

(A)

Comparator String	S1	S2	S3	S4	S5
NN	5	6	4	4	0
NG	4	0	8	4	0
GN	0	0	0	4	0
GG	5	5	0	4	0

(B)

Comparator String	S1	S2	S3	S4	S5
NN	1	1	1	1	0
NG	1	0	1	1	0
GN	0	0	0	1	0
GG	1	1	0	1	0

(C)

Figure 6.1: Example showing (A) Unprocessed, (B) Thresholded and (C) Binarized data.

6.3 Tumor phylogenetic tree-based statistics and subgrouping of samples

We next investigated whether either of the two different types of statistics computed based on tumor phylogenetic trees inferred from FISH copy number data are associated with disease-free and overall survival time. We first built tumor progression trees for each of the 65 patients using an evolution model incorporating single gene gain/loss (denoted SD) and whole genome duplication (denoted GD), as developed in Chapter 3 and implemented in the software FISHtrees [35]. The model also allows for gains/losses of entire chromosomes, but that capability is not relevant here because each probe is on a distinct chromosome.

For each gene in each tree, we counted the total number of edges across which that gene's copy number is increased (SD gene gain or a GD event) and decreased (SD gene lost event) and normalized counts by the total number of SD and GD events, collectively producing an eight-dimensional array of event frequencies for each sample across the four genes. Next, we calculated the Simpson Index (SI) and Shannon Entropy(MI) for each sample. SI and MI are computed by taking sum of squared frequency values and sum of frequency-weighted negative logarithms of frequencies, respectively. For example, if there were only two genes and the (gain, loss) counts for gene 1 and gene 2 were (10,20) and (30,40), then SI of this patient would be: $(10/100)^2+(20/100)^2+(30/100)^2+(40/100)^2 = 0.3$ and MI is: $-0.1\log(0.1)-0.2\log(0.2)-0.3\log(0.3)-0.4\log(0.4) = 0.56$. A lower SI or a higher MI indicates that the underlying gene count distribution is more diverse (closer to uniform) across the four genes tested. The two measures of diversity are standard in genetics (SI) or information theory (MI), and they have been previously used to study the diversity of tumor cell populations [130]. We divided the patients into two groups for each measure, with those above versus below the mean SI (mean MI) assigned to different subgroups.

Alternative tree-based statistics and subgrouping procedures based on tree topology were also

considered. For this second tree-based statistic, we extracted various types of features based on the distributions of cells across the tree levels and performed KM analysis to evaluate whether these features are associated with survival time and disease-free survival time. We used the following combinations of features and patient subgrouping procedures:

1. We first calculated the weighted mean tree depth (average distance from the diploid root of all cells in the tree) for each patient and then used the mean depth value across all the patients for dividing them into two subgroups, one group above the mean and one group below the mean. We calculated the weighted mean tree depth by computing the weighted average of cell distribution across the tree levels.
2. We calculated the weighted mean tree depth for each patient and then used the K-means clustering with Euclidean distance for dividing the patients into two subgroups.
3. Using the distribution of cells, we calculated MI (Shannon Entropy) for each patient and then used mean MI across all the patients for dividing them into two subgroups.
4. Using the distribution of cells, we calculated SI (Simpson's Index) for each patient and then used mean SI across all the patients for dividing them into two subgroups.
5. Using the distribution of cells, we calculated SI for each patient and then used the K-means clustering with euclidean distance for dividing the patients into two subgroups.
6. Using the distribution of cells, we calculated MI for each patient and then used the K-means clustering with Euclidean distance for dividing the patients into two subgroups.
7. We used K-means clustering with euclidean distance on the vector of distribution of cells across different tree levels data for subgrouping of patients.
8. We used K-means clustering with rectilinear distance on the vector of distribution of cells across different tree levels data for subgrouping of patients.
9. We used K-means clustering with cosine distance on the vector of distribution of cells across different tree levels data for subgrouping of patients.

10. We calculated weighted cell distribution across the tree levels for each patient and then used the mean weighed value across all the patients for subgrouping purpose. For calculating the weighted cell distribution, we computed the sum of linear combination of cell distribution and tree levels.
11. We calculated weighted cell distribution across tree levels 1 to 3 for each patient and then used the mean weighed value across all the patients for subgrouping purpose.
12. We calculated weighted cell distribution across tree levels 1 to 4 for each patient and then used the mean weighed value across all the patients for subgrouping purpose.
13. We calculated weighted cell distribution across tree levels 1 to 5 for each patient and then used the mean weighed value across all the patients for subgrouping purpose.
14. We calculated weighted cell distribution across tree levels 1 to 6 for each patient and then used the mean weighed value across all the patients for subgrouping purpose.

6.4 Two-means clustering and survival analysis based on sample-based statistics

We examined nine combinations of representations of probe count G(ain), N(ormal), L(oss) how data are represented (raw, binarized, or thresholded) and similarity measure (L1/Rectilinear, Euclidean, Cosine). Six of the nine combinations yielded significant ($P<0.05$) differences in disease-free survival, with all yielding similar patient assignments to high-risk and low-risk clusters (Figure 6.5). Binarizing the data (thresholded so that 60% of values are not 0) with rectilinear distance yielded the most significant difference (P=0.0054) in overall survival between the two patient clusters (Figure 6.2(A)) as well as a significant difference for disease-free survival (P=0.0107). The KM curves for the five remaining cluster analyses showing significant survival differences are provided in Figure 6.3. Multivariate survival analysis using both cluster assignment and tumor stage showed that cluster assignment is not a significant predictor of survival

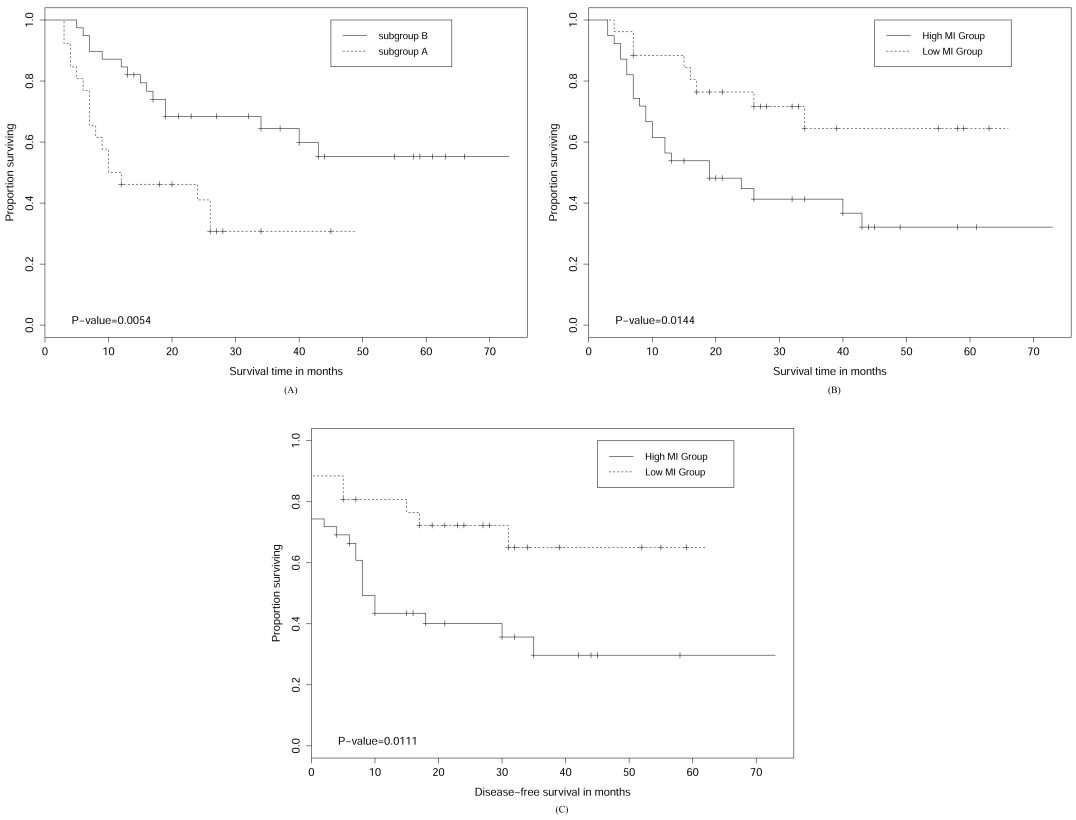

Figure 6.2: (A) KM curve for test of association between overall survival time and subgrouping of patients based on binarized data and L1 distance based K-means clustering. (B) KM curve for the test of association between overall survival time and MI-based subgrouping of patients. (C) KM curve for the test of association between disease-free survival time and MI-based subgrouping of patients. All the P-values in (a), (b) and (c) are two-sided.

independent of tumor stage.

The survival analysis results for the six clustering choices that gave statistically significant results are as follows:

1. Experiment 1: L1/Rectilinear distance. No binarization or thresholding is performed on the dataset. P = 0.0436.
2. Experiment 2: L1/Rectilinear distance. Data are binarized. The threshold is fixed at a value so that 60% of the total entries are 0. P = 0.0054.
3. Experiment 3: Cosine distance. Data are binarized. Threshold is fixed at a value so that

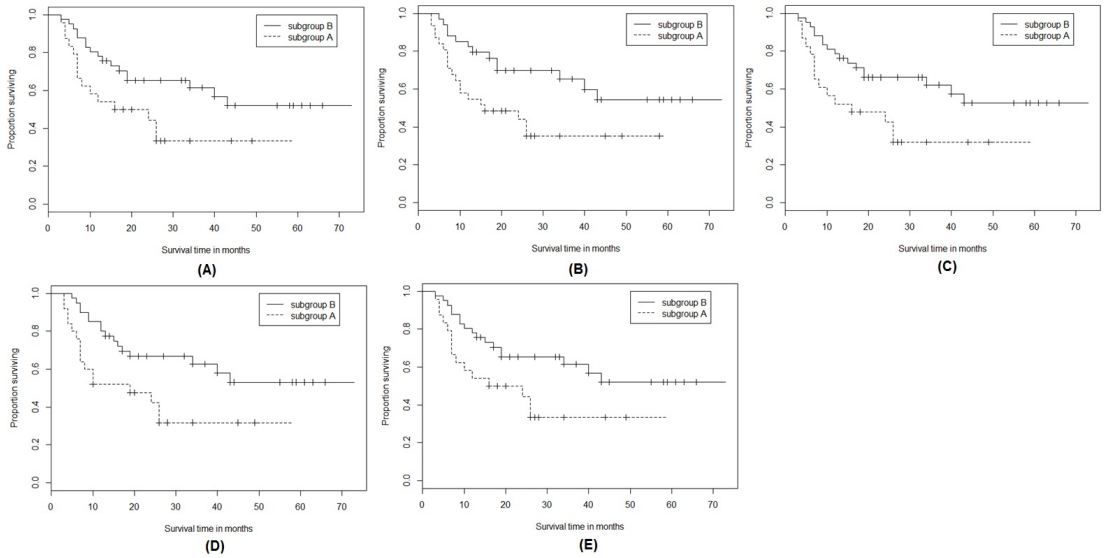

Figure 6.3: KM curves for test of association between overall survival time and subgrouping of patients for five different choices of original/ binarization/ thresholding of the sample-based comparator frequencies and of three distance measures used in the K-means clustering: (A) no binarization or thresholding of the dataset and L1/Rectilinear distance for K-means clustering, (B) binarized data and Cosine distance for K-means clustering, (C) thresholded data and Euclidean distance for K-means clustering, (D) binarized data and Euclidean distance for K-means clustering, (E) thresholded data and L1/Rectilinear distance for K-means clustering.

60% of the total entries are 0. P = 0.0226.

4. Experiment 4: Euclidean distance. Data are thresholded. The threshold is fixed at a value so that 60% of the total entries are 0. P = 0.0246.

5. Experiment 5: Euclidean distance. Data are binarized. The threshold is fixed at a value so that 60% of the total entries are 0. P = 0.0185.

6. Experiment 6: L1/Rectilinear distance. Data are thresholded. The threshold is fixed at a value so that 60% of the total entries are 0. P = 0.0436.

	Cluster Center		Nearest Sample	
String	Cluster B	Cluster A	Cluster B	Cluster A
NNNN	11	2	9	2
NNNL	2	1	0	1
NNGN	2	1	2	1
NNGL	1	1	1	1
NGNN	3	2	2	3
NGNL	1	3	0	3
NGGN	3	7	5	8
NGGL	1	4	1	3
GNNN	4	1	4	0
GNNL	1	1	0	0
GNGN	2	2	1	2
GNGL	1	2	0	2
GGNN	4	5.5	3	9
GGNL	1	4	1	3
GGGN	6	18.5	5	16
GGGL	2	11	3	11

Figure 6.4: Cluster centers for subgrouping of patients from two-means clustering using the original FISH data (no binarization or thresholding) and rectilinear distance.

To better characterize features predictive of survival, we further clustered the raw G, N, L data to identify short survival (subgroup A) and long survival (subgroup B) clusters. Figure 6.4 provides descriptive statistics about the cluster centers, i.e., the averages of the elements of each cluster. The center of cluster B is heavily biased towards the "No change of any gene" string while cluster A shows bias towards the "GGGN" and "GGGL" strings. In the same table, we show the two representative samples closest to each cluster center. String "NNNN" is most populated across the cluster B representatives and "GGGN" and "GGGL" are most populated in the cluster A representatives. Cluster A thus appears to favor patients with many cells with gain of all three oncogenes while cluster B tends to favor patients with many cells exhibiting no gain

or loss of any of the genes.

Figure 6.5 reports the subgroup ID of each sample belonging to two different clusters for the six choices of data representation and similarity measure listed above as Experiments 1-6. The clustering is robust to the methodology, with 50 of the 65 patients assigned to the same clusters for all six experiments, but because the P-values are close to 0.05 significance threshold, small changes in the clustering assignments can affect whether the KM results are statistically significant. Cluster membership exhibits a clear association with tumor stage, which is unsurprising since stage is known to be associated with survival. For example, in the experiment of Figure 6.4, the long-survival cluster B is strongly associated with tumor stages 1 or 2 (P = 0.0045, Fishers exact test). The cluster assignment largely predicts tumor stage and thus is not an independent predictor of survival.

6.5 Two-means clustering and survival analysis, taking into account smoking or tumor stage

We next split the samples based on smoker/non-smoker status and performed survival analysis on these two sets of samples separately. 19 patients are non-smokers and 34 are smokers, with the remaining 12 lacking smoking status information and thus not considered in this analysis. Figure 6.6 reports the P-values from the survival analysis for clustering using the choices in Experiments 1-6, and based on subgrouping of smokers/non-smokers. 2-means clustering is associated with survival for the patients belonging to the smoker category, with five of the six P-values$<$0.05 and for experiment 2, a P-value of 0.00342.

We then performed COXPH analysis to test the independent predictive power of cluster-based subgrouping, smoking behavior, and tumor stage as explanatory variables. Figure 6.7 reports the results of COXPH analysis using clustering and smoking behavior, showing no significant difference in disease-free (overall) survival for the two clusters after controlling for smoking behavior

ID	Age	Gender	Smoke habit	Stage	Current Status	Survival Time (months)	Disease-Free Survival Time (months)	Experiment Number					
								1	2	3	4	5	6
T1	59	female	Smoker	1	Alive	63	52	B	B	B	B	B	B
T2	56	male	Smoker	1	Alive	58	58	B	B	B	B	B	B
T3	67	female	Smoker	1	Alive	15	15	B	B	B	B	B	B
T4	52	female	Smoker	1	Alive	61	22	B	B	B	B	B	B
T5	57	female	Smoker	1	Dead	40	30	B	B	B	B	B	B
T6	73	female	Smoker	1	Alive	58	7	B	B	A	B	A	B
T7	75	male	Smoker	1	Dead	43	35	B	B	B	B	B	B
T8	25	female	Non-Smoker	1	Alive	44	42	A	B	B	A	B	A
T9	62	female	.	1	Alive	45	45	B	A	A	B	A	B
T10	66	female	Smoker	1	Alive	43	2	B	B	B	B	B	B
T11	46	female	.	1	Alive	39	39	B	B	B	B	B	B
T12	58	male	Smoker	1	Alive	33	29	B	B	B	B	B	B
T13	68	female	Non-Smoker	1	Alive	28	28	A	A	A	A	A	A
T14	52	male	.	1	Alive	22	5	A	A	A	A	A	A
T15	61	female	Non-Smoker	1	Alive	19	19	B	B	B	B	B	B
T16	66	male	Smoker	2	Alive	73	73	B	B	B	B	B	B
T17	54	male	Smoker	2	Dead	19	10	B	B	B	B	B	B
T18	71	male	Smoker	2	Alive	66	62	B	B	B	B	B	B
T19	48	male	Smoker	2	Dead	7	0	B	B	B	B	B	B
T20	30	female	.	2	Alive	59	59	B	B	B	B	B	B
T21	88	male	Non-Smoker	2	Dead	34	31	B	B	B	B	B	B
T22	33	male	Non-Smoker	2	Dead	26	0	A	A	A	A	A	A
T23	38	female	Non-Smoker	2	Alive	59	59	A	B	A	A	B	A
T24	58	male	Non-Smoker	2	Alive	55	55	B	B	B	B	B	B
T25	55	male	Smoker	2	Dead	9	7	B	B	B	B	B	B
T26	53	male	Non-Smoker	2	Dead	9	0	B	A	A	B	B	B
T27	32	female	Non-Smoker	2	Alive	20	16	A	A	A	B	A	A
T28	80	male	.	2	Dead	16	15	A	B	A	A	B	A
T29	96	female	.	2	Dead	3	2	B	A	A	B	A	B
T30	89	female	Non-Smoker	2	Dead	26	18	A	A	A	A	A	A
T31	49	female	Non-Smoker	2	Alive	49	44	A	A	A	A	A	A
T32	58	male	Smoker	2	Alive	37	7	B	B	B	B	B	B
T33	82	male	Smoker	2	Dead	5	0	B	B	B	B	B	B
T34	41	male	Non-Smoker	2	Alive	34	34	B	A	A	B	A	B
T35	62	male	Smoker	2	Alive	32	32	B	B	B	B	B	B
T36	40	female	Smoker	2	Dead	12	4	B	B	B	B	B	B
T37	56	male	.	2	Alive	32	32	B	B	B	B	B	B
T38	42	male	Smoker	2	Dead	12	8	A	A	A	A	B	A
T39	59	male	Smoker	2	Alive	14	10	B	B	B	B	B	B
T40	61	male	Non-Smoker	2	Dead	13	8	B	B	B	B	B	B
T41	80	male	Smoker	2	Alive	27	24	A	A	A	A	B	A
T42	21	male	Non-Smoker	2	Dead	15	5	B	B	A	B	B	B
T43	57	male	Smoker	2	Alive	27	27	B	B	B	B	B	B
T44	75	female	Smoker	2	Alive	21	21	B	B	A	B	B	B
T45	68	male	Non-Smoker	2	Alive	23	23	B	B	B	B	B	B
T46	41	female	Smoker	3	Alive	21	21	B	B	B	B	B	B
T47	57	male	Smoker	3	Dead	6	6	B	B	B	B	B	B
T48	76	male	Non-Smoker	3	Dead	10	8	B	A	A	B	A	B
T49	43	male	Non-Smoker	3	Dead	7	0	B	B	B	B	B	B
T50	29	male	Smoker	3	Dead	24	8	A	A	A	A	A	A
T51	70	female	Smoker	3	Dead	3	0	A	A	A	A	A	A
T52	62	female	Smoker	3	Dead	8	0	A	A	A	A	A	A
T53	73	male	.	4	Dead	5	0	A	A	A	A	A	A
T54	74	female	Smoker	4	Dead	17	17	B	B	B	B	B	B
T55	76	female	Smoker	4	Dead	7	5	A	A	A	A	A	A
T56	61	male	.	4	Dead	19	0	B	B	B	B	A	B
T57	51	male	Smoker	4	Dead	4	0	A	A	A	A	A	A
T58	33	male	Non-Smoker	4	Dead	7	7	A	A	A	A	A	A
T59	89	female	.	4	Dead	4	0	A	A	A	A	A	A
T60	64	male	Smoker	4	Alive	55	55	B	B	B	B	B	B
T61	79	female	Smoker	4	Dead	6	0	A	A	A	A	A	A
T62	41	male	Non-Smoker	4	Alive	18	18	A	A	A	A	A	A
T63	76	female	Smoker	4	Alive	34	34	A	A	A	A	A	A
T64	67	female	.	4	Dead	10	10	A	A	A	A	A	A
T65	63	male	.	4	Dead	7	0	A	A	A	A	A	A

Figure 6.5: Clinical information and subgroup IDs of the patients. Subgroup IDs are shown for six experimental cases that exhibited statistically significant association between overall survival time and different subgrouping of patients due to choices of original/ binarization/ thresholding of the sample-based comparator frequencies and of three distance measures used in the K-means clustering.

	P-values	
Experiment	Non Smoker	Smoker
1	0.20	3.42E-03
2	0.73	3.42E-03
3	0.78	7.14E-02
4	0.30	3.42E-03
5	0.82	1.71E-02
6	0.20	3.42E-03

Figure 6.6: P-values from the analysis for six different subgrouping of patients based on sample-based comparator frequencies and taking into account smoking history.

Multivariate Survival Analysis with Cox Proportional Hazards Model									
Experiment		Disease-free survival				Overall survival			
		Global P value	HR	95% CI	P value	Global P value	HR	95% CI	P value
1	Cluster subgrouping	5.06E-01	1.673	0.704-3.980	2.44E-01	4.60E-01	1.768	0.730-4.279	2.06E-01
	Smoking behavior		1.329	0.549-3.221	5.28E-01		1.318	0.536-3.237	5.48E-01
2	Cluster subgrouping	8.54E-02	2.641	1.132-6.165	2.47E-02	6.49E-02	2.867	1.201-6.842	1.77E-02
	Smoking behavior		1.533	0.641-3.667	3.37E-01		1.543	0.638-3.732	3.36E-01
3	Cluster subgrouping	2.11E-01	2.142	0.916-5.010	7.88E-02	2.64E-01	2.036	0.869-4.768	1.01E-01
	Smoking behavior		1.459	0.607-3.508	3.98E-01		1.377	0.572-3.316	4.76E-01
4	Cluster subgrouping	3.95E-01	1.815	0.775-4.249	1.70E-01	3.57E-01	1.909	0.802-4.545	1.44E-01
	Smoking behavior		1.332	0.558-3.178	5.19E-01		1.322	0.547-3.194	5.35E-01
5	Cluster subgrouping	2.17E-01	2.233	0.920-5.421	7.58E-02	2.85E-01	2.043	0.861-4.851	1.05E-01
	Smoking behavior		1.495	0.609-3.668	3.80E-01		1.342	0.557-3.235	5.12E-01
6	Cluster subgrouping	5.06E-01	1.673	0.704-3.980	2.44E-01	4.60E-01	1.768	0.730-4.279	2.06E-01
	Smoking behavior		1.329	0.549-3.221	5.28E-01		1.318	0.536-3.237	5.48E-01

Figure 6.7: Results of COXPH survival analysis using six sample-based comparator frequencies-dependent subgrouping of patients and smoking behavior as explanatory variable.

Multivariate Survival Analysis with Cox Proportional Hazards Model									
Experiment		Disease-free survival				Overall survival			
		Global P value	HR	95% CI	P value	Global P value	HR	95% CI	P value
1	Cluster subgrouping	6.00E-04	1.259	0.592-2.678	5.49E-01	6.87E-05	1.405	0.677-2.915	3.62E-01
	Tumor stage		1.809	1.287-2.542	6.00E-04		1.967	1.413-2.737	6.10E-05
2	Cluster subgrouping	3.00E-04	1.625	0.757-3.488	2.13E-01	2.87E-05	1.833	0.874-3.843	1.09E-01
	Tumor stage		1.729	1.228-2.435	1.70E-03		1.903	1.364-2.656	1.55E-04
3	Cluster subgrouping	3.00E-04	1.586	0.752-3.345	2.26E-01	3.16E-05	1.761	0.851-3.644	1.27E-01
	Tumor stage		1.772	1.271-2.469	7.00E-04		1.968	1.419-2.727	4.82E-05
4	Cluster subgrouping	5.00E-04	1.351	0.631-2.892	4.38E-01	5.90E-05	1.492	0.716-3.109	2.85E-01
	Tumor stage		1.783	1.265-2.512	9.00E-04		1.946	1.396-2.714	8.67E-05
5	Cluster subgrouping	4.00E-04	1.529	0.701-3.335	2.86E-01	6.99E-05	1.416	0.659-3.042	3.73E-01
	Tumor stage		1.733	1.224-2.454	1.90E-03		1.927	1.367-2.715	1.79E-04
6	Cluster subgrouping	6.00E-04	1.259	0.592-2.678	5.49E-01	6.87E-05	1.405	0.677-2.915	3.62E-01
	Tumor stage		1.809	1.287-2.542	6.00E-04		1.967	1.413-2.737	6.10E-05

Figure 6.8: Results of COXPH survival analysis using six sample-based comparator frequencies-dependent subgrouping of patients and tumor stage as explanatory variable.

Experiment Number	Survival Time	Disease Free Survival Time
1	0.51	0.50
2	0.42	0.46
3	0.30	0.33
4	0.17	0.27
5	0.37	0.52
6	0.30	0.33
7	0.08	0.14
8	0.16	0.21
9	0.27	0.32
10	0.20	0.20
11	0.24	0.39
12	0.02	0.03
13	0.07	0.08
14	0.34	0.34

Figure 6.9: P-values showing association between overall, disease-free survival time and sub-grouping of patients based on 14 tree level cell distribution based features.

and vice versa. None of the six experiments yields a P-value < 0.05 and the confidence intervals include 1 (not significant) except for experiment 2. Figure 6.8 reports results of COXPH analysis on clustering and tumor staging, showing that the cluster-based subgrouping is not significantly associated with disease-free (overall) survival, independent of tumor stage. In every experiment, the HRs using tumor stage exclude 1 in the confidence interval and are significant, while the HRs using cluster subgrouping include 1 and are not significant.

6.6 KM survival analysis using tree-based statistics

Based on our previous work with other FISH data sets [82, 83], we reasoned that the weakness of the above clustering analysis is that it fails to take into account likely evolutionary relationships between cells with similar combinations of FISH probe counts. Therefore, we used the software FISHtrees [34, 35] to construct evolutionary models of how each tumor progressed based on the observed FISH patterns of the four genes (not using CEP4) and tabulated the Simpsons Index

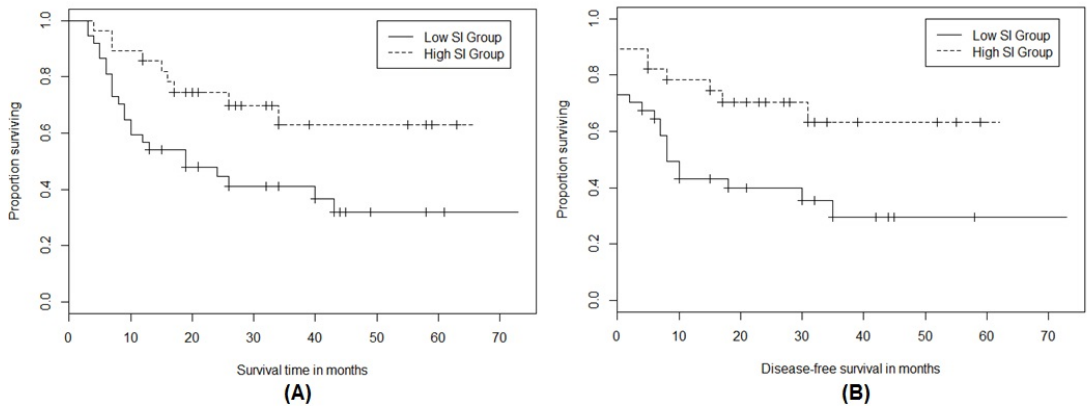

Figure 6.10: KM curves for test of association between (A) overall and (B) disease-free survival time and SI-based subgrouping of patients.

(SI) and Shannon Entropy (MI) test statistics, as described in Methods.

After computing SI and MI from the tumor phylogenetic trees, we subdivided the patients into two groups based on mean SI, mean MI and computed KM curves based on overall (disease-free) survival time. For MI-based subgrouping, the P-values obtained from the KM analysis for overall (disease-free) survival is 0.0144 (0.0111), and the KM curve is shown in Figure 6.2(B) (Figure 6.2(C)). Patients with higher MI value, who have similar frequencies of gene gain and loss counts across genes, have shorter overall and disease-free survival time compared to those with preferential gain or loss of specific genes. SI-based subgrouping, shown as Figure 6.10, yielded comparable results, with P-values from KM analysis of overall (disease-free) survival of 0.0168 (0.0117). Additional KM analysis based on tree topological features, shown in Figure 6.9, failed to yield statistical significance in all but one of the fourteen combinations considered.

6.7 Multivariate COXPH survival analysis using tree-based statistics, tumor stage, and smoking

Next, we performed COXPH tests of the relationship between overall (disease-free) survival time and combinations of the SI/MI-based subgrouping, tumor stage, and smoking status as ex-

Multivariate Survival Analysis with Cox Proportional Hazards Model								
	Disease-free survival				Overall survival			
	Global P value	HR	95% CI	P value	Global P value	HR	95% CI	P value
SI Subgrouping	6.08E-05	2.313	1.058-5.059	3.57E-02	1.15E-05	2.223	1.018-4.853	4.50E-02
Tumor stage		1.794	1.307-2.462	2.99E-04		1.968	1.431-2.707	3.16E-05

(A)

Multivariate Survival Analysis with Cox Proportional Hazards Model								
	Disease-free survival				Overall survival			
	Global P value	HR	95% CI	P value	Global P value	HR	95% CI	P value
MI Subgrouping	4.68E-05	2.467	1.096-5.553	2.91E-02	8.14E-06	2.411	1.072-5.422	3.33E-02
Tumor stage		1.812	1.319-2.488	2.41E-04		1.985	1.422-2.731	2.56E-05

(B)

Figure 6.11: Results of COXPH survival analysis using (A) Simpsons Index- (SI) and (B) Shannon Entropy (MI) based subgrouping of patients and tumor stage as explanatory variable.

Multivariate Survival Analysis with Cox Proportional Hazards Model								
	Disease-free survival				Overall survival			
	Global P value	HR	95% CI	P value	Global P value	HR	95% CI	P value
SI Subgrouping	5.46E-02	2.771	1.149-6.680	2.31E-02	6.13E-02	2.749	1.133-6.672	2.53E-02
Smoking behavior		0.988	0.434-2.247	9.77E-01		0.885	0.386-2.026	7.72E-01

(A)

Multivariate Survival Analysis with Cox Proportional Hazards Model								
	Disease-free survival				Overall survival			
	Global P value	HR	95% CI	P value	Global P value	HR	95% CI	P value
MI Subgrouping	4.46E-02	2.467	1.096-5.553	2.91E-02	4.52E-02	2.984	1.179-7.549	2.09E-02
Smoking behavior		1.812	1.319-2.488	2.41E-04		0.903	0.396-2.061	8.09E-01

(B)

Figure 6.12: Results of COXPH survival analysis using (A) SI and (B) MI based subgrouping of patients and smoking behavior as explanatory variable.

ID	Age	Gender	Smoke habit	Stage	Current Status	Survival Time (months)	Disease-Free Survival Time (months)	SI Based Subgrouping	MI Based Subgrouping
T1	59	female	Smoker	1	Alive	63	52	B	B
T2	56	male	Smoker	1	Alive	58	58	A	A
T3	67	female	Smoker	1	Alive	15	15	A	A
T4	52	female	Smoker	1	Alive	61	22	A	A
T5	57	female	Smoker	1	Dead	40	30	A	A
T6	73	female	Smoker	1	Alive	58	7	B	B
T7	75	male	Smoker	1	Dead	43	35	A	A
T8	25	female	Non-Smoker	1	Alive	44	42	A	A
T9	62	female	.	1	Alive	45	45	A	A
T10	66	female	Smoker	1	Alive	43	2	A	A
T11	46	female	.	1	Alive	39	39	B	B
T12	58	male	Smoker	1	Alive	33	29	B	B
T13	68	female	Non-Smoker	1	Alive	28	28	B	B
T14	52	male	.	1	Alive	22	5	A	A
T15	61	female	Non-Smoker	1	Alive	19	19	B	B
T16	66	male	Smoker	2	Alive	73	73	A	A
T17	54	male	Smoker	2	Dead	19	10	A	A
T18	71	male	Smoker	2	Alive	66	62	B	B
T19	48	male	Smoker	2	Dead	7	0	A	A
T20	30	female	.	2	Alive	59	59	B	B
T21	88	male	Non-Smoker	2	Dead	34	31	B	B
T22	33	male	Non-Smoker	2	Dead	26	0	B	B
T23	38	female	Non-Smoker	2	Alive	59	59	B	B
T24	58	male	Non-Smoker	2	Alive	55	55	B	B
T25	55	male	Smoker	2	Dead	9	7	A	A
T26	53	male	Non-Smoker	2	Dead	9	0	A	A
T27	32	female	Non-Smoker	2	Alive	20	16	B	A
T28	80	male	.	2	Dead	16	15	B	B
T29	96	female	.	2	Dead	3	2	A	A
T30	89	female	Non-Smoker	2	Dead	26	18	A	A
T31	49	female	Non-Smoker	2	Alive	49	44	A	A
T32	58	male	Smoker	2	Alive	37	7	A	A
T33	82	male	Smoker	2	Dead	5	0	A	A
T34	41	male	Non-Smoker	2	Alive	34	34	B	B
T35	62	male	Smoker	2	Alive	32	32	B	B
T36	40	female	Smoker	2	Dead	12	4	A	A
T37	56	male	.	2	Alive	32	32	A	A
T38	42	male	Smoker	2	Dead	12	8	B	A
T39	59	male	Smoker	2	Alive	14	10	B	B
T40	61	male	Non-Smoker	2	Dead	13	8	A	A
T41	80	male	Smoker	2	Alive	27	24	B	B
T42	21	male	Non-Smoker	2	Dead	15	5	B	B
T43	57	male	Smoker	2	Alive	27	27	B	B
T44	75	female	Smoker	2	Alive	21	21	B	B
T45	68	male	Non-Smoker	2	Alive	23	23	B	B
T46	41	female	Smoker	3	Alive	21	21	A	A
T47	57	male	Smoker	3	Dead	6	6	A	A
T48	76	male	Non-Smoker	3	Dead	10	8	A	A
T49	43	male	Non-Smoker	3	Dead	7	0	B	B
T50	29	male	Smoker	3	Dead	24	8	A	A
T51	70	female	Smoker	3	Dead	3	0	A	A
T52	62	female	Smoker	3	Dead	8	0	A	A
T53	73	male	.	4	Dead	5	0	A	A
T54	74	female	Smoker	4	Dead	17	17	B	B
T55	76	female	Smoker	4	Dead	7	5	B	B
T56	61	male	.	4	Dead	19	0	A	A
T57	51	male	Smoker	4	Dead	4	0	A	A
T58	33	male	Non-Smoker	4	Dead	7	7	A	A
T59	89	female	.	4	Dead	4	0	B	B
T60	64	male	Smoker	4	Alive	55	55	B	B
T61	79	female	Smoker	4	Dead	6	0	A	A
T62	41	male	Non-Smoker	4	Alive	18	18	A	A
T63	76	female	Smoker	4	Alive	34	34	A	A
T64	67	female	.	4	Dead	10	10	A	A
T65	63	male	.	4	Dead	7	0	A	A

Figure 6.13: Subgroup IDs of the patients for SI and MI based subgrouping.

	Disease-free survival				Overall survival			
	Global P-value	HR	95% CI	P-value	Global P-value	HR	95% CI	P-value
SI Subgrouping		2.633	1.093-6.343	3.09E-02		2.705	1.124-6.508	2.63E-02
Tumor stage	9.55E-03	1.567	1.090-2.252	1.53E-02	1.51E-03	1.843	1.274-2.665	1.16E-03
Smoking behavior		1.029	0.454-2.333	9.45E-01		0.899	0.395-2.049	8.01E-01

(A)

	Disease-free survival				Overall survival			
	Global P-value	HR	95% CI	P-value	Global P-value	HR	95% CI	P-value
MI Subgrouping		2.881	1.143-7.264	2.49E-02		3.022	1.199-7.611	1.90E-02
Tumor stage	7.12E-03	1.587	1.102-2.285	1.31E-02	9.87E-04	1.868	1.289-2.707	9.59E-04
Smoking behavior		1.052	0.464-2.386	9.03E-01		0.918	0.403-2.089	8.38E-01

(B)

Figure 6.14: Results of COXPH survival analysis using (A) SI and (B) MI based subgrouping of patients, smoking behavior and tumor stage as explanatory variable.

planatory variables. Figure 6.11(A) and Figure 6.11(B) show results for SI-based or MI-based subgrouping and tumor stage. For each subgrouping, the combined P-value is statistically significant, showing that the two covariates are independently associated with overall (disease-free) survival time. The HR is higher for either SI-based subgrouping or MI-based subgrouping than tumor stage-based subgrouping, with statistically significant association for the SI-based and MI-based subgroupings, meaning that there is a significant difference in overall (disease-free) survival for the two subgroups after adjusting for tumor stage. MI-based subgrouping shows a higher HR and lower P-values (0.0029, 0.0033) for both disease-free and overall survival in comparison to SI-based subgrouping (0.0036,0.0045).

Figure 6.12 shows results from comparable experiments taking into account SI/MI-based subgrouping and patients smoking behavior. The P-values are statistically significant for both SI- and MI-based subgrouping, after adjusting for smoking behavior. The global P-values are statistically significant for the MI-based subgrouping experiment, and the HR confidence intervals are greater than 1 for both SI- and MI-based subgrouping and less than 1 for smoking-pattern-based subgrouping, except for prediction of disease-free survival when considered in combination with MI subgrouping. Therefore, the tree-based statistics provide additional information regarding the overall and disease-free patient survival beyond what can be achieved by looking at smoking behavior alone.

Figure 6.14 shows results of COXPH analysis between all three variables: SI/MI-based subgrouping, smoking behavior and tumor stage. SI- and MI-based subgroupings yield statistically significant results for prediction of both disease-free and overall survival after adjusting for both smoking behavior and tumor stage. The HRs are higher for SI and MI subgrouping compared to the other two.

Figure 6.13 shows the subgroup assignment of the patients based on SI and MI. Low SI and high MI patients are labeled as group "A", and high SI and low MI patients as group "B". The two experiments had the same subgroup assignments for all patients except T27 and T38, who were assigned to the short-surviving group by the MI subgrouping but the long-surviving group in the SI subgrouping.

6.8 Discussion

In this chapter, we built evolutionary models of tumor progression for each patient using the gene probes for *TERC, EGFR, CCND1* and *TP53* in tongue cancer. The concept that cancer is an evolutionary process and can be modeled using phylogenetic algorithms has been repeatedly demonstrated [8]. Using the tumor progression trees, we derived test statistics that are significantly associated with disease-free and overall survival in multivariate analysis, even after taking into account tumor stage and smoking.

We have developed a general analysis framework for FISH data sets using several gene markers on hundreds (here 250-450 cells per tumor) of single cells from numerous samples. In the previous chapter, the prostate cancer study compared non-paired progressing and non-progressing tumors, and one analysis objective was to derive a test statistic that could distinguish between these two categories. In this study, the patients were selected to have similar clinical parameters (e.g., almost identical Gleason scores), so that when testing for associations with prognosis, the clinical covariates would not be confounding factors.

In contrast, in the OTSCC study, we analyzed 65 oral tongue cancer samples with known

tumor stage, tobacco usage, and survival. The tumor stages ranged from 1 to 4 and the study included both smokers and non-smokers. Therefore, a principal objective of the data analysis was to derive a test statistic T that would (i) partition the cases into two or more categories and (ii) be associated with significantly different (disease-free) survival after taking into account tumor stage and smoking status. Survival analysis using single gene markers or two gene markers simultaneously showed no significant association with prognosis, contrary to some previous studies of the same genes, especially *CCND1* [2, 15, 24, 57, 77, 94, 98, 112, 114, 115, 116] but consistent with others [112, 128, 142, 181, 186]. One limitation of some of the previous studies claiming associations between some of the four genes and poor prognosis is that tumor stage was not considered as a covariate by these studies [2, 57, 61, 77, 94, 116].

Despite the different study designs, we have shown that the same phylogenetic analysis framework is effective in both prostate and tongue studies. In Chapters 2, 3 and 4, we have developed machine learning methods for identifying tree based features for predicting tumor stages. The work in this chapter and in Chapter 5 develops new directions in analyzing tumor phylogenetic data and finding associations between tumor progression characteristics and clinical parameters. The work in this chapter demonstrates that the tree structures capture important characteristics of tumor progression mechanisms and that characteristic features quantifying the basic tree topologies can be important predictors of patient survival time.

Chapter 7

Generalized Matching Distance and Its Application to Tumor Phylogenetics

Phylogenetic trees are used to infer evolutionary relationships among taxa representing biological species. Formal tree comparison measures are widely used to identify and to quantify commonalities and differences between trees. Such comparisons among phylogenies are also used for a variety of tasks that are not obviously tree-construction problems, such as studying symbiosis between host and pathogens [36], identifying horizontal gene transfer [1], and predicting protein-protein interactions [64].

A number of methods have been proposed to calculate distances between phylogenetic trees built on a common set of taxa. The most popular approach is the Robinson-Foulds (RF) [140] method, which measures the total number of bipartitons present in one tree but not in the other and vice versa. RF, however, returns poorly distributed distance values that do not discriminate well [27]. RF also lacks robustness in the face of small changes to the topology [102]. An alternative approach is to calculate edit distances between trees, meaning the minimum number of tree rearrangement operations needed to transform one tree into another [80, 184]. Three common tree operations are Nearest Neighbor Interchange (NNI), Subtree Pruning and Regrafting (SPR), and Tree Bisection and Reconnection (TBR). Computing edit distance using all three op-

erations is NP-hard [3, 39, 85], limiting their practical use. RF, by contrast, can be computed in polynomial time.

Lin *et al.* [102] proposed the Matching Distance (MD), a breakthrough in phylogenetic tree comparison because it is computable in polynomial time and is a metric, like RF, but is empirically well-distributed and robust to small changes like the intractable edit-distance methods. To compute MD(T_1, T_2), one starts with a complete bipartite graph $B(T_1, T_2)$ whose vertices are the bipartitions of T_1 and T_2 induced by removing one internal edge of either tree. Each bipartition b_i can be represented by two complementary 0/1 vectors indicating whether each taxon is present(1)/absent(0) in the first/second set of the bipartition. The weight of the edge (b_1, b_2) is the minimum Hamming distance between one of the two vectors representing b_1 and one of the two vectors representing b_2. Then, MD(T_1, T_2) is the weight of the minimum bipartite matching on $B(T_1, T_2)$. MD can be computed in polynomial time.

Another generalized version of the Robinson-Foulds (GRF) metric was proposed in [19]. The RF metric is absolute in the sense that it adds 0 or 1 if a particular bipartition is present in one tree or both trees. GRF relaxes this restriction by computing a matching between the bipartitions of two trees, and by assigning a cost function that measures the dissimilarity between the matched bipartitions based on the presence and absence of subsets of taxa. Computation of the GRF is NP-hard and an Integer Linear Program was proposed [19].

All these approaches, however, are limited in the kinds of trees to which they apply. RF and GRF assume that trees being compared are defined on identical sets of taxa and that no observed taxon is a tree ancestor of another observed taxon. These restrictions are inherited by MD and needed in the proof that MD is a metric [102]. Some of the edit distance methods are defined only for binary trees and the MD metric proof relies on the trees being binary. These tree structure assumptions break down in some important applications.

A motivation of the work in the present chapter is tumor phylogenetics, in which one builds phylogenies on tumor cell populations to understand tumor progression, including metastasis [8, 34, 55, 100, 134, 156]. In evaluating phylogenetic models of tumor evolution, one paradigm is to

start with a ground-truth tree T and an evolutionary model that is simulated on data related to T. The modeled method generates a tree S on the simulated data, and one compares S to T (over many replicates) to quantify the quality of the tumor progression model [71]. Many observed tumor states can evolve from another observed tumor state and create nodes of arbitrarily high degree due to intra-tumor heterogeneity, which is extensive in various real data sets [83, 153, 156]. Similar problems would be expected to arise in other applications of phylogenetics using small marker sets or short time scales, such as phylogenetics of viral strains or intraspecies phylogenetics arising in population genetic studies.

It would be desirable to generalize MD so that 1) the two taxa sets need not be identical, 2) a taxon can be at a non-leaf node and 3) the trees need not be binary, while retaining the polynomial time property. In Chapter 2, we established a generalized version of MD to meet the three criteria and applied it to compare tumor phylogenies inferred on single cell gene copy number data. In this chapter, we refer to the distance measure proposed in Chapter 2 as F^1. That generalization, though, resulted in a measure that was no longer a metric and could result in poor ability to distinguish even very different topologies in some situations. In this chapter, we address these problems by proposing a novel generalized matching distance function called F^2, which differs from F^1 in accounting for unmatched bipartitions in the graph B ignored by F^1 and MD. We show that F^2 improves upon the prior art by retaining the generality of F^1 over MD and RF, and improving discrimination of different trees relative to F^1.

Our principal contributions in this chapter are as follows:

1. We propose a novel weighted matching based formulation, $F^2(T_i, T_j)$, for measuring dissimilarities between two phylogenetic trees T_i and T_j. Our proposed measure, F^2, allows the trees to have different sets of taxa, allows trees to be non-binary, and allows taxa to be at non-leaf nodes.
2. We propose a pre-processing step to remove the taxa not in common to the two trees and show that this step helps avoid potential over-penalization of the F^2 distance measure.

3. We demonstrate empirically, through extensive analyses using the tree reconstruction paradigm [71], that F^2 better captures underlying dissimilarities between the trees in comparison to F^1 or to a similar generalization of RF.

7.1 Methods

In this section, we first define our proposed distance measure between a pair of trees, F^2. We then establish some key mathematical properties of F^2 in comparison to the earlier F^1 measure. Finally, we summarize data and methods used to empirically compare the performance of F^1, F^2, and RF.

7.1.1 Distance formulations

Removing an edge from a tree T_i creates two components, partitioning the taxa into two sets, called a *bipartition*. T_i can be uniquely represented by the set of bipartitions P_i. Let X_i denotes the set of taxa on which T_i is built; let the total number of nodes in T_i be n_i. The bipartitions derived by removing leaf edges have one singleton set and are called "trivial bipartitions" and are usually ignored for tree comparison purposes [102]. From here onward, we use the single word "bipartition" to mean "nontrivial bipartition," i.e., a bipartition arising from removal of an internal edge. The total number of bipartitions in T_i is denoted by r_i. In a leaf-labeled binary tree, $r_i = |X_i| - 2$, but the present work allows for trees that are not binary. The shared set of taxa between two trees T_i and T_j is X_{ij} ($X_{ij} = X_i \cap X_j$), and the cardinality of X_{ij} is n_{ij}.

To illustrate the notation, we present an example using two hypothetical trees T_i and T_j in Fig. 7.1(A) and Fig. 7.1(B), respectively. Here, $X_i = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $r_i = 3$, $n_i = 9$, $X_j = \{1, 2, 3, 6, 7, 8, 9\}$, $r_j = 2$, $n_j = 7$, $X_{ij} = \{1, 2, 3, 6, 7, 8, 9\}$ and $n_{ij} = 7$. The set of bipartitions P_i in T_i is

$$\{\{\{1, 3, 6, 7, 8, 9\}, \{2, 4, 5\}\}, \{\{3, 6, 7, 8, 9\}, \{1, 2, 4, 5\}\}, \{\{6, 8, 9\}, \{1, 2, 3, 4, 5, 7\}\}\}$$

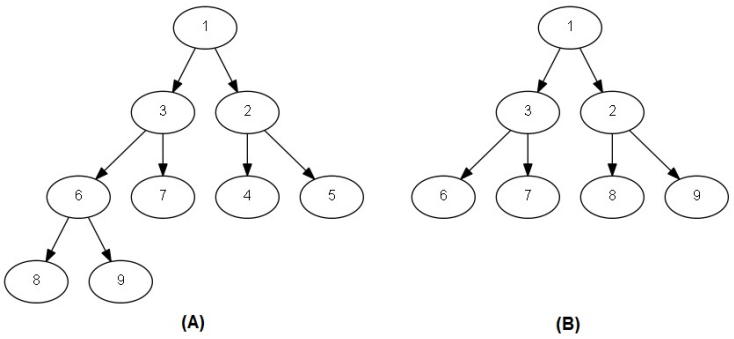

Figure 7.1: Trees (A) T_i and (B) T_j used to illustrate terminology used in the paper.

and the set of bipartitions P_j in T_j is $\{\{\{1,3,6,7\},\{2,8,9\}\},\{\{1,2,8,9\},\{3,6,7\}\}\}$.

In Chapter 2, we proposed the distance measure F^1, a generalization of MD to arbitrary sets of taxa, computed as follows. Given two trees T_i and T_j on any sets of taxa, the algorithm first computes the set of bipartitions P_i and P_j. It then builds a complete weighted bipartite graph $B(P_i, P_j)$, where each node in B represents a bipartition. For a pair of bipartitions $p_i \in P_i$ and $p_j \in P_j$, the weight of the edge connecting p_i and p_j is $\max(|p_i^1 \cap p_j^1| + |p_i^2 \cap p_j^2|, |p_i^1 \cap p_j^2| + |p_i^2 \cap p_j^1|)$, where p_i^1, p_i^2 and p_j^1, p_j^2 denote the sets of taxa in the bipartitions p_i and p_j respectively. Third, it computes the maximum weighted matching M_{ij} on $B(P_i, P_j)$. In the MD formulation for identical taxa sets, minimum-weight matching using the Hamming distance [102] and maximum-weight matching using $|X_i| -$ Hamming distance are equivalent. Here, we prefer the maximum-weight matching version. The F^1 distance, expressed as a percent disagreement between T_i and T_j using the weight of the maximum matching W_{ij}, is:

$$F_{ij}^1 = F^1(T_i, T_j) = \left(1 - \frac{W_{ij}}{n_{ij} \min(r_i, r_j)}\right) \times 100 = (1 - S_{ij}^1) \times 100 \tag{7.1}$$

If we apply the matching algorithm on P_i and P_j computed from the trees in Fig. 7.1, the first and second bipartitions in P_i are matched with the first and second bipartitions in P_j, respectively. The weight W_{ij} of the matching is 10. Plugging these values in the equation for F^1, we get $F_{ij}^1 = 28.57\%$.

F^1 can perform poorly on dissimilar trees because it considers only the bipartitions that

are members of M_{ij} and is oblivious to unmatched bipartitions. Trees built on arbitrary taxa sets may be very different on the unmatched bipartitions and F^1 then underestimates the tree dissimilarities. Unmatched bipartitions do not arise for MD because limiting to leaf-labeled binary trees on identical taxa sets [102] implies that the numbers of bipartitions in the two trees are equal.

To address this issue, we propose here a novel formula F^2 that considers all the bipartitions while calculating the distance:

$$F^2_{ij} = F^2(T_i, T_j) = \left(1 - \frac{W_{ij}}{n_{ij}(r_i + r_j) - W_{ij}}\right) \times 100 = (1 - S^2_{ij}) \times 100 \tag{7.2}$$

For the example trees presented in Fig. 7.1, $F^2_{ij} = 60\%$.

7.1.2 Theoretical properties of F^1 and F^2

Here, we describe some key theoretical properties of F^1 and F^2 that collectively show that F^2 better captures dissimilarity of trees with distinct taxa sets than F^1.

We first note that when taxa sets are dissimilar, it is possible to have arbitrarily different trees that nonetheless have zero F^1 distance. This observation is formalized in the following lemma:

Lemma 32. *For any integer q and tree T_i, there exists a tree T_j which differs from T_i in q edges, but for which $F^1_{ij} = 0$.*

Proof. We prove the lemma by construction, defining T_j to be any tree for which T_i is a subtree of T_j and the remainder of T_j has q arbitrary additional edges. Then, $W_{ij} = n_{ij} \min(r_i, r_j) = n_{ij} r_i$ and so $F^1_{ij} = 0$, although T_i and T_j differ from each other on the q edges. □

F^1 compares only the taxa in common and because of the term $\min(r_i, r_j)$ in the formulation of F^1, one can add arbitrarily many taxa to the larger of the two trees without changing F^1. In contrast F^2 uses the term $(r_i + r_j)$, so adding taxa to either tree, except as leaf children of the root, decreases F^2.

There is a complementary problem for F^2. Two trees can have arbitrarily high F^2 distance despite perfect agreement in all shared taxa. This assertion can be proven by construction and formally stated as follows:

Lemma 33. *For any $0 \leq \epsilon < 100$, there exist trees T_i and T_j such that $F^2_{ij} \geq \epsilon$ even though $\min(r_i, r_j)$ number of bipartitions of T_i and T_j completely match each other on the common set of nodes $X_{ij} = X_i \cap X_j$.*

Proof. Without loss of generality, suppose, $r_i = \min(r_i, r_j)$. If r_i bipartitions of T_i and T_j completely match with each other on X_{ij}, we have $W_{ij} = n_{ij} r_i$. Now,

$$\begin{aligned} F^2_{ij} &= \left(1 - \frac{W_{ij}}{n_{ij}(r_i + r_j) - W_{ij}}\right) \times 100 = \left(1 - \frac{W_{ij}}{n_{ij} r_i + n_{ij} r_j - W_{ij}}\right) \times 100 \\ &= \left(1 - \frac{n_{ij} r_i}{n_{ij} r_j}\right) \times 100 = \left(1 - \frac{r_i}{r_j}\right) \times 100 \end{aligned} \tag{7.3}$$

Now, if $r_j \geq \frac{100 r_i}{100 - \epsilon}$, then from the above, we have,

$$F^2_{ij} \geq 100 \left(1 - \frac{r_i}{\frac{100 r_i}{100 - \epsilon}}\right) \geq 100 \left(1 - \frac{r_i(100 - \epsilon)}{100 r_i}\right) \geq 100 \left(\frac{100 - 100 + \epsilon}{100}\right) \geq \epsilon \tag{7.4}$$

□

Depending on the application, the potential for high dissimilarity established by Lemma 33 may or may not be desired. To reduce the imbalance between the trees, one can remove the taxa not shared via a pruning procedure applied before computing the distance. We illustrate this procedure in Fig. 7.2. Fig. 7.2(A) and Fig. 7.2(B) show two trees T_i and T_j where the shared set of taxa is $\{1, 4, 5, 6, 7, 9\}$. In the pruning step, the nodes present in one tree but not in the other are selected for removal. We denote this set of nodes Y, where $Y = (X_i \cup X_j) \backslash (X_i \cap X_j)$. For each $y \in Y$, we remove y and associated edges and connect the children of y to the parent of

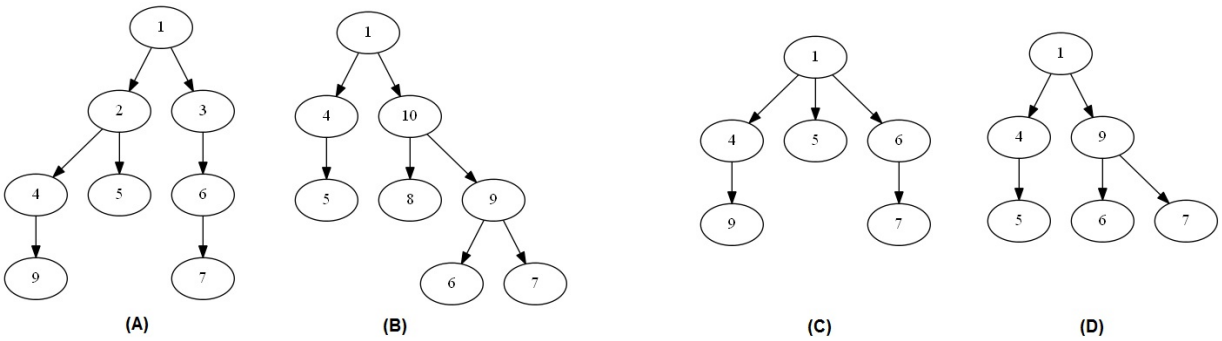

Figure 7.2: Example of pruning. Two trees (A) T_i and (B) T_j before and after ((C) T_i and (D) T_j) removal of the nodes not present in both of the trees.

y. The final pruned tree is the same for any order of node pruning. Fig. 7.2(C) and Fig. 7.2(D) show the two resulting trees after the pruning step on the trees in Fig. 7.2(A) and Fig. 7.2(B), respectively.

A theoretical disadvantage of F^1 is that it is not a metric. We can establish that F^1 is not a metric by showing that it is possible to construct trees T_i, T_j, and T_k such that $F^1(T_i, T_k) > F^1(T_i, T_j) + F^1(T_j, T_k)$, leading to the following formal statement:

Lemma 34. *There exist trees T_i, T_j, and T_k for which F^1 does not satisfy the triangle inequality, and hence F^1 is not a metric.*

Proof. We prove the lemma by construction. Suppose T_j is a subtree of both T_i and T_k. Then $W_{ij} = n_{ij} \min(r_i, r_j) = n_{ij} r_j$ and $W_{jk} = n_{jk} \min(r_j, r_k) = n_{jk} r_j$. As a consequence, $F^1_{ij} = 0$ and $F^1_{jk} = 0$. Now, suppose on X_{ik}, the set of taxa in common to T_i and T_k, each node has the same parent except for one edge where the parent-child relationship is reversed in T_k, meaning for a pair of node u and v, u is the parent of v in T_i, but v is the parent of u in T_k. Because of this, $W_{ik} = n_{ik} \min(r_i, r_k) - 2$, resulting in $F^1_{ik} > 0$. So, we have $F^1_{ik} > F^1_{ij} + F^1_{jk}$ and F^1 does not satisfy the triangle inequality. □

7.1.3 Tree generation methods

To evaluate empirically the performance of our proposed distance measure, we used tumor phylogenetic trees built on single-cell gene copy number data collected using fluorescence *in situ* hybridization (FISH), which allows one to count copy numbers of several genes in hundreds of single cells. For inference of tumor phylogenies, we used the FISHtrees software [34] which builds a maximum parsimony model of tumor progression using a combination of single gene duplication (SD) (gain or loss of a single gene), single chromosome duplication (CD) (gain or loss of all the genes residing on the same chromosome), or whole genome duplication (GD) (doubling of all gene copy numbers) events. The input to FISHtrees is three-dimensional; the first dimension is a set of files, one per sample. Each file is a two-dimensional matrix, where the columns are FISH probes and rows represent cells. Entry (c, p) is the number of copies of probe p in cell c. Cells with identical probe counts are combined into a "count pattern" and represented by a single node in the tree.

Neighbor Joining (NJ) and Maximum Parsimony (MP) are two of the most widely used algorithms for inferring single-tumor phylogenies [118, 161]. We performed experiments to compare the performances of F^1, F^2 and RF on phylogenetic trees built using NJ, MP and FISHtrees. In NJ, we treated the individual copy number profiles as taxa and used Euclidean distance. For MP, we considered copy numbers as arbitrary characters. We used implementations of NJ and MP in MEGA version 6 [167]. For MP, we used the "Close-Neighbor-Interchange on Random Trees" search method, and for the parameters "Number of Initial Trees" and "MP search level", we used the default values of 10 and 1 respectively.

7.1.4 Implementation of F^1, F^2 and RF

We wrote Matlab programs to compute F^1, F^2 and RF. For computing maximum matchings, we used the Hungarian algorithm [97] (function"Hungarian" by Alexander Melin from the Matlab Central File exchange).

We used the following procedure to calculate the distance between two particular trees T_i and T_j using RF:

1. Set $matched = 0$.
2. For each of the r_i bipartitions in T_i, if any of the r_j bipartitions in T_j matches completely with that bipartition, then increment $matched$ by 1.
3. For each of the r_j bipartitions in T_j, if any of the r_i bipartitions in T_i matches completely with that bipartition, then increment $matched$ by 1.
4. Calculate RF distance as a percentage = $(1 - \frac{matched}{r_i + r_j}) \times 100$.

7.2 Results

In this section, we present experimental results to evaluate our proposed formulation F^2 to measure the distance between phylogenetic trees in comparison to F^1 and the Robinson-Foulds (RF) metric. We first present results on the simulated dataset, which also reveal the importance of performing the pruning step before applying the F^2 formulation. Next, we present results on the real tumor datasets to evaluate F^1, F^2 and RF.

7.2.1 Generating trees for comparisons

To evaluate the performances of F^1, F^2 and RF, we used both simulated and real tumor datasets. We generated a simulated dataset of 100 trees with six probes, two of which were considered as being on the same chromosome and the remaining four residing on separate chromosomes. Each tree was generated by starting from a diploid (all probes have 2 copies) root node and executing a branching process in which each node was recursively assigned a number of children drawn from a geometrically distributed random variable with mean 0.50. Each child was generated from its parent by selecting a single gene duplication (SD), single chromosome duplication (CD), or whole genome duplication (GD) event with probability 11.67% for each of the six possible SD

events, 18% of a CD event, and 12% of a GD event. These probabilities were selected manually to approximately reflect apparent frequencies observed in real breast cancer data sets. This process terminated when all leaf nodes were assigned zero children by the sampling. We then generated simulated FISH input data derived from each tree by uniformly sampling 300 cells from the nodes. The simulated data correspond to probe counts for each sampled cell in the tree. We applied FISHtrees [34] to reconstruct a minimum-cost tree for the SD+CD+GD event model.

We also used single-cell copy number data collected from Cervical Cancer (CC) [183] and Tongue Cancer (TC) patients. These datasets consist of the following samples:

1. The CC [183] FISH data consist of 47 samples organized into 16 primary samples of metastatic patients, 16 paired metastasis samples from the same patients, and 15 primary samples from patients who did not progress to metastasis. Each sample has 223-250 cells profiled on four FISH probes: *LAMP3*, *PROX1*, *PRKAA1* and *CCND1*. The number of unique cell count patterns in a sample ranges between 46-213. All of the four genes are oncogenes, which typically show copy number gains in tumor cells.
2. The TC FISH data were collected from 65 patients whose tumors were at one of four clinical stages. Each sample consists of 241-250 cells and profiled on four FISH probes: *TERC*,*CCND1*, *EGFR* and *P53*. The number of unique cell count patterns is 73-244. *TERC*,*CCND1* and *EGFR* are oncogenes, and *P53* is a tumor suppressor.

7.2.2 Comparison of F^1, F^2, and RF using simulated data

We performed experiments to evaluate F^1, F^2, and RF as the number of sampled cells varied. We sampled 20, 50, 100, 150, and 200 cells from each of the 100 simulated replicates and inferred tumor phylogenies using FISHtrees [34]. Then, we used F^1, F^2 and RF to calculate the distance between each pair of simulated and inferred trees according to the tree reconstruction paradigm.

Fig. 7.3 and Table 7.1 show F^1 and F^2 mean values and standard deviations as functions of size of simulated subtrees. Table 7.1 shows the mean number of bipartitions in the simulated and

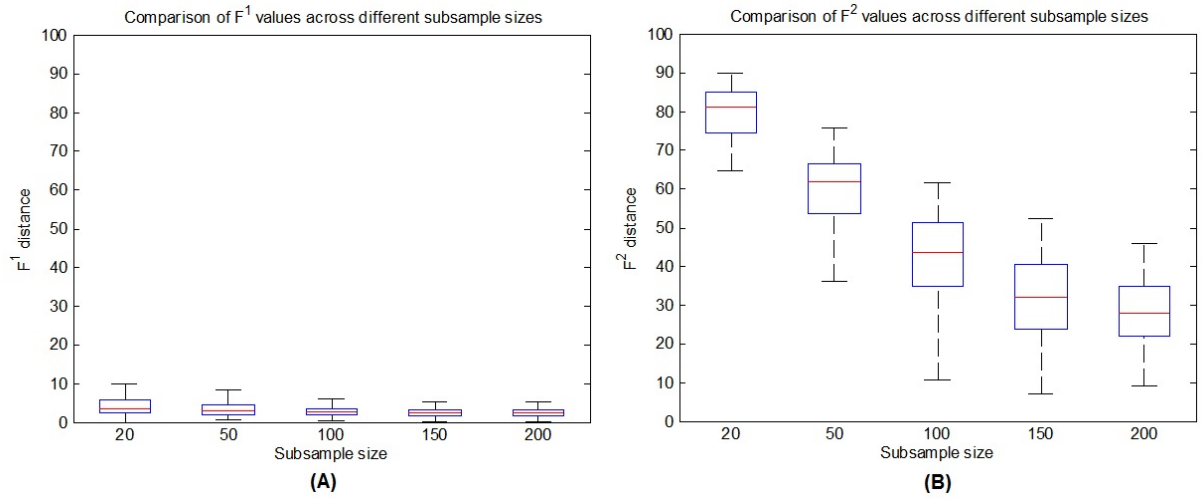

Figure 7.3: Comparison of (A) F^1 and (B) F^2 distance values between simulated and inferred trees without pruning taxa. Comparisons are made for subsample sizes of 20, 50, 100, 150 and 200.

inferred trees (mean r_S and mean r_I), mean cardinality of the common taxa set between each pair of simulated and inferred trees (mean n_{SI}), mean F^1, and mean F^2 values with standard deviation (s.d.) for simulated tree comparisons without pruning. Since the simulated and inferred trees have taxa not in common between them, RF is not defined on these data and is therefore omitted from the figure and the table. We would expect that increasing sample sizes would lead to more accurate inferences and thus to lower distances between simulated and inferred trees. The figure shows that both mean F^1 and mean F^2 distances decrease with increasing sample size, as expected. F^2 values decrease at a substantially higher rate than mean F^1 values, though, and thus provide clearer discrimination between different levels of similarity.

Table 7.1: F^1 and F^2 distance values for trees inferred from simulated data on different subsamples of the simulated dataset without pruning taxa.

# cells	mean r_S	mean r_I	mean n_{SI}	mean F^1(s.d.)	mean F^2(s.d.)
20	57.66	11.42	21.25	4.14 (2.23)	79.59 (6.27)
50	57.66	22.63	40.88	3.49 (1.79)	60.20 (9.57)
100	57.66	33.08	57.55	3.09 (1.58)	42.60 (10.87)
150	57.66	39.41	69.00	2.68 (1.21)	32.04 (10.34)
200	57.66	42.32	74.16	2.79 (1.69)	27.88 (8.80)

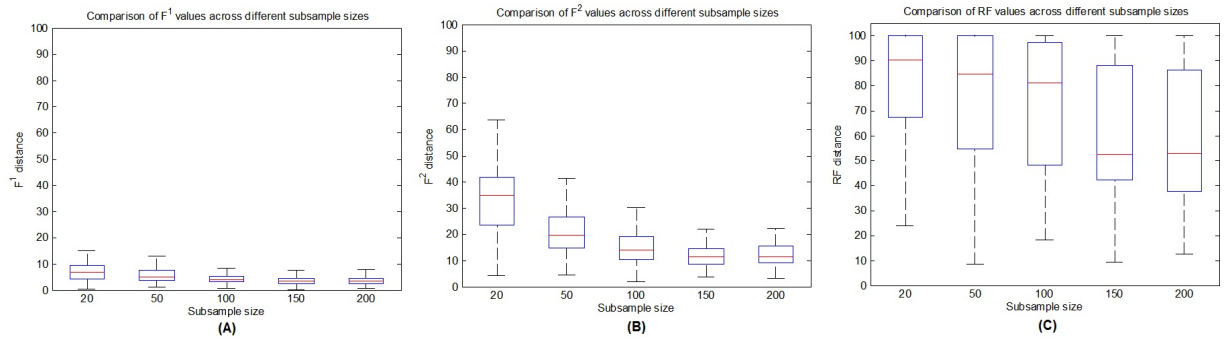

Figure 7.4: Comparison of (A) F^1, (B) F^2, and (C) RF distance values between simulated and inferred trees after pruning trees in each pairwise comparison to contain only their shared taxa. Comparisons are made for subsample sizes of 20, 50, 100, 150 and 200.

One important contributing factor to the improved F^2 performance may be the high imbalance in the mean number of bipartitions between the simulated and inferred trees, which is factored into F^2 but not F^1 scores. To test whether imbalance alone explains the better separation by F^2 scores, we pruned the two trees in each comparison to a common set of taxa as described in Methods and repeated the pairwise comparisons. This pruning also makes it possible to compute RF scores and these were included in comparison with F^2 and F^1 scores. Fig. 7.4 and Table 7.2 shows the results. Pruning reduces the imbalance in the mean number of bipartitions between the simulated and inferred trees. As a consequence, the mean F^2 values decrease in comparison to no pruning, but still show substantially better ability to separate trees than F^1. RF is substantially inferior to both methods, showing much higher variance at all sample sizes and essentially no ability to identify the trend of increasing similarity with sample size.

Table 7.2: F^1 and F^2 distance values for trees inferred on simulated data after pruning is performed on the trees.

# cells	mean r_S	mean r_I	mean n_{SI}	mean F^1 (s.d.)	mean F^2 (s.d.)	mean RF (s.d.)
20	8.7	11.42	21.25	7.04 (3.57)	33.66 (14.33)	81.05 (23.66)
50	20.48	22.63	40.88	5.72 (2.6)	20.43 (7.93)	71.94 (25.32)
100	31.49	33.08	57.55	4.53 (2.04)	15.28 (6.35)	72.16 (26.46)
150	39.32	39.41	69.00	3.72 (1.57)	11.80 (4.01)	59.96 (25.16)
200	42.55	42.32	74.16	3.86 (1.98)	11.91 (4.34)	59.05 (25.21)

Fig. 7.5 provides a scatter plot of the F^1 vs. F^2 distance values for trees inferred on samples

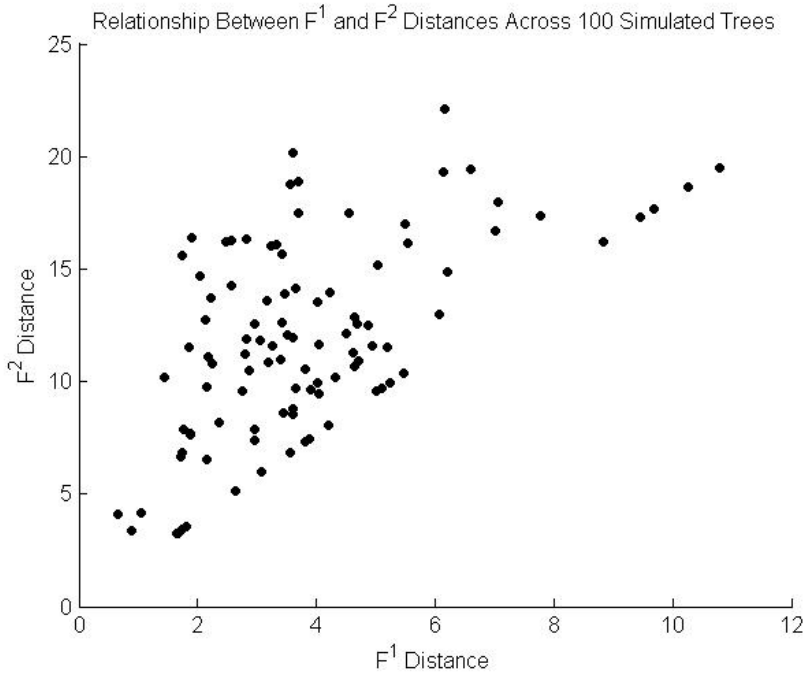

Figure 7.5: Scatter plot showing relationship of F^1 and F^2 distances between simulated and inferred trees.

consisting of 200 cells. The figure shows that for most values of F^1, F^2 exhibits a wide range of values and thus can discriminate the trees better that F^1. The correlation coefficient of F^1 and F^2 is 0.58. From this figure, it can also be seen that F^1 has a narrower distribution, concentrated in a narrow region near 0 (with mean 3.72 and s.d. 1.83) compared to F^2. The larger spread of the distribution of F^2 (s.d. of 11.98) again indicates that it provides better discrimination between more or less similar pairs of trees.

7.2.3 Comparison of performances of F^1, F^2, and RF in assessing tree-building algorithms

One motivation of the present work is to provide a sound basis for comparing qualities of inferences of different tree-building algorithms on common data sets. Neighbor Joining (NJ) and Maximum Parsimony (MP) are two of the most widely used algorithms for inferring single-tumor phylogenies [118, 161]. We therefore performed experiments to compare F^1, F^2 and RF

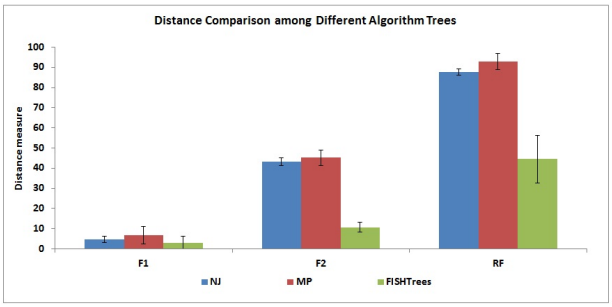

Figure 7.6: Mean F^1, F^2, and RF distance values illustrating the performances of NJ, MP, and FISHtrees algorithms to infer trees on the simulated data.

in measuring the distances among phylogenetic trees built using NJ, MP and FISHtrees. We applied the three algorithms to infer tumor phylogenies on the simulated data, deriving trees for each method on 100 simulated replicates.

The mean distances are shown in Fig. 7.6. F^1 tends to infer low distance values for all three methods, RF higher values, and F^2 values intermediate between the two. F^1 shows the worst performance in distinguishing the methods, showing variations between the three methods within the noise level of comparisons within methods. Both F^2 and RF show statistically indistinguishable perfomance for NJ and MP but substantially better performance for FISHrees. We quantified the degree of separation of the FISHtrees vs. NJ/MP trees by the "fold change" ratio between mean NJ/MP distance values and mean FISHtrees distance values for each of the distance measures. F^2 produced fold changes of 4.02 and 4.2 for FISHtrees vs. NJ and FISHtrees vs MP trees, respectively. RF produced fold changes of 2.18 and 2.08 for FISHtrees vs. NJ and FISHtrees vs. MP trees. F^1 produced fold changes of 1.51 and 1.97 for FISHtrees vs. NJ and FISHtrees vs. MP trees, respectively. F^2 is thus the most effective at bringing out differences in accuracy between the three methods, with F^1 the least effective and RF intermediate.

7.2.4 Comparison of F^1, F^2, and RF using real cancer data

We next performed experiments to compare F^1, F^2, and RF in distinguishing trees from tumor samples in the real CC and TC datasets in order to demonstrate the applicability of the proposed

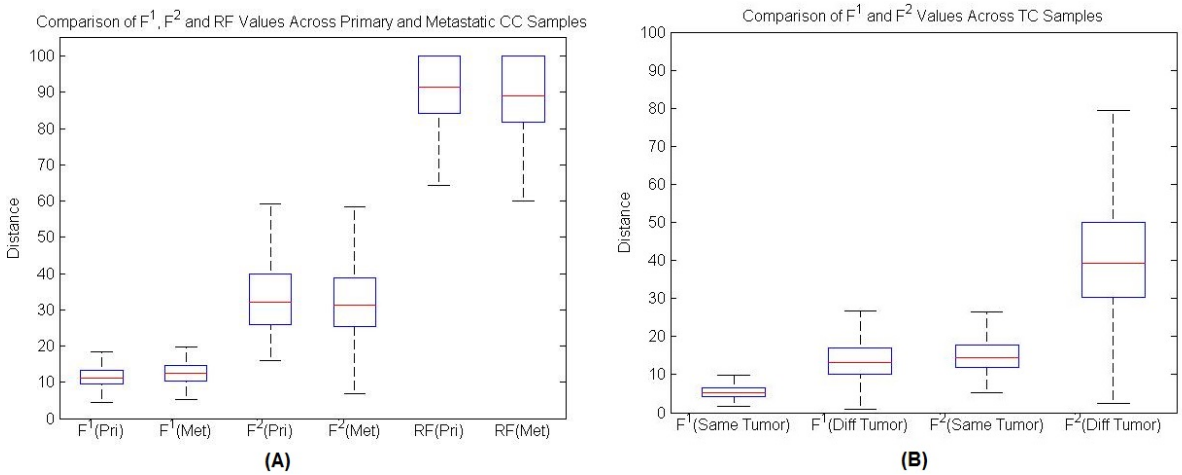

Figure 7.7: Comparison of different distance formulations based on trees inferred from real tumor data derived from (A) CC and (B) TC samples.

method to real biological data.

For each of the 16 pairs of primary and metastatic samples in the CC dataset, we used FISHtrees to infer the tumor phylogenetic trees and then measured the pairwise distances between the trees for each pair of primary samples and for each pair of metastatic samples. Fig. 7.7(A) shows distributions of scores for each of the comparisons. We know of no reason to expect primary trees to be more similar to one another than metastatic trees are to one another or vice versa, and indeed the two situations yield comparable distributions of scores for all three methods. F^2 appears to show the greatest range across sample pairs, however, with F^1 measures concentrated at small values and RF at large values.

We next examined the ability of F^1 and F^2 to distinguish trees inferred on subsamples of the same tumor versus different tumors. We created 30 random subsamples of 200 cells from each of the 65 tongue cancer (TC) samples, inferred tumor phylogenetic trees using FISHtrees, and calculated pairwise F^1 and F^2 distances between all pairs of trees inferred on different tumors and all pairs inferred on different subsamples of the same tumors. The box plots in Fig. 7.7(B) show a comparison of the distances for trees inferred from subsamples of the same tumor versus subsamples from different tumors. F^2 better highlights the lower distances observed for trees created from the same tumor versus those created from distinct tumors than does F^1. Further-

more, F^2 exhibits a greater spread than F^1 within each class, again suggesting that F^2 is better able to discriminate among subtle differences in levels of agreement.

7.3 Discussion

We introduced a formulation, F^2, that computes a distance between two phylogenetic trees built on arbitrary sets of taxa, where the taxa can be non-leaf nodes. F^2 generalizes the Matching Distance [102], which provided a more sensitive measure of tree distance than the prevailing Robinson-Foulds measure by finding optimum matching in a weighted bipartite graph of tree bipratitions. Like MD and RF, F^2 is computable in polynomial time and is therefore scalable to large trees. We further proposed a pruning-based preprocessing step that removes the taxa not shared between two trees in order to avoid potential over-penalization of unmatched bipartitions. We tested F^2 by comparing the inferred and ground truth trees in a simulated dataset and have shown that it exhibits better discrimination in comparison to RF and the earlier F^1 generalization of MD. The proposed formulation also better distinguishes the phylogenetic trees built on real FISH gene copy number data collected from tumors. Besides tumor phylogenetics, the proposed formulation can be used to measure distance between phylogenies in other contexts where the trees do not share taxa. One potential application is microbial metagenomics, where small sampling from some environment (e.g., gastrointestinal tract) misses some taxa and repeated samples are unlikely to have the exact same set of taxa. The method may also be useful in statistical applications of phylogenetics involving leave-one-out cross validation or bootstrapping over taxa.

Chapter 8

Conclusions and Future Directions

Inference of tumor phylogenies reflecting the underlying mechanisms of tumor progression is an important but challenging problem, for which, to the best of our knowledge, no general, reliable algorithms have yet been developed. In this thesis, our goal was to develop novel theory and algorithms specifically tuned to the underlying mechanisms of tumor evolution. An evolutionary perspective has long been recognized as important for explaining the intratumor and intertumor heterogeneity and understanding practical challenges to cancer treatment, such as the emergence of disease resistance after treatment. In this thesis, we started with a simple single gene duplication model of tumor progression in Chapter 2 and developed an exact and an heuristic algorithm for inference of tumor phylogenies. These algorithms allowed us, for the first time, to develop a procedure that can build tumor progression models considering gene copy number data collected across hundreds to thousands of cells, and thus amenable to capturing realistic levels of intratumor heterogeneity. We also developed a framework for identification of characteristic features of tumor progression allowing us to highlight the different selective pressures working on different stages of cancer. Finally, we developed machine learning algorithms to identify tree-based features that are informative about future tumor progression, providing a new strategy for developing phylogenies for use in cancer diagnosis and treatment.

In Chapter 3, we extended our phylogeny inference model from Chapter 2 and derived novel

theory and algorithms for building a model of tumor evolution incorporating changes at the scale of full chromosomes and all probes in the genome, in addition to single gene probes. We derived algorithms to reconstruct maximum parsimony sequences of single gene, single chromosome and whole genome duplication events, and thus estimates of evolutionary distance, between pairs of cells assayed by FISH probes. We incorporated these inferences into our method for building phylogenies of hundreds of cells in single tumors. Simulation experiments showed that addition of chromosome and genome level duplication events improve phylogeny inference accuracy by the comprehensive model relative to the the simpler single gene duplication model. Applying the new method on the real tumor data resulted in inference of more parsimonious models of tumor progression and improvement of tumor diagnosis and prognosis.

The major limitation of the models developed in Chapters 2 and 3 was that all the gain/loss events at gene, chromosome and genome level were assumed to occur at equal rates. In cancer cells, however, these events occur at different rates, and thus it is necessary to take this into account for building a realistic model of tumor progression. In Chapter 4, we developed algorithms for inferring tumor-specific mutation parameters and applied these to improve single-tumor phylogenetic tree inference at the cellular level. We proposed novel theory and algorithms to develop a weighted parsimony model of copy number evolution. We also proposed an iterative EM algorithm for joint inference of parameters and tree topologies. We showed that the resulting models provide insight into tumor progression mechanisms and lead to improved prediction of future tumor progression in multiple tumor types.

In Chapters 5 and 6, our goal was to demonstrate further utility of the trees in understanding progression mechanisms and predicting patient outcomes across different types of tumors. In Chapter 5, we derived test statistics that are significantly associated with prostate cancer progression. In Chapter 6, we performed extensive survival analyses on tongue cancer patients to show that the tree based features improve prediction of overall and disease-free survival time, even after taking into account tumor stage and smoking information. These results show that the tumor phylogenetic models can help us better understand the genetic alterations in diverse types

of cancers.

In Chapter 7, we proposed a novel distance measure to compute the dissimilarities between two phylogenetic trees built on arbitrary sets of taxa, where the taxa can be non-leaf nodes. There is no method in the literature that can compare phylogenies with these two characteristics. To avoid potential over-penalization due to unmatched bipartitions, we also proposed a pruning method to remove taxa not shared between two phylogenies. Experiments on simulated and real tumor data showed that the proposed distance measure exhibits better discrimination of phylogenetic trees in comparison to the widely used Robinson Foulds method and a simpler extension of the matching distance proposed by Lin *et al.* [102].

8.1 Future directions

The work in this thesis makes an important step towards scalable algorithms for inferring cancer-specific evolution in single tumors. The theory and algorithms presented here provide initial developments in inferring large scale tumor phylogenies. However, further improvements of the models are possible and likely to be needed to take full advantage of emerging new data on tumor heterogeneity. One important direction is improving upon the theory of the heuristic approximations used in Steiner tree inference. In this thesis, we have not performed any theoretical analysis to provide an upper bound on the performance guarantee of the heuristic algorithm. Also, the algorithms developed in this thesis currently consider only three mechanisms of copy number evolution. However, gene copy number change events can happen with other mechanisms too, such as breakage fusion bridge cycle, chromothripsis, chromoplexy etc. The evolutionary model might be further extended to consider these and other mutational mechanisms by which copy number profiles of tumor cells may evolve beyond the three that have been discussed.

The work presented in this thesis focused on FISH data. Although FISH is currently the only technology to reliably profile gene copy numbers across hundreds to thousands of cells per patient in sizable patient populations, it has its own limitations. One limitation is that only a

limited number of genetic markers can be profiled using FISH. Another limitation is that FISH can capture only copy number variations. Other types of variations, such as single nucleotide variations, which are also important in the evolution of tumors, cannot be captured using FISH. Although single-cell sequencing technology is still far from becoming practical for the number of cells needed for the questions we examine, one can reasonably anticipate that it will eventually become the dominant technology to perform comparable cross-tumor studies. There would thus be value in extending the theory developed here to single-cell sequencing data. However, there are substantial algorithmic challenges as a result of the much larger number and variety of markers single-cell sequencing can identify and the novel sources of noise arising from single-cell amplification.

In this thesis, we have taken some first steps towards using tumor phylogenetics as a source of features to predict future tumor progression. While these results show the promise of these directions, they also suggest many avenues for improvement. We have considered three different feature types: i) Tree based features (e.g., distribution of cells across different tree levels), ii) Raw cell count based features (e.g.,average copy number of each gene across all the cells) and iii) Diversity based features (e.g., shannon entropy of gene copy number distribution). We have used each feature types separately for prediction of tumor progression. To improve the prediction power, we can consider aggregating these different classes of features. We can also consider various machine learning methods, such as LASSO [172] to identify the set of most informative features from these aggregated features. We have used limited feature selection in Chapter 2, which resulted in improved classification performance. There we had smaller number of features which allowed us enumeration of all subsets of features to identify the most informative feature subset. But with the aggregate features, such exhaustive enumeration is not possible and we need to use statistical measures, such as LASSO to identify the best feature set.

The survival analyses that we performed in this thesis focused on identification of tree features for clustering patients into short and long survival groups. One prediction task that may have important clinical significance is prediction of survival time as a quantitative measure, as

opposed to clustering long versus short survival time as binary states. Using regression, we can fit survival time for the patients with mortality during follow-up and can treat the remainder as unknown variables that we can model as exponential random variables restricted to exceed follow-up time.

Successful cancer treatment depends on genetic diversity of the patients and future evolutionary trajectory of a cancer. The current state of cancer may not dictate how cancer should be treated, rather the focus should be on how it is likely to evolve in future. The work in this thesis makes advances in predicting future evolutionary trajectories of individual tumors using a strategy based on computational inference of those tumors' likely evolutionary trajectories to date. The tumor phylogenies developed by our algorithms have been shown to be predictive of multiple measures of progression across multiple tumor types, and to a greater degree than raw measures of average tumor state or cruder measures of tumor heterogeneity. The phylogenies capture important feature of the process how tumor is evolving and this process is the major driver of future tumor progression. In summary, the work in this thesis represents a crucial step forward in developing models of tumor progression as a new source of clinical guidance for individualized patients.

Bibliography

[1] Addario-Berry, L., Hallett, M.T., Lagergren, J.: Towards identifying lateral gene transfer events. In: Pacific Symposium on Biocomputing. vol. 8, pp. 279–290 (2003) 7

[2] Åkervall, J.A., Michalides, R.J., Mineta, H., Balm, A., Borg, Å., Dictor, M.R., Jin, Y., Loftus, B., Mertens, F., Wennerberg, J.P.: Amplification of cyclin d1 in squamous cell carcinoma of the head and neck and the prognostic value of chromosomal abnormalities and cyclin d1 overexpression. Cancer 79(2), 380–389 (1997) 6.8

[3] Allen, B.L., Steel, M.: Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics 5(1), 1–15 (2001) 7

[4] Almendro, V., Cheng, Y., Randles, A., Itzkovitz, S., Marusyk, A., Amettler, E., Gonzalez-Farre, X., Muñoz, M., Rusness, H.G., Helland, Å., Rye, I.H., Borresen-Dale, A., Maruyama, R., van Oudenaarden, A., Dowsett, M., Jones, R.L., Reis-Filho, J., Gascon, P., Gönen, M., Michor, F., Polyak, K.: Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. Cell Reports 6(3), 514–527 (2014) 3.3

[5] Ambatipudi, S., Gerstung, M., Pandey, M., Samant, T., Patil, A., Kane, S., Desai, R.S., Schäffer, A.A., Beerenwinkel, N., Mahimkar, M.B.: Genome-wide expression and copy number analysis identifies driver genes in gingivobuccal cancers. Genes, Chromosomes and Cancer 51(2), 161–173 (2012) 6

[6] Anderson, A.R., Weaver, A.M., Cummings, P.T., Quaranta, V.: Tumor morphology and

phenotypic evolution driven by selective pressure from the microenvironment. Cell 127(5), 905–915 (2006) 5

[7] Anderson, K., Lutz, C., Van Delft, F.W., Bateman, C.M., Guo, Y., Colman, S.M., Kempski, H., Moorman, A.V., Titley, I., Swansbury, J., et al.: Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature 469(7330), 356–361 (2011) 1.1, 1.6

[8] Attolini, C.S.O., Michor, F.: Evolutionary theory of cancer. Annals of the New York Academy of Sciences 1168, 23–51 (2009) 2, 3, 6.8, 7

[9] Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al.: Punctuated evolution of prostate cancer genomes. Cell 153(3), 666–677 (2013) 1.3

[10] Bandelt, H., Forster, P., Röhl, A.: Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution 16(1), 37–48 (1999) 1.6, 2.1.1, 1, 2

[11] Beaumont, M.A.: Approximate bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics 41, 379–406 (2010) 1.6, 4

[12] Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., Lengauer, T.: Learning multiple evolutionary pathways from cross-sectional data. Journal of Computational Biology 12(6), 584–598 (2005) 1.5, 2, 3

[13] Beerenwinkel, N., Rahnenführer, J., Kaiser, R., Hoffmann, D., Selbig, J., Lengauer, T.: Mtreemix: a software package for learning and using mixture models of mutagenetic trees. Bioinformatics 21(9), 2106–2107 (2005) 3

[14] Beerenwinkel, N., Schwarz, R.F., Gerstung, M., Markowetz, F.: Cancer evolution: mathematical models and computational inference. Systematic biology 64(1), e1–e25 (2015) 1.3, 1.5, 4

[15] Bellacosa, A., Almadori, G., Cavallo, S., Cadoni, G., Galli, J., Ferrandina, G., Scambia,

G., Neri, G.: Cyclin d1 gene amplification in human laryngeal squamous cell carcinomas: prognostic significance and clinical implications. Clinical Cancer Research 2(1), 175–180 (1996) 6.8

[16] Bilke, S., Chen, Q.R., Westerman, F., Schwab, M., Catchpoole, D., Khan, J.: Inferring a tumor progression model for neuroblastoma from genomic data. Journal of Clinical Oncology 23(29), 7322–7331 (2005) 3

[17] Birchmeier, W., Behrens, J.: Cadherin expression in carcinomas: role in the formation of cell junctions and the prevention of invasiveness. Biochim Biophys Acta (BBA)-Reviews on Cancer 1198(1), 11–26 (1994) 1.7, 2.1, 3.2.2

[18] Bleyer, A., Welch, H.G.: Effect of three decades of screening mammography on breast-cancer incidence. New England Journal of Medicine 367(21), 1998–2005 (2012) 1.2, 2

[19] Böcker, S., Canzar, S., Klau, G.W.: The generalized Robinson-Foulds metric. In: Algorithms in Bioinformatics, pp. 156–169. Springer (2013) 7

[20] Boffetta, P., Hecht, S., Gray, N., Gupta, P., Straif, K.: Smokeless tobacco and cancer. The lancet oncology 9(7), 667–675 (2008) 6

[21] Bogojeska, J., Alexa, A., Altmann, A., Lengauer, T., Rahnenführer, J.: Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. Bioinformatics 24(20), 2391–2392 (2008) 3

[22] Bogojeska, J., Lengauer, T., Rahnenführer, J.: Stability analysis of mixtures of mutagenetic trees. BMC Bioinformatics 9, 165 (2008) 3

[23] Bouchardy, C., Rapiti, E., Fioretta, G., Laissue, P., Neyroud-Caspar, I., Schäfer, P., Kurtz, J., Sappino, A.P., Vlastos, G.: Undertreatment strongly decreases prognosis of breast cancer in elderly women. Journal of Clinical Oncology 21(19), 3580–3587 (2003) 1.2

[24] Bova, R.J., Quinn, D.I., Nankervis, J.S., Cole, I.E., Sheridan, B.F., Jensen, M.J., Morgan, G.J., Hughes, C.J., Sutherland, R.L.: Cyclin d1 and p16ink4a expression predict reduced

survival in carcinoma of the anterior tongue. Clinical Cancer Research 5(10), 2810–2819 (1999) 6.8

[25] Brawley, O.W.: Prostate cancer epidemiology in the united states. World journal of urology 30(2), 195–200 (2012) 5.1

[26] Brennan, J.A., Boyle, J.O., Koch, W.M., Goodman, S.N., Hruban, R.H., Eby, Y.J., Couch, M.J., Forastiere, A.A., Sidransky, D.: Association between cigarette smoking and mutation of the p53 gene in squamous-cell carcinoma of the head and neck. New England Journal of Medicine 332(11), 712–717 (1995) 6.1

[27] Bryant, D., Steel, M.: Computing the distribution of a tree metric. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6, 420–426 (2009) 7

[28] Buckley, C., Beards, C., Fox, H.: Pathological prognostic indicators in cervical cancer with particular reference to patients under the age of 40 years. BJOG: An International Journal of Obstetrics & Gynaecology 95(1), 47–56 (1988) 2

[29] Cahill, D.P., Kinzler, K.W., Vogelstein, B., Lengauer, C.: Genetic instability and darwinian selection in tumours. Trends in cell biology 9(12), M57–M60 (1999) 1.6, 2, 5

[30] Campbell, P.J., Pleasance, E.D., Stephens, P.J., Dicks, E., Rance, R., Goodhead, I., Follows, G.A., Green, A. R. Futreal, P.A., Stratton, M.R.: Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proceedings of the National Academy of Sciences 105(35), 13081–13086 (2008) 3

[31] Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.L., et al.: The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467(7319), 1109–1113 (2010) 1.1

[32] Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., et al.: Improved survival with vemurafenib in

melanoma with braf v600e mutation. New England Journal of Medicine 364(26), 2507–2516 (2011) 1.2

[33] Cheng, Y.K., Beroukhim, R., Levine, R.L., Mellinghoff, I.K., Holland, E.C., Michor, F.: A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. PLoS Computational Biology 8(1), e1002337 (2012) 2

[34] Chowdhury, S.A., Shackney, S.E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A.A., Schwartz, R.: Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. Bioinformatics 29(13), i189–i198 (2013) 1, 3, 2, 3.1.4, 3.2, 2, 3.2.1, 3.2.2, 3.2.2, 3.3, 4.2.1, 4.2.3, 4.2.4, 5.3, 6.6, 7, 7.1.3, 7.2.1, 7.2.2

[35] Chowdhury, S.A., Shackney, S.E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A.A., Schwartz, R.: Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. PLoS Computational Biology 10(7), e1003740 (2014) 1, 4, 4.1.2, 4.2.3, 4.2.4, 4.3, 6.3, 6.6

[36] Currie, C.R., Wong, B., Stuart, A.E., et al.: Ancient tripartite coevolution in the attine ant-microbe symbiosis. Science 299(5605), 386–388 (2003) 7

[37] Dahlgren, L., Dahlstrand, H.M., Lindquist, D., Högmo, A., Björnestål, L., Lindholm, J., Lundberg, B., Dalianis, T., Munck-Wikland, E.: Human papillomavirus is more common in base of tongue than in mobile tongue cancer and is a favorable prognostic factor in base of tongue cancer patients. International Journal of Cancer 112(6), 1015–1019 (2004) 6.1

[38] Das, P.M., Singal, R.: Dna methylation and cancer. Journal of Clinical Oncology 22(22), 4632–4642 (2004) 1.3

[39] DasGupta, B., He, X., Jiang, T., et al.: On distances between phylogenetic trees. In: Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 427–436. Society for Industrial and Applied Mathematics (1997) 7

[40] De Bono, J., Ashworth, A.: Translating cancer research into targeted therapeutics. Nature

467(7315), 543–549 (2010) 1.2

[41] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) pp. 1–38 (1977) 1.5, 4

[42] Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., Schäffer, A.A.: Inferring tree models of oncogenesis from comparative genomic hybridization data. Journal of Computational Biology 6(1), 37–51 (1999) 1.5, 2, 3, 4

[43] Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., Schäffer, A.A.: Distance-based reconstruction of tree models for oncogenesis. Journal of Computational Biology 7(6), 789–803 (2000) 1.5, 1.6, 3, 5

[44] Di Noia, J., Neuberger, M.: Molecular mechanisms of antibody somatic hypermutation. Annual Review of Biochemistry 76, 1–22 (2007) 1.3, 4

[45] Ding, L., Raphael, B.J., Chen, F., Wendl, M.C.: Advances for studying clonal evolution in cancer. Cancer Letters 340(2), 212–219 (2013) 3

[46] Dreyfus, S.E., Wagner, R.A.: The steiner problem in graphs. Networks 1(3), 195–207 (1971) 2.1.2

[47] Druker, B.J.: Sti571 (gleevec) as a paradigm for cancer therapy. Trends in Molecular Medicine 8(4), S14–S18 (2002) 1.2

[48] Edmonds, J.: Optimum branchings. Journal of Research of the National Bureau of Standards B 71(4), 233–240 (1967) 1.5

[49] Elledge, R.M., McGuire, W.L.: Prognostic factors and therapeutic decisions in axillary node-negative breast cancer. Annual Review of Medicine 44(1), 201–210 (1993) 2

[50] Fearon, E., Vogelstein, B.: A genetic model for colorectal tumorigenesis. Cell 61(5), 759–767 (1990) 1.1, 3

[51] Felsenstein, J., Felenstein, J.: Inferring phylogenies, vol. 2. Sinauer Associates Sunderland

(2004) 1.4

[52] Fischer, A., Vázquez-García, I., Illingworth, C.J., Mustonen, V.: High-definition reconstruction of clonal composition in cancer. Cell Reports (2014) 1.5

[53] Fisher, R., Pusztai, L., Swanton, C.: Cancer heterogeneity: implications for targeted therapeutics. British Journal of Cancer 108(3), 479–485 (2013) 1.2, 4

[54] Flaherty, K.T., Puzanov, I., Kim, K.B., Ribas, A., McArthur, G.A., Sosman, J.A., O'Dwyer, P.J., Lee, R.J., Grippo, J.F., Nolop, K., et al.: Inhibition of mutated, activated braf in metastatic melanoma. New England Journal of Medicine 363(9), 809–819 (2010) 1.2

[55] Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G., Shapiro, E.: Cell lineage analysis of a mouse tumor. Cancer Research 68(14), 5924–5931 (2008) 3, 7

[56] Fu, M., Wang, C., Li, Z., Sakamaki, T., Pestell, R..: Minireview: Cyclin D1: normal and abnormal functions. Endocrinology 145(12), 5439–5447 (2004) 1.7, 2.1, 3.2.2

[57] Fujii, M., Ishiguro, R., Yamashita, T., Tashiro, M.: Cyclin d1 amplification correlates with early recurrence of squamous cell carcinoma of the tongue. Cancer Letters 172(2), 187–192 (2001) 6.8

[58] Garey, M.R., Johnson, D.S.: The rectilinear Steiner tree problem is NP-complete. SIAM Journal on Applied Mathematics 32(4), 826–834 (1977) 1.6, 2.1.1

[59] Garnis, C., Campbell, J., Zhang, L., Rosin, M.P., Lam, W.L.: Ocgr array: an oral cancer genomic regional array for comparative genomic hybridization analysis. Oral Oncology 40(5), 511–519 (2004) 6

[60] Garraway, L.A., Jänne, P.A.: Circumventing cancer drug resistance in the era of personalized medicine. Cancer Discovery 2(3), 214–226 (2012) 1.2

[61] Gebhart, E., Ries, J., Wiltfang, J., Liehr, T., Efferth, T.: Genomic gain of the epidermal

growth factor receptor harboring band 7p12 is part of a complex pattern of genomic imbalances in oral squamous cell carcinomas. Archives of Medical Research 35(5), 385–394 (2004) 6.8

[62] Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al.: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine 366(10), 883–892 (2012) 1.1, 2, 3, 3.3

[63] Gerstung, M., Baudis, M., Moch, H., Beerenwinkel, N.: Quantifying cancer progression with conjunctive bayesian networks. Bioinformatics 25(21), 2809–2815 (2009) 2, 3

[64] Goh, C.S., Bogan, A.A., Joachimiak, M., et al.: Co-evolution of proteins with their interaction partners. Journal of Molecular Biology 299(2), 283–293 (2000) 7

[65] Greaves, M., Maley, C.C.: Clonal evolution in cancer. Nature 481(7381), 306–313 (2012) 1.2

[66] Greenblatt, M.S., Bennett, W.P., Hollstein, M., Harris, C.C.: Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Research 54(18), 4855–4878 (1994) 1.3, 4

[67] Greenman, C.D., Pleasance, E.D., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., Jones, D., Lau, K.W., Carter, N., Edwards, P.A., Futreal, P.A., Stratton, M.R., Campbell, P.J.: Estimation of rearrangement phylogeny for cancer genomes. Genome Research 22(2), 346–361 (2010) 1.5, 2, 3, 4

[68] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al.: Patterns of somatic mutation in human cancer genomes. Nature 446(7132), 153–158 (2007) 1.3

[69] Guenthoer, J., Diede, S.J., Tanaka, H., Chai, X., Hsu, L., Tapscott, S.J., Porter, P.L.: Assessment of palindromes as platforms for dna amplification in breast cancer. Genome

Research 22(2), 232–245 (2012) 1.3

[70] Höglund, M., Gisselsson, D., Mandahl, N., Johansson, B., Mertens, F., Mitelman, F., T., S.: Multivariate analyses of genomic imbalances in solid tumors reveal distinct and converging pathways of karyotypic evolution. Genes Chromosomes Cancer 31(2), 156–171 (2001) 3

[71] Hainke, K., Rahnenführer, J., Fried, R.: Cumulative disease progression models for cross-sectional data: a review and comparison. Biometrical Journal 54(5), 617–640 (2012) 7, 3

[72] Hamaguchi, M., Meth, J.L., von Klitzing, C., Wei, W., Esposito, D., Rodgers, L., Walsh, T., Welcsh, P., King, M., Wigler, M.H.: *DBC2*, a candidate for a tumor suppressor gene involved in breast cancer. Proceedings of the National Academy of Sciences 99(21), 13647–13652 (2002) 1.7, 2.1, 3.2.2

[73] Hanahan, D., Weinberg, R.A.: The hallmarks of cancer. cell 100(1), 57–70 (2000) 1.2

[74] Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. Cell 144(5), 646–674 (2011) 1.1, 1.2

[75] Hanan, M.: On Steiner's problem with rectilinear distance. SIAM Journal on Applied Mathematics 14(2), 255–265 (1966) 2.1.1, 2.1.2

[76] Harris, R.S., Petersen-Mahrt, S.K., Neuberger, M.S.: Rna editing enzyme apobec1 and some of its homologs can act as dna mutators. Molecular Cell 10(5), 1247–1253 (2002) 1.3, 4

[77] Hassan, N.M.M., Tada, M., Hamada, J.i., Kashiwazaki, H., Kameyama, T., Akhter, R., Yamazaki, Y., Yano, M., Inoue, N., Moriuchi, T.: Presence of dominant negative mutation of tp53 is a risk of early recurrence in oral cancer. Cancer Letters 270(1), 108–119 (2008) 6.8

[78] Hastings, P., Lupski, J.R., Rosenberg, S.M., Ira, G.: Mechanisms of change in gene copy

number. Nature Reviews Genetics 10(8), 551–564 (2009) 1.3, 1.6

[79] Heah, K.G., Hassan, M., Huat, S.C.: p53 expression as a marker of microinvasion in oral squamous cell carcinoma. Asian Pacific Journal of Cancer Prevention 12, 1017–22 (2011) 6.2

[80] Hein, J., Jiang, T., Wang, L., Zhang, K.: On the complexity of comparing evolutionary trees. Discrete Applied Mathematics 71(1), 153–169 (1996) 7

[81] Heng, H.H., Bremer, S.W., Stevens, J.B., Horne, S.D., Liu, G., Abdallah, B.Y., Karen, J.Y., Christine, J.Y.: Chromosomal instability (cin): what it is and why it is crucial to cancer evolution. Cancer and Metastasis Reviews 32(3-4), 325–340 (2013) 1.6, 4.3

[82] Heselmeyer-Haddad, K., Berroa Garcia, L.Y., Bradley, A., Hernandez, L., Hu, Y., Habermann, J.K., Dumke, C., Thorns, C., Perner, S., E.Pestova, Burke, C., Chowdhury, S.A., Schwartz, R., Schäffer, A.A., Paris, P., Ried, T.: Single-cell genetic analysis reveals insights into clonal development of prostate cancers and indicates loss of pten as a marker of poor prognosis. American Journal of Pathology 184, 2671–2686 (2014) 1, 6.6

[83] Heselmeyer-Haddad, K., Berroa Garcia, L.Y., Bradley, A., Ortiz-Melendez, C., Lee, W.J., Christensen, R., Prindiville, S.A., Calzone, K.A., Soballe, P.W., Hu, Y., Chowdhury, S.A., Schwartz, R., Schäffer, A.A., Ried, T.: 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. American Journal of Pathology 181(5), 1807–1822 (2012) 1.3, 1.7, 2, 2.1, 2.2.2, 2.2.3, 2.2.3, 3, 5, 3.2, 3.2.2, 3.2.2, 4, 4.2, 4.2.2, 5.1, 6.6, 7

[84] Heselmeyer-Haddad, K., Chaudhri, N., Stoltzfus, P., Cheng, J.C., Wilber, K., Morrison, L., Auer, G., Ried, T.: Detection of chromosomal aneuploidies and gene copy number changes in fine needle aspirates is a specific, sensitive, and objective genetic test for the diagnosis of breast cancer. Cancer Research 62(8), 2365–2369 (2002) 2, 3

[85] Hickey, G., Dehne, F., Rau-Chaplin, A., Blouin, C.: Spr distance computation for unrooted trees. Evolutionary Bioinformatics Online 4, 17 (2008) 7

[86] Hjelm, M., Höglund, M., Lagergren, J.: New probabilistic network models and algorithms for oncogenesis. Journal of Computational Biology 13(4), 853–865 (2006) 3, 4

[87] Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X., Wang, J.: Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell 148(5), 873–885 (2012) 1.5, 3

[88] Howe, L., Subbaramaiah, K., Brown, A., Dannenberg, A.: Cyclooxygenase-2: a target for the prevention and treatment of breast cancer. Endocrine-related Cancer 8(2), 97–114 (2001) 1.7, 2.1, 3.2.2

[89] Huang, F.Y., Chiu, P.M., Tam, K.F., Kwok, Y.K.Y., Lau, E.T., Tang, M.H.Y., Ng, T.Y., Liu, V.W.S., Cheung, A.N.Y., Ngan, H.Y.S.: Semi-quantitative fluorescent PCR analysis identifies *prkaa1* on chromosome 5 as a potential candidate cancer gene of cervical cancer. Gynecologic Oncology 103(1), 219–225 (2006) 1.7, 2.1, 3.2.2

[90] Huang, X., Gollin, S., Raja, S., Godfrey, T.: High-resolution mapping of the 11q13 amplicon and identification of a gene, *TAOS1*, that is amplified and overexpressed in oral cancer cells. Proceedings of the National Academy of Sciences 99(17), 11369–11374 (2002) 3.2.2

[91] Huelsenbeck, J.P., Ronquist, F., et al.: Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics 17(8), 754–755 (2001) 1.4

[92] Janocko, L.E., Brown, K.A., Smith, C.A., Gu, L.P., Pollice, A.A.: Distinctive patterns of Her-2/neu, c-myc, and cyclin D1 gene amplification by fluorescence in situ hybridization

in primary breast cancers. Cytometry 46(3), 136–149 (2001) 2, 3

[93] Jiang, F., Desper, R., Papadimitriou, C.H., Schäffer, A.A., Kallioniemi, O.P., Richter, J., Schraml, P., Sauter, G., Mihatsch, M.J., Moch, H.: Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data. Cancer Research 60(22), 6503–6509 (2000) 1.5

[94] Kaminagakura, E., Werneck da Cunha, I., Soares, F.A., Nishimoto, I.N., Kowalski, L.P.: Ccnd1 amplification and protein overexpression in oral squamous cell carcinoma of young patients. Head & Neck 33(10), 1413–1419 (2011) 6.8

[95] Kanao, H., Enomoto, T., Kimura, T., Fujita, M., Nakashima, R., Ueda, Y., Ueno, Y., Miyatake, T., Yoshizaki, T., Buzard, G.S., Tanigami, A., Yoshino, K., Murata, Y.: Overexpression of *LAMP3/TSC403/DC-LAMP* promotes metastasis in uterine cervical cancer. Cancer Research 65(19), 8640–8645 (2005) 1.7, 2.1, 3.2.2

[96] Koch, T., Martin, A.: Solving steiner tree problems in graphs to optimality. Networks 32(3), 207–232 (1998) 2.1.1

[97] Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Research Logistics Quarterly 2(1-2), 83–97 (1955) 2, 7.1.4

[98] Kyomoto, R., Kumazawa, H., Toda, Y., Sakaida, N., Okamura, A., Iwanaga, M., Shintaku, M., Yamashita, T., Hiai, H., Fukumoto, M.: Cyclin-d1-gene amplification is a more potent prognostic factor than its protein over-expression in human head-and-neck squamous-cell carcinoma. International Journal of Cancer 74(6), 576–581 (1997) 6.8

[99] Lengauer, C., Kinzler, K.W., Vogelstein, B.: Genetic instabilities in human cancers. Nature 396(6712), 643–649 (1998) 1.6

[100] Letouzé, E., Allory, Y., Bollet, M.A., Radvanyi, F., Guyon, F.: Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. Genome Biology 11(7), R76 (2010) 1.5, 3, 7

[101] Li, R., Faden, D.L., Fakhry, C., Langelier, C., Jiao, Y., Wang, Y., Wilkerson, M.D., Pedamallu, C.S., Old, M., Lang, J., et al.: Clinical, genomic, and metagenomic characterization of oral tongue squamous cell carcinoma in patients who do not smoke. Head & Neck (2014) 6, 6.1

[102] Lin, Y., Rajan, V., Moret, B.M.E.: A metric for phylogenetic trees based on matching. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(4), 1014–1022 (2012) 2.1.5, 2, 7, 7.1.1, 7.1.1, 7.1.1, 7.3, 8

[103] Liu, J., Bandyopadhyay, N., Ranka, S., Baudis, M., Kahveci, T.: Inferring progression models for CGH data. Bioinformatics 25(17), 2208–2215 (2009) 3

[104] Loeb, L.A.: Mutator phenotype may be required for multistage carcinogenesis. Cancer Research 51(12) (1991) 4

[105] Loeb, L.A.: A mutator phenotype in cancer. Cancer Research 61(8), 3230–3239 (2001) 1.1, 1.3

[106] Loeb, L.A., Springgate, C.F., Battula, N.: Errors in dna replication as a basis of malignant changes. Cancer Research 34(9), 2311–2321 (1974) 1.3

[107] Logothetis, C.J., Gallick, G.E., Maity, S.N., Kim, J., Aparicio, A., Efstathiou, E., Lin, S.H.: Molecular classification of prostate cancer progression: foundation for marker-driven treatment of prostate cancer. Cancer Discovery 3(8), 849–861 (2013) 5.1

[108] Maasland, D.H., van den Brandt, P.A., Kremer, B., Goldbohm, R.A., Schouten, L.J.: Alcohol consumption, cigarette smoking and the risk of subtypes of head-neck cancer: results from the netherlands cohort study. BMC Cancer 14(1), 187 (2014) 6

[109] Martins, F.C., De, S., Almendro, V., Gönen, M., Park, S.Y., Blum, J.L., Herlihy, W., Ethington, G., Schnitt, S.J., Tung, N., Garber, J.E., Fetten, K., Michor, F., Polyak, K.: Evolutionary pathways in BRCA1-associated breast tumors. Cancer Discovery 2(6), 503–511 (2012) 1.5, 1.6, 2, 3, 4

[110] Marusyk, A., Polyak, K.: Tumor heterogeneity: causes and consequences. Biochim Biophys Acta (BBA)-Reviews on Cancer 1805(1), 105–117 (2010) 3, 4

[111] McGlynn, K.A., Edmonson, M.N., Michielli, R.A., London, W.T., Lin, W.Y., Chen, G.C., Shen, F.M., Buetow, K.H.: A phylogenetic analysis identifies heterogeneity among hepatocellular carcinomas. Hepatology 36(6), 1341–1348 (2002) 3

[112] Michalides, R.J., van Veelen, N.M., Kristel, P.M., Hart, A.A., Loftus, B.M., Hilgers, F.J., Balm, A.J.: Overexpression of cyclin d1 indicates a poor prognosis in squamous cell carcinoma of the head and neck. Archives of Otolaryngology–Head & Neck Surgery 123(5), 497–502 (1997) 6.8

[113] Miller, C.A., White, B.S., Dees, N.D., Griffith, M., Welch, J.S., Griffith, O.L., Vij, R., Tomasson, M.H., Graubert, T.A., Walter, M.J., et al.: Sciclone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Computational Biology 10(8), e1003665 (2014) 1.5

[114] Miyamoto, R., Uzawa, N., Nagaoka, S., Nakakuki, K., Hirata, Y., Amagasa, T.: Potential marker of oral squamous cell carcinoma aggressiveness detected by fluorescence in situ hybridization in fine-needle aspiration biopsies. Cancer 95(10), 2152–2159 (2002) 6.8

[115] Myo, K., Uzawa, N., Miyamoto, R., Sonoda, I., Yuki, Y., Amagasa, T.: Cyclin d1 gene numerical aberration is a predictive marker for occult cervical lymph node metastasis in tnm stage i and ii squamous cell carcinoma of the oral cavity. Cancer 104(12), 2709–2716 (2005) 6.8

[116] Nakata, Y., Uzawa, N., Takahashi, K.I., Sumino, J., Michikawa, C., Sato, H., Sonoda, I., Ohyama, Y., Okada, N., Amagasa, T.: ¡ i¿ egfr¡/i¿ gene copy number alteration is a better prognostic indicator than protein overexpression in oral tongue squamous cell carcinomas. European Journal of Cancer 47(15), 2364–2372 (2011) 6.8

[117] Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepan-

sky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, R., McCombie, W.R., Hicks, J., Wigler, M.: Tumour evolution inferred by single-cell sequencing. Nature 472(7341), 90–94 (2011) 1.1, 1.6, 2, 3, 4.3

[118] Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Må nér, S., Zetterberg, A., Hicks, J., Wigler, M.: Inferring tumor progression from genomic heterogeneity. Genome Research 20(1), 68–80 (2010) 1.1, 1.5, 2, 3.1.5, 3.2.1, 7.1.3, 7.2.3

[119] Navin, N.E., Hicks, J.: Tracing the tumor lineage. Molecular Oncology 4(3), 267–283 (2010) 1.5

[120] Newton, M.A.: Discovering combinations of genomic aberrations associated with cancer. Journal of the American Statistical Association 97(460), 931–942 (2002) 3, 4

[121] Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al.: Mutational processes molding the genomes of 21 breast cancers. Cell 149(5), 979–993 (2012) 1.1, 1.3, 1.5

[122] Nonet, G.H., Stampfer, M.R., Chin, K., Gray, J.W., Collins, C.C., Yaswen, P.: The *ZNF217* gene amplified in breast cancers promotes immortalization of human mammary epithelial cells. Cancer Research 61(4), 1250–1254 (2001) 1.7, 2.1, 3.2.2

[123] Nordfors, C., Vlastos, A., Du, J., Ährlund-Richter, A., Tertipis, N., Grün, N., Romanitan, M., Haeggblom, L., Roosaar, A., Dahllöf, G., et al.: Human papillomavirus prevalence is high in oral samples of patients with tonsillar and base of tongue cancer. Oral Oncology 50(5), 491–497 (2014) 6.1

[124] Notta, F., Mullighan, C.G., Wang, J.C., Poeppl, A., Doulatov, S., Phillips, L.A., Ma, J., Minden, M.D., Downing, J.R., Dick, J.E.: Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. Nature 469(7330), 362–367 (2010) 3

[125] Nowak, M.A., Komarova, N.L., Sengupta, A., Jallepalli, P.V., Shih, I.M., Vogelstein, B.,

Lengauer, C.: The role of chromosomal instability in tumor initiation. Proceedings of the National Academy of Sciences 99(25), 16226–16231 (2002) 1.3, 1.6

[126] Nowell, P.C.: The clonal evolution of tumor cell populations. Science 194(4260), 23–28 (1976) 1.1, 2, 3, 4, 4.2.5

[127] Oesper, L., Mahmoody, A., Raphael, B.J.: Theta: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biology 14(7), R80 (2013) 3

[128] Ögmundsdóttir, H., Björnsson, J., Holbrook, W.: Role of tp53 in the progression of pre-malignant and malignant oral mucosal lesions. a follow-up study of 144 patients. Journal of Oral Pathology & Medicine 38(7), 565–571 (2009) 6.8

[129] Paris, P.L., Andaya, A., Fridlyand, J., Jain, A.N., Weinberg, V., Kowbel, D., Brebner, J.H., Simko, J., Watson, J.V., Volik, S., et al.: Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. Human Molecular Genetics 13(13), 1303–1313 (2004) 5.1

[130] Park, S.Y., Gönen, M., Kim, H.J., Michor, F., Polyak, K.: Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. The Journal of Clinical Investigation 120(2), 636–644 (2010) 3, 4, 3, 4, 5.2, 5.4, 6.3

[131] Pegram, M.D., Konecny, G., Slamon, D.J.: The molecular and cellular biology of her2/neu gene amplification/overexpression and the clinical development of herceptin (trastuzumab) therapy for breast cancer. In: Advances in Breast Cancer Management, pp. 57–75. Springer (2000) 1.2

[132] Pennington, G., Smith, C.A., Shackney, S., Schwartz, R.: Cancer phylogenetics from single-cell assays. Tech. rep., Carnegie Mellon University (2006) 1.5, 1.6, 3

[133] Pennington, G., Smith, C.A., Shackney, S., Schwartz, R.: Expectation-maximization method for reconstructing tumor phylogenies from single-cell data. In: Computational Systems Bioinformatics Conference. pp. 371–380 (2006) 1.5, 1.6

[134] Pennington, G., Smith, C.A., Shackney, S., Schwartz, R.: Reconstructing tumor phylogenies from heterogeneous single-cell data. Journal of Bioinformatics and Computational Biology 5(2a), 407–427 (2007) 1.5, 1.6, 2, 2.3, 3, 4, 4.1.1, 7

[135] Perner, S., Demichelis, F., Beroukhim, R., Schmidt, F.H., Mosquera, J.M., Setlur, S., Tchinda, J., Tomlins, S.A., Hofer, M.D., Pienta, K.G., et al.: Tmprss2: Erg fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Research 66(17), 8337–8341 (2006) 5.1

[136] Polzin, T., Daneshmand, S.V.: Improved algorithms for the steiner problem in networks. Discrete Applied Mathematics 112(1), 263–300 (2001) 2.1.1

[137] Poste, G.: Bring on the biomarkers. Nature 469(7329), 156–157 (2011) 1.2

[138] Purdom, E., Ho, C., Grasso, C.S., Quist, M.J., Cho, R.J., Spellman, P.: Methods and challenges in timing chromosomal abnormalities within cancer samples. Bioinformatics 29(24), 3113–3120 (2013) 3, 4

[139] Rahnenführer, J., Beerenwinkel, N., Schulz, W.A., Hartmann, C., Deimling, A.V., Wullich, B., Lengauer, T.: Estimating cancer survival and clinical outcome based on genetic tumor progression scores. Bioinformatics 21(10), 2438–2446 (2005) 6

[140] Robinson, D., Foulds, L.R.: Comparison of phylogenetic trees. Math Biosci 53(1), 131–147 (1981) 3.2.1, 7

[141] Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4(4), 406–425 (1987) 1.4, 1.6, 3

[142] Sathyan, K., Sailasree, R., Jayasurya, R., Lakshminarayanan, K., Abraham, T., Nalinakumari, K., Abraham, E.K., Kannan, S.: Carcinoma of tongue and the buccal mucosa represent different biological subentities of the oral carcinoma. Journal of Cancer Research and Clinical Oncology 132(9), 601–609 (2006) 6.8

[143] Schwartz, R., Shackney, S.E.: Applying unmixing to gene expression data for tumor phy-

logeny inference. BMC Bioinformatics 11(1), 42 (2010) 1.5

[144] Schwarz, R.F., Fletcher, W., Förster, F., Merget, B., Wolf, M., Schultz, J., Markowetz, F.: Evolutionary distances in the twilight zonea rational kernel approach. PloS One 5(12), e15788 (2010) 1.5

[145] Schwarz, R.F., Trinh, A., Sipos, B., Brenton, J.D., Goldman, N., Markowetz, F.: Phylogenetic quantification of intra-tumour heterogeneity. PLoS Computational Biology 10(4), e1003535 (2014) 1.5

[146] Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., et al.: Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 461(7265), 809–813 (2009) 1.1, 1.6

[147] Shahrabi Farahani, H., Lagergren, J.: Learning oncogenetic networks by reducing to mixed integer linear programming. PLoS One 8(6), e65773 (2013) 3

[148] Shapiro, E., Biezuner, T., Linnarsson, S.: Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics 14(9), 618–630 (2013) 1.5

[149] Shiboski, C.H., Schmidt, B.L., Jordan, R.C.: Tongue and tonsil carcinoma. Cancer 103(9), 1843–1849 (2005) 6

[150] Shlush, L.I., Chapal-Ilani, N., Adar, R., Pery, N., Maruvka, Y., Spiro, A. Shouval, R., Rowe, J., Tzukerman, M., Bercovich, D., Izraeli, S., Marcucci, G., Bloomfield, C., Zuckerman, T. Skorecki, K., Shapiro, E.: Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. Blood 120(3), 603–612 (2012) 2, 3

[151] Siegel, R., Naishadham, D., Jemal, A.: Cancer statistics, 2013. CA: A Cancer Journal for Clinicians 63(1), 11–30 (2013) 5.1

[152] Smith, R.A., Cokkinides, V., von Eschenbach, A.C., Levin, B., Cohen, C., Runowicz,

C.D., Sener, S., Saslow, D., Eyre, H.J.: American cancer society guidelines for the early detection of cancer. CA: a cancer journal for clinicians 52(1), 8–22 (2002) 1.2

[153] Snuderl, M., Fazlollahi, L., Le, L.P., Nitta, M., Zhelyazkova, B.H., Davidson, C.J., Akhavanfard, S., Cahill, D.P., Aldape, K.D., Betensky, R.A., Louis, D.N., Iafrate, A.J.: Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. Cancer Cell 20(6), 810–817 (2011) 3, 4, 7

[154] Snyder, T.L.: On the exact location of steiner points in general dimension. SIAM Journal on Computing 21(1), 163–180 (1992) 2.1.1, 2.1.2, 1

[155] Soder, A.I., Hoare, S.F., Muir, S., Going, J.J., Parkinson, E.K., Keith, W.N.: Amplification, increased dosage and in situ expression of the telomerase rna gene in human cancer. Oncogene 14(9), 1013–1021 (1997) 6

[156] Sottoriva, A., Spiteri, I., Piccirillo, S.G., Touloumis, A., Collins, V.P., Marioni, J.C., Curtis, C., Watts, C., Tavaré, S.: Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proceedings of the National Academy of Sciences 110(10), 4009–4014 (2010) 7

[157] Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C., Tavaré, S.: Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. Cancer Research 73(1), 41–49 (2013) 1.6, 3, 3.2.1

[158] Sprouffske, K., Pepper, J.W., Maley, C.C.: Accurate reconstruction of the temporal order of mutations in neoplastic progression. Cancer Prevention Research 4(7), 1135–1144 (2011) 3, 4

[159] Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al.: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144(1), 27–40 (2011) 1.3

[160] Strino, F., Parisi, F., Micsinai, M., Kluger, Y.: Trap: a tree approach for fingerprinting subclonal tumor composition. Nucleic Acids Research 41(17), e165–e165 (2013) 1.5

[161] Subramanian, A., Shackney, S., Schwartz, R.: Inference of tumor phylogenies from genomic assays on heterogeneous samples. Journal of Biomedicine and Biotechnology p. 797812 (2012) 1.5, 2, 3, 3.1.5, 3.2.1, 7.1.3, 7.2.3

[162] Subramanian, A., Shackney, S., Schwartz, R.: Novel multisample scheme for inferring phylogenetic markers from whole genome tumor profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10(6), 1422–1431 (2013) 1.5

[163] Swanton, C.: Intratumor heterogeneity: evolution through space and time. Cancer Research 72(19), 4875–4882 (2012) 1.1, 1.2

[164] Szabo, A., Boucher, K.: Estimating an oncogenetic tree when false negatives and positives are present. Mathematical Biosciences 176(2), 219–236 (2002) 1.5, 3

[165] Szerlip, N.J., Pedraza, A., Chakravarty, D., Azim, M., McGuire, J., Fang, Y., Ozawa, T., Holland, E.C., Huse, J.T., Jhanwar, S., Leversha, M.A., Mikkelsen, T., Brennan, C.W.: Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. Proceedings of the National Academy of Sciences 109(8), 3041–3046 (2012) 3, 4

[166] Tableman, M., Kim, J.S.: Survival analysis using S: analysis of time-to-event data. CRC press (2003) 6.1

[167] Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S.: MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution 30(12), 2725–2729 (2013) 3.1.5, 7.1.3

[168] Tan, M., Yu, D.: Molecular mechanisms of erbB2-mediated breast cancer chemoresistance. In: Breast Cancer Chemosensitivity, pp. 119–129. Springer (2007) 1.7, 2.1, 3.2.2

[169] Tao, Y., Ruan, J., Yeh, S.H., Lu, X., Wang, Y., Zhai, W., Cai, J., Ling, S., Gong, Q.,

Chong, Z., Qu, Z., Li, Q., Liu, J., Yang, J., Zheng, C., Zeng, C., Wang, H., Zhang, J., Wang, S., Hao, L., Dong, L., Li, W., Sun, M., Zou, W., Yu, C., Li, C., Liu, G., Jiang, L., Xu, J., Huang, H., Li, C., Mi, S., Zhang, B., Chen, B., Zhao, W., Hu, S., Zhuang, S., Shen, Y., Shi, S., Brown, C., White, K.P., Chen, D.C.P., Wu, C.: Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. Proceedings of the National Academy of Sciences 108(29), 12042–12047 (2011) 3

[170] Taylor, B.J., Nik-Zainal, S., Wu, Y.L., Stebbings, L.A., Raine, K., Campbell, P.J., Rada, C., Stratton, M.R., Neuberger, M.S.: Dna deaminases induce break-associated mutation showers with implication of apobec3b and 3a in breast cancer kataegis. Elife 2 (2013) 1.3

[171] Therneau, T.M.: Modeling survival data: extending the Cox model. Springer (2000) 6.1, 6.2

[172] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996) 8.1

[173] Timmermann, B., Kerick, M., Roehr, C., Fischer, A., Isau, M., Boerno, S.T., Wunderlich, A., Barmeyer, C., Seemann, P., Koenig, J., et al.: Somatic mutation profiles of msi and mss colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PloS One 5(12), e15661 (2010) 1.3, 4

[174] Tofigh, A., Sjlund, E., Hglund, M., Lagergren, J.: A global structural em algorithm for a model of cancer progression. In: Advances in Neural Information Processing Systems. pp. 163–171 (2011) 1.5

[175] Tolliver, D., Tsourakakis, C., Subramanian, A., Shackney, S., Schwartz, R.: Robust unmixing of tumor states in array comparative genomic hybridization data. Bioinformatics 26(12), i106–i114 (2010) 1.5

[176] Urbschat, S., Rahnenführer, J., Henn, W., Feiden, W., Wemmert, S., Linsler, S., Zang,

K.D., Oertel, J., Ketter, R.: Clonal cytogenetic progression within intratumorally heterogeneous meningiomas predicts tumor recurrence. International Journal of Oncology 39(6), 1601–1608 (2011) 3, 4

[177] von Heydebreck, A., Gunawan, B., Füzesi, L.: Maximum likelihood estimation of oncogenetic tree models. Biostatistics 5(4), 545–556 (2004) 2

[178] Von Heydebreck, A., Gunawan, B., Füzesi, L.: Maximum likelihood estimation of oncogenetic tree models. Biostatistics 5(4), 545–556 (2004) 1.5

[179] Vousden, K.H., Lane, D.P.: *P53* in health and disease. Nature Reviews Molecular Cell Biology 8(4), 275–283 (2007) 1.7, 2.1, 3.2.2

[180] Waggoner, S.E.: Cervical cancer. The Lancet 361(9376), 2217–2225 (2003) 1.7

[181] Wang, L., Liu, T., Nishioka, M., Aguirre, R.L., Win, S.S., Okada, N.: Activation of erk1/2 and cyclin d1 expression in oral tongue squamous cell carcinomas: relationship between clinicopathological appearances and cell proliferation. Oral oncology 42(6), 625–631 (2006) 6.8

[182] Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al.: Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature 512(7513), 155–160 (2014) 4

[183] Wangsa, D., Heselmeyer-Haddad, K., Ried, P., Eriksson, E., Schäffer, A.A., Morrison, L.E., J., Luo, J., Auer, G., Munck-Wikland, E., Ried, T., Lundqvist, E.Å.: Fluorescence in situ hybridization markers for prediction of cervical lymph node metastases. American Journal of Pathology 175(6), 2637–2645 (2009) 1.7, 2, 2.1, 2.2.2, 2.2.2, 2.2.3, 5, 3.2, 3.2.2, 4.2, 6.2, 7.2.1, 1

[184] Waterman, M.S., Smith, T.F.: On the similarity of dendrograms. Journal of Theoretical Biology 73(4), 789–800 (1978) 7

[185] Weinberg, R.: The biology of cancer. Garland Science (2013) 1, 1.1

[186] Werkmeister, R., Brandt, B., Joos, U.: Clinical relevance of¡ i¿ erb¡/i¿ b-1 and-2 oncogenes in oral carcinomas. Oral Oncology 36(1), 100–105 (2000) 6.8

[187] Wigle, J.T., Oliver, G.: *PROX1* function is required for the development of the murine lymphatic system. Cell 98(6), 769–778 (1999) 1.7, 2.1, 3.2.2

[188] Wolfer, A., Ramaswamy, S.: *MYC* and metastasis. Cancer Research 71(6), 2034–2037 (2011) 1.7, 2.1, 3.2.2

[189] Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y., Wang, J.: Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 148(5), 886–895 (2012) 1.5, 1.6, 2, 3

[190] Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al.: Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467(7319), 1114–1117 (2010) 1.1

[191] Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., Song, C., Witten, D., Blau, C.A., Noble, W.S.: Inferring clonal composition from multiple sections of a breast cancer. PLoS Computational Biology 10(7), e1003703 (2014) 1.5