

Abstract of “Tumor Phylogeny Reconstruction from DNA Sequencing Data” by Gryte Satas, Ph.D.,  
Brown University, February 2020.

Cancer is an evolutionary process where cells acquire somatic mutations over time. As a result of this process, tumors are often highly heterogeneous — containing cell populations with different sets of mutations. While tumor evolution cannot be directly observed, high-throughput DNA sequencing provides a measure of tumor heterogeneity. In this thesis, we introduce three algorithms that leverage tumor heterogeneity to infer the phylogenetic history of cancer. These algorithms account for specific issues in the two technologies for cancer DNA sequencing — bulk sequencing, where thousands of cells from a single tumor are sequenced simultaneously, and single-cell sequencing. First, we introduce an algorithm, PASTRI, that simultaneously deconvolves bulk sequencing data and constructs a tumor phylogeny based on somatic single-nucleotide variants (SNVs); PASTRI uses a hybrid probabilistic-combinatorial approach to more accurately model sequencing noise compared to existing combinatorial approaches. In addition to SNVs, tumors often contain copy-number aberrations (CNAs), a structural variation where larger regions of the genome are duplicated or deleted. Most existing phylogenetic inference algorithms use either SNVs or CNAs alone; however CNAs often overlap SNVs and failure to model these interactions between the two types of mutations can lead to inaccurate phylogenies. We introduce two algorithms, SPRUCE and SCARLET, that integrate SNVs and CNAs using bulk sequencing data and single-cell sequencing data, respectively. Finally, we investigate technical biases in single-cell DNA sequencing data and demonstrate how these biases can be used to phase genomic variants.

Tumor Phylogeny Reconstruction from DNA Sequencing Data

by

Gryte Satas

B.A., Washington University in St. Louis, 2012

Sc.M., Brown University, 2017

A dissertation submitted in partial fulfillment of the  
requirements for the Degree of Doctor of Philosophy  
in the Department of Computer Science at Brown University

Providence, Rhode Island

February 2020

© Copyright 2020 by Gryte Satas

This dissertation by Gryte Satas is accepted in its present form by  
the Department of Computer Science as satisfying the dissertation requirement  
for the degree of Doctor of Philosophy.

Date \_\_\_\_\_  
Benjamin J. Raphael, Director

Recommended to the Graduate Council

Date \_\_\_\_\_  
Sohini Ramachandran, Reader

Date \_\_\_\_\_  
Sohrab Shah, Reader

Approved by the Graduate Council

Date \_\_\_\_\_  
Andrew G. Campbell  
Dean of the Graduate School



# Acknowledgements

This dissertation has my name on it, but it wouldn't be possible without a community of people supporting me. I would like to extend my deepest gratitude to all those who have been there for the last five years.

First, my advisor, Ben Raphael, who shaped me into the researcher I am today. The co-authors of the work presented here today: Layla Oesper, Mohammed El-Kebir, Simone Zaccaria, Geoffrey Mon and Ben Raphael. The current and former Raphael group members for their support and guidance: Layla Oesper, Max Leiserson, Mohammed El-Kebir, Matt Reyna, Simone Zaccaria, Fabio Vandin, Hsin-ta Wu, Iman Hajirasouliha, Jonathan Eldridge, Dora Erdos, Sergey Aganezov, Rebecca Elyanow, Ashley Conard, Matt Myers, Tyler Park, Ron Zeira, Uthsav Chitra. In particular I'd like to acknowledge Layla Oesper, who taught me so much when I was just starting, and whose influence I feel to this day in every page I write, every figure I make, and every talk I give; Mohammed El-Kebir, collaborating with whom was some of the best fun I had during this whole process and pushed me to grow as a computer scientist; my office-mate, lab-mate, and house-mate Rebecca Elyanow, who put up with me and was always there to eat the cookies I stress-baked; and Simone Zaccaria and Matt Reyna, who could always put a smile on my face when I was having a long day.

All the administrative staff that made my life easier behind the scenes. In particular, I'd like to thank Lauren Clarke, Jane McIlmail, Kathy Kirman Billings, and especially Nathaniel Gill for

always having cookies when I needed them most.

My parents, who worked hard to give me every opportunity possible.

Finally, all of my friends who supported me and brought me joy: Eamon, for giving me the initial push and convincing me that I was capable; Jon and Rebecca, Ryan and Allison, and everyone in the Nordic Jedi, Brown GWICS, doggo chat, and the Tuesday afternoon tea group, for keeping me sane and keeping things lighthearted; Chris, for helping me find the light at the end of the tunnel.

# Contents

|                                                                                                                   |           |
|-------------------------------------------------------------------------------------------------------------------|-----------|
| <b>List of Figures</b>                                                                                            | <b>x</b>  |
| <b>1 Introduction</b>                                                                                             | <b>1</b>  |
| 1.1 Somatic Mutations in Cancer . . . . .                                                                         | 2         |
| 1.2 DNA Sequencing of Cancer Genomes . . . . .                                                                    | 4         |
| 1.2.1 Bulk DNA Sequencing . . . . .                                                                               | 4         |
| 1.2.2 Single-cell DNA Sequencing . . . . .                                                                        | 4         |
| 1.3 Tumor Phylogenetics . . . . .                                                                                 | 5         |
| 1.3.1 Bulk Sequencing Data . . . . .                                                                              | 5         |
| 1.3.2 Phylogeny Inference from Single-Cell Sequencing Data . . . . .                                              | 8         |
| 1.4 Contributions . . . . .                                                                                       | 9         |
| 1.4.1 A hybrid combinatorial-probabilistic approach to phylogeny inference from<br>bulk sequencing data . . . . . | 9         |
| 1.4.2 Integrating SNVs and CNAs for phylogeny inference from bulk sequencing data                                 | 10        |
| 1.4.3 Integrating SNVs and CNAs for phylogeny inference from single-cell data . .                                 | 10        |
| 1.4.4 Leveraging technical biases in single-cell DNA sequencing to phase genomic<br>variants . . . . .            | 10        |
| <b>2 Tumor Phylogeny Inference Using Tree-Constrained Importance Sampling</b>                                     | <b>11</b> |
| 2.1 Introduction . . . . .                                                                                        | 11        |

|          |                                                                                        |           |
|----------|----------------------------------------------------------------------------------------|-----------|
| 2.2      | Methods . . . . .                                                                      | 12        |
| 2.2.1    | Model . . . . .                                                                        | 12        |
| 2.2.2    | Probabilistic Model . . . . .                                                          | 16        |
| 2.2.3    | Tree Inference . . . . .                                                               | 18        |
| 2.3      | Results . . . . .                                                                      | 23        |
| 2.3.1    | Comparison to Phylogeny Reconstruction Algorithms . . . . .                            | 23        |
| 2.3.2    | Chronic Lymphocytic Leukemia . . . . .                                                 | 26        |
| 2.4      | Discussion . . . . .                                                                   | 27        |
| <b>3</b> | <b>Inferring the Mutational History of a Tumor using Multi-state Perfect Phylogeny</b> |           |
|          | <b>Mixtures</b>                                                                        | <b>30</b> |
| 3.1      | Introduction . . . . .                                                                 | 30        |
| 3.2      | Results . . . . .                                                                      | 35        |
| 3.2.1    | The Perfect Phylogeny Mixture Deconvolution Problem . . . . .                          | 35        |
| 3.2.2    | Multi-State Model for the Somatic Mutational Process in Cancer . . . . .               | 37        |
| 3.2.3    | Experimental Evaluation on Simulated and Real Data . . . . .                           | 43        |
| 3.3      | Discussion . . . . .                                                                   | 49        |
| 3.4      | Methods . . . . .                                                                      | 53        |
| 3.4.1    | Mathematical Characterization of Solutions to the PPMDP . . . . .                      | 54        |
| 3.4.2    | SPRUCE Algorithm for the Cladistic-PPMDP . . . . .                                     | 57        |
| <b>4</b> | <b>Single-cell Tumor Phylogeny Inference with Supported Mutation Losses</b>            | <b>59</b> |
| 4.1      | Introduction . . . . .                                                                 | 59        |
| 4.2      | Results . . . . .                                                                      | 63        |
| 4.2.1    | SCARLET algorithm for Loss-supported Phylogeny Model . . . . .                         | 63        |
| 4.2.2    | Simulated Data . . . . .                                                               | 65        |
| 4.2.3    | Single-cell phylogeny of metastatic colorectal cancer . . . . .                        | 67        |
| 4.3      | Discussion . . . . .                                                                   | 71        |
| 4.4      | Methods . . . . .                                                                      | 72        |

|          |                                                                                      |            |
|----------|--------------------------------------------------------------------------------------|------------|
| 4.4.1    | Loss-supported phylogeny model . . . . .                                             | 72         |
| 4.4.2    | Maximum Likelihood Loss-supported Refinement Problem . . . . .                       | 78         |
| 4.4.3    | SCARLET Algorithm for Maximum-Likelihood Loss-Supported Refinement Problem . . . . . | 79         |
| <b>5</b> | <b>Haplotype Phasing in Single-Cell DNA Sequencing Data</b>                          | <b>81</b>  |
| 5.1      | Introduction . . . . .                                                               | 81         |
| 5.2      | Methods . . . . .                                                                    | 83         |
| 5.2.1    | Quantifying Concurrent Dropout . . . . .                                             | 85         |
| 5.2.2    | Augmenting Haplotype Assembly with Amplification Fragments . . . . .                 | 86         |
| 5.3      | Results . . . . .                                                                    | 87         |
| 5.3.1    | Whole-genome Single Cell Sequencing of Neurons . . . . .                             | 87         |
| 5.3.2    | Whole-exome Single-cell Sequencing of Breast Cancer . . . . .                        | 90         |
| 5.4      | Discussion . . . . .                                                                 | 92         |
| <b>6</b> | <b>Conclusions</b>                                                                   | <b>95</b>  |
| 6.1      | Future Work . . . . .                                                                | 96         |
| <b>A</b> | <b>Supplementary Methods and Results for Chapter 2</b>                               | <b>119</b> |
| A.1      | Supplementary Methods . . . . .                                                      | 119        |
| A.1.1    | Sequential Importance Sampling . . . . .                                             | 119        |
| A.1.2    | Comparison to Other Models . . . . .                                                 | 120        |
| A.2      | Supplementary Results . . . . .                                                      | 122        |
| A.2.1    | Additional Simulation Results . . . . .                                              | 122        |
| <b>B</b> | <b>Supplementary Methods and Results for Chapter 3</b>                               | <b>125</b> |
| B.1      | Supplementary Methods . . . . .                                                      | 125        |
| B.1.1    | The Perfect Phylogeny Mixture Deconvolution Problem . . . . .                        | 126        |
| B.1.2    | Uniqueness of $U$ given $F$ and $T$ . . . . .                                        | 131        |
| B.1.3    | Combinatorial Characterization of the PPMDP . . . . .                                | 137        |

|          |                                                                                                   |            |
|----------|---------------------------------------------------------------------------------------------------|------------|
| B.1.4    | Problem Complexity . . . . .                                                                      | 142        |
| B.1.5    | Cladistic Perfect Phylogeny Mixture Deconvolution Problem . . . . .                               | 143        |
| B.1.6    | Algorithm for the Cladistic-PPMDP . . . . .                                                       | 144        |
| B.1.7    | Multi-State Model for the Somatic Mutational Process in Cancer . . . . .                          | 148        |
| B.2      | Supplemental Results . . . . .                                                                    | 152        |
| B.2.1    | Simulations . . . . .                                                                             | 153        |
| B.2.2    | Comparison against PhyloWGS . . . . .                                                             | 154        |
| <b>C</b> | <b>Supplementary Methods and Results for Chapter 4</b>                                            | <b>163</b> |
| C.1      | Supplementary Methods . . . . .                                                                   | 163        |
| C.2      | Supplementary Tables and Figures . . . . .                                                        | 163        |
| C.3      | Supplementary Methods . . . . .                                                                   | 163        |
| C.3.1    | Solving ML-LSR Problem: Likelihood model . . . . .                                                | 163        |
| C.3.2    | Solving ML-LSR Problem: Finding max-likelihood subtree roots . . . . .                            | 164        |
| C.3.3    | Solving ML-LSR Problem: Integer Linear Program (ILP) to solve maximum<br>likelihood IDP . . . . . | 167        |
| C.3.4    | Simulation Details . . . . .                                                                      | 169        |
| C.3.5    | Copy-number analysis of CRC2 . . . . .                                                            | 170        |
| C.4      | Proofs . . . . .                                                                                  | 170        |
| C.4.1    | Proof of Theorem 1 . . . . .                                                                      | 170        |
| C.4.2    | Proof of Lemma 1 . . . . .                                                                        | 171        |
| C.4.3    | Hardness of ML-LSR . . . . .                                                                      | 173        |
| <b>D</b> | <b>Supplementary Methods and Results for Chapter 5</b>                                            | <b>175</b> |
| D.1      | Supplementary Methods . . . . .                                                                   | 175        |
| D.1.1    | Creating Synthetic Diploid Simulations . . . . .                                                  | 175        |
| D.2      | Supplementary Results . . . . .                                                                   | 176        |
| D.2.1    | Comparison to Reference-Based Phasing . . . . .                                                   | 176        |

# List of Figures

|     |                                                                                        |    |
|-----|----------------------------------------------------------------------------------------|----|
| 1.1 | Tumor Evolution . . . . .                                                              | 2  |
| 1.2 | Somatic mutations in cancer . . . . .                                                  | 3  |
| 2.1 | Tree Constraint on Cluster Cell Fractions . . . . .                                    | 13 |
| 2.2 | Generative model for mutation allele counts from a tumor . . . . .                     | 16 |
| 2.3 | Overview of PASTRI algorithm . . . . .                                                 | 18 |
| 2.4 | Frequency matrix permutations. . . . .                                                 | 23 |
| 2.5 | Comparison of PASTRI to other phylogenetic reconstruction algorithms. . . . .          | 24 |
| 2.6 | PASTRI results on CLL Patient 5. . . . .                                               | 27 |
| 3.1 | Overview of multi-state perfect phylogeny problem. . . . .                             | 34 |
| 3.2 | Cancer cell fraction and state trees. . . . .                                          | 41 |
| 3.3 | Overview of SPRUCE algorithm . . . . .                                                 | 42 |
| 3.4 | SPRUCE results on simulated data. . . . .                                              | 45 |
| 3.5 | SPRUCE results on CLL077 . . . . .                                                     | 47 |
| 3.6 | Multiple compatible state trees lead to distinct phylogenetic reconstructions. . . . . | 50 |
| 4.1 | Loss-supported model of tumor evolution . . . . .                                      | 62 |
| 4.2 | Overview of SCARLET algorithm. . . . .                                                 | 64 |
| 4.3 | SCARLET results on simulated data . . . . .                                            | 65 |
| 4.4 | SCARLET results on metastatic colorectal cancer scDNA data. . . . .                    | 68 |

|      |                                                                                 |     |
|------|---------------------------------------------------------------------------------|-----|
| 5.1  | Overview of SC-HAP algorithm. . . . .                                           | 84  |
| 5.2  | Validating SC-Hap on Whole-genome DNA Sequencing Data. . . . .                  | 88  |
| 5.3  | Validating SC-Hap on Whole-genome DNA Sequencing Data. . . . .                  | 91  |
|      |                                                                                 |     |
| A.1  | Comparison of phylogenetic mixture models with binomial observations. . . . .   | 121 |
| A.2  | PASTRI runtime . . . . .                                                        | 122 |
| A.3  | PASTRI ancestral relationship accuracy . . . . .                                | 123 |
| A.4  | PASTRI cluster frequency accuracy . . . . .                                     | 124 |
|      |                                                                                 |     |
| B.1  | Complete perfect phylogeny tree correspondence. . . . .                         | 127 |
| B.2  | Concepts of the PPMDP . . . . .                                                 | 128 |
| B.3  | Example of the usage equation. . . . .                                          | 136 |
| B.4  | Example PPMDP instance and solution. . . . .                                    | 141 |
| B.5  | PPMDP reduction from Subset Sum. . . . .                                        | 143 |
| B.6  | Example cladistic PPMDP instance and solution. . . . .                          | 145 |
| B.7  | SPRUCE results on simulated data for $n = 5$ instances with noisy VAFs. . . . . | 155 |
| B.8  | SPRUCE results on simulated data for larger instances with noisy VAFs. . . . .  | 156 |
| B.9  | SPRUCE results on simulated data with ISA violations. . . . .                   | 157 |
| B.10 | SPRUCE solution space for CLL077. . . . .                                       | 157 |
| B.11 | A22 tree reported by Gundem et al. . . . .                                      | 158 |
| B.12 | A22 characters with no compatible state trees. . . . .                          | 159 |
| B.13 | Maximal cliques in the compatibility graph of A22. . . . .                      | 160 |
| B.14 | SPRUCE results on A22. . . . .                                                  | 161 |
| B.15 | Relationship between T and T-bar. . . . .                                       | 162 |
|      |                                                                                 |     |
| C.1  | CRC2 Data and SCARLET Mutation Matrix. . . . .                                  | 165 |
|      |                                                                                 |     |
| D.1  | ScHap comparison to EAGLE2 on whole-exome data. . . . .                         | 178 |



# Chapter 1

## Introduction

Tumors develop through the accumulation of mutations during the lifetime of an individual through the process of clonal evolution [114] (Figure 1.1A). As a result of this process, many tumors are heterogeneous, containing multiple populations of cells (or *clones*), each with its own unique combination of somatic mutations. This *intratumor* heterogeneity complicates the diagnosis and is considered to be a significant cause for resistance to treatment and relapse [50, 146]. The field of *cancer phylogenetics* aims to reconstruct and analyze the evolutionary history of tumors. Phylogenetic analyses of tumors, fueled by advances in high-throughput sequencing, have already been used to identify mutations that drive cancer progression [72, 100], analyze the interaction between the immune system and the evolution of tumors [89, 167], and identify common evolutionary patterns of tumorigenesis and metastasis [150, 151]. In this thesis, we present several algorithms that expand the field of cancer phylogenetics by using improved models for high-throughput sequencing data, and modeling the complex interactions between different mutation types. In this introduction, we present a primer on high-throughput sequencing of cancer genomes, different types of somatic mutations, and existing algorithms for reconstructing tumor phylogenies.

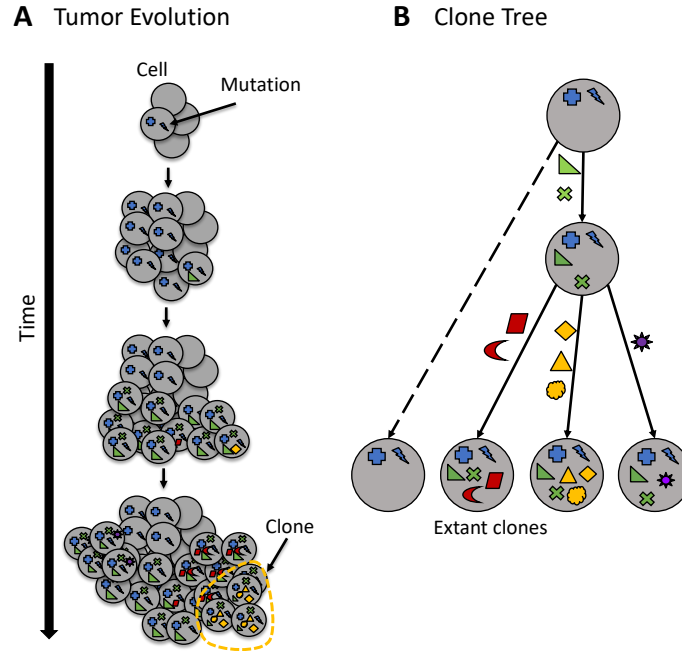


Figure 1.1: **Tumor Evolution.** (A) Cancer is an evolutionary process, during which cells accumulate mutations over time. Within a tumor, different cells may contain distinct sets of mutations, resulting in intratumor heterogeneity. (B) We model the evolutionary process as a phylogenetic tree. Here, vertices correspond to clones – groups of cells with the same set of mutations. Edges correspond to direct ancestral relationships between clones. Edges are labeled by sets of mutations that separate clones.

## 1.1 Somatic Mutations in Cancer

Cancer is a disease of the genome. Cancer cells have acquired a set of genomic alterations that lead to abnormal cell growth and the potential to spread to other areas of the body. There are several classes of somatic mutations that we see in cancer genomes. Here, we focus on two types of variants (Figure 1.2). *Single-nucleotide variants* (SNVs) affect a single base in the genome. Even such a small change in the genetic code can have a significant impact on cellular function. For example BRAF V600E, an SNV in codon 600 of exon 15 of the BRAF gene, has been identified in a large number of tumor samples across multiple cancer types including melanoma, colorectal carcinoma, papillary thyroid carcinoma, hairy cell leukemia, and Langerhans cell histiocytosis [88].

*Copy-number aberrations* (CNAs) are a type of variation where larger regions of the genome are

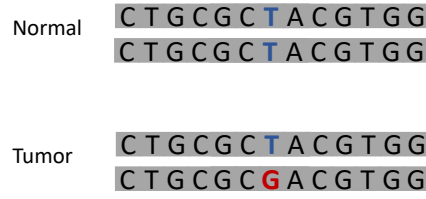
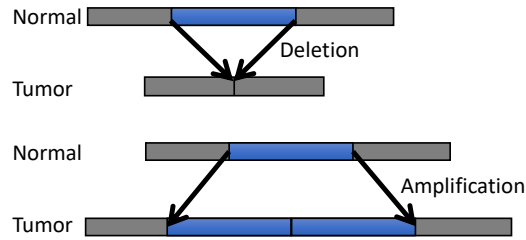
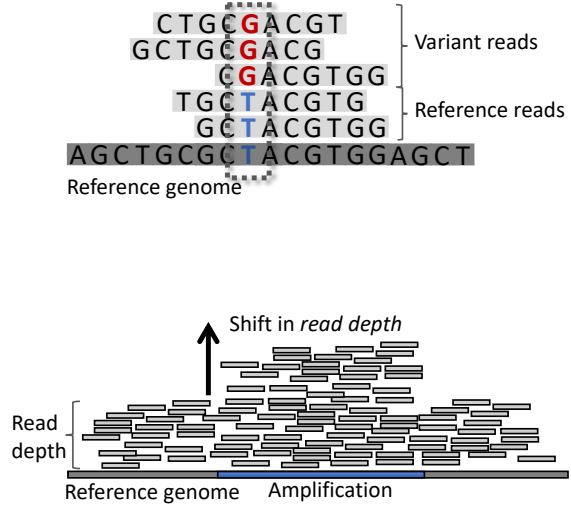
**A Single-Nucleotide Variants (SNVs)****Copy-Number Aberrations (CNAs)****B**

Figure 1.2: **Somatic mutations in cancer.** (A) In this thesis, we focus on two types of somatic variants found in cancer genomes. Single-nucleotide variants (SNVs) are characterized by a change in a single base of the genome. Here, a *T* allele changes to a *G* allele in the tumor genome. Copy-number aberrations (CNAs) involve changes in larger regions of the genome, where intervals of the genome are either deleted or amplified. (B) We identify SNVs and CNAs using DNA sequencing data. SNVs in the sequenced genome appear as variant reads aligning to the locus in the reference genome. CNVs in the sequenced genome appear as shifts in read depth – the number of reads aligning to a region. Amplifications result in increases of read depth and deletions result in decreases.

deleted or duplicated. CNAs can range in size from hundreds of base pairs to whole chromosomes and duplications of the whole genome. Copy-number aberrations are very common in cancers, particularly in solid tumors, with a recent study across multiple cancer types identifying a median of 39 aberrations per tumor sample [162]. Gene copy-number affects the level of gene expression, and tumors often show patterns of amplification of oncogenes (e.g. in the *MYC*, *EGFR*, and *RAS* gene families [6]). In addition, deletions may remove functional copies of tumor suppressor genes – for example, deletion of the p-arm of chromosome 17, which contains tumor suppressor gene *TP53*, is seen recurrently across multiple cancer types. When seen in combination with an SNV in the remaining copy of *TP53*, this combination of deletion+SNV is strongly associated with decreased survival times of patients in multiple cancer types including acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), and chronic lymphoblastic leukemia (CLL) [142].

## 1.2 DNA Sequencing of Cancer Genomes

DNA sequencing of cancer genomes has played a large role in increasing our understanding of the process of cancer development at the genetic level. The human genome is roughly three billion base pairs long, and comprises 23 pairs of chromosomes. Currently we lack the technology to sequence individual chromosomes from end to end. Instead, most cancer sequencing is done using short-read DNA sequencing, where we measure snippets of the genome, or *reads*, that are approximately 100 base pairs long. There are two main types of DNA sequencing for cancer genomics that we focus on in this thesis – *bulk* and *single cell*.

### 1.2.1 Bulk DNA Sequencing

In bulk DNA sequencing, each sample is composed of a mixture of thousands of tumor cells. This complicates analysis of tumors, as we expect a high level of heterogeneity amongst individual tumor cells, and we do not observe the mutational profiles of component clones directly. Instead, we observe a mixed signal of all the genetic material present in the sample. Thus, the analysis of bulk DNA sequencing requires specialized algorithms that deconvolve the mixed data into constituent clones. The vast majority of cancer sequencing performed to date, including in large scale projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), has use bulk sequencing data. where each sample is composed of a mixture of thousands of tumor cells.

### 1.2.2 Single-cell DNA Sequencing

Single-cell DNA (scDNA) sequencing allows researchers to measure the heterogeneity of cancer genomes at the level of individual cells. While scDNA sequencing avoids the problems of deconvolving mixed samples, a number of other technical issues and biases complicate data analysis.

Because an individual cell contains only a single copy of the genome, whole-genome amplification

(WGA) is an important step for scDNA-seq in order to obtain sufficient DNA to sequence. Several WGA methods have been introduced, with the three most common for cancer sequencing being degenerate oligonucleotide primed PCR (DOP-PCR), multiple displacement amplification (MDA), and multiple annealing and looping based amplification cycles (MALBAC) [56]. These approaches present trade-offs in the quality of signal measured for different mutation types. In general, MDA is able to achieve much more complete coverage of the genome (in one study,  $\sim 84\%$  covered with at least one read, versus  $\sim 72\%$  for MALBAC and  $\sim 39\%$  for DOP-PCR [70]). This makes MDA better-suited for identifying single-nucleotide variants than the other technologies. However, MDA also has highly non-uniform coverage which makes the identification of copy-number aberrations challenging. DOP-PCR have much higher uniformity and thus have in general been used to profile cancer cells for copy-number aberrations.

## 1.3 Tumor Phylogenetics

Like any evolutionary process, the somatic mutational process giving rise to a tumor can be described by a phylogenetic tree, whose leaves correspond to present clones, and whose edges describe the ancestral relationships between clones (Figure 1.1B). We typically do not directly observe an evolutionary process; instead we observe the present day variation that is the result of the evolutionary process. There is a large body of work on phylogenetics for reconstructing the evolutionary history of species or populations [156]. However, these algorithms are not readily applicable to cancer sequencing data. Bulk sequencing and single-cell DNA sequencing provide different challenges for the reconstruction of tumor phylogenies.

### 1.3.1 Bulk Sequencing Data

In traditional phylogenetics, we observe leaves of a phylogeny. For each leaf, we measure the set of variants present within that leaf. The goal is to reconstruct a tree that relates these leaves. In

tumors, these leaves correspond to extant cells at the time of sampling. However, for bulk sequencing data, we do not individually observe these cells. Instead a sequencing sample corresponds to a mixture of cells.

Thus, algorithms from bulk tumor sequencing data must simultaneously deconvolve the mixed sample and reconstruct a tumor phylogeny.

### SNV based phylogenetic reconstruction

In bulk sequencing data, for each mutation, we observe a *variant allele frequency* (VAF), which is defined as the proportion of reads at a locus that contains the variant. In the absence of copy-number aberrations, we expect the VAF to be directly proportional to the fraction of cells in the mixture that have the mutation.

There are many algorithms that reconstruct the evolutionary history of a tumor based on the SNVs that are present in a tumor sample. Many of these algorithms rely on the *infinite sites assumption* (ISA) [90] which states that mutations occur at most once in the history of a tumor, and never revert back to the germline state. Under the ISA, there is a nice relationship between the cell fractions of mutations in a sample and the relationship of mutations in a phylogenetic tree. Let  $f_a, f_b$  be the cell fractions of mutations  $a$  and  $b$  in a sample. The *ancestry rule* [38] states that if mutation  $a$  is ancestral to mutation  $b$  in a phylogeny, then  $f_a \geq f_b$ , because all cells that have mutation  $b$  must have mutation  $a$ . The *sum rule* [38, 122] states that the cell fraction of a mutation  $a$  must be equal to or exceed the sum of fractions of child mutations –  $f_a \geq \sum_{v_b \in \text{CHILD}(v_a)} f_b$  where  $v_a$  is the first vertex in a phylogeny that has mutation  $a$ . A number of algorithms, including TrAP [144], Rec-BTP [68], LICHeE [122], AncesTree [38], CITUP [34, 92], MACHINA [42] and CALDER [106] have described a combinatorial framework based on properties like the ones above to infer tumor phylogenies based on mutation cell fraction.

However, mutation cell fractions are derived from sequencing data which has error. Thus, mutations with similar cell fractions may be indistinguishable. Moreover, in the presence of selection, tumors

often consist of multiple clones distinguished by sets of mutations. We expect these sets of mutations to have very similar mutation cell fractions. Thus, instead of operating on individual mutations, many phylogenetic reconstruction algorithms operate on sets of mutations.

We thus divide the task of characterizing the phylogenetic history of a tumor from bulk sequencing data into two problems: (A) clustering mutations and (B) identifying the tree structure. Methods such as PyClone [130], SciClone [103], and Clomial [166] focus on (A) and cluster mutations, without requiring that these clusters are generated by a tree. These algorithms use a probabilistic model for the data to estimate the number of clones, the assignments of mutations into clones, and the cell fraction of clusters of mutations. The combinatorial class of methods mentioned above typically take in clusters of mutations as input.

There is, however, a circular dependence between clustering and identifying the tree. The cell fractions of mutation clusters are used to construct the tree, but an underlying phylogenetic tree constrains mutation cell fractions. This constraint is not accounted for in the clustering algorithms and clustering algorithms may assign cell fractions that do not allow a tree. As such, treating the two problems sequentially may fail, especially when data is noisy. Methods such as PhyloSub [74] have been developed to cluster mutations and infer the tree simultaneously. These methods combine a robust error model for the data, and a tree constraint on the clusters in the generative model. Thus, the resulting clusters necessarily respect the tree constraint. These algorithms use MCMC to sample trees, cluster cell fractions and cluster assignments from the posterior to find the most likely set of parameters.

### **SNV and CNA-based phylogenetic reconstruction**

In the previous section, we describe the relationship between mutation cell fractions and phylogenetic structure under the assumption of no CNAs. However, we do not directly observe the cell fraction of a mutation, we observe the variant allele frequency. When there are no copy-number aberrations, the variant allele frequency is directly proportional to the cell fraction. However,

copy-number aberrations result in violations of this proportionality. For example, an amplification of a mutation increases the number of copies of the mutation in the mixture, and thus increases its variant allele frequency without affecting its cell fraction. Algorithms like PhyloWGS [31] and Canopy [73] use observed copy-number aberrations to correct variant allele frequency and infer a joint phylogeny for SNVs and CNAs.

### 1.3.2 Phylogeny Inference from Single-Cell Sequencing Data

Unlike bulk sequencing, in single-cell sequencing we directly measure individual cells from a tumor. While this avoids the challenges of deconvolving a tumor mixture, phylogeny reconstruction from single-cell sequencing faces its own challenges. Because each cell contains a very small amount of genetic material, scDNA-seq remains prone to errors.

#### SNV based phylogenetic reconstruction

Due to highly uneven coverage resulting from whole genome amplification, detecting SNVs in scDNA-seq is challenging. In particular, when detecting SNVs, we see high rates of false negative errors (observing no reads that contain a variant when one is present in the cell) as well as total dropout (observing no reads at a locus). In addition, false positive errors (observing reads that contain a variant when the cell does not contain one) may result from errors introduced during amplification or sequencing. Single-cell phylogenetic reconstruction algorithm must account for the high error rates in the data.

Several algorithms tackle the problem of phylogenetic reconstruction from single-cell data, using the infinite sites model including SCITE [71], Sci $\phi$  [138] and OncoNEM [129].



## SNV and CNA-based phylogenetic reconstruction

There are presently no algorithms that directly integrate SNV and CNA data into a phylogenetic algorithm based on single-cell data. This has been largely due to the fact that it is challenging to measure both SNVs and CNAs from the same set of cells (as described in Section 1.2.2). However, a number of algorithms attempt to model the affect that copy-number aberrations have on observed SNVs without directly modeling copy-number aberrations. Copy-number aberrations, in particular deletions, may result in losses of mutations. These algorithms rely on either the Dollo parsimony model (SPhyR [37] and SASC [27]) in which mutations are gained at most once but may be lost recurrently, or the finite sites model (SiFit [163] and SiCloneFit [164]) in which mutations may be gained and lost recurrently.

## 1.4 Contributions

### 1.4.1 A hybrid combinatorial-probabilistic approach to phylogeny inference from bulk sequencing data

In Chapter 2, we introduce PASTRI (Probabilistic Algorithm for Somatic Tree Inference), which performs both clustering and tree inference. It uses a hybrid approach, modeling observed data with a probabilistic model, and integrating a combinatorial tree enumeration algorithm into inference. By taking advantage of the combinatorial framework, PASTRI is able to perform inference more efficiently, and as a result obtains improved accuracy and reduced runtime over probabilistic methods. Due to the probabilistic model, PASTRI is more robust to noise in data than combinatorial approaches, and demonstrates improved accuracy on simulated data.

### 1.4.2 Integrating SNVs and CNAs for phylogeny inference from bulk sequencing data

In Chapter 3, we introduce SPRUCE (Somatic Phylogeny Construction Using Combinatorial Enumeration), an algorithm which integrates both single nucleotide variant (SNV) and CNA information to reconstruct tumor phylogenetic trees [40]. We show that by explicitly considering the combinatorial structure of the problem, SPRUCE achieves higher recall values on both small and large instances compared to existing methods.

### 1.4.3 Integrating SNVs and CNAs for phylogeny inference from single-cell data

In Chapter 4, we introduce SCARTLET (Single Cell Algorithm for Reconstructing the Loss-supported Evolution of Tumors). SCARLET is a tool to reconstruct phylogenies from single-cell DNA sequencing data. We introduce a loss-supported model for tumors, where copy-number aberrations may result in losses of SNVs. We show that integrating these two signals yields a more informed and consistent phylogenetic reconstruction.

### 1.4.4 Leveraging technical biases in single-cell DNA sequencing to phase genomic variants

In Chapter 5, we investigate amplification biases in single-cell data. In particular, we show that long-range read depth correlation can be used to phase genomic variants.

## Chapter 2

# Tumor Phylogeny Inference Using Tree-Constrained Importance Sampling

### 2.1 Introduction

In this chapter, we introduce PASTRI (Probabilistic Algorithm for Somatic Tree Inference), a new algorithm that simultaneously clusters mutations into clones and infers a phylogenetic tree that relates the clones. PASTRI separates inference into two parts by exploiting the conditional independence of the observed allele counts from the latent phylogenetic tree given the cluster cell fractions and uses an importance sampling approach to infer the most likely tree. First, we sample likely cluster cell fractions from an informed proposal distribution which comes from a clustering algorithm without the tree constraint [103, 130, 166], and calculate the data likelihood given these cell fractions. Second, given samples of cancer cell fractions, we use the combinatorial algorithm

described in [38, 122] to exactly enumerate the set of possible trees for the given cell fractions. This allows us to efficiently compute the likelihood of all trees that respect the tree constraint, under a realistic noise model for the data. Using clustering estimates allows the algorithm to focus in on areas of high probability and search the space more efficiently than previous MCMC approaches. Moreover, PASTRI samples only cell fractions, and forgoes sampling from the space of trees and cluster assignments. As a result, PASTRI is faster and has better convergence properties than previous MCMC approaches.

We show on simulated data that PASTRI outperforms both combinatorial and probabilistic methods in accuracy and runtime. We then look at a chronic lymphocytic leukemia (CLL) dataset introduced in [128] which used PhyloSub to classify patients into those with linear or complex branching phylogenies. We reanalyze data from a patient that was originally classified as having a complex branching phylogeny. We show that PASTRI is able to find a better solution that is linear instead of branching, under both our model and the PhyloSub model.

## 2.2 Methods

First, in Section 2.2.1, we introduce our model for tumor evolution and sequencing mixtures. We conclude this section by describing the tree constraint on cell fractions (Figure 2.1). In Section 2.2.2, we describe the generative model that we use for trees, clusters of mutations, and observed allele accounts (Figure 2.2). Then, in Section 2.2.3, we describe the PASTRI algorithm, and how we calculate the posterior probability of phylogenetic trees (Figure 2.3).

### 2.2.1 Model

As a first step, we model tumor evolution using single-nucleotide variants (SNVs), leaving the extension to copy-number aberrations as future work. As in previous work [31, 38, 68, 73, 74, 92, 122, 145], we assume that each locus mutates at most once during the lifetime of the tumor, an

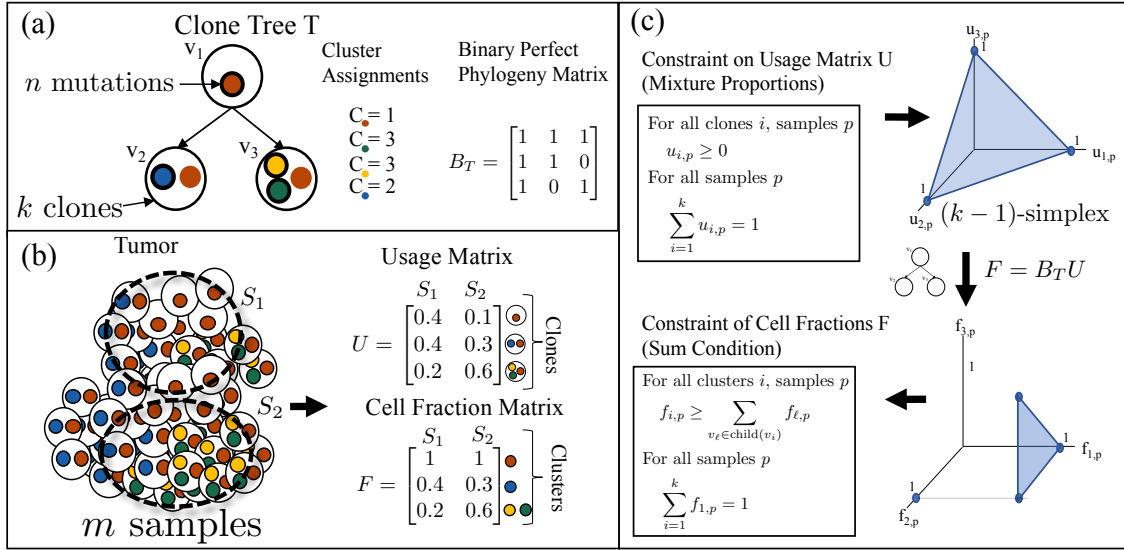


Figure 2.1: **Tree Constraint on Cluster Cell Fractions.** (a) We model the evolution of a tumor as a clone tree  $T$ , with  $k$  vertices corresponding to clones in the tumor. A mutation is assigned to the clone in which it originates. Under the infinite sites assumption, a mutation occurs once, and is never lost. Thus, if a mutation occurs in clone  $v_i$ , all descendent clones of  $v_i$  will also contain that mutation. A clone tree  $T$  can be described by a binary perfect phylogeny matrix  $B_T$ . (b) From a heterogeneous tumor, we take  $m$  samples. Each sample is made up of a mixture of clones. The usage matrix  $U$  describes the proportion of each clone in each sample. The cell fraction  $F$  describes the proportion of cells that contain a given cluster of mutations. For example, here the clone  $v_1$  occurs in  $u_{1,1} = 40\%$  of sample  $S_1$ , but the red mutation from clone  $v_1$  occurs in  $f_{1,1} = 100\%$  of all cells. (c) As the usage matrix  $U$  describes mixture proportions, the columns of  $U$  are constrained to be non-negative and on the  $(k-1)$ -simplex. For a tree  $T$ , we relate  $F$  and  $U$  as  $F = B_T U$ . The set of allowed cell fractions for a tree  $T$  is a linear transformation of the  $(k-1)$ -simplex and is unique for every distinct tree  $T$ . This set can be described by the Sum Condition.

assumption known as the *infinite-sites assumption* (ISA). As such, we encode the state of a locus as a binary character, with a 0 indicating the germline state and a 1 indicating a mutation. If  $n$  loci are affected by mutations, we model the genome  $\mathbf{g}$  of a cell as a binary string of length  $n$ . We model cancer evolution as a *clone tree*  $T = (V(T), E(T))$ , a directed tree with  $|V(T)| = k$  vertices (Figure 2.1(a)). Vertex  $v_i$  corresponds to a clone  $i$  in the tumor, and a directed edge  $(v_i, v_j)$  encodes the evolutionary relationship that clone  $j$  is a direct descendant of clone  $i$ . Alternatively we can represent a tree  $t$  as a  $k \times k$  perfect-phylogeny matrix  $B_T$  [62]. Column  $i$  of matrix  $B_T$  corresponds to vertex  $v_i$ , such that  $b_{j,i} = 1$  if vertex  $v_j$  is on the unique path from  $v_i$  to the root

and 0 otherwise.

We use *cluster* to refer to the set of mutations that occur in a particular clone. Then, let *cluster assignment* vector  $\vec{c} = (c_1, \dots, c_n)$  be a vector mapping each mutation  $j$  to a clone  $v$ , such that  $c_j = v$  indicates that  $v$  is the first clone in which mutation  $j$  occurs. By the ISA, the genomes of all descendant clones of  $v_i$  also contain mutation  $j$ . The genome of a clone  $v_i$  is then defined by the set of mutations assigned to vertices on the path from the root of the tree to  $v_i$ .

We obtain one or more DNA sequencing samples from a tumor, separated spatially or temporally (Figure 2.1(b)). Each of these samples contain mixtures of the  $k$  clones in different proportions. Let *usage matrix*  $U$  be the  $k \times m$  matrix of clone proportions across  $m$  samples, such that  $u_{i,p}$  is the proportion of cells in sample  $p$  belonging to clone  $i$ . For all clones  $i$  and samples  $p$ , the entries of  $U$  are non-negative, i.e.  $u_{i,p} \geq 0$  for all  $i, p$ , and the columns of  $U$  are on the  $(k - 1)$ -simplex,  $\sum_{i=1}^k u_{i,p} = 1$  for all  $p$ . Let clone 1 correspond to normal (non-tumor) cells, such that  $u_{1,p}$  correspond to the proportion of normal admixture in each sample  $p$ . For a vertex  $v_i$  and sample  $p$ , the *cell fraction*  $f_{i,p}$  of the mutations assigned to  $v_i$  is the proportion of the cells in the sample that contain those mutation. As a result of the ISA, all descendants of  $v_i$  will also contain any mutation that occurred at  $v_i$ . Thus, we relate the clone usage matrix  $U$  to the cluster cell fraction matrix  $F = f_{i,p}$  as follows. Let  $\delta(v_i)$  be the set of children of  $v_i$  in  $T$ , and let  $\Delta(v_i)$  be the set of all descendants of  $v_i$ . Then, in sample  $p$ ,  $f_{i,p} = u_{i,p} + \sum_{v_j \in \Delta(v_i)} u_{j,p} = u_{i,p} + \sum_{v_j \in \delta(v_i)} f_{j,p}$ . Thus, we have the following constraint of cell fractions given a tree  $T$ , described in previous works as the *Sum Condition* [38, 74, 91, 122, 145].

$$f_{i,p} \geq \sum_{v_\ell \in \delta(v_i)} f_{\ell,p} \text{ for all vertices } i \text{ and samples } p. \quad (\text{Sum Condition})$$

As clone 1 corresponds to normal cells and all tumor cells are descendants of a normal cell, we also have the constraint that  $f_{1,p} = 1$  for all samples  $p$ .

We can equivalently relate  $F$  to  $U$  and a tree  $T$  as  $F = B_T U$ , where  $B_T$  is the binary perfect phylogeny matrix corresponding to  $T$ . As  $B_T$  is invertible, we also have that  $U = B_T^{-1} F$ . Allowed

cell fractions  $F$  for a tree  $T$  are then those for which  $U = B_T^{-1}F$  is a valid usage matrix, i.e. the entries are non-negative and the columns are on the  $(k - 1)$ -simplex. As described in [38], a necessary and sufficient condition for this is that  $F$  respects the Sum Condition, and the normal clone has a frequency of  $f_{1,p} = 1$  for all samples  $p$  (Figure 2.1(c)).

For each detected mutation  $j$  in sample  $p$ , we measure the number of variant reads  $a_{j,p}$  and the total number of reads  $d_{j,p}$  that align to the locus. Suppose we observe data for  $n$  mutations across all samples. Let  $A$  and  $D$  be  $n \times m$  matrices corresponding to the observed number of variant and total reads for each mutation. The *variant-allele frequency*  $a_{j,p}/d_{j,p}$  of a mutation  $j$  in sample  $p$  is proportional to amount of genomic material containing the variant in the mixture. Under the infinite sites assumption with a diploid genome, this is  $\frac{1}{2}f_{c_j,p}$ , as each cell containing the mutation has one mutated and one unmutated copy.

### 2.2.2 Probabilistic Model

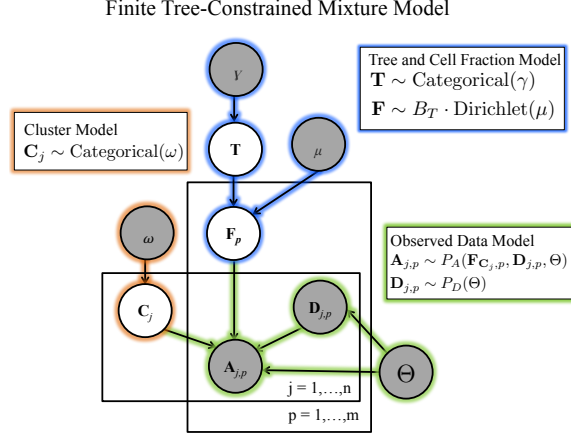


Figure 2.2: **Generative model for mutation allele counts from a tumor.** A latent clone tree  $\mathbf{T}$  generates  $m$  samples consisting of mixtures of cells with different mutations, where cluster  $i$  of mutations in sample  $p$  is found found in cell fraction  $\mathbf{F}_{i,p}$ . Each mutation is assigned to a cluster  $\mathbf{C}_j$ . Variant read counts  $\mathbf{A}$  are generated for each mutation with a binomial likelihood model, given an observed total read counts  $\mathbf{D}$ .

We model the data using a finite tree-constrained mixture model (Figure 2.2). We divide the model into three components: the tree and cell fraction model (highlighted in blue), the cluster assignment model (orange), and the observed data model (green).

We first describe the tree and cell fraction model. Let  $\mathbf{T}$  be the random variable corresponding to the latent true clone tree.  $\mathbf{T}$  follows a categorical distribution which selects tree  $T$  with weight  $\gamma_T$ . For the results in this paper, we set  $\gamma$  such that the probability of all trees is uniform. As the columns of a usage matrix lie on the  $(k-1)$ -simplex, i.e.  $\sum_{i=1}^k u_{i,p} = 1$  for all samples  $p$ , we model the usages for sample  $p$ ,  $U_p \sim \text{Dirichlet}(\mu_1, \dots, \mu_k)$  with vector of hyperparameters  $\mu$  of length  $k$ . In terms of  $F_p$ , this is  $F_p \sim B_T \cdot \text{Dirichlet}(\mu_1, \dots, \mu_k)$ , where  $B_T$  is the perfect phylogeny matrix corresponding to tree  $T$ . As described in Section 2.2.1,  $F = B_T U$  will respect the Sum Condition for tree  $T$  for valid usage matrix  $U$  (i.e. a usage matrix where all columns are on the  $(k-1)$ -simplex). Any usage matrix  $U$  whose columns are not on the  $(k-1)$ -simplex will have a probability of 0 under the Dirichlet model. Thus, a cell fraction matrix  $F$  will only have non-zero probability if it respects the Sum Condition across all samples.



As in a standard mixture model, the observed data is composed of mixtures of  $k$  clusters. Let  $\mathbf{C} = (\mathbf{C}_1, \dots, \mathbf{C}_n)$  be the random variable corresponding to cluster assignments where  $\mathbf{C}_j$  follows a categorical distribution parameterized by weights  $\omega$ . Under this model, the cluster assignments are conditionally independent given the fixed hyperparameter  $\omega$ . This choice allows us to easily marginalize over possible cluster assignments during inference, described below. This model contrasts with Dirichlet Process clustering approaches, where the number of clusters is not fixed, and the cluster assignments have a complex dependence on each other.

Let  $\mathbf{A}$  and  $\mathbf{D}$  be the random variables corresponding to the number of variant and total reads across  $n$  mutations and  $p$  samples. Under the ISA, we expect the variant-allele frequency  $\mathbf{A}_{j,p}/\mathbf{D}_{j,p}$  of a mutation  $j$  in sample  $p$  to be proportional to the cell fraction of the cluster containing the mutation,  $\mathbf{F}_{\mathbf{C}_{j,p}}$ . That is  $E[\mathbf{A}_{j,p}/\mathbf{D}_{j,p}] = \frac{1}{2}\mathbf{F}_{\mathbf{C}_{j,p}}$ . A simple model for allele counts is  $\mathbf{A}_{j,p} \sim \text{Binomial}(\mathbf{D}_{j,p}, \frac{1}{2}\mathbf{F}_{\mathbf{C}_{j,p}})$ , and we use this model in the Section 2.3. However, we allow for more sophisticated observation models, involving additional parameters  $\Theta$ , that can model sequencing error, over-dispersion, or copy-number aberrations. The key feature of this model is that the observed allele counts do not depend directly on the tree  $\mathbf{T}$ , only on the cell fractions  $\mathbf{F}$ . This allows us to sample cell fractions and compute the data likelihood and model parameter likelihoods separately.

In summary, the complete data likelihood of our model is

$$\begin{aligned} \Pr(\mathbf{A}, \mathbf{C}, \mathbf{F}, \mathbf{T} | \mathbf{D}, \omega, \gamma, \mu) &= \Pr(\mathbf{A} | \mathbf{C}, \mathbf{F}, \mathbf{D}) \Pr(\mathbf{F} | \mathbf{T}, \mu) \Pr(\mathbf{T} | \gamma) \Pr(\mathbf{C} | \omega) \\ &= \prod_{p=1}^m \left( \text{Dirichlet}(B_t^{-1} \mathbf{F}_p | \mu) \prod_{j=1}^n \text{Binomial}(\mathbf{A}_{j,p} | \mathbf{F}_{\mathbf{C}_{j,p}}, \mathbf{D}_{j,p}) \right) \gamma_{\mathbf{T}} \prod_{j=1}^n \omega_{\mathbf{C}_j}. \end{aligned} \quad (2.1)$$

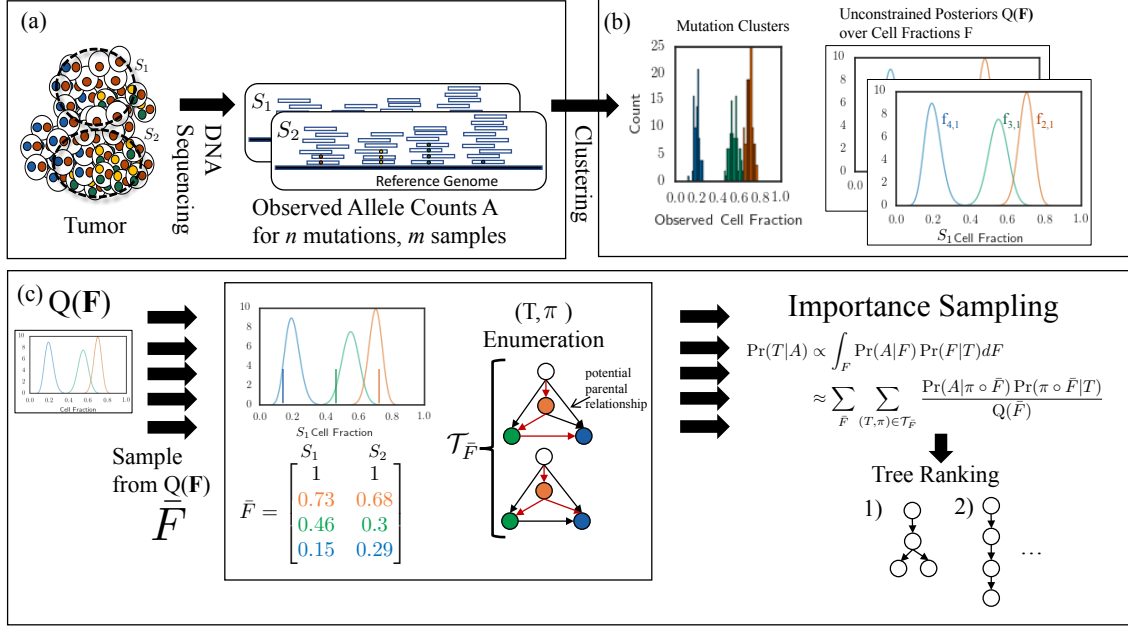


Figure 2.3: **Overview of Pastr Algorithm.** (a) We observe variant-allele and total read counts for  $n$  mutations  $m$  samples of the tumor. (b) A clustering algorithm that does not require that the data is generated by a phylogenetic tree gives an estimate of the posterior distribution  $Q(\mathbf{F})$  over cluster cell fractions. (c) PASTRI draws samples  $\bar{\mathbf{F}}$  from  $Q(\mathbf{F})$ . For each sample, we enumerate the set  $\mathcal{T}_{\bar{\mathbf{F}}}$  of trees  $T$  and cell fraction assignments (permutations)  $\pi$  that respect the Sum Condition across all samples. All trees/permutation pairs not in  $\mathcal{T}_{\bar{\mathbf{F}}}$  have a probability of 0 due to the tree constraint. We estimate the posterior probability of each tree with importance sampling, thus giving us a ranking of which trees are most likely to have generated the data.

### 2.2.3 Tree Inference

Given variant-allele counts  $\mathbf{A} = A$  and total read counts  $\mathbf{D} = D$ , we want to compute posterior probability of a tree  $\mathbf{T} = T$  given the observed data and hyperparameters,

$$\Pr(\mathbf{T} = T \mid \mathbf{A} = A, \mathbf{D} = D, \omega, \gamma, \mu) \propto \Pr(\mathbf{A} = A \mid \mathbf{T} = T, \mathbf{D} = D, \omega, \mu) \Pr(\mathbf{T} = T \mid \gamma) \quad (2.2)$$

In order to calculate this posterior probability from the complete data likelihood given in Equation 2.1, we must marginalize over latent cluster assignments  $\mathbf{C}$  and cluster cell fractions  $\mathbf{F}$ ,

$$\begin{aligned} \Pr(\mathbf{A} = A \mid \mathbf{T} = T, \mathbf{D} = D, \omega, \mu) \Pr(\mathbf{T} = T \mid \gamma) &= \int_F \sum_{\vec{c}} \Pr(\mathbf{A} = A \mid \mathbf{C} = \vec{c}, \mathbf{F} = F, \mathbf{D} = D) \\ &\times \Pr(\mathbf{C} = \vec{c} \mid \omega) \Pr(\mathbf{F} = F \mid \mathbf{T} = T, \mu) \Pr(\mathbf{T} = T \mid \gamma) dF \end{aligned} \quad (2.3)$$

We first show how we analytically marginalize over cluster assignments  $\mathbf{C}$ . Marginalizing over cell fractions is more complicated as  $\mathbf{F}$  is continuous, high-dimensional, and the entries of matrix  $\mathbf{F}$  are not independent. In Section 2.2.3, we demonstrate how we numerically integrate over cell fractions using importance sampling.

In Equation 2.1, the terms that depend on  $\mathbf{C}$  are  $\Pr(\mathbf{A} \mid \mathbf{C}, \mathbf{F}, \mathbf{D}) \Pr(\mathbf{C} \mid \omega)$ . We marginalize over all vector assignments  $\mathbf{C} = \vec{c}$  such that  $c_i \in \{1, \dots, k\}$  to obtain a variant-allele count likelihood  $\Pr(\mathbf{A} \mid \mathbf{F}, \mathbf{D}, \omega)$  that does not depend on  $\mathbf{C}$ . Because the cluster assignments are conditionally independent given  $\omega, A, F$ , we can marginalize over them independently for each mutation.

$$\Pr(\mathbf{A} = A \mid \mathbf{F} = F, \mathbf{D} = D, \omega) = \sum_{\vec{c}} \Pr(\mathbf{A} = A \mid \mathbf{C} = \vec{c}, F = F, D = D) \Pr(\mathbf{C} = \vec{c} \mid \omega) \quad (2.4)$$

$$= \sum_{\vec{c}} \Pr(\mathbf{C} = \vec{c} \mid \omega) \prod_{j=1}^n \prod_{p=1}^m \text{Binomial}(\mathbf{A}_{j,p} \mid \mathbf{F}_{c_j,p}, \mathbf{D}_{j,p}) \quad (2.5)$$

$$= \prod_{j=1}^n \sum_{i=1}^k \left( \Pr(\mathbf{C}_j = i \mid \omega) \prod_{p=1}^m \text{Binomial}(\mathbf{A}_{j,p} \mid \mathbf{F}_{\mathbf{C}_j,p}, \mathbf{D}_{j,p}) \right) \quad (2.6)$$

$$= \prod_{j=1}^n \sum_{i=1}^k \left( \omega_j \prod_{p=1}^m \text{Binomial}(\mathbf{A}_{j,p} \mid \mathbf{F}_{\mathbf{C}_j,p}, \mathbf{D}_{j,p}) \right) \quad (2.7)$$

We refer to this term,  $\Pr(\mathbf{A} = A \mid \mathbf{F} = F, \mathbf{D} = D, \omega)$ , as the *unconstrained data likelihood*, as it does not depend on the tree  $T$ . Thus we have

$$\Pr(\mathbf{T} = T \mid \mathbf{A} = A, \mathbf{D} = D, \omega, \gamma, \mu) \quad (2.8)$$

$$\propto \gamma_T \int_F \Pr(\mathbf{A} = A \mid \mathbf{F} = F, \mathbf{D} = D, \omega) \Pr(\mathbf{F} = F \mid \mathbf{T} = T, \mu) dF. \quad (2.9)$$

In the following section, we describe how we integrate over cell fractions  $\mathbf{F}$ .

### Importance Sampling

We cannot analytically calculate the posterior probability (Equation 2.9) due to the integral over cell fractions  $\mathbf{F}$ . In this section, we describe how to calculate this distribution using importance sampling [149] in combination with a combinatorial tree enumeration algorithm.

Importance sampling uses a proposal distribution  $Q(\mathbf{X})$  over a random variable or set of random variables  $\mathbf{X}$  which approximates distribution of interest  $P(\mathbf{X})$ . We numerically calculate an integral  $\int P(\mathbf{X})d\mathbf{X}$  as follows. Let  $\bar{X}_1, \dots, \bar{X}_N$  be samples from  $Q(\mathbf{X})$ . Then,

$$\int P(\mathbf{X})d\mathbf{X} = \int \frac{P(\mathbf{X})}{Q(\mathbf{X})}Q(\mathbf{X})d\mathbf{X} = E_Q \left[ \frac{P(\mathbf{X})}{Q(\mathbf{X})} \right] \approx \frac{1}{N} \sum_{i=1}^N \frac{P(\bar{X}_i)}{Q(\bar{X}_i)}. \quad (2.10)$$

In our case, the distribution of interest is  $P(\mathbf{F}) = \Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = F, \omega) \Pr(\mathbf{F} = F \mid \mathbf{T} = T, \gamma)$ . We use a clustering method such as SciClone [103] or PyClone [130] which gives an estimate of the posterior probability over  $\mathbf{F}$ , without the tree constraint,  $Q(\mathbf{F}) = \Pr_Q(\mathbf{F} = F \mid \mathbf{D} = D, \mathbf{A} = A, \bullet) \propto \Pr_Q(\mathbf{A} = A \mid \mathbf{D} = S, \mathbf{F} = F, \bullet) \Pr_Q(\mathbf{F} = F, \bullet)$  where we use  $\Pr_Q(\mathbf{X})$  to denote the probability under model used by the clustering method and where  $\bullet$  indicates the distribution may depend on a number of hyperparameters. Different clustering methods may use slightly different likelihoods for observed allele counts, such as parameters that describe the rate of sequencing errors. However, we assume that  $\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = F, \omega) \approx \Pr_Q(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = F, \bullet)$ . This assumption should hold as long as sequencing error is not a significant factor. Thus,  $Q(\mathbf{F})$  and  $P(\mathbf{F})$  differ primarily in the generative model for cell fractions: a tree constraint,  $\Pr(\mathbf{F} = F \mid \mathbf{T} = T, \mu)$  versus no constraint  $\Pr_Q(\mathbf{F} = F, \bullet)$ . However, as there is an underlying tree that generated the data, we expect that the true  $F$  respects a tree constraint on cell fractions. Thus, for the true tree  $T$ , we expect a significant portion of posterior probability mass respects the tree constraint and using the unconstrained  $Q(\mathbf{F})$  is an effective approximation of constrained  $P(\mathbf{F})$ .

Sampling from the unconstrained posterior  $Q(\mathbf{F})$ , cell fractions are not yet assigned to vertices of the tree. A permutation  $\pi$  of the rows of a sample  $\bar{F}$  correspond to assignments of cell fractions to vertices of the tree. We denote by  $\pi_\ell \circ \bar{F}$  the  $\ell^{th}$  permutation of the rows of  $\bar{F}$ . All permutations result in the same data likelihood  $\Pr(\mathbf{A} = A \mid \mathbf{F} = \bar{F}, \mathbf{D} = D, \omega)$ . However, not all permutations respect the Sum Condition (Figure 2.4). In general, the number of cell fraction assignments that respect the Sum Condition is small relative to the total number of permutations. For example, consider a linear tree  $T$ , as in Figure 2.4. If there are  $k$  vertices, there is a single permutation (a descending order of frequencies) that respects the Sum Condition out of  $k!$  total permutations. Remember, cell fraction samples  $\bar{F}$  that do not respect the Sum Condition for a tree  $T$  have a probability  $\Pr(\mathbf{F} = \bar{F} \mid \mathbf{T}, \mu) = 0$ . Thus, if we enumerate all  $k!$  permutations for each sample  $\bar{F}$ , most permutations would have a probability of 0.

Here we introduce an approach where, given a sample  $\bar{F}$ , we can enumerate exactly the set of trees  $T$  and permutations  $\pi$  for which  $\pi \circ \bar{F}$  respects the Sum Condition. Indeed, this problem – given a set of cell fractions  $\bar{F}$ , find a tree  $T$  and an assignment  $\pi$  of cell fractions to vertices which respect the Sum Condition – is the exact problem described by combinatorial methods. [38, 92, 122] describe the solutions to this problem as finding a constrained set of spanning trees of a particular graph. [40, 122] use a specialized version of the Gabow-Myers algorithm [53] to enumerate only this specific set of trees.

We then calculate the posterior probability of a tree  $T$  as follows. First, generate  $N$  samples  $\bar{F}^{(1)}, \dots, \bar{F}^{(N)} \sim Q(\mathbf{F})$ . For each sample  $\bar{F}^{(i)}$ , enumerate the set  $\mathcal{T}_{\bar{F}^{(i)}}$  of tree/permutation pairs

$(T, \pi_\ell)$  that respect the Sum Condition across all vertices. Then,

$$\Pr(\mathbf{T} = T \mid \mathbf{A} = A, \mathbf{D} = D, \omega, \gamma, \mu) \propto \gamma_T \int_F \Pr(\mathbf{A} = A \mid \mathbf{F} = F, \mathbf{D} = D, \omega) \Pr(\mathbf{F} = F \mid \mathbf{T} = T, \mu) dF \quad (2.11)$$

$$\approx \frac{\gamma_T}{N \cdot k!} \sum_{i=1}^N \sum_{\ell=1}^{k!} \frac{\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \bar{F}^{(i)}, \omega) \Pr(\mathbf{F} = \pi_\ell \circ \bar{F}^{(i)} \mid \mathbf{T} = T, \mu)}{Q(\bar{\mathbf{F}}^{(i)})} \quad (2.12)$$

$$= \frac{\gamma_T}{N \cdot k!} \sum_{i=1}^N \sum_{\substack{(T', \pi_\ell) \in \mathcal{T}_{\bar{F}^{(i)}} \\ \text{s.t. } T=T'}} \frac{\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \bar{F}^{(i)}, \omega) \Pr(\mathbf{F} = \pi_\ell \circ \bar{F}^{(i)} \mid \mathbf{T} = T, \mu)}{Q(\bar{\mathbf{F}}^{(i)})}. \quad (2.13)$$

Note that any pair  $(T, \pi_\ell) \notin \mathcal{T}_{\bar{F}^{(i)}}$  will have a probability  $\Pr(\mathbf{T} \mid \mathbf{F} = \pi_\ell \circ \bar{F}^{(i)}, \mu) = 0$ . Thus, summing over just  $(T, \pi_\ell) \in \mathcal{T}_{\bar{F}^{(i)}}$  (Equation 2.13) is equivalent to summing over all permutations (Equation 2.12). In practice, instead of calculating the likelihood for each tree  $T$  separately, we can use the same set of samples. Thus, for each sample  $\bar{F}^{(i)}$ , we only need to enumerate  $\mathcal{T}_{\bar{F}^{(i)}}$  once.

In Section 2.3, we report optimal cluster cell fractions  $F^*$  and cluster assignments  $\vec{c}^*$  as follows. Let  $T^*$  be the optimal tree, the one with the highest posterior probability.

$$F^* = \operatorname{argmax}_{\bar{F}} \max_{(T^*, \pi) \in \mathcal{T}_{\bar{F}}} \Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \bar{F}, \omega) \Pr(\mathbf{F} = \pi \circ \bar{F} \mid \mathbf{T} = T^*, \mu) \quad (2.14)$$

As cluster assignments are conditionally independent given  $\omega, F^*$  and allele counts, we can optimize them independently.

$$c_j^* = \operatorname{argmax}_{i \in \{1, \dots, k\}} \omega_i \Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = F^*, \mathbf{C}_j = i) \quad (2.15)$$

In Supplementary Methods A.1, we describe how we generalize importance sampling to use multiple proposal distributions. This allows us to sequentially adjust the proposal distribution to find more samples that respect the Sum Condition, using the unconstrained posterior as a starting

point.

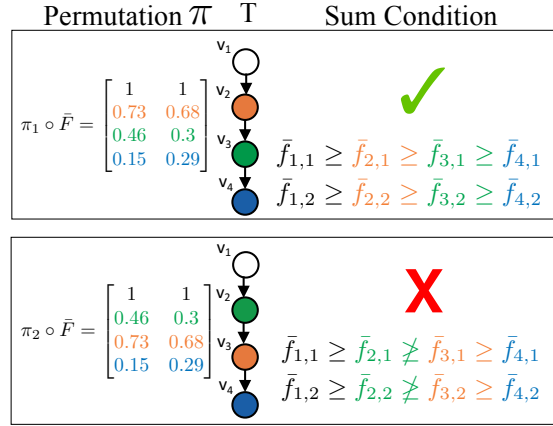


Figure 2.4: **Importance Sampling.** The data likelihood  $\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \pi_i \circ \bar{F}, \omega)$  is the same for all permutations  $\pi$  of  $\bar{F}$  for this tree  $T$ . However,  $\pi_1 \circ \bar{F}$  satisfies the Sum Condition, and  $\pi_2 \circ \bar{F}$  does not. Thus  $\pi_2 \circ \bar{F}$  has a probability  $\Pr(\mathbf{F} = \pi_2 \circ \bar{F} \mid \mathbf{T} = T, \mu) = 0$ .

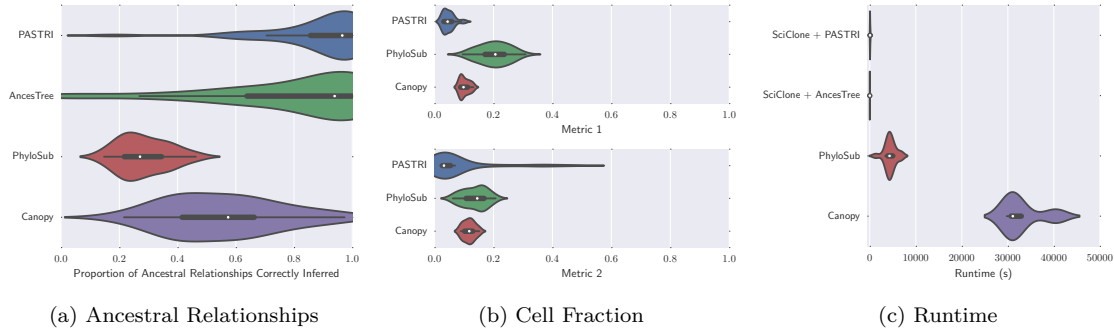
## 2.3 Results

### 2.3.1 Comparison to Phylogeny Reconstruction Algorithms

We compare PASTRI to three other methods for constructing tumor phylogenies: PhyloSub [74], Canopy [73] and AncesTree [38]. PhyloSub and Canopy both employ a Bayesian non-parametric model to simultaneously infer the number of clones, the most likely clusters and cluster cell fractions, and the phylogeny. AncesTree is a combinatorial method which takes as input clusters of SNVs and infers the largest tree with the set of clusters. We use the same set of SciClone results as input to both PASTRI and AncesTree.

We generate 50 instances each of 3, 4 and 5 vertex trees with 20 SNVs. In all instances, we use 5 samples with a read depth of 200X. PhyloSub and Canopy were both run using default parameters. PASTRI was run with hyperparameters  $\omega, \gamma$ , and  $\mu$  set to be uniform for model parameters, and 10,000 samples. We compare the methods on three metrics – accuracy in recovering ancestral

relationships, accuracy in cluster cell fractions and runtime.



**Figure 2.5: Comparison to phylogenetic reconstruction algorithms.** We simulate 50 trees with 5 nodes, 5 samples, and 20 mutations. (a) To quantify the accuracy of the tree inference, we measure for what proportion of pairs of mutations  $c, d$  did the algorithms correctly recover the ancestral relationship –  $c$  is ancestral to  $d$ ,  $d$  is ancestral to  $c$ ,  $c$  and  $d$  are in the same cluster, or  $c$  and  $d$  are on separate branches. (b) To measure how accurately the algorithms recover the true cell fractions, we report two metrics, given by Equations 2.16 and 2.17 to measure accuracy in recovering cluster cell fractions. Both metrics measure the average distance between entries the true cell fraction matrix and the reported. Metric 1 penalizes for overestimating the number of clusters, and Metric 2 penalizes for underestimating. (c) We report the runtime in seconds of each method.

### Recovering the Phylogeny

To assess the methods' abilities to recover the true phylogeny, we measured the proportion of ancestral relationships between SNVs that were correctly reported in the best reported tree by each method. A pair of SNVs  $c$  and  $d$  can have one of four relationships in a tree:  $c$  and  $d$  may be in the same cluster,  $c$  may be ancestral to  $d$ ,  $d$  may be ancestral to  $c$ , or  $c$  and  $d$  may be on distinct branches of the tree. For all pairs of distinct SNVs in each sample, we measure whether the reported relationship matched the relationship in the true tree. The results for 5 vertex trees are shown in Figure 2.5(a), with 3 and 4 vertex trees in Supplementary Figure 3. We see that PASTRI outperforms all three methods. Indeed on three vertex trees, PASTRI correctly infers all ancestral relationships on 58% of instances, while neither PhyloSub nor Canopy have a single instance on which this happens.



## Recovering Cluster Frequencies

To assess the methods' ability to recover true cluster frequencies, we compare the reported clone frequencies  $F_{PT}^*$ ,  $F_{PS}^*$  and  $F_{CP}^*$  – for PASTRI, PhyloSub and Canopy respectively – to the true values  $F$ . We do not report AncesTree in this section as AncesTree does not perform its own clustering. As the algorithms may return a different number of clusters than the true model, we use two different metrics to measure accuracy. Let  $\delta(\mathbf{f}_i, \mathbf{f}_j) = \frac{1}{\ell} \sum_{\ell=1}^n |f_{i,\ell} - f_{j,\ell}|$ , the average per-entry distance between two rows.

1. Metric 1 – A measure of sensitivity, matches the *true* clones to the nearest reported clones.

$$M_1(F, F^*) = \sum_{i=1}^k \operatorname{argmin}_{j \in [1 \dots k^*]} \delta(\mathbf{f}_i, \mathbf{f}_j^*) \quad (2.16)$$

2. Metric 2 – A measure of specificity, matches the *reported* clones to the nearest *true* clones.

$$M_2(F, F^*) = \sum_{j=1}^{k^*} \operatorname{argmin}_{i \in [1 \dots k]} \delta(\mathbf{f}_i, \mathbf{f}_j^*) \quad (2.17)$$

Results using both metrics for 5 vertex trees are shown in Figure 2.5(b). Results with 3 and 4 vertex trees in Supplementary Figure 4. PASTRI consistently outperforms both methods. Note that in all cases, the median PASTRI result is below the first quartile of results for both PhyloSub and Canopy.

## Runtime

We measure the runtime for all three algorithms. For PASTRI and AncesTree, the combined runtime of clustering and inference is used. The results for 5 vertex trees are shown in Figure 2.5(c). Results for 3 and 4 vertex trees in Supplementary Figure 2. Note that both PhyloSub and Canopy are sampling methods and the runtime is determined by how many samples each method uses within their default parameters. As such, we expect increasing the number of samples improves the accuracy

of the methods. However, we see that PASTRI achieves improved accuracy with significantly lower runtimes than either PhyloSub or Canopy. In general, we found runtimes ranged from on the order of seconds for SciClone and AncesTree, less than a minute for PASTRI, hours for PhyloSub, and days for Canopy.

### 2.3.2 Chronic Lymphocytic Leukemia

Rose-Zerilli et al [128] presents a dataset from a cohort of 13 patients with chronic lymphocytic leukemia (CLL). Each patient had 2-5 samples taken longitudinally over the course of their disease and these were subject to a number of analyses including targeted deep sequencing. The authors classify the patients into those with linear phylogenies (4/13 patients) and those with complex branching phylogenies (9/13 patients) on the basis of PhyloSub [74] analysis. Branching phylogenies are used as evidence of sub-clonal competition prior to therapy.

Here we investigate Patient 5 in the study. In [128], this patient was classified as having a complex branching phylogeny (Figure 2.6) using PhyloSub analysis. Running PASTRI results in an optimal tree with 8 nodes that is mostly linear. The fully linear tree was ranked 3/115. We calculate the likelihood of the data under both the PASTRI allele count observation model, and the PhyloSub observation model, which allows for sequencing error in observations. Under both models, the likelihood for both trees found by PASTRI are better than the branching phylogeny reported in [128]. There are two possible explanations for the discrepancy. First, it is possible that PhyloSub’s sampling procedure never found either the optimal tree or the linear tree. Second, the tree-structured stick breaking prior over trees used by PhyloSub has hyperparameters that influence the width and the depth of the trees. As a result, the branching phylogeny presented may have been preferred to a linear phylogeny. However, if an intention of a study is to classify patients by phylogenetic tree structure, one should keep the priors over trees as neutral as possible.

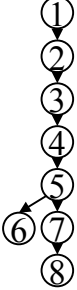
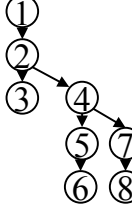
| Optimal tree<br>found by PASTRI                                                                                                                                                                                                                                                         | Linear tree<br>found by PASTRI                                                                                                                                                                                                                                                          | Reported branching tree<br>in Rose-Zerilli et. al                                                                                                                                                                                                                                 |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  $F^{\text{OPT}} = \begin{bmatrix} .97 & .95 & .94 \\ .84 & .72 & .65 \\ .66 & .61 & .63 \\ .58 & .49 & .53 \\ .53 & .44 & .52 \\ .21 & .23 & .27 \\ .12 & .13 & .05 \\ .05 & .11 & .18 \end{bmatrix}$ |  $F^{\text{LIN}} = \begin{bmatrix} .97 & .95 & .94 \\ .85 & .72 & .65 \\ .66 & .60 & .64 \\ .58 & .49 & .54 \\ .53 & .41 & .52 \\ .23 & .23 & .27 \\ .09 & .15 & .21 \\ .04 & .11 & .18 \end{bmatrix}$ |  $F^{\text{PS}} = \begin{bmatrix} 1 & 1 & 1 \\ .92 & .81 & .89 \\ .06 & .14 & .23 \\ .77 & .64 & .58 \\ .13 & .13 & .04 \\ .10 & .05 & .03 \\ .58 & .41 & .52 \\ .41 & .02 & .37 \end{bmatrix}$ |
| Data Log-Likelihood under PASTRI observation model:                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                   |
| -1134.07                                                                                                                                                                                                                                                                                | -1135.12                                                                                                                                                                                                                                                                                | -1161.98                                                                                                                                                                                                                                                                          |
| Data Log-Likelihood under PhyloSub observation model:                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                   |
| -1133.37                                                                                                                                                                                                                                                                                | -1134.45                                                                                                                                                                                                                                                                                | -1140.18                                                                                                                                                                                                                                                                          |

Figure 2.6: **Patient 5** from Rose-Zerilli et al [128]. This patient was classified as having a complex branching phylogeny using PhyloSub analysis (right). Running PASTRI finds an optimal phylogeny that is mostly linear (left). Restricting to linear phylogenies results in the center tree, which was the third most likely phylogeny out of 115 possible. We calculate the likelihood of the data under both the PASTRI allele count observation model, and the PhyloSub observation model, which allows for sequencing error in observations. Under both models, the likelihood under both trees found by PASTRI are better than the branching phylogeny reported in [128].

## 2.4 Discussion

We introduced PASTRI, a new method that simultaneously clusters mutations and infers tumor phylogenies from bulk-sequencing data. PASTRI exploits the conditional independence of the observed allele counts from the latent phylogenetic tree given the cluster cell fractions. Because of this conditional independence, we are able to exploit combinatorial constraints [38, 74, 91, 122, 145] to efficiently marginalize over cluster cell fractions, using a combinatorial tree enumeration algorithm. At the same time, we utilize a probabilistic model for the observed allele counts that models errors in sequencing data. By leveraging combinatorial structure into the probabilistic inference, we obtain improved accuracy over prior combinatorial algorithms – due to better modeling of noise in sequence

data – and improved runtime over probabilistic methods – due to more efficient inference.

By using importance sampling, we direct the computation toward regions of the sample space with highest probability. As a result, PASTRI will more precisely calculate the posterior probability of the trees that have the highest probability as compared to low probability trees. In our application, this is an acceptable tradeoff, as we are most interested in recovering the highest probability trees, and generally are not concerned with the exact ordering of very unlikely trees. Note that our importance sampling approach can use any algorithm that computes a posterior distribution over clusters; we have used SciClone [103] in this work, but PyClone [130], Clomial [166] or other algorithms can also be used.

In Section 2.3.2, we examined data from a study that aimed to classify patients as having either a linear or a complex branching phylogeny. We showed a case where an existing algorithm failed to find a linear phylogeny that had higher likelihood than the reported branching phylogeny. In cases with few samples or low coverage, there is significant ambiguity in the tree structure, and many trees may have similar likelihood. Thus, reporting a single solution with highest likelihood not accurately recover the underlying phylogeny. In particular, when distinguishing between linear and branching phylogenies, there may be many branching trees, but only a single linear tree. Thus, it is easy to overfit phylogeny reconstruction, and select a branching phylogeny when a linear is nearly as likely, in the presence of noisy data.

There are a number of directions for future work. First, is to improve the model selection problem of choosing the number of clones/clusters. In this work, we relied on the model selection procedure performed by the clustering algorithm. However, it is plausible that in some scenarios the phylogenetic tree constraint would shift cluster cell fractions in such a way to affect the choice of the number of clones. Second, while we have demonstrated PASTRI using single-nucleotide mutations, the hybrid combinatorial-probabilistic approach generalizes to the analysis of copy-number aberrations (CNAs), which are widespread in solid tumors. As CNAs affect the observed allele counts of single-nucleotide mutations in a predictable way (as described in [31, 40]), the interaction between observed CNAs and allele counts can be accounted for in the observation model. Finally, the hybrid

inference algorithm presented here could generalize to other problems where there is conditional independence between a combinatorial structure (here a tree) and a probabilistic model.

## Chapter 3

# Inferring the Mutational History of a Tumor using Multi-state Perfect Phylogeny Mixtures

1. Joint work with Mohammed El-Kebir

### 3.1 Introduction

Cancer is an evolutionary process, where somatic mutations accumulate in a population of cells [114]. Accumulation and selection of somatic mutations means that tumors are a heterogeneous mixture of cells with different complements of somatic mutations. In the past few years, researchers have begun to quantify this *intra-tumor heterogeneity* using DNA sequencing from one or multiple samples from the same tumor. Recent studies have demonstrated the importance of intra-tumor heterogeneity in cancer treatment and resistance [49, 101]. A simple characterization of intra-tumor heterogeneity classifies mutations as *clonal* (present in all tumor cells) vs. *subclonal* (present in a subset of tumor cells). However, one can obtain more information more information about the relationship between

tumor clones by applying phylogenetic techniques to reconstruct the evolutionary history of the tumor [31, 38, 57, 74, 92, 113, 122, 144, 160].

Importantly, the process of clonal evolution in a tumor occurs at the level of single cells. In phylogenetic terminology, the somatic evolutionary process is modeled by a phylogenetic tree, whose leaves correspond to extant entities called *taxa*, and whose edges describe the ancestral relationships between the taxa. The taxa are the individual cells in a tumor. Yet, due to technical and financial constraints, the majority of cancer sequencing projects do not sequence individual cells [107–109, 154, 168], but rather bulk tumor samples containing thousands to millions of cells. All of the datasets from The Cancer Genome Atlas (TCGA) [20] and nearly all datasets from the International Cancer Genome Consortium (ICGC) [170] measure mutations in a single tumor sample. More recently, sequencing of multiple bulk samples from the same tumor have been undertaken [12, 14, 57, 61, 85, 112, 131, 133]. In some of these studies, researchers have used phylogenetic trees to represent the relationships between these individual samples. However, the underlying data in these sequencing studies is not the presence/absence of a mutation in an individual cell, but rather the fraction of sequence reads that indicate a mutation, a quantity that is proportional to the fraction of cells that contain the mutation. In phylogenetic terminology, we do not measure the individual taxa, but rather *mixtures* of taxa. Thus, proper phylogenetic analysis of bulk cancer sequencing data demands specialized techniques that handle such mixtures.

Since the goal of cancer phylogenetic studies is to understand ancestral relationships between mutations, character-based phylogenies are desirable. In a character-based phylogeny, taxa are represented by a collection of *characters*, each of which have one of several distinct *states*. The basic problem of character-based phylogenetic tree reconstruction is the following: given the states for  $n$  characters on  $m$  taxa, find a vertex-labeled tree whose leaf labels are the taxa and whose internal nodes are labeled by an ancestral state for each character such that the resulting tree maximizes an objective function (e.g. maximum parsimony or maximum likelihood) over all such labeled trees. Motivated by bulk cancer sequencing data, we consider a phylogenetic tree mixture problem, where the input is not the set of states for each taxon, but rather *mixtures* of these states. The goal

is to explain these given mixtures as a mixing of the leaves of an unknown phylogenetic tree in unknown proportions (Figure 3.1). Stated another way: if one is given  $m$  mixtures of the leaves of a phylogenetic tree, can one reconstruct the tree and the mixing proportions?

The difficulty of the phylogenetic tree mixture problem depends on the evolutionary model. The simplest such model assumes that characters have only two states and that characters change state at most once; i.e. the characters evolve with no homoplasy. The latter condition is called the *infinite sites assumption* and gives rise to a *perfect phylogeny*. Driven by the application to cancer sequencing, there has been a surge of interest in solving the phylogenetic tree mixture problem for a two-state perfect phylogeny, relying on the idea that the infinite sites assumption is a reasonable model for somatic single nucleotide variants (SNVs) [38, 68, 74, 92, 113, 122, 144, 160].

While the two-state perfect phylogeny model might be reasonable for somatic SNVs, it is fairly restrictive for modeling the somatic mutational process in cancer. For instance, somatic copy-number aberrations (CNAs) are ubiquitous in solid tumors [162], and these mutations generally have more than two states. Moreover, the observed frequency of an SNV is confounded by CNAs affecting the locus [36]. Thus far, most analysis methods have either excluded genomic regions with CNAs from tree reconstruction [38, 74, 92, 122, 144] or have made restrictive assumptions on the mutational process. Based on these assumptions, methods estimate the *cancer cell fraction* (CCF), the proportion of tumor cells containing an SNV from the observed SNV frequencies [12, 61, 113]. However, these restrictions are not necessarily true for all tumors and relying on these restrictions may result in incomplete reconstructions or rule out alternative valid phylogenies. An alternative desirable approach is to use a multi-state model for cancer phylogenies. The difficulty of the general multi-state phylogenetic mixture problem is unknown. However, as in the two-state case, we can make the simplifying assumption of no homoplasy. For multi-state characters, the no-homoplasy assumption is referred to as the *infinite alleles assumption* in population genetics, or the *multi-state perfect phylogeny* [47, 65]. In this model, a character may change state more than once on the tree, but changes *to the same state at most once*.

Here, we introduce a perfect phylogeny mixture problem in the case of multi-state characters that



evolve without homoplasy under the infinite alleles assumption. We define this problem formally as the Multi-State Perfect Phylogeny Mixture Deconvolution Problem (PPMDP) and derive a mathematical characterization of its solutions as a restricted class of spanning trees of a labeled multi-graph. We specialize the PPMDP to analyze bulk-sequencing data from multiple samples of a tumor by introducing a multi-state model that captures the interplay between SNVs and CNAs. We apply the resulting algorithm SPRUCE (Somatic Phylogeny Reconstruction Using Combinatorial Enumeration) to simulated data. We find that there is considerable ambiguity in the solution with sequencing data from few samples, but that the number of phylogenetic trees decreases dramatically as the number of samples increases. We show that by explicitly considering the combinatorial structure of the problem, SPRUCE achieves higher recall values on both small and large instances compared to existing methods [31]. On two cancer datasets, a chronic lymphocytic leukemia [133] and a prostate tumor [61], we infer phylogenetic trees that contain CNAs and SNVs. On the latter dataset, we demonstrate the importance of jointly inferring cancer cell fractions and phylogenies.

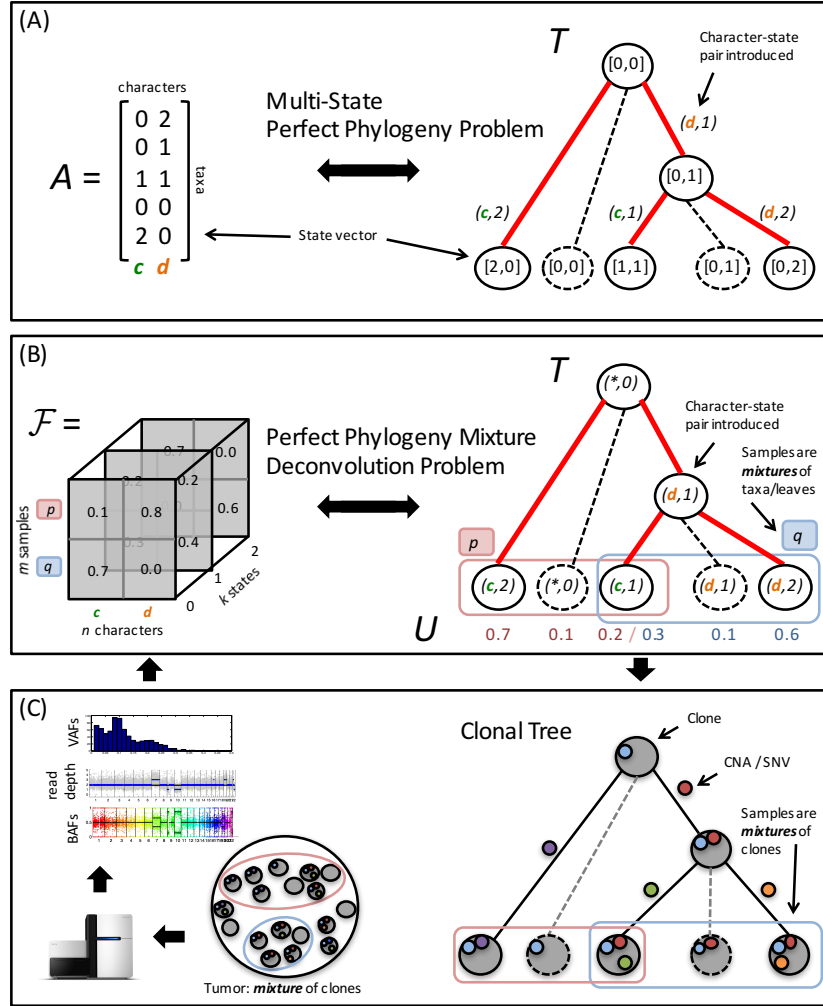


Figure 3.1: **Overview:** (A) In the multi-state perfect phylogeny problem, we are given a matrix  $A$  composed of state vectors for each taxon. We seek to find a tree  $T$  whose leaves correspond to the taxa that satisfies the infinite alleles assumption. (B) In the Perfect Phylogeny Mixture Deconvolution Problem (PPMDP), we do not observe the taxa directly. Instead, our measurements  $\mathcal{F}$  correspond to mixtures of the taxa of a tree  $T$  according to proportions  $U$ . The goal is to infer  $T$  and  $U$  given  $\mathcal{F}$ . (C) Tumors consist of mixtures of distinct cellular populations (*clones*). From bulk DNA sequencing data of one or more samples of a tumor we obtain: variant allele frequencies (VAFs) of single-nucleotide variants (SNVs), and read depth ratios and  $B$ -allele frequencies (BAFs) from copy number aberrations (CNAs). These measurements are superpositions of the mutational signals from the clones present in each sample, mixed according to the prevalence. We show that with a few simplifying assumptions one can obtain  $\mathcal{F}$  from these measurements. The evolution of the clones is described by a clonal tree, which we model as a multi-state perfect phylogeny.

## 3.2 Results

Below, we formulate the Perfect Phylogeny Mixture Deconvolution Problem (PPMDP) (Section 3.2.1), a general problem of constructing phylogenetic trees from mixtures, under the infinite alleles model. We then show how measurements of SNVs and CNAs in bulk tumor sequencing data yield instances of the PPMDP, which we solve using an algorithm that we call SPRUCE (Section 3.2.2). We demonstrate the performance of SPRUCE on simulated and real cancer sequencing data (Section 3.2.3). Throughout this section taxa correspond to distinct tumor populations, characters correspond to genomic loci, and the state of a character encodes the mutational state and copy number of the corresponding locus.

### 3.2.1 The Perfect Phylogeny Mixture Deconvolution Problem

In the multi-state perfect phylogeny model, one is given a *state vector* for each of  $m$  taxa; the entries of this vector give the state of each of  $n$  characters. We label each of the  $k$  states by an integer from 0 to  $k - 1$ . Equivalently, we can represent the input by a matrix  $A$  whose rows are the state vectors of the taxa (Figure 3.1A). The goal is to infer a tree  $T$  whose leaves are labeled by the rows of  $A$  and whose internal nodes are labeled by state vectors satisfying the infinite alleles assumption. That is, a character may change state multiple times on the tree but changes to the same state at most once. Throughout this paper, we restrict our attention to the case of rooted, or directed, perfect phylogenies, where the state vector for the root vertex is the zero vector. This restricted version is consistent with the assumptions for our cancer sequencing application below. Moreover, perfect phylogeny algorithms can generally be extended to the non-zero rooted case or unrooted case with some additional work [65].

In the case of bulk-sequencing data, we do not measure  $A$ , or equivalently the leaves of  $T$  directly, but rather we are given  $m$  samples that are *mixtures* of the rows (taxa) of  $A$ , or equivalently mixtures of the leaves of  $T$ , in unknown proportions (Figure 3.1B). Formally, for each sample  $p$ , character  $c$ , and state  $i \in \{0, \dots, k - 1\}$  we are given the frequency  $f_{p,(c,i)}$ , which is the proportion

of taxa in sample  $p$  that have state  $i$  for character  $c$ . Note that frequencies  $f_{p,(c,i)}$  are non-negative and sum up to 1 across all states. Our input measurements are thus described by a  $k \times m \times n$  frequency tensor  $\mathcal{F} = [[f_{p,(c,i)}]]$ . While the frequencies  $\mathcal{F}$  are obtained by mixing the leaves of  $T$ , the number of leaves in  $T$  is typically unknown. We define a *complete perfect phylogeny tree* as having exactly  $n(k-1) + 1$  vertices such that each edge is labeled by a unique character-state pair. Thus each vertex of a complete perfect phylogeny tree is a taxon, instead of just the leaves. We can assume without loss of generality that  $\mathcal{F}$  is obtained by mixing all the vertices of this complete perfect phylogeny tree (Figure B.1), and in the following we restrict our attention to complete perfect phylogeny trees  $T$ .

We denote the root vertex of  $T$  by  $v_{(*,0)}$  and the remaining vertices by  $v_{(c,i)}$  where  $i$  is the new state of character  $c$  that occurs on the incoming edge. For each vertex  $v_{(c,i)}$  and each sample  $p$ , we define the *usage*  $u_{p,(c,i)}$  to be the proportion of  $v_{(c,i)}$  in sample  $p$ . The usages  $u_{p,(c,i)}$  are non-negative and that the usages of all taxa in a sample sum to 1. Note that  $f_{p,(c,i)}$  denotes the frequency of character-state pair  $(c,i)$  in sample  $p$ , whereas  $u_{p,(c,i)}$  denotes the proportion of the taxon  $v_{(c,i)}$  in sample  $p$ . The *usage matrix*  $U = [u_{p,(c,i)}]$  represents the usages of all taxa across all samples. It follows that the frequency of a character-state pair is the sum of the usages of taxa that have state  $i$  for character  $c$ . That is,  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ , where  $T_{(c,i)}$  is the subtree consisting of vertices whose corresponding state vectors have state  $i$  for character  $c$ . Thus, given a tree  $T$  and usage matrix  $U$ , the frequencies  $f_{p,(c,i)}$  for each sample, character, and state are determined. Our problem is the converse.

*Perfect Phylogeny Mixture Deconvolution Problem (PPMDP)*. Given an  $k \times m \times n$  frequency tensor  $\mathcal{F} = [[f_{p,(c,i)}]]$ , find a complete perfect phylogeny tree  $T$  and a usage matrix  $U = [u_{p,(c,i)}]$  such that  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$  for all character-state pairs  $(c,i)$  and all samples  $p$ .

In Section 3.4.1 we derive a mathematical characterization of solutions to PPMDP and show that these correspond to constrained spanning trees in a directed multi-graph  $G_{\mathcal{F}}$ . We say that a tree  $T$  *generates*  $\mathcal{F}$  if there exists a usage matrix  $U$  such that  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ . See Supplemental Methods B.1.1 for further details and formal definitions of  $T$  and  $A$ .

For our application to cancer genomes described below, we consider a special case of the PPMDP where in addition to the frequencies  $\mathcal{F}$ , we are given a set  $\mathcal{S} = \{S_c \mid c = 1, \dots, n\}$  of state trees, one for each character. The state trees describe the order of states for a character; such an order-constrained character is called a *cladistic* character [47]. Formally, the vertex set of a *state tree*  $S_c$  is the set  $\{0, \dots, k-1\}$  of states, and the edges describe the relationships of the states of character  $c$ . A perfect phylogeny  $T$  is *consistent* with  $\mathcal{S}$  if the states for each character  $c$  occur in the same order as prescribed by the state tree  $S_c \in \mathcal{S}$ . That is, for all characters  $c$  and states  $i, j$ ,  $(c, i)$  is ancestral to  $(c, j)$  in  $T$  if and only if  $i$  is ancestral to  $j$  in  $S_c$ . In the Cladistic-PPMDP, we are given frequencies  $\mathcal{F}$  and state trees  $\mathcal{S}$  and seek to find a complete perfect phylogeny tree  $T$  that generates  $\mathcal{F}$  and is consistent with  $\mathcal{S}$ . In Section 3.4.2 we introduce an enumeration algorithm that exploits the specific combinatorial structure of solutions to the Cladistic-PPMDP. See Supplemental Methods B.1.5) for further details.

### 3.2.2 Multi-State Model for the Somatic Mutational Process in Cancer

Our motivating application for the PPMPD is to analyze multi-sample cancer sequencing data. A tumor is a collection of cells that have evolved from the normal, founder cell by a series of somatic mutations. Here,  $m$  distinct samples from a tumor have been sequenced, each such sample being a mixture of cells with different somatic mutations. Our aim is to reconstruct the phylogenetic tree describing the evolutionary relationships between subpopulations of cells (or *clones*) within the tumor (Figure 3.1C). Earlier work on this problem[38, 68, 74, 92] has focused on the problem of single-nucleotide variants (SNVs) that have changed at most once in the progression from normal to tumor, and thus the phylogenetic tree is a two-state perfect phylogeny. Other approaches have considered SNVs in regions of copy number aberrations by scaling VAFs into *cancer cell fractions* (CCFs), which is the proportion of tumor cells in a sample that have the variant. This quantity has been broadly used in analyzing intra-tumor heterogeneity [5, 12–14, 61, 99, 113, 131]. Here,

we consider copy number aberrations by modeling positions in the genome as multi-state characters.

More specifically, the characters in our model are positions in the genome whose states we model with a triple  $(x, y, z)$  of integers. Here,  $x$  and  $y$  are the number of maternal and paternal copies of the position respectively and  $z$  is the number of mutated copies. We call  $(x, y)$  the *copy-number state*. In general, we lack information to phase mutations to the maternal or paternal chromosome. Thus, we assume without loss of generality that  $x \geq y$ . Unfortunately, for a character  $c$  and sample  $p$ , we cannot derive the frequencies  $f_{p,(c,i)}$  for states  $(x_i, y_i, z_i)$  directly from DNA sequencing data. Instead, alignment of DNA sequence reads from a tumor and matched normal genome gives us measurements of three quantities: *variant allele frequencies* (VAFs) for SNVs, which is the proportion of sequence reads that contain the mutant allele; *B-allele frequencies* for germline single-nucleotide polymorphisms which reveal loss of heterozygosity events; *read-depth ratios* for larger regions, which are the total number of the reads of the region in the tumor sample divided by the total number of reads in the same region of a matched normal [126]. We will show that with a few simplifying assumptions about the allowed state transitions, we can derive frequencies  $\mathcal{F} = [[f_{p,(c,i)}]]$  and the state trees  $\mathcal{S}$  for each character directly from DNA sequencing data. This gives us input to the Cladistic-PPMDP.

We define the matrix  $H = [h_{p,c}]$  of variant allele frequencies, where  $h_{p,c}$  is the VAF of character  $c$  in sample  $p$ . By running a copy-number caller [48, 67, 113, 115, 117] on the B-allele frequencies and the read-depth ratios, we obtain the tensor  $\mathcal{M} = [[\mu_{p,c,(x,y)}]]$  where  $\mu_{p,c,(x,y)}$  is the *mixing proportion* of the copy-number state  $(x, y)$  of character  $c$  in sample  $p$  (Figure 3.2A). Note that  $\sum_{(x,y)} \mu_{p,c,(x,y)} = 1$ . The following linear system relates mixing proportions  $\mathcal{M}$  and variant allele frequencies  $H$  to state frequencies  $\mathcal{F} = [[f_{p,(c,(x,y,z))}]]$ .

$$\sum_z f_{p,(c,(x,y,z))} = \mu_{p,c,(x,y)} \quad \text{for all copy-number states } (x, y) \quad (3.1)$$

$$\sum_{(x,y,z)} z \cdot f_{p,(c,(x,y,z))} = h_{p,c} \cdot \sum_{(x,y)} (x + y) \cdot \mu_{p,c,(x,y)} \quad (3.2)$$

Importantly, this system of equations is under-determined, i.e. given  $\mathcal{M}$  and  $H$  there are many possible  $\mathcal{F} = [[f_{p,(c,(x,y,z))}]]$  that satisfy (3.1) and (3.2). We resolve the ambiguity in  $\mathcal{F}$  by imposing biologically-motivated restrictions on the set of state trees  $\mathcal{S}$ . We define a *state graph* to be a directed graph whose nodes are character states, and whose edges are allowed transitions between states. Figure 3.2B shows a partial representation of this graph including the three types of edges: mutation edges (incrementing  $z$ ), amplification edges (incrementing either  $x$  or  $y$ ) and deletion edges (decrementing either  $x$  or  $y$ ). We define a *valid state tree*  $S$  as a rooted subtree of the state graph such that (i)  $S$  is rooted at  $(1, 1, 0)$ ; (ii)  $S$  has at most one mutation edge; and (iii) vertices of  $S$  with identical copy-number state form a connected subtree of  $S$ . Intuitively, the first condition requires that the root state is the non-mutated heterozygous diploid state (assuming autosomes), the second condition is the infinite sites assumption on SNVs and the third condition is the infinite alleles assumption on copy-number states. Using these constraints, the system (3.1) and (3.2), restricted to the states in  $S$ , is fully determined (Supplemental Methods B.1.7). Note that the conditions above do allow for SNVs to be affected by multiple copy-number events as long as the same copy-number state does not recur.

A valid state tree  $S_c$  is *compatible* with character  $c$  provided that the solution to the system consists of non-negative frequencies  $f_{p,(c,i)}$  (Figure 3.2). There may be several compatible state trees, which has implications for calculating the CCF. Formally, given a sample  $p$  with purity  $\rho$  (proportion of tumor cells in the sample), the CCF of an SNV  $c$  is given by

$$\text{CCF}(p, c) = \frac{1}{\rho} \sum_{\substack{(x,y,z) \\ z > 0}} f_{p,(c,(x,y,z))}. \quad (3.3)$$

Note that the CCF is not uniquely determined from the VAF and the copy-number mixing proportions (Figure 3.2C). To overcome this ambiguity, previous work [12, 61] makes the simplifying assumption that all tumor cells have a fixed number  $n_{\text{chr}}$  of copies carrying the SNV. The quantity  $n_{\text{chr}}$  is computed as the maximum likelihood estimate under a binomial model, resulting in a single CCF choice for each SNV. In contrast, we do not fix the number of mutated copies in distinct

tumor populations. By considering all compatible state trees, we allow for multiple alternative explanations with possibly distinct CCFs.

While the above discussion considered a single VAF  $h_{p,c}$  for each mutation in each sample, in practice VAFs are measured with uncertainty. In Supplemental Methods B.1.7, we describe how to extend the approach using a confidence interval  $[\underline{h}_{p,c}, \bar{h}_{p,c}]$  to account for noise in VAFs. Given VAF confidence intervals and copy-number mixing proportions, our algorithm SPRUCE (Figure 3.3) solves the Cladistic-PPMDP by enumerating combinations of state trees. Instead of naively considering all such combinations, we examine only those state trees that are compatible or pairwise compatible with the input (Section 3.4.2).



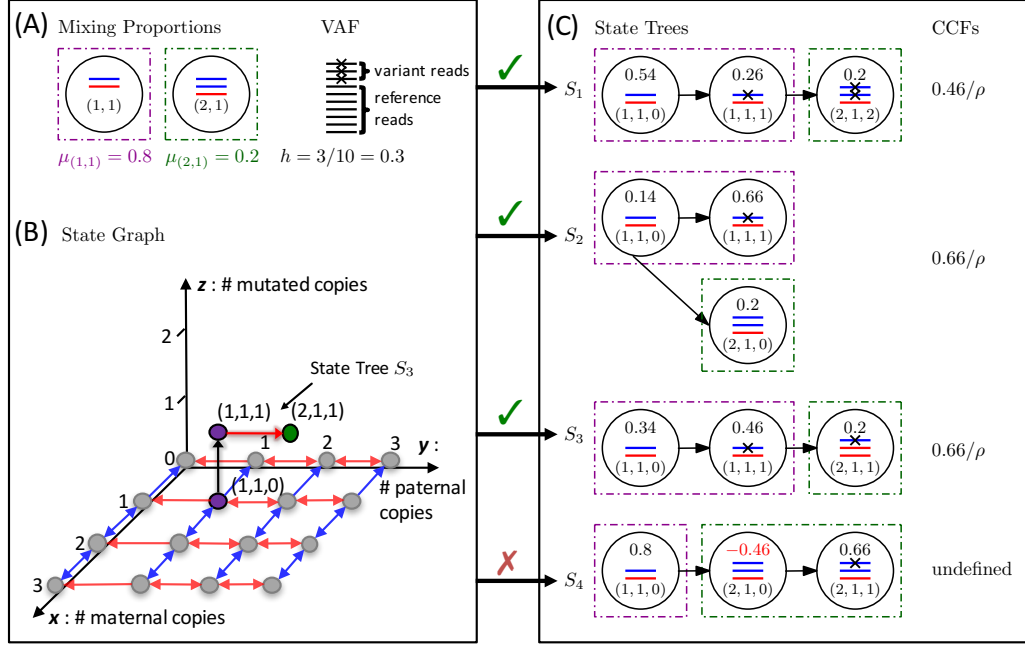


Figure 3.2: **The cancer cell fraction (CCF) of an SNV depends on its copy-number states, variant allele frequency and state tree.** (A) For each character in each sample, we observe copy-number states  $(x, y)$  with mixing proportions  $\mu_{(x,y)}$  as well as a VAF  $h$ . (B) Our multi-state model encodes the somatic mutational process in cancer in the form of a state graph whose vertices  $(x, y, z)$  denote  $x$  maternal copies,  $y$  paternal copies and  $z$  mutated copies of the locus. The edges of the state graph correspond to mutation events, amplifications and deletions. A state tree of a character models the evolutionary history of its mutational states and corresponds to a constrained subtree of the state graph. (C) The observed data of an SNV and the state graph determine a set of compatible state trees, which have non-negative frequencies for each state. Here, states trees  $S_1$ ,  $S_2$  and  $S_3$  are compatible with the input, whereas state tree  $S_4$  is incompatible because it has a negative frequency for state  $(2, 1, 0)$ . The CCF of state trees  $S_2$  and  $S_3$  is  $0.66/\rho$ , where  $\rho$  is the purity of the sample. State tree  $S_1$  has different numbers of copies of the mutation resulting in a distinct CCF of  $0.46/\rho$ .

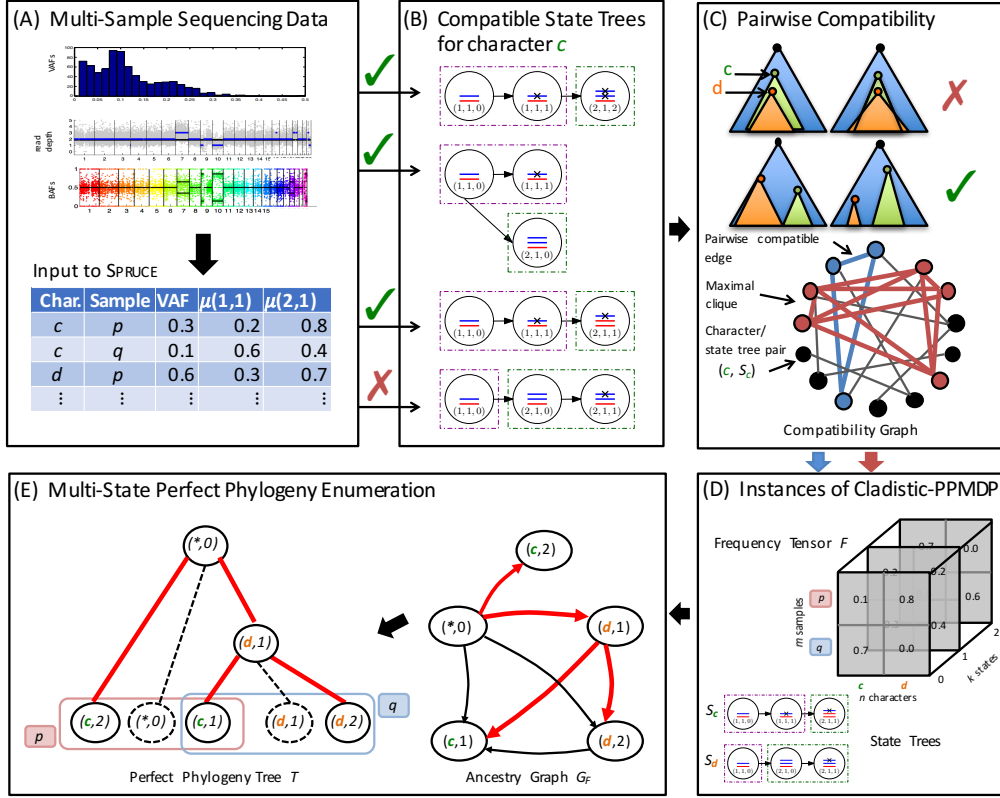


Figure 3.3: **Overview of SPRUCE:** (A) Input is derived from VAFs of SNVs, and copy numbers and mixing proportions determined from analysis of read depth and B-allele frequencies. (B) Combining the input with our multi-state model for the somatic mutational process in cancer results in compatible state trees for every character. (C) Two character-state tree pairs are compatible if there exists a perfect phylogeny tree that contains both. We construct the pairwise compatibility graph by considering all such pairs. (D) A maximal clique in the compatibility graph yields a frequency tensor  $\mathcal{F}$  and collection  $\mathcal{S}$  of state trees that are an input to Cladistic-PPMDP. (E) For each instance, we construct the multi-state ancestry graph  $G_{\mathcal{F}}$ , which encodes potential ancestral relationships between character-state pairs. We then enumerate all multi-state perfect phylogeny trees with maximum size in this graph and obtain their corresponding usage matrices  $U$ .

### 3.2.3 Experimental Evaluation on Simulated and Real Data

We start by running our method SPRUCE on simulated data and analyze its performance, study the effect of violations of the infinite alleles assumption, and compare against PhyloWGS [31]. We then analyze a chronic lymphocytic leukemia tumor from [133] and a prostate cancer tumor from [61].

#### Simulated data

We construct simulated instances by randomly generating a tree on  $n = 5$  characters, from which we obtain  $m \in \{2, 5, 10\}$  samples. This process yields VAFs and copy-number mixing proportions for each character in each sample. For each value of  $m \in \{2, 5, 10\}$  we generate 20 instances with different simulated tree topologies, resulting in a total of 60 instances. In addition to considering error-free VAFs, we introduce noise using sequencing coverage values ranging from 50x to 10,000x. We refer to Supplemental Results B.2.1 for more details on the generation of simulated data.

We run SPRUCE on each instance, in each case recording trees that contain the largest subset of characters. Figure 3.4a shows the number of solutions of the PPMDP, i.e. the number of trees for all 60 instances, as the sequencing coverage is varied. Not surprisingly, we see that increased coverage results in a decrease in the number of solutions, as the uncertainty in the VAFs of each SNV decreases as coverage increases. In fact, a coverage of 10,000x approaches the error-free data. Correspondingly, we see a drop in running time (Figures B.7b). However, we find that a more efficient way to deal with ambiguity (and running time) is to increase the number of samples. For instance, a target coverage of 50x with  $m = 5$  samples has a similar number of solutions as a coverage of 10,000x with  $m = 2$  samples. This demonstrates how the combinatorial constraints (specifically the Multi-State Sum condition (MSSC) below) become stronger with increasing numbers of samples, and constrain the number of possible trees. We note that in the case of error-free VAFs, the true tree is always contained in the solution space (Figure 3.4b shows one such solution space).

Without further prior knowledge about tree topologies, all enumerated trees are equally likely to be

the true tree under a uniform prior. We define the *representative* tree as the tree whose edges occur most frequently across all enumerated trees. Thus, in expectation, the representative tree has the largest set of correct parental relationships given the uniform prior. The *recall* of a reported tree  $T$  is the fraction of edges in the simulated tree  $T^*$  that are recovered in  $T$ . Note that the precision of a tree  $T$ , which is the fraction of edges in the tree  $T$  that occur in  $T^*$ , is a constant multiple of the recall in this case. We thus only report recall values. The recall of the representative tree increases with increasing number of samples and coverage (Figure B.7c) and is higher than the median recall value (Figure B.7d).

Next, we test the effect of the presence of violations of the infinite alleles assumption in the data. In our setting, violations correspond to a copy-number state occurring more than once on the tree. We simulate 20 instances ( $n = 5$ ;  $m = 10$ ; 1,000x) with varying numbers of violations, each violation corresponding to a repeated edge label. Figure B.9 shows that the recall of the representative tree decreases with an increasing number of violations. However, for small numbers of violations, recall remains high. In moderately-rearranged tumors with few or no overlapping copy-number aberrations, the number of violations is likely to be small.

As the number of characters increases, exhaustive enumeration of the space of solutions becomes infeasible. Thus, we analyzed how well we reconstruct the true tree by limiting the running time for each combination of compatible state trees to  $N$  seconds. We considered 20 simulated instances with  $n = 15$  characters,  $m = 10$  samples and a target coverage of 1,000x. Figure 3.4d shows that the recall of the representative tree increases with increasing values of  $N$ , but that the gains are modest. Thus, SPRUCE obtains good solutions within a reasonable amount of computation time.

## Comparison to Existing Methods

There has been a recent explosion of interest in methods for reconstructing phylogenetic trees on SNVs given mixed bulk measurements [38, 68, 74, 92, 113, 122, 144]. Yet, only recently has the attention shifted to modeling both SNVs and CNAs simultaneously. A probabilistic model

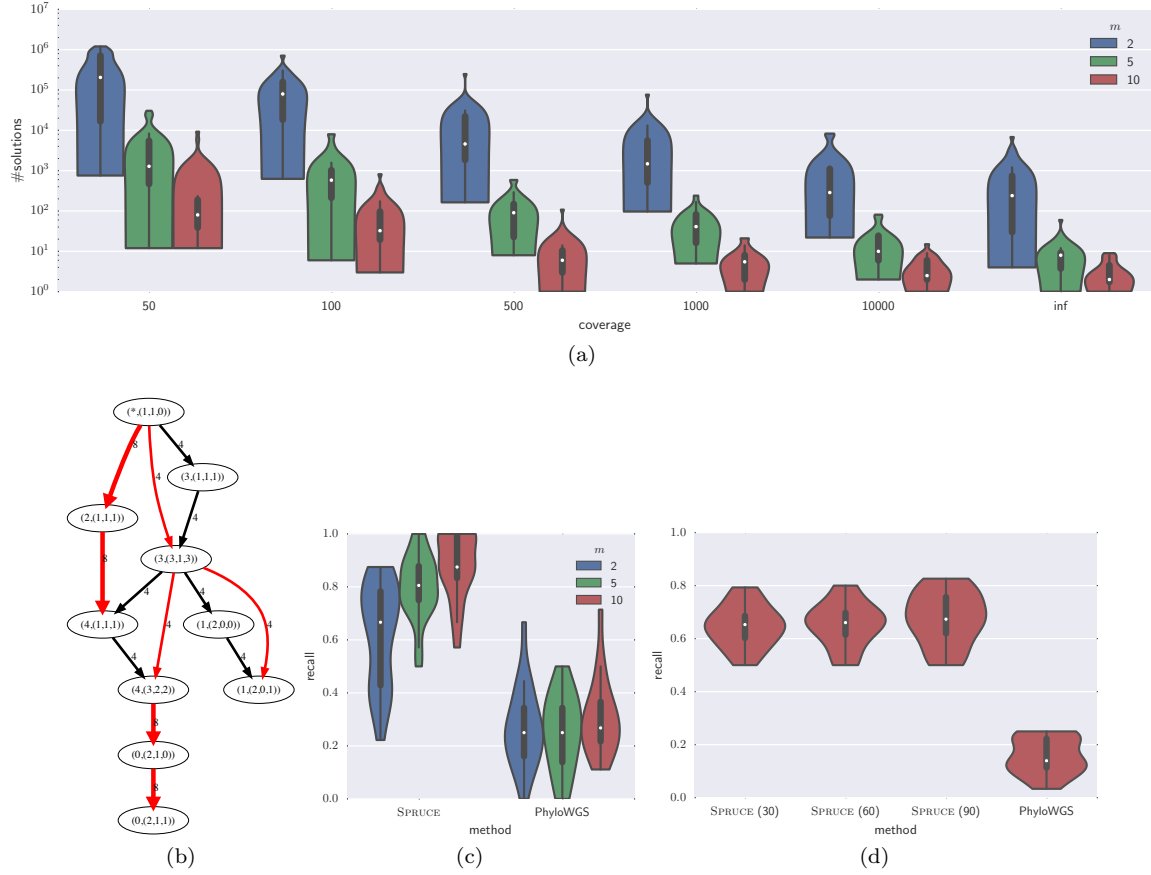


Figure 3.4: **Simulation results.** (A) Number of solutions (log-scale) for instances with  $n = 5$  characters,  $m \in \{2, 5, 10\}$  samples and varying sequencing coverages. A coverage of ‘inf’ corresponds to error-free VAFs. (B) Solution space of an  $n = 5$ ,  $m = 5$  instance with error-free VAFs. Vertices are from the solution trees and each edge is labeled by the number of solutions in which it occurs. Red edges indicate the simulated tree. (C) Recall of SPRUCE’s representative trees compared to PhyloWGS’s maximum likelihood trees on 60 instances with  $n = 5$  and error-free VAFs (shown in Figure B.7a). (D) Recall values on 20 instances with noisy VAFs (1,000x),  $n = 15$  and  $m = 10$ . We run SPRUCE with varying running time limit ( $N$ ) per state tree combination in seconds.

introduced by Li and Li [84] considers SNVs and CNAs to infer the genomic composition of tumor populations but does so without inferring a tree describing their evolutionary relationships. In another line of work, Chowdhury et al. [24] have developed rich phylogenetic models for copy-number aberrations, but these are for single-cell data and do not address the phylogenetic tree mixture problem. In the following, we compare our method against PhyloWGS [31], a method

based on a probabilistic graphical model, which to our knowledge is currently the only method that jointly infers a phylogenetic tree on SNVs and CNAs from bulk-sequencing data.

Unlike our multi-state model, PhyloWGS has a two-state perfect phylogeny model where characters correspond to either CNA events or SNV events. The vertices of the returned trees contain clusters of the two-state characters. To assess their trees, we remap the multi-state characters of the simulated trees to two-state characters as used by PhyloWGS. We use an adjusted measure of recall that accounts for clustering (Supplemental Results B.2.2). We run PhyloWGS with default parameters on the 60 simulated  $n = 5$  instances and 20 simulated  $n = 15$  instances described above and compare the recall of the representative tree found by our method to the maximum likelihood tree reported by PhyloWGS. Our method achieves higher recall across varying number of samples  $m \in \{2, 5, 10\}$  on error-free VAFs with  $n = 5$  characters (Figure 3.4c), and on noisy VAFs with  $n = 15$  characters (Figure 3.4d). We believe that the lower recall values of PhyloWGS, especially those from the larger  $n = 15$  instances, indicate difficulties in MCMC sampling from the complicated space of possible trees. In contrast, because SPRUCE explicitly models the combinatorial structure of the problem, it is able to effectively find good solutions.

## Results on chronic lymphocytic leukemia

We applied SPRUCE to a liquid chronic lymphocytic leukemia tumor (sample CLL077 from [133]) that has previously been analyzed by other phylogenetic tree reconstruction methods [31, 38, 74, 92]. Four time-separated samples (b, c, d, e) were sequenced using targeted and whole-genome sequencing technology. The targeted data includes 14 SNVs, one of which (*SAMHD1*) is classified as an copy-neutral loss of heterozygosity (CN-LOH) in all four samples. Two SNVs (in genes *BCL2CB* and *NAMPTL*) were classified as being unaffected by CNAs, but in some of the samples the lower bound of the VAF confidence interval was greater than 0.5. As such these SNVs were incompatible with all state trees. The twelve remaining characters each had only one compatible state tree associated with them. We run SPRUCE until completion. The entire solution space consisted of 20 trees of 9 vertices (Figure B.10). Figure 3.5a shows one tree from the solution space.

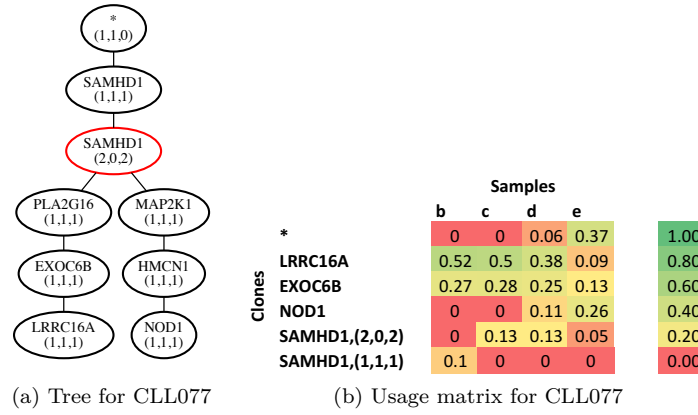


Figure 3.5: **CLL077 results.** (A) The tree computed by SPRUCE for CLL077 with a CN-LOH event in red. (B) Usage matrix for CLL077 shows that the samples (columns) are mixed and consist of many clones (rows) as indicated by the coloring.

A similar tree with two branches is also reported by PhyloSub [74], PhyloWGS [31], CITUP [92] and AncesTree [38] for this dataset. However, the tree reported here and the one reported by AncesTree predict the order of all the mutations on each branch, while PhyloSub, PhyloWGS and CITUP group some mutations together. Additionally, AncesTree did not consider the SNV in gene *SAMHD1*, as its VAF > 0.5. Here, we reconstruct a tree containing the CN-LOH event on *SAMHD1*.

By enumerating the entire search space, we can detect ambiguities in the input data. For instance, in our tree *LRRC16A* is a child of *EXOC6B* whereas there are solutions which assign *LRRC16A* as a child of either *OCA2* or *DAZAP1* (which is absent in the shown tree). Without additional data or further assumptions, there is not enough information to distinguish between these ancestral relationships. In contrast, by only providing one solution, AncesTree and CITUP give an incomplete picture that does not reflect the true uncertainty inherent to the data.

## Results on prostate cancer

Next, we analyzed a solid prostate cancer tumor (sample A22 from [61]) where 11 spatial samples were taken from the primary tumor and different metastases. Ten of these samples underwent

whole-genome sequencing followed by targeted sequencing resulting in a set of 114 SNVs. The authors construct a phylogeny for this sample as follows: First, they assume that all tumor cells have a fixed number of mutated copies of each variant and rescale the VAF of each SNV to a CCF accordingly. Next, they cluster SNVs according to the calculated CCFs using a Dirichlet process. The reported tree (Figure B.11) is obtained by manually applying the Pigeon-Hole Principle [113], a special case of the Sum Condition (described below in Section 3.4.1). This principle requires that the sum of CCFs of the children of each vertex in the tree does not exceed 1. Finally, copy-number aberrations are manually placed on the tree.

We apply SPRUCE to the same dataset. We consider only SNVs on autosomes that have a VAF greater than 0.05 in at least one sample. This results in a set of 89 SNVs. We use the copy-number calls and mixing proportions reported in [61], which were obtained using the Battenberg algorithm [113]. For each SNV in each sample, we obtain 99.9% confidence interval on the VAF using a binomial model. Thirteen characters admit no compatible state tree (Figure B.12). This may be a result of violations of the infinite alleles model or errors in the input data. We then cluster SNVs that are compatible with a single, shared state tree, and whose VAF confidence intervals overlapped in all samples. This yields 68 characters, four of which correspond to four clusters containing a total of 12 SNVs (Supplemental Table 1). The compatibility graph (see Section 3.4.2 for a description) consists of 5183 edges, which encode pairwise compatibility between character and state tree pairs. There are 310,029 maximal cliques, with a maximum size of 30 (Figure B.13). Thus, the largest possible tree we can recover has at most 30 characters. The reason why the maximum cliques do not contain all characters is likely for the same reasons outlined above: errors in the input data or violations of the infinite alleles model.

As described in Section 3.2.2, the cancer cell fraction (CCF) of an SNV depends on the assigned state tree. For each compatible state tree, we compute the CCF interval using the purity values reported in [61]. Figure 3.6A shows that there are characters with multiple compatible state trees that yield distinct CCF intervals. Notably, *PPFIA1* admits two compatible state trees with distinct CCF intervals for four samples (Figure 3.6B). *PPFIA1* is reported as undergoing a single-copy

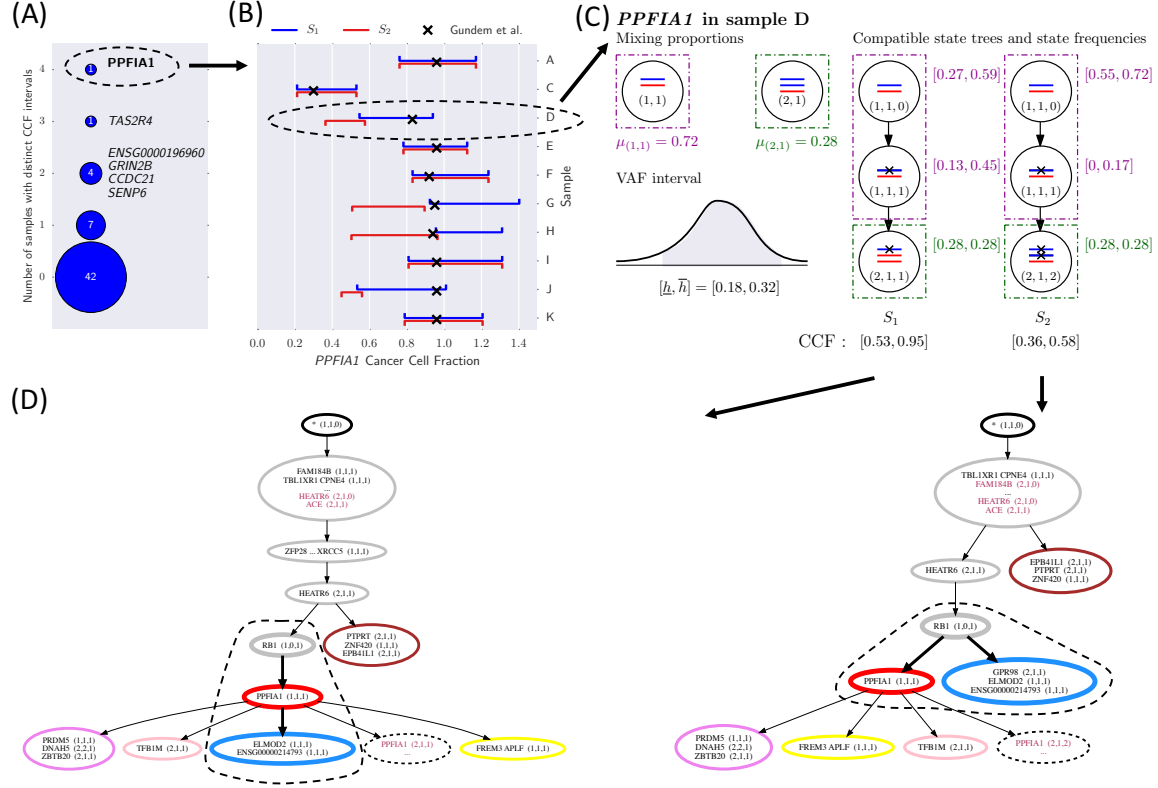


amplification. The two state trees ( $S_1$  and  $S_2$ ) for this character both have the SNV occur prior to the amplification, but in  $S_1$  the non-mutated copy is amplified whereas in  $S_2$  the mutated copy is amplified (Figure 3.6C). Gundem et al. [61] assume that the number of copies of the mutation is constant across all cells that have the mutation, and report a single CCF for this mutation. Since their assumption rules out  $S_2$ , it is not surprising that we observe that their reported CCFs are more consistent with the CCFs resulting from  $S_1$  than those from  $S_2$ . However, we emphasize that there is no signal in the data that requires one to assume that the number of copies of the mutation is constant across all cells that have the mutation.

We run SPRUCE on all maximal cliques that contain either  $S_1$  or  $S_2$  for *PPFIA1*. Figure 3.6D shows two resulting trees: one using state tree  $S_1$  and the other using state tree  $S_2$ . The topology of the two trees is largely similar, except for the location of the cluster containing *ELMOD2* (colored blue). In the first tree, *PPFIA1* (colored red) is the parent of the blue cluster. This is the same relationship reported by the authors (Figure B.11). In contrast, in the second tree *PPFIA1* and the blue cluster are siblings. As the second tree allows for significantly lower CCF values across samples (Figure 3.6B), *RB1* (colored gray) can accommodate both *PPFIA1* and the blue cluster as children. In summary, we have shown that different state trees lead to different CCFs and ultimately different phylogenetic reconstruction of the observed data. There is increasing evidence of the importance of intra-tumor heterogeneity in informing treatment [49] including immunotherapy [101]. Thus, it is essential to have methods to reconstruct phylogenetic trees as accurately as possible from data, and to eliminate unnecessary assumptions about the evolutionary process.

### 3.3 Discussion

We introduce SPRUCE, a method that reconstructs phylogenetic trees using both SNVs and CNAs measured in bulk-sequencing data from multiple samples of a tumor. In contrast to existing approaches, SPRUCE is grounded in an explicit phylogenetic model. Specifically, we formulate the Perfect Phylogeny Mixture Deconvolution Problem (PPMDP) that follows from the no homoplasy,



***PPFIA1* is compatible with two state trees leading to distinct CCF intervals, which in turn lead to two different phylogenetic reconstructions.** (A) For the set of compatible state trees of each character, we show the number of samples in which different state trees lead to distinct CCFs. Sizes of circles correspond to the number of characters. *PPFIA1* has distinct CCF intervals in four samples. (B) *PPFIA1* is compatible with two state trees  $S_1$  and  $S_2$ . We show for each sample the CCF interval for  $S_1$  (blue intervals),  $S_2$  (red intervals), and the cluster CCF (marked  $\times$ ) reported in [61]. The Gudem et al. cluster CCFs are largely consistent with  $S_1$ . (C) We zoom in on sample D. Given the 99.9% VAF confidence interval and the mixing proportions for sample D (purity  $\rho = 0.772$ ), we observe that the two compatible state trees  $S_1$  and  $S_2$  lead to different CCF intervals. Note that  $S_1$  adheres to the Gudem et al. assumption of a fixed number of mutated copies. (D) Condensed representation of the resulting perfect phylogeny trees using  $S_1$  (left) and  $S_2$  (right). Vertices are labeled by the introduced character-state pairs, and colors correspond to clusters reported in [61]. In the left tree the blue cluster is a child of *PPFIA1*, whereas in the right tree it is a sibling of *PPFIA1*. See Figure B.14 for the complete trees.

or infinite alleles, assumption for multi-state characters. We derive a combinatorial characterization of the solutions to this problem, and develop an enumeration algorithm to solve it in the presence of measurement errors. This theoretical framework allows us to evaluate whether a dataset (in whole

or in part) satisfies the phylogenetic model. We applied SPRUCE to both real and simulated cancer sequencing datasets. In all cases, we find that there is substantial ambiguity in bulk-sequencing data, with many alternative phylogenetic reconstructions that are consistent with the data. This observation has several important implications.

First, the ambiguity in phylogenetic reconstructions that we observe in real data suggests caution in deriving strong conclusions about the somatic mutation process for a single solution. On a prostate cancer tumor [61], we find that only a subset of characters are compatible with a perfect phylogeny tree. The observed incompatibilities could be a result of inaccuracies in the input data – including uncertainties in VAFs or incorrect copy-number calls and mixing proportions – or violation of the infinite alleles model. We attempted to account for uncertainties in the VAFs using a binomial model with wide 99.9% confidence intervals. In general, a binomial model underestimates the variance as it does not account for biases introduced in PCR amplification, sequencing, or read mapping. A more tailored error-model that accounts for these biases may lead to better VAF estimates. Copy-number callers make restrictions and assumptions about the number of copy states present in the sample. For example, Battenberg [113] and THetA [116] allow for at most two tumor copy-number states for each interval in each sample. This may be too restrictive of an assumption. In fact, looking at multiple samples from A22 we observe intervals that exhibit up to six copy-number states (e.g. chromosome 8q). Better copy-number callers that allow for more than two copy-number states per interval in a sample, and simultaneously consider all samples would improve our results.

The multiple sources of uncertainty in each step of tumor phylogenetic inference – including SNV calling, copy number segmentation, and estimation of CCFs – strongly argues against treating these steps independently, which is the current practice. One specific example of an issue that can arise in analysis pipelines is the computation of the cancer cell fraction (CCF). This quantity has been used extensively in analysis of intra-tumor heterogeneity, such as the inference of developmental patterns of metastases [13, 14], the timing and fitness effects of cancer driver mutations [5, 99] and the reconstruction of tumor phylogenies [12, 61, 113]. Many of these studies compute CCFs using single values for tumor purity and mutation copy number. However, the presence of multiple

clones (i.e. intra-tumor heterogeneity) implies that the data may not unambiguously identify a single CCF value for an SNV. Instead, one must consider the composition of each tumor sample, as revealed by the phylogenetic tree and usages, to derive the CCF. By jointly deriving CCFs and a phylogeny, SPRUCE is able to report alternative explanations of the input data, explanations that we proved are solutions to the PPMDP. We demonstrated one such example of two trees with distinct topologies that include different CCFs for *PPFIA1* mutations in the prostate cancer dataset from [61]. Extending such ideas further, one could incorporate the determination of copy-number states into the process of phylogenetic tree reconstruction; for example, one could select copy number states to maximize pairwise compatibility between SNVs.

A second implication of the large space of possible phylogenetic trees is the difficulty it may present for probabilistic approaches that sample from this space. We showed that SPRUCE outperforms PhyloWGS [31] on simulated datasets. We attribute this success to the difficulty of sampling the huge solution space directly, as PhyloWGS does, instead of exploiting the combinatorial structure, as SPRUCE does. We envision that the combinatorial characterization of the PPMDP that we derived could be useful for the development of better probabilistic models, which have proved useful both in clustering mutations [113] and in simultaneous clustering and tree inference [8, 31, 74, 160]. For example, rather than considering generic tree-structured priors, one could use priors that are informed by the combinatorial structure of the multi-state ancestry graph.

A third implication of the large number of possible solutions to the PPMDP from bulk-sequencing data is the design of tumor sequencing strategies. In any sequencing approach, one is faced with a tradeoff in cost vs. information gain. In tumor sequencing, two parameters that can be varied are the number of spatial or temporal samples [57, 85, 111] and the sequencing depth per sample [60]. Several papers have noted that a linear phylogeny can often be fit to data from a single tumor sample [38, 68, 74, 92, 113, 144], and that multiple samples are needed to accurately recover branching phylogenies. Our simulations show that increasing the number of samples is far more effective than increasing coverage per sample to reduce the number of possible phylogenies as well as to increase the accuracy of the inferred phylogenies. Indeed, low-coverage data with many samples

can yield orders of magnitude fewer solutions than even perfectly noise-free data with a small number of samples. Multi-sample sequencing has other benefits, including the ability to assess spatial heterogeneity [57, 85, 141] and decrease the likelihood that sample-specific batch effects affect conclusions of studies. While single-cell genome sequencing provides the highest resolution for studies of intratumor heterogeneity [107–109], technical and financial considerations have limited its use in large-scale studies. In this work, we have shown that the combination of multi-sample sequencing and phylogenetic mixture algorithms can provide high-fidelity phylogenetic trees.

Importantly, SPRUCE relies on the infinite alleles, or no homoplasy, assumption, which is a reasonable simplifying assumption, but may be violated in some tumors, particularly those with extensive copy number aberrations. It is an interesting question whether more permissive phylogenetic models such as the maximum parsimony model [51, 52] may more accurately model the evolutionary process. Extending maximum parsimony to the case of phylogenetic mixtures is an intriguing open computational problem. Finally, while we focused here on applications to cancer genome sequencing, the introduced methodology and framework have applications in other cases of mixed samples, including metagenomics [77, 135] and studying the process of somatic hypermutation in the immune system, as explored by [144] using a single-sample, two-state perfect phylogeny model.

In summary, we have derived the first theoretical results for a phylogenetic tree mixture problem under the infinite alleles models, and used these results to obtain new insights into intra-tumor heterogeneity.

### 3.4 Methods

In this section, we describe the combinatorial structure of solutions to the PPMDP and Cladistic-PPMDP by generalizing previous results on the two-state case. We then present our algorithm SPRUCE (Somatic Phylogeny Reconstruction Using Combinatorial Enumeration), which is outlined in Figure 3.3. Further details and proofs are presented in Supplemental Methods.

### 3.4.1 Mathematical Characterization of Solutions to the PPMDP

We first review results for the perfect phylogeny reconstruction problem with unmixed samples. Let  $A \in \{0, \dots, k-1\}^{m \times n}$  be a matrix whose rows correspond to  $m$  taxa and whose columns represent  $n$  characters, each of which has at most  $k$  states. If  $k = 2$  then deciding whether  $A$  admits a perfect phylogeny is solvable in polynomial time [44, 62]. In contrast, if  $k > 2$ , the reconstruction of a multi-state perfect phylogeny is NP-complete in general [10], but is fixed-parameter tractable in the number of states per character [2, 76]. There is an elegant connection between multi-state perfect phylogeny and restricted triangulations of chordal graphs [18], which was recently exploited by Gusfield and collaborators to obtain combinatorial conditions for the multi-state perfect phylogeny [64, 66].

In the case of mixed samples, the perfect phylogeny reconstruction problem becomes more complicated. In a two-state perfect phylogeny a character changes to state 1 only once, and thus the mutated state persists in the tree. Therefore, the subtree  $\bar{T}_{(c,1)}$  rooted at vertex  $v_{(c,1)}$  is equal to the subtree  $T_{(c,1)}$  consisting of all vertices with state 1 for character  $c$  (Figure B.15A). This observation provides strong constraints on the possible ancestral relationships between any pair of characters, given their observed frequencies across all samples. Previously, [38, 122] showed that the phylogenetic trees that produce the observed frequencies correspond to constrained spanning trees of a certain graph. In [38], we defined the *ancestry graph*, which has a vertex for every character-state pair  $(c, 1)$  and its edges relate two character-state pairs that are in a potential ancestral relationship as captured by the *Ancestry Condition* (AC):

$$f_{p,(c,1)} \geq f_{p,(d,1)} \quad \text{for all samples } p \text{ and pairs of characters satisfying } (c, 1) \prec_T (d, 1). \quad (\text{AC})$$

We proved in [38] that solutions to the two-state problem with mixed samples correspond to constrained spanning trees of the ancestry graph that satisfy the *Sum Condition* (SC):

$$f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \quad \text{for all samples } p \text{ and characters } c, \quad (\text{SC})$$

where  $\delta(c, i)$  denotes the children of  $v_{(c,i)}$ . Note that this condition is a generalization of the Pigeon-Hole-Principle [113] and appeared in other forms elsewhere [74, 92, 122, 144].

In a multi-state perfect phylogeny the state of a character  $c$  can change more than once on the tree, but never changes back to a previous state. Thus, in general  $T_{(c,i)}$  is not equal to  $\bar{T}_{(c,i)}$  as a character can change state again in  $\bar{T}_{(c,i)}$  (Figure B.15B). This makes the situation much more complicated, since we must consider not only the children of  $v_{(c,i)}$ , but also the relationships between  $T_{(c,i)}$  and subtrees  $T_{(c,j)}$  for  $j \neq i$ . We capture these relationships using the notion of a descendant set for each character-state pair  $(c, i)$ . We write  $(c, i) \prec_T (d, j)$  if  $v_{(c,i)}$  is ancestral to  $v_{(d,j)}$  in  $T$ . Note that  $\prec_T$  is a reflexive relation, i.e. a vertex  $v_{(c,i)}$  is ancestral to itself. Formally, given a complete perfect phylogeny tree  $T$ , we define the *descendant set*  $D_{(c,i)} = \{j \mid (c, i) \prec_T (c, j)\}$  as the set of states for character  $c$  that are descendants of character-state pair  $(c, i)$  in  $T$ . The descendant set of a character precisely determines the relationship between  $\bar{T}_{(c,i)}$  and  $T_{(c,i)}$ ; namely we have  $\bar{T}_{(c,i)} = \bigcup_{l \in D_{(c,i)}} T_{(c,l)}$  (Figure B.15B). Hence, to obtain the usages of all vertices in  $\bar{T}_{(c,i)}$  we must consider *cumulative frequencies*  $f_p^+(D_{(c,i)}) = \sum_{l \in D_{(c,i)}} f_{p,(c,l)}$ , i.e. the sum of the frequencies of all descendant states of the character-state pair. It turns out that the positivity of values  $u_{p,(c,i)}$  is a necessary and sufficient condition for  $T$  to generate  $\mathcal{F}$  as captured by the *Multi-State Sum Condition* (MSSC) in the following theorem.

**Theorem 1.** *A complete perfect phylogeny tree  $T$  generates  $\mathcal{F}$  if and only if*

$$f_p^+(D_{(c,i)}) \geq \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) \quad \text{for all character-state pairs } (c, i) \text{ and samples } p. \quad (\text{MSSC})$$

Note that (MSSC) is a generalization of (SC). In the two-state case,  $D_{(c,1)} = \{1\}$  for all characters

$c$ , and thus the cumulative frequencies  $f_p^+(D_{(c,1)})$  are equal to the input frequencies  $f_{p,(c,1)}$ . Thus, in the two-state case, the relative order of the frequencies of the mutated ( $=1$ ) states of characters constrain the ancestral relationships between characters, as described above. In contrast, in the multi-state case, we must consider the relative order of cumulative frequencies  $f_p^+(D_{(c,i)})$ , but these depend on descendant sets which are *a priori* unknown. Thus, we must consider all combinations of descendant sets  $D_{(c,i)}$  and  $D_{(d,j)}$  that satisfy the *Multi-State Ancestry Condition* (MSAC).

**Proposition 1.** *Let  $T$  be a complete perfect phylogeny tree that generates  $\mathcal{F}$ . If  $(c, i) \prec_T (d, j)$  then there exist a valid descendant set pair  $(D_{(c,i)}, D_{(d,j)})$  such that*

$$f_p^+(D_{(c,i)}) \geq f_p^+(D_{(d,j)}) \quad \text{for all } p, (c, i), (d, j) \text{ and descendant sets } D_{(c,i)}, D_{(d,j)}. \quad (\text{MSAC})$$

Note that (MSAC) is a generalization of (AC). For two character-state pairs there are potentially many descendant set pairs satisfying (MSAC). As such, the *multi-state ancestry graph*  $G_{\mathcal{F}}$  is a multi-graph whose vertices correspond to character-state pairs and whose multi-edges are labeled by pairs of descendant sets that respect (MSAC). In the Cladistic-PPMDP, the descendant sets  $D_{(c,i)}$  are determined by the input set  $\mathcal{S}$  of state trees. Thus, the *cladistic multi-state ancestry graph*  $G_{(\mathcal{F}, \mathcal{S})}$  is a simple graph, similarly to the  $k = 2$  case. Solutions to a PPMDP instance  $\mathcal{F}$  correspond to *threaded* spanning trees of the multi-state ancestry graph  $G_{\mathcal{F}}$  that satisfy (MSSC) and that respect the descendant sets of the edges, as shown in the following theorem.

**Theorem 2.** *A complete perfect phylogeny tree  $T$  generates  $\mathcal{F}$  if and only if  $T$  is a threaded spanning tree of the ancestry graph  $G_{\mathcal{F}}$  such that (MSSC) holds.*

Similarly, we show in Supplemental Proposition 2 that solutions to a Cladistic-PPMDP instance  $(\mathcal{F}, \mathcal{S})$  correspond to threaded spanning trees of the cladistic multi-state ancestry graph  $G_{(\mathcal{F}, \mathcal{S})}$ .

We have previously shown a hardness result for the PPMDP in the case with  $k = 2$  states and  $m = O(n)$  samples [38]. In Supplemental Methods B.1.4, we prove a stronger hardness result with  $m = 2$  samples.



### 3.4.2 SPRUCE Algorithm for the Cladistic-PPMDP

The SPRUCE algorithm (Figure 3.3) takes as input VAF confidence intervals and copy-number mixing proportions for each character per sample, and outputs multi-state perfect phylogeny trees. The algorithm consists of the following steps: (1) Computing a collection of compatible state trees  $\{S_c\}$  for each character  $c$ . (2) Deriving a compatibility graph whose edges are pairs  $(S_c, S_d)$  of pairwise compatible state trees for pairs  $(c, d)$  of characters. (3) Computing the set of maximal cliques in the compatibility graph. Each maximal clique corresponds to an instance  $(\mathcal{F}, \mathcal{S})$  of the Cladistic-PPMDP where  $\mathcal{S}$  is the set of state trees encoded by the vertices of the clique and  $\mathcal{F}$  is the corresponding frequency tensor. (4) Solving the instance by enumerating all multi-state perfect phylogeny trees on the largest subset of characters. We briefly describe these steps.

We described step (1) in Section 3.2.2. In order to obtain instances  $(\mathcal{F}, \mathcal{S})$  of the Cladistic-PPMDP, one could naively consider all combinations of state tree to character assignments. Instead in step (2), we restrict the combinations that we need to consider by introducing the compatibility graph. We denote by  $\mathcal{F}[X]$  the subtensor of  $\mathcal{F}$  that is restricted to characters  $X \subseteq [n]$ . A set of state trees on characters  $X \subseteq [n]$  is *compatible* if there exists a tree that generates  $\mathcal{F}[X]$ . Now consider an instance  $(\mathcal{F}, \mathcal{S})$  and let  $T$  be a tree that generates  $\mathcal{F}$  and is consistent with  $\mathcal{S}$ . Then, for any pair of characters  $c, d$ , it holds that there exists a smaller tree that generates  $\mathcal{F}[\{c, d\}]$  and is consistent with  $S_c$  and  $S_d$ . We say that  $S_c$  and  $S_d$  are *pairwise compatible*. We restrict the combinations we consider to those that consist of pairwise compatible state trees. More precisely, we define a *compatibility graph* with edges between two vertices  $(c, S_c)$  and  $(d, S_d)$  if and only if  $S_c$  and  $S_d$  are pairwise compatible.

In step (3), we use the compatibility graph to obtain instances of the Cladistic-PPMDP. Following the discussion above, each instance  $(\mathcal{F}, \mathcal{S})$  that admits a consistent tree  $T$  and generates  $\mathcal{F}$  must be composed of a set of pairwise compatible state tree and character pairs. Such sets correspond to cliques in the compatibility graph. In our application, we are not guaranteed to find trees that contain all the characters. Instead, we aim to find trees on the largest subset of characters that

form a maximal clique in the compatibility graph. We enumerate all maximal cliques using the Bron-Kerbosch algorithm [15, 43].

Finally, for step (4), we describe an algorithm to find all threaded spanning trees that satisfy (MSSC) in  $G_{(\mathcal{F}, \mathcal{S})}$  by adapting the Gabow-Myers algorithm [53] for enumerating spanning trees. We extend the approach used in [122] for the two-state problem with noisy frequencies. The crucial observation is that any subtree of a solution  $T$  must also be a consistent, threaded tree and satisfy (MSSC). Here, a subtree  $T'$  is *consistent* if it is rooted at  $v_{(*,0)}$ , and for each character  $c$  the set of states  $\{i \mid v_{(c,i)} \in V(T')\}$  induces a connected subtree in  $S_c$ . Since the Gabow-Myers algorithm constructively grows spanning trees, we can arrive at the desired threaded spanning trees by maintaining the following invariant: The partial tree  $T'$  must be consistent with  $\mathcal{S}$ , and (MSSC) must hold for each vertex  $v_{(c,i)} \in V(T')$ . In addition, we maintain a subset of edges  $H \subseteq E(G_{(\mathcal{F}, \mathcal{S})})$  called the *frontier* that can be used to extend the partial tree  $T'$ . More precisely, the second invariant requires that any edge  $(v_{(c,i)}, v_{(d,j)}) \in H$  can be used to extend  $T'$  without introducing a cycle or violating the first invariant. Supplemental Algorithm 1 maintains the two invariants and reports all solutions of a Cladistic-PPMDP instance  $(\mathcal{F}, \mathcal{S})$ . The running time is the same as the original Gabow-Myers algorithm:  $O(|V| + |E| + |E| \cdot K)$  where  $K$  is the number of spanning trees in  $G_{(\mathcal{F}, \mathcal{S})} = (V, E)$  (disregarding (MSSC)). Further details describing the enumeration algorithm as well as an extension to noisy frequencies are in Supplemental Methods B.1.6.

SPRUCE is available for download at <http://compbio.cs.brown.edu/projects/spruce>.

## Chapter 4

# Single-cell Tumor Phylogeny Inference with Supported Mutation Losses

### 4.1 Introduction

Cancer arises from an evolutionary process during which somatic mutations accumulate in a population of cells. Different cells within a tumor acquire distinct complements of somatic mutations, resulting in a heterogeneous tumor. Quantifying this intra-tumor heterogeneity and reconstructing the evolutionary history of a tumor is crucial for diagnosis and treatment of cancer [19, 146]. The evolution of a tumor is typically described by a *phylogenetic tree*, or *phylogeny*, whose leaves represent the cells observed at the present time and whose internal nodes represent ancestral cells. Tumor phylogenies are challenging to reconstruct using DNA sequencing data from bulk tumor samples, since this data contains mixtures of mutations from thousands–millions of heterogeneous cells in the

sample [4, 32, 39, 40, 73, 75, 93, 105, 106, 123, 125, 132, 161]. Recently, single-cell DNA sequencing (scDNA-seq) of tumors has become more common, and new technologies such as those from 10X Genomics [1], Mission Bio [104], and others [56, 110, 165] are improving the efficiency and lowering the costs of isolating, labeling, and sequencing individual cells. While scDNA-seq overcomes the difficulties of phylogeny reconstruction from bulk samples, it introduces a new challenge of higher rates of missing data and errors due to DNA amplification errors, undersampling, and sequencing errors [56].

Early work in phylogeny inference from scDNA-seq data uses single-nucleotide variants (SNVs) as phylogenetic markers. A particular challenge for SNV-based analysis is high rates (up to 30% for high-depth scDNA-seq [56]) of *allele dropout* errors, where only one of two alleles is observed at a heterozygous site. Methods address this challenge by using an evolutionary model to infer a phylogeny while simultaneously imputing missing data and correcting errors in the observed SNVs. Algorithms such as SCITE [71], OncoNEM [129], SciPhi [138], and B-SCITE [94] use the simplest phylogenetic model for SNVs, the *infinite sites model*. In this model, a locus in a cell has one of two states: an SNV (or mutation) is either present at the locus (state 1) or absent (state 0). Transitions between states are constrained in the phylogeny such that each mutation is gained ( $0 \rightarrow 1$ ) at most once during evolution, and never subsequently lost ( $1 \rightarrow 0$ ). A phylogeny that respects the infinite sites model is known as a *perfect phylogeny* and the state of mutations in the leaves of the phylogeny is summarized by a *mutation matrix* whose binary entries indicate the presence (state 1) or absence (state 0) of every mutation in each observed cell (Figure 4.1(A)). On error-free data, the perfect phylogeny is unique [63]. However, on typical scDNA-seq data, errors in the mutation matrix must be corrected to yield a perfect phylogeny model. Because many such corrections are possible, multiple phylogenies are typically equally consistent with the data (Figure 4.1(B)).

An additional challenge in inferring phylogenies from cancer sequencing data is that somatic mutations in tumors occur across all genomic scales from SNVs to copy-number aberrations (CNAs), which amplify or delete larger genomic regions. CNAs may overlap SNVs and affect the state of SNVs in cells; e.g., a deletion that overlaps an SNV may result in a *mutation loss* ( $1 \rightarrow 0$ ). The

infinite sites model does not allow mutation losses and therefore may yield incorrect phylogenies when applied to SNVs in regions containing CNAs. One solution is to exclude regions containing CNAs and build phylogenies from SNVs in diploid, or copy-neutral, regions. However,  $\approx 90\%$  of solid tumors are highly aneuploid [147], containing extensive CNAs, and  $\approx 30\%$  of solid tumors have whole-genome duplications [9]. Identifying collections of SNVs with no possibility of overlapping CNAs during evolution of such tumors may be challenging.

Recently, several methods [27, 37, 95, 102, 163, 164] have been introduced for single-cell phylogeny inference that allow loss of mutations. SPHYR [37], SASC [27], and PyDollo [102] use the *Dollo model* [33], which relaxes the infinite sites model. In the Dollo model, a mutation may be gained ( $0 \rightarrow 1$ ) at most once, but may be lost ( $1 \rightarrow 0$ ) multiple times. SiFit [163], SiCloneFit [164], and PhiSCS [95] use the finite sites model, a further relaxation that allows mutation to be gained more than once. A challenge in using these less stringent evolutionary models is that they increase the ambiguity in phylogenetic reconstruction (Figure 4.1(C)). Even in simple cases with no error, multiple phylogenies are consistent with the data and the number of phylogenies further increases when there are errors and uncertainty in the mutation matrix. Both the errors in scDNA-seq data and the mutation losses in the phylogeny conspire to yield considerable challenges and ambiguity in the single-cell phylogeny inference problem. This ambiguity is further amplified because both sequencing errors and losses result in the same signal in the observed data: an observed ‘0’ in the mutation matrix instead of a ‘1’. Thus, it is particularly difficult to distinguish between errors in the data and potential mutation losses.

A major limitation in using the Dollo or finite sites models to allow mutation losses is that neither of these models consider evidence from CNAs that support or refute a mutation loss at a locus. While more general multi-state models of tumor evolution have been used to infer phylogenies from bulk tumor sequencing data [32, 40, 73], these approaches neither model the errors in scDNA-seq data nor scale to hundreds–thousands of observed cells. Since mutation losses are the major complication in SNV evolution and responsible for most of the violations of the infinite sites model in scDNA-seq data [79, 102], the full generality of a multi-state model may not be necessary to obtain accurate

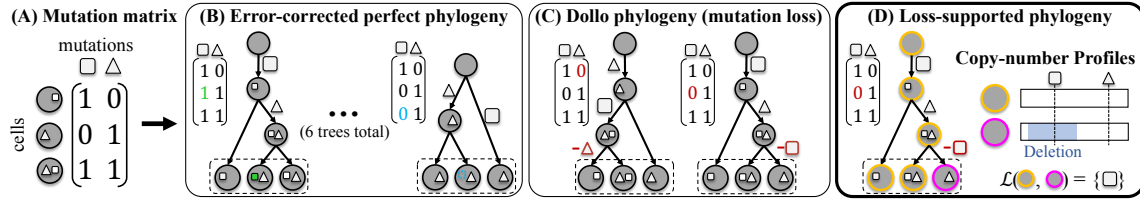


Figure 4.1: **Loss-supported model reduces ambiguity in inference of phylogeny from single-cell sequencing data.** (A) A mutation matrix with two mutations in three cells does not admit a perfect phylogeny. This may be due to either errors or mutation losses. (B) Under the infinite sites model, existing methods correct errors in the observed matrix to yield a perfect phylogeny. (C) Under the Dollo model, existing methods identify mutation losses to explain violations of the infinite sites model. Both the infinite sites and Dollo models yield multiple equally-plausible phylogenies. (D) The loss-supported model overcomes this ambiguity by using copy-number data to constrain mutation losses.

phylogenies from scDNA-seq data. Rather, we describe an approach that constrains mutation losses by using copy-number data from the same cells.

We introduce SCARLET (Single-Cell Algorithm for Reconstructing the Loss-supported Evolution of Tumors), an algorithm that infers phylogenies from scDNA-seq data by integrating SNVs and copy-number data. SCARLET is based a novel evolutionary model, the *loss-supported phylogeny*, that constrains mutation losses to loci where the copy-number data has evidence of a deletion (Figure 4.1(D)). The loss-supported phylogeny generalizes the infinite sites and Dollo models. SCARLET also relies on a probabilistic model of the read counts for each SNV to address errors and missing data that are common in scDNA-seq. On simulated data, we show that SCARLET infers more accurate phylogenies compared to existing methods. We then use SCARLET to analyze scDNA-seq data from a metastatic colorectal cancer patient from Leung et al. (2017) [83]. We find that the published phylogeny – constructed from SNVs under the infinite sites model – has the implausible conclusion that genome-wide copy-number profiles evolved twice independently during the evolution of this tumor. In contrast, SCARLET infers a loss-supported phylogeny that has three mutation losses, with each loss supported by a copy-number change at the locus. Moreover, the SCARLET phylogeny supports the hypothesis of a single migration between the colon primary tumor and liver metastasis (*monoclonal seeding*). In contrast, previous published phylogenies [83, 164] reported a more complex origin of the metastasis with multiple migrations (*polyclonal seeding*). By integrating

information from both SNVs and CNAs, SCARLET obtains more accurate reconstructions of tumor evolution at single-cell resolution.

## 4.2 Results

### 4.2.1 SCARLET algorithm for Loss-supported Phylogeny Model

We developed a new algorithm, SCARLET (Single-Cell Algorithm for Reconstructing the Loss-supported Evolution of Tumors) to infer phylogenetic trees from single-cell DNA sequencing (scDNA-seq) data by integrating data from both single-nucleotide variants (SNVs) and copy-number aberrations (CNAs). SCARLET has three important features (Figure 4.2): (1) a novel evolutionary model, the *loss-supported phylogeny*, which constrains mutation losses to loci where there is a corresponding decrease in copy number; (2) an algorithm to compute a loss-supported phylogeny by *refinement* of a coarse phylogenetic tree derived from copy-number data alone; (3) maximum-likelihood inference of SNVs using a probabilistic model of observed read counts in scDNA-seq data. We describe each of these key features below.

The loss-supported model is a model of SNV evolution where mutation gains ( $0 \rightarrow 1$ ) occur at most once, but mutation losses ( $1 \rightarrow 0$ ) are constrained by sets  $\mathcal{L}$  of *supported losses* that are defined by CNAs in the same cells (Figure 4.1(D)). Specifically, we assume that for each cell we measure both a mutation profile  $\mathbf{b}$  of SNVs and a *copy-number profile*  $c$ . For each pair  $(c, c')$  of copy-number profiles, we define the *supported loss set*  $\mathcal{L}(c, c')$  as the set of SNVs at loci where there is a *decrease* in copy number (e.g., due to a deletion or loss-of-heterozygosity (LOH) event) between profiles  $c$  and  $c'$ . In the loss-supported phylogeny, a mutation loss at an SNV loci  $a$  is allowed between cells  $v$  and  $w$  only if  $a$  is in  $\mathcal{L}(c_v, c_w)$ . The loss-supported model can thus be viewed as a generalization of other models for SNV evolution: the perfect phylogeny model is the special case where  $\mathcal{L} = \emptyset$ , while the Dollo model and finite sites model corresponds to  $\mathcal{L}$  being the complete set of all mutations. In contrast to these extremes, the loss-supported model allows for intermediate values of  $\mathcal{L}$  derived

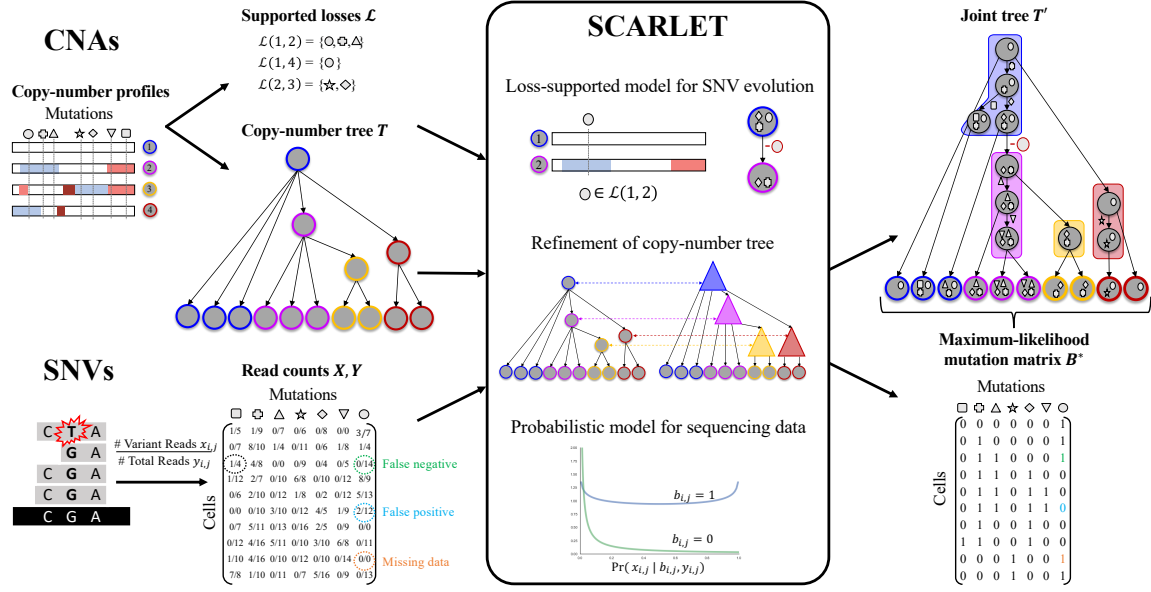


Figure 4.2: **Scarlet algorithm for the Maximum Likelihood Loss-Supported Refinement problem.** SCARLET integrates single-nucleotide variants (SNVs) and copy-number aberrations (CNAs) for tumor phylogeny inference. For CNAs, observed copy-number profiles indicate amplified (red) or deleted (blue) genomic regions along the entire genome and are used to obtain two inputs for SCARLET. First, supported loss sets  $\mathcal{L}(c, c')$  for pairs of copy-number profiles (empty sets are not shown) indicate mutations that are affected by deletions. Second, a copy-number tree  $T$  which describes the ancestral relationships between observed cells (leaves) as determined by copy-number profiles. For SNVs, variant  $\mathbf{X}$  and total  $\mathbf{Y}$  read counts are provided to SCARLET for every cell and every mutation. SCARLET computes a joint tree  $T'$  on the observed cells and a maximum-likelihood mutation matrix  $B^*$  by constraining mutation losses to the supported loss sets  $\mathcal{L}$ , computing a refinement  $T'$  of  $T$ , and selecting the maximum-likelihood  $B^*$  using a probabilistic model for the presence ( $b_{i,j} = 1$ ) or absence ( $b_{i,j} = 0$ ) of each SNV in each cell.

from copy-number data.

The loss-supported model depends on the copy-number profiles of both the observed and ancestral cells. However, we do not directly measure the copy-number profiles of the ancestral cells. To overcome this limitation, SCARLET uses a *copy-number tree*  $T$  which is derived from the copy-number profiles of the observed cells [25, 41, 134, 161] (Figure 4.2). SCARLET computes the supported loss sets  $\mathcal{L}$  from the copy-number profiles of the observed cells (leaves of  $T$ ) and the copy-number profiles of the ancestral cells (internal vertices of  $T$ ). Typically, scDNA-seq data of SNVs (e.g., from targeted sequencing) measures copy-number profiles with low resolution, and thus tumor cells



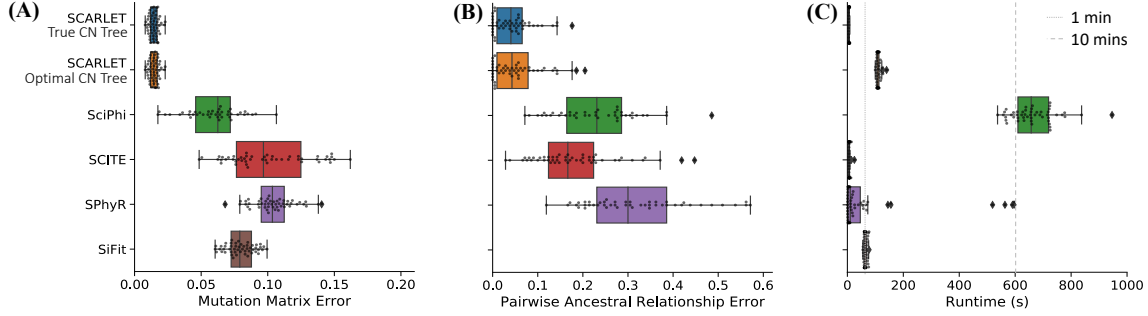


Figure 4.3: **Scarlet outperforms existing methods for phylogeny inference on simulated single-cell data.** (A) Mutation matrix error, (B) Pairwise ancestral relationship error, and (C) Runtime for each method. SCARLET was run either knowing (‘True CN Tree’) or not knowing (‘Optimal CN Tree’) the true copy number tree.

share a limited number of distinct copy-number profiles. Consequently, the copy-number tree  $T$  has many multifurcations, or unresolved ancestral vertices with more than two children. SCARLET finds a joint tree  $T'$  that is a loss-supported phylogeny and a *refinement*[153] of  $T$  by resolving multifurcations in  $T$  using the mutation profiles of the observed cells (Figure 4.2).

Data from scDNA-seq typically has high error rates in identifying SNVs, and particularly high rates of false negatives and missing data due to amplification bias and allele dropout [56]. SCARLET models these errors using a beta-binominal distribution [138] of the observed read counts. As such, SCARLET computes the loss-supported refinement  $T'$  that maximizes the likelihood of the observed sequencing data under this probabilistic model (Figure 4.2).

#### 4.2.2 Simulated Data

We compared SCARLET to four existing algorithms that build phylogenies from single-cell sequencing data, SCITE [71], SciPhi [138], SPhyR [37], and SiFit [163], on simulated data. We simulated 50 trees, each with 20 mutations, 4 copy-number profiles, and 1-8 mutation losses per tree. From these trees, we simulated 100 observed cells with each cell having equal probability of being a child of any vertex in the simulated tree, and simulated sequencing data with an expected sequencing depth of  $100\times$  and allelic dropout rate of 0.15. Additional details of simulated data and parameters

of each method are in Section SC.3.4.

We evaluated the phylogenies output by the methods by two measures that have been previously used in tumor evolution studies [27, 37, 59, 106, 132, 138]. First, the *mutation matrix error*  $M(\mathbf{B}, \hat{\mathbf{B}}) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n |b_{i,j} - \hat{b}_{i,j}|$  is the normalized Hamming distance between the inferred binary mutation matrix  $\hat{\mathbf{B}}$  and the true binary mutation matrix  $\mathbf{B}$  and assesses the accuracy of the error-corrected mutation profiles for each observed cell. Second, the *pairwise ancestral relationship error*  $E(T, \hat{T})$  is the proportion of pairwise ancestral relationships between mutations in the inferred tree  $\hat{T}$  that differ from the ancestral relationships in the true tree  $T$ . Specifically, every pair  $a, a'$  of mutations has one of four possible ancestral relationships in  $\hat{T}$  and in  $T$ : (1)  $a$  and  $a'$  occur on the same branch; (2)  $a$  is ancestral to  $a'$ ; (3)  $a'$  is ancestral to  $a$ ; (4)  $a$  and  $a'$  are incomparable. Note that we do not calculate the tree error for SiFit because it uses a finite sites model which allows mutations to recur and consequently pairs of mutations may not have a unique relationship.

SCARLET outperforms all other methods on both mutation matrix error and ancestral relationship error (Figure 4.3(A)-(B)). The high errors of SCITE and SciPhi were expected since these methods use an infinite sites model while the simulations include mutation losses which violates the model assumptions. However, the methods that do allow mutation losses, SPhyR (based on the  $k$ -Dollo model) and SiFit (based on the finite sites model), do not exhibit improvement over the other methods and perform worse than SCARLET. These results confirm that models that include unconstrained mutation losses have significant ambiguity as it is difficult to distinguish between true mutation losses and false positives/negatives in the data (Figure 4.1). By using copy-number information to constrain mutation losses, SCARLET overcomes the ambiguity in phylogeny reconstruction and obtains lower error in the inferred mutation matrix and phylogeny.

We evaluated the effect of the input copy number tree on SCARLET's accuracy by running SCARLET in two modes: when the true copy-number tree is either known ('SCARLET True CN Tree') or unknown ('SCARLET Optimal CN Tree'). In this latter case, we enumerated all copy-number trees, ran SCARLET once for each copy-number tree, and output the solution with the highest likelihood. In both cases, we provided SCARLET with the true copy-number profiles of each cell and the true set

$\mathcal{L}$  of supported losses. SCARLET exhibited comparable performance when running with or without knowledge of the copy-number tree (Figure 4.3(A)-(B)). Notably, in 46/50 simulated instances, the maximum likelihood solution obtained when running SCARLET with unknown copy-number tree was identical to the solution found when providing the true copy-number tree. Clearly, running SCARLET with all possible copy-number trees (16 copy-number trees in this simulation) increases the runtime (Figure 4.3(C)), but the runtime remains reasonable when the number of copy-number profiles is small, which is the case for many real datasets (see below).

### 4.2.3 Single-cell phylogeny of metastatic colorectal cancer

We used SCARLET to analyze single-cell DNA sequencing of a metastatic colorectal cancer patient CRC2 from Leung et al [83]. This data set included targeted sequencing of 1000 genes in 141 cells from a primary colon tumor and 45 cells from a matched liver metastasis (Figure C.1(A)). The authors identified 36 single-nucleotide variants (SNVs) and used SCITE [71] to derive a perfect phylogeny from these SNVs (Figure 4.4(A)). This perfect phylogeny tree shows two distinct branches of metastatic cells, and Leung et al. [83] concluded that this was evidence of *polyclonal seeding* of the liver metastasis; i.e., two distinct cells (or groups of cells) with different complements of mutations migrated from the primary colon tumor to the liver metastasis. Examining the copy-number data, one finds a curious discrepancy between the SCITE tree and the single-cell copy-number profiles. Whole-genome sequencing of 42 single cells from the same patient reveals that all metastatic cells share losses of chromosomes 2, 3p, 4, 7, 9, 16, 22 relative to the cells in the primary tumor (Figure 4.4(B)). According to the SCITE tree, all of these large copy-number aberrations (CNAs) would had to have occurred *twice independently* in the two distinct branches of metastatic cells. Although CNAs can exhibit homoplasy, this high rate of occurrence of the exact same events seems highly unlikely. Thus, we observe an inconsistency between the copy-number data and the SCITE tree constructed using only SNV data. Notably, this same dataset was recently analyzed by SiCloneFit [164] using a finite sites model. The SiCloneFit tree also showed two branches of metastatic cells and concluded that there was polyclonal seeding of the metastases. Thus, the

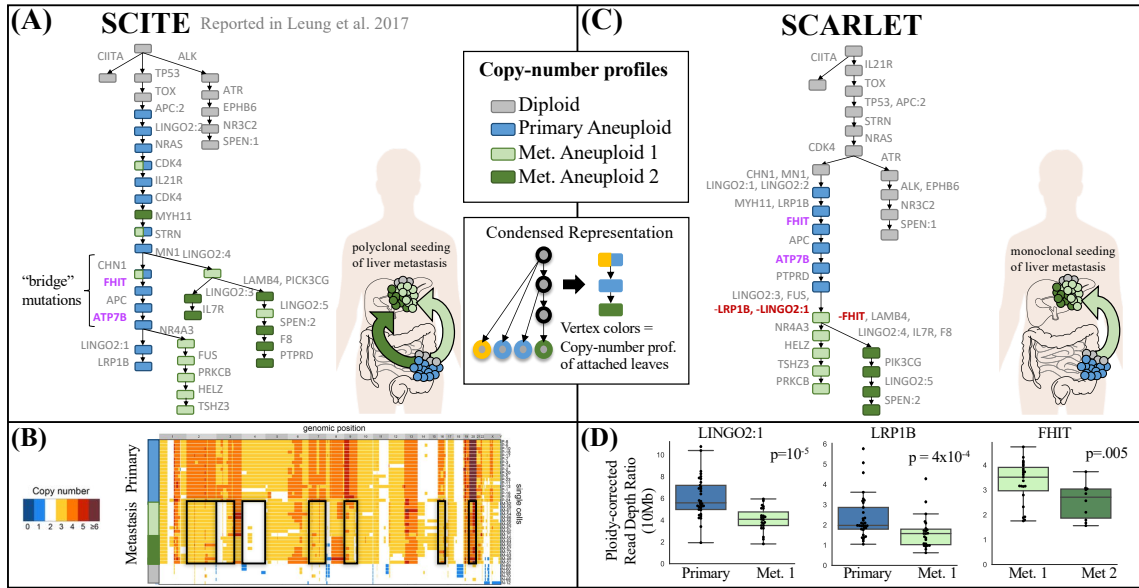


Figure 4.4: **Scarlet infers a loss-supported phylogeny consistent with copy-number profiles from a metastatic colorectal cancer patient.** (A) Published perfect phylogeny tree [83] of 141 single cells from the primary colon tumor (blue) and 45 single-cells from the liver metastasis (green) of patient CRC2. Two distinct branches of metastatic cells — suggesting polyclonal seeding of the liver metastasis — are separated by four “bridge mutations” occurring in cells of the primary tumor. (B) Published copy-number profiles from DOP-PCR whole-genome sequencing of 42 single cells from both the primary tumor and metastasis of CRC2 (figure adapted from Leung et al. [83]). All metastatic cells share deletions of six chromosomes (black boxes), but are separated into two groups (light and dark green) by a small number of additional CNAs. (C) SCARLET infers a loss-supported phylogeny from the same data. The SCARLET tree has a single branch containing all metastatic cells — suggesting monoclonal seeding of the liver metastasis and consistent with the similar copy-number profiles of all metastatic cells. SCARLET identifies mutation losses (red) in *LINGO2*, *LRP1B*, and *FHIT*. (D) Significant decreases in read depths are observed at the loci of the three mutation losses identified by SCARLET.

SiCloneFit phylogeny also has the same inconsistency between the SNV phylogeny and copy-number data.

We analyzed this dataset using SCARLET to see whether joint analysis of SNVs and CNAs data could help resolve the inconsistency between the tree derived from SNVs and the observed copy number profiles. We first derived four distinct copy-number profiles by hierarchical clustering of ploidy-corrected read-depth ratios from the targeted single-cell sequencing data. These copy-number profiles included an aneuploid profile for all primary cells (P), two different aneuploid

profiles for metastatic cells (M1 and M2), and the profile of diploid cells (D); Leung et al. [83] similarly derived four copy-number profiles from whole-genome sequencing of a different set of 42 cells from the same patient. Since four copy-number profiles is a small number to infer a tree using a copy-number evolution model, we instead ran SCARLET in the ‘Optimal CN Tree’ setting selecting the copy-number tree that produced the highest likelihood. Specifically, we ran SCARLET on all nine possible rooted copy-number trees with the root having the diploid profile (D), and internal vertices labeled by one of the three aneuploid copy profiles (P, M1, M2). For each copy-number tree, we derived the set  $\mathcal{L}$  of supported losses as the mutation loci that exhibited significant decreases in read depth (i.e., number of aligned sequencing reads). Additional details are in Section SC.3.4.

SCARLET constructed a tree (Figure 4.4(C) and Figure C.1(B)) with a single clade containing all metastatic cells. This is consistent with the copy-number data, since the shared chromosomal losses could have occurred once in a common ancestor of all metastatic cells. Moreover, this tree suggests that the liver metastasis was the product of *monoclonal seeding*; i.e., a single cell (or small group of cells) with the same somatic mutations migrated from the primary colon tumor to the metastasis and all metastatic cells descended from the founder cells present in this single migration. This result contradicts previous results [83, 164] of a more complicated polyclonal seeding of the metastasis. The SCARLET tree contains three mutation losses: in genes **FHIT**, **LRP1B**, and **LINGO2**. Each of these losses is supported by a significant decrease in read depth (Figure 4.4(D)), providing evidence that the loci containing these mutation were likely affected by deletions. Notably **FHIT** and **LRP1B** are located in fragile sites in the genome [140], which are known regions of genomic instability. In addition, the loss of the mutation **LINGO2:1** in **LINGO2** is further supported by a shift in the variant allele frequency of another mutation, **LINGO2:2**, in the same gene. Specifically, the variant allele frequency of **LINGO2:2** is  $\approx 1$  in the metastatic cells (Figure C.1(A)), suggesting that this mutant allele is homozygous, consistent with a deletion or loss of heterozygosity event where the **LINGO2:1** mutations was lost.

We examined further the evidence for polyclonal seeding in the initial study of this patient. Leung et al.[83] included a statistical analysis of the variant read counts of the four “bridge mutations”,

ATP7B, FHIT, APC and CHN1 that occurred between the first and second metastatic branches in the SCITE tree. This analysis showed that mutations in ATP7B and FHIT were present in a subset of primary tumor cells and in the second metastatic branch (detected in 10/13 and 13/13 cells respectively) while being absent in the second metastatic branch (detected in 1/15 and 1/15 cells respectively). Under the infinite sites model used by SCITE, mutation loss is not allowed and thus polyclonal seeding is necessary to explain the absence of these mutations. The same analysis found high uncertainty regarding the placement of mutations in APC and CHN1 and thus these were not cited as evidence for polyclonal seeding.

The loss-supported model used by SCARLET provides an alternate explanation for the absence of FHIT and ATP7B. SCARLET identifies a supported mutation loss to explain the presence of the mutation in FHIT only in a subset of metastatic cells (M1). This loss is supported by a shift in read depth ( $p = 0.005$ ) in the 10Mb region containing the locus (Fig 4.4D). SCARLET does not identify a supported mutation loss to similarly explain ATP7B as we did not observe a significant decrease in read depth for the corresponding locus ( $p = 0.34$ ). However, this lack of a significant decrease in read depth at the ATP7B locus does not necessarily imply that there was no mutation loss. In particular, because targeted sequencing was performed for only 1000 genes, the copy number data is fairly low resolution and we calculated read depth in 10Mb bins. Thus, we may lack the statistical power to identify a shorter deletion, especially a deletion present in only the 10 metastatic cells with copy number profile M2. In summary, we argue that the sequencing data provides stronger evidence for the phylogeny constructed by SCARLET, which is consistent with both SNV and copy-number data, and supports a more parsimonious explanation of monoclonal seeding of the liver metastasis.

### 4.3 Discussion

Somatic mutations in tumors range across all genomic scales, from single-nucleotide variants (SNVs) through large copy-number aberrations (CNAs). To date, most methods for constructing phylogenies from single-cell DNA sequencing (scDNA-seq) data [27, 37, 71, 94, 95, 102, 129, 138, 163, 164] used only SNVs, ignoring CNAs and thus throwing out important information for phylogenetic inference. Here, we introduced SCARLET, the first algorithm – to our knowledge – that uses measurements of *both* SNVs *and* CNAs to reconstruct tumor phylogenies from scDNA-seq data. SCARLET is based on a novel loss-supported evolutionary model, which constrains mutation losses to loci containing evidence of a CNA. By using the information about CNAs that is readily available in scDNA-seq data, the loss-supported model has less ambiguity in the phylogeny inference than the Dollo and finite sites models which allow mutation losses to occur anywhere on the tree. In scDNA-seq data, where there is often considerable uncertainty in the mutations present in each cell, this reduction in ambiguity enables more accurate phylogeny inference. On simulated scDNA-seq data, we find that SCARLET outperforms existing methods that do not utilize copy-number data. On targeted scDNA-seq data from a metastatic colorectal cancer patient, we showed that SCARLET found a phylogeny containing three mutation losses. Notably, SCARLET’s tree was both more consistent with the copy-number data and provided a simpler explanation of monoclonal seeding of the liver metastasis, compared to the more complex phylogenies reported previously [83, 164]. Thus, accurate modeling of mutations losses results in different conclusions regarding the migration patterns of metastasis.

There are a number of directions for future improvement. First, the current implementation of SCARLET either requires the copy-number tree in input or enumerates all possible copy-number trees and selects the maximum likelihood result. This approach is applicable when the number of distinct copy-number profiles is small; e.g., in the case of targeted scDNA-seq data [82, 104, 159] where copy-number data typically is lower resolution. However, with higher-quality copy-number data, extensions to larger numbers of copy-number profiles is needed. One approach is to use copy-number evolution models [25, 41, 134, 161] to identify a modest number of copy-number trees

that summarize the uncertainty in the copy-number evolutionary history. Second, one could extend the loss-supported model into a unified evolutionary model for SNVs and CNAs. Indeed, the loss-supported model provides a natural framework to integrate SNVs directly with evolutionary models of CNAs. As single-cell sequencing technologies continue to improve, higher quality measurements of both SNVs and CNAs from the same sets of cells will become available. We anticipate that SCARLET and the loss-supported model will play a crucial role in the analysis of these data.

## 4.4 Methods

### 4.4.1 Loss-supported phylogeny model

We model the evolutionary history of a tumor as a rooted, directed phylogenetic tree  $T = (V(T), E(T))$ , whose vertex set  $V(T) = L(T) \cup I(T)$  consists of a set  $L(T)$  of  $n$  leaves corresponding to *observed cells* and a set  $I(T)$  of inner vertices corresponding to *ancestral cells*. A directed edge  $(v, w) \in E(T)$  indicates that cell  $v$  is an ancestor of cell  $w$ . We do not directly observe  $T$  but rather we measure a set of phylogenetic markers for every observed cell  $v \in L(T)$ . In the case where the markers are somatic single-nucleotide variants (SNV), the measurements correspond to a binary *mutation profile*  $\mathbf{b}_v \in \{0, 1\}^m$  for each observed cell  $v$ , where  $b_{v,a} = 1$  indicates that cell  $v$  has a somatic mutation at locus  $a$  and  $b_{v,a} = 0$  indicates that cell  $v$  does not have a somatic mutation at locus  $a$ . We assume that the mutation profile  $\mathbf{b}_r$  of the root  $r$  is  $\mathbf{b}_r = \vec{0}$  since the root represents the normal cell that preceded the tumor. We define the *mutation matrix*  $\mathbf{B} = [\mathbf{b}_v]_{v \in L(T)}$  to be the matrix whose rows are the mutation profiles of leaves  $v \in L(T)$ .

The problem of phylogenetic tree inference is to find a tree  $T$  and an *augmented mutation matrix*  $\mathbf{B}' = [\mathbf{b}'_v]_{v \in V(T)}$  whose rows correspond to binary mutation profiles of the vertices of  $T$  and where the submatrix  $[\mathbf{b}'_v]_{v \in L(T)}$  is equal to  $\mathbf{B}$ . Since there are many possible trees that relate the observed cells, methods for phylogeny inference find  $T$  and  $\mathbf{B}'$  that best fit a specific evolutionary model. The simplest evolutionary model for SNVs is the *infinite sites*, or *perfect phylogeny* model. In this model,



each mutation is gained ( $0 \rightarrow 1$ ) at most once, and is never subsequently lost. A more general model the Dollo model allows mutations to be gained ( $0 \rightarrow 1$ ) at most once, but lost ( $1 \rightarrow 0$ ) multiple times. Formally, the Dollo model is defined as follows.

**Definition 1.** *A phylogenetic tree  $T$  is a Dollo phylogeny with respect to augmented mutation matrix  $\mathbf{B}'$  provided that for every locus  $a$ , there is at most one edge  $(v, w) \in E(T)$  such that  $b'_{v,a} = 0$  and  $b'_{w,a} = 1$ .*

In contrast to the perfect phylogeny model, under the Dollo model there are often multiple phylogenies that are consistent with input data (Figure 4.1).

DNA sequencing data often contains additional information about the genomic locations where mutation losses are possible. Specifically, we assume that for each cell  $v$ , we also observe a copy-number profile  $\mathbf{p}_v = [p_{v,1}, \dots, p_{v,N}]$  where  $p_{v,i}$  indicates the number of copies of genomic segment  $i$  in cell  $v$ . For simplicity, we label the unique copy-number profiles observed for all the cells by integers  $\{1, \dots, k\}$ , such that the vector  $\mathbf{c} = [c_v]$  represents the copy-number profile assignment  $c_v \in \{1, \dots, k\}$  of every cell  $v$ . The copy-number profiles of cells provide constraints on mutation losses. In particular, we allow mutation losses only at loci where an overlapping deletion or loss-of-heterozygosity (LOH) distinguishes the copy-number profiles. We record the information about the loci where losses are allowed in a collection  $\mathcal{L}$  of *supported loss sets*. For each pair  $c, c'$  of distinct copy-number profiles we define the set  $\mathcal{L}(c, c') \subseteq \{1, \dots, m\}$  of *supported losses* to be the set of all the mutation loci located in genomic regions with a decrease in copy number (indicating possible deletion or LOH) between  $c$  and  $c'$ . We define  $\mathcal{L}(c, c) = \emptyset$  for all  $c$ . We denote the collection of supported losses as  $\mathcal{L} = \{\mathcal{L}(c, c') : (c, c') \in \{1, \dots, k\} \times \{1, \dots, k\}\}$ . We define a *loss-supported phylogeny* as a Dollo phylogeny where all mutation losses are supported.

**Definition 2.** *Given copy number profiles  $\mathbf{c}' = [c_v]_{v \in V(T)}$  and supported losses  $\mathcal{L}$ , a phylogenetic tree  $T$  is a loss-supported phylogeny with respect to augmented mutation matrix  $\mathbf{B}'$  provided that: (1)  $T$  is a Dollo phylogeny; (2) If  $b'_{v,a} = 1$  and  $b'_{w,a} = 0$  for edge  $(v, w)$  then  $a \in \mathcal{L}(c'_v, c'_w)$ .*

The loss-supported phylogeny inference problem is to infer a loss-supported phylogeny  $T$  given a mutation matrix  $\mathbf{B}$  and copy-number profile vector  $\mathbf{c}$  that label the leaves of  $T$ , as well as a set  $\mathcal{L}$  of

supported losses. However, this general problem has a major complication: the copy-number profiles of the ancestral cells are unknown. Without knowledge of ancestral copy-number profiles, the loss sets  $\mathcal{L}$  cannot be used to constrain mutation losses. Ideally, one might infer copy-number profiles of ancestral cells (e.g., using a copy-number evolution model [25, 41, 134, 161]) while *simultaneously* inferring a loss-supported phylogeny on the SNVs. The derivation of a score/likelihood for such joint model is not straightforward, and is left for future work. Instead, in the next section, we describe an algorithm that infers a loss-supported phylogeny by *refining* a *copy-number tree* given in input.

### Loss-supported Refinement Problem

In this section, we introduce the *Loss-Supported Refinement (LSR) problem*, a special case of the loss-supported phylogeny inference problem, where we have additional information about the evolutionary relationships between copy-number profiles. In particular, we assume that we are given a *copy-number tree*  $T = (V(T), E(T))$  and a copy-number profile vector  $\mathbf{c} = [c_v]_{v \in V(T)}$  for all vertices in  $T$ . As single-cell DNA sequencing data of SNVs typically measures copy-number profiles with low-resolution, this copy-number tree typically has many multifurcations (i.e., unresolved ancestral vertices with more than two children). We use the mutation matrix  $\mathbf{B} = [\mathbf{b}_v]$  for all  $v \in L(T)$  to refine vertices in  $T$ , which results in a *joint tree*  $T'$  that reflects the evolutionary history of both the SNVs and CNAs. This sequential approach is inspired by an asymmetry between SNVs and CNAs in the loss-supported model: CNAs affect the observed state transitions of SNVs as deletions result in SNV loss, but SNVs do not result in changes in copy-number state. The joint tree  $T'$  is a *refinement* [158] of  $T$ ; i.e.,  $L(T') = L(T)$  and  $T$  may be obtained by contracting edges in  $T'$ . A refinement is formalized as a mapping  $\gamma : V(T) \rightarrow 2^{V(T')}$ , where for all  $v \in V(T)$ ,  $\gamma(v)$  is a rooted subtree  $T'[\gamma(v)]$  of  $T'$ . Then, tree  $T'$  is a refinement of tree  $T$  provided that contracting  $T'[\gamma(v)]$  into a single vertex  $v$  for all  $v \in V(T)$  yields the tree  $T$ . We refer to the set of subtrees defined by  $\gamma$  as the *refinement subtrees*.

We define the LSR problem as the problem of finding a refinement  $T'$  of a copy-number tree  $T$  such

that  $T'$  is a loss-supported phylogeny.

**Problem 1. Loss-Supported Refinement (LSR) problem**

Given a copy-number tree  $T$ , a copy-number profile vector  $\mathbf{c} = [c_v]_{v \in V(T)}$ , a mutation matrix  $\mathbf{B} = [\mathbf{b}_v]_{v \in L(T)}$ , and supported losses  $\mathcal{L}$ , find a refinement  $T'$  of  $T$ , a copy-number profile vector  $\mathbf{c}' = [c_{v'}]_{v' \in V(T')}$ , and an augmented mutation matrix  $\mathbf{B}' = [\mathbf{b}'_{v'}]_{v' \in V(T')}$  with  $\mathbf{b}_{v'} = \mathbf{b}'_{v'}$  for all  $v' \in L(T')$ , such that

1.  $c'_{v'} = c_v$  for all  $v \in V(T)$  and  $v' \in \gamma(v)$ , and
2.  $T'$  is a loss-supported phylogeny with respect to  $\mathbf{B}'$ ,  $\mathbf{c}'$ , and  $\mathcal{L}$ .

We provide four sufficient and necessary conditions for a solution  $T', \mathbf{c}', \mathbf{B}'$  to the LSR problem. Briefly, these characterize a solution by giving conditions on the set of refinement subtrees defined by  $\gamma$ . The four conditions state that (1) each mutation occurs at most once, (2) mutations are not lost *within* refinement subtrees, (3) all mutation losses *between* refinement subtrees are supported, and (4) refinement subtree copy-number profiles are preserved. We formally define these four conditions as follows. Let  $r(v)$  be the root of subtree  $T'[\gamma(v)]$  and  $p(r(v))$  be the parent of  $r(v)$ .

**Theorem 3.** *Given copy-number tree  $T$ , copy-number profile vector  $\mathbf{c}$ , mutation matrix  $\mathbf{B}$ , and supported losses  $\mathcal{L}$ , a refinement  $T'$  of  $T$ , copy-number profile vector  $\mathbf{c}'$ , and augmented mutation matrix  $\mathbf{B}'$  are a solution to the LSR problem if and only if*

1. *For all loci  $a$ , there exists exactly one edge  $(v', w') \in E(T')$  with  $b'_{v',a} = 0$  and  $b'_{w',a} = 1$ ;*

*And for all  $v \in V(T)$ :*

2. *There does not exist any edge  $(v', w') \in E(T'[\gamma(v)])$  with  $b'_{v',a} = 1$  and  $b'_{w',a} = 0$ ;*
3. *If  $b'_{r(v),a} = 1$  and  $b'_{p(r(v)),a} = 0$ , then  $a \in \mathcal{L}(c'_{p(r(v))}, c'_{r(v)})$ ;*
4.  $c'_{v'} = c_v$  for all  $v' \in \gamma(v)$ .

Note that, taken together, conditions (1) and (2) imply that each of these subtrees  $T'[\gamma(v)]$  is a perfect phylogeny with respect to submatrix  $\mathbf{B}'[\gamma(v)]$ . We use this structure to solve the LSR problem in the next section.

## Solving the Loss-Supported Refinement problem

In this section, we derive an efficient algorithm to solve the LSR problem. This algorithm decomposes the LSR problem into  $k = |I(T)|$  instances – one for each copy-number profile – of the Incomplete Directed Perfect Phylogeny (IDP) problem [119], using the characterization of LSR solutions given in Theorem 3. Specifically, Theorem 3 characterizes LSR solutions by giving a set of conditions on the set of refinement subtrees of  $T$ . We design an algorithm to find a set  $\mathcal{T} = \{T'_v : v \in V(T)\}$  of subtrees, augmented mutation matrix  $\mathbf{B}'$ , and copy-number profiles  $\mathbf{c}'$  that satisfy Theorem 1. Using  $\mathcal{T}$  and  $\mathbf{B}'$ , we then construct a refinement  $T'$  such that  $T'[\gamma(v)] = T'_v$  and  $c'_{v'} = c_v$  for all vertices  $v' \in V(T'_v)$ .

We present a recursive algorithm that refines  $T$  from the leaves to the root. The algorithm relies on three additional constraints on the solution  $T'$ ,  $\mathbf{c}'$  and  $\mathbf{B}'$  that do not effect the existence of a solution, described in the following lemma.

**Lemma 1.** *If there exists a solution to the LSR problem for a given  $T$ ,  $\mathbf{c}$ ,  $\mathbf{B}$ ,  $\mathcal{L}$ , then there exists a solution  $T'$ ,  $\mathbf{c}'$ ,  $\mathbf{B}'$  that satisfies the following conditions.*

- i For all  $(v, w) \in E(T)$ ,  $p(r(w))$  is a leaf of subtree  $T'[\gamma(v)]$ .
- ii For all  $v \in V(T) \setminus \{r\}$ , if  $b'_{v',a} = 1$  for all  $v' \in L(T'[\gamma(v)])$  then  $b'_{r(v),a} = 1$ .
- iii For all  $v \in V(T)$  and all loci  $a$ ,  $b'_{p(r(v)),a} \geq b'_{r(v),a}$ .

Our recursive algorithm is composed of a base and recursive step.

### Base step

The base step determines  $T'_v$  and  $\mathbf{B}'[V(T'_v)]$  for leaf vertices  $v \in L(T)$ . For any leaf in a refinement  $T'$ ,  $\gamma(v) = \{v\}$ . Thus the subtree  $T'_v \in \mathcal{T}$  is composed of a single vertex  $v$ , with mutation profile  $\mathbf{b}'_v = \mathbf{b}_v$  and copy-number profile  $c'_v = c_v$ .

### Recursive step

The recursive step aims to find  $T'_v$  and  $\mathbf{B}'[V(T'_v)]$  for internal vertices  $v \in I(T)$ . We approach this by first identifying the set of constraints on the leaves  $L(T'_v)$  of  $T'_v$  given by Theorem 1. Given these

constraints, we find mutation profiles  $\mathbf{B}'[L(T'_v)]$  for the leaves that respect a perfect phylogeny, as required by condition (1) and (2). These mutation profiles uniquely determine the structure of  $T'_v$  as  $T'_v$  is a perfect phylogeny [63].

By condition (i) of Lemma 1,  $T'_v$  has a leaf for every  $(v, w) \in E(T)$ ; thus  $L(T'_v) = \{p(r(w)) : (v, w) \in E(T)\}$ . We first recursively solve for  $T'_w$  and  $\mathbf{B}'[L(T'_w)]$  for every vertex  $w$  such that  $(v, w) \in E(T)$ . Thus, we know the mutation profile  $\mathbf{b}'_{r(w)}$  of the root of each child subtree. We do not directly observe  $\mathbf{B}'[L(T'_v)]$ , but the mutation profile of a vertex is constrained by condition (3) of Theorem 1 and constraint (iii) of Lemma 1 given the mutation profile of a child. Specifically, the parent has the same mutation profile as the child, except if there is a mutation loss. The mutation profiles are further constrained by condition (1) of Theorem 1 as each mutation occurs at most once across all subtrees. We respect this condition by minimizing the number of mutation gains per locus, by only having a mutation gain at locus  $a$  in a subtree if a strict subset of the leaves have the mutation.

We summarize these leaf constraints on  $\mathbf{B}'[L(T'_v)]$  as a ternary matrix  $\bar{\mathbf{B}}'_v = [\bar{\mathbf{b}}'_v]_{v \in L(T_v)}$  where  $\bar{\mathbf{b}}'_v = [\bar{b}'_{v,a}] \in \{0, 1, ?\}^m$ . The first constraint fixes the values for some entries of  $\mathbf{B}'[L(T'_v)]$  such that  $\bar{b}'_{p(r(w)),a} = 1$  when  $b'_{r(w),a} = 1$  and  $\bar{b}'_{p(r(w)),a} = 0$  when  $b'_{r(w),a} = 0$  and  $a \in \mathcal{L}(c_v, c_w)$ . The second constraint further sets some of the previously non-fixed entries in  $\mathbf{B}'[L(T'_v)]$  to minimize the total number of mutation gains in  $T'_v$ . To achieve the minimum number of mutation gains, we thus maximize the number of all-zero and all-one columns of  $\bar{\mathbf{B}}'_v$ : we set to 0 any previously undetermined entries  $\bar{b}'_{v',a}$  for columns of  $\bar{\mathbf{B}}'_v$  that only have '0' ('1', resp.) entries (setting of  $\bar{b}'_{v',a} = 0$ ,  $\bar{b}'_{v',a} = 1$  resp.). At last, we set any remaining undetermined entry of  $\bar{\mathbf{B}}'_v$  to be '?'.

Finally, we aim to find  $\mathbf{B}'[L(T_v)]$  by filling the '?' entries of  $\bar{\mathbf{B}}'_v$ . More specifically, given  $\bar{\mathbf{B}}'_v$ , we seek  $\mathbf{B}'[L(T_v)]$  such that if  $\bar{b}'_{v',a} \in \{0, 1\}$  then  $b'_{v',a} = \bar{b}'_{v',a}$  for all mutations  $a$  and  $\mathbf{B}'[L(T_v)]$  is a perfect phylogeny matrix. This problem is known as the Incomplete Directed Perfect Phylogeny (IDP) problem and has been shown to be solvable in  $O(n^2m)$  time [119]. In our case  $n = |L(T_v)| = d_v$  where  $d_v$  is the out-degree of vertex  $v$  in  $T$ . Solving an instance of the IDP problem yields a perfect phylogeny mutation matrix  $\mathbf{B}'[L(T_v)]$ , which in turn determines the perfect phylogeny tree  $T'_v$  and

mutation matrix  $\mathbf{B}'[L(T'_v)]$ .

#### 4.4.2 Maximum Likelihood Loss-supported Refinement Problem

The LSR problem assumes that the mutation matrix  $\mathbf{B}$  is error-free. In practice, we do not observe this mutation matrix  $\mathbf{B}$ , but instead we observe read counts from a sequencing experiment. Specifically, we measure a variant read count matrix  $\mathbf{X} = [\mathbf{x}_v]_{v \in L(T)}$  and a total read count matrix  $\mathbf{Y} = [\mathbf{y}_v]_{v \in L(T)}$ , where  $x_{v,a} \in \mathbb{N}$  is the number of variant reads at locus  $a$  in cell  $v$  and  $y_{v,a} \in \mathbb{N}$  is the total number of reads. Whole-genome amplification [56], which typically precedes single-cell DNA sequencing, introduces a considerable amount of error into these read count matrices. Specifically, single-cell sequencing SNV data has high rates of false negative errors (i.e.,  $x_{v,a} = 0$  when  $b_{v,a} = 1$ ) and missing data (i.e.,  $y_{v,a} = 0$ ). In addition, sequencing and whole-genome amplification introduce false positive errors (i.e.,  $x_{v,a} > 0$  when  $b_{v,a} = 0$ ) as well. Most existing methods [37, 71, 94, 95, 129, 163, 164] for single-cell phylogeny inference discretize read counts into an observed mutation matrix  $\tilde{\mathbf{B}}$ , using either two or three genotypes in addition to missing data (i.e.,  $\tilde{b}_{v,a} \in \{0, 1, ?\}$  or  $\tilde{b}_{v,a} \in \{00, 01, 11, ?\}$ ). However, discretizing the mutation data loses information about the likelihood of errors. For example, a locus with a single variant read is far more likely to be a false positive error than a locus with hundreds of variant reads, but a discretized mutation matrix does not distinguish between these cases.

Here, we adopt a maximum-likelihood approach that models the observed variant and total read counts. A similar approach was used in SciPhi [138] with the infinite sites model for SNVs. Our approach aims to find the mutation matrix  $\mathbf{B}^* = \operatorname{argmax} \Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B})$ ; i.e., the mutation matrix that admits a solution  $T', \mathbf{B}', \mathbf{c}'$  to the LSR problem and maximizes the likelihood of the observed variant read counts  $\mathbf{X}$  given the total read counts  $\mathbf{Y}$ . This formulation is not specific to a particular likelihood model for read counts but does assume that variant read counts  $\mathbf{X}$  are independent of each other across cells and loci given  $\mathbf{Y}$  and  $\mathbf{B}$  – i.e., a likelihood of the form  $\Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B}) = \prod_{v=1}^n \prod_{a=1}^m \Pr(x_{v,a} \mid y_{v,a}, b_{v,a})$ . Let  $\mathcal{B}_{T, \mathbf{c}, \mathcal{L}}$  be the set of mutation matrices  $\mathbf{B}$  such that there exists a solution  $T', \mathbf{c}', \mathbf{B}'$  to the LSR problem given  $T$ ,  $\mathbf{c}$ , and  $\mathcal{L}$ . We formulate the problem as

follows.

**Problem 2. *Maximum Likelihood Loss-Supported Refinement (ML-LSR) problem***

*Given variant read counts  $\mathbf{X} = [\mathbf{x}_v]_{v \in L(T)}$ , total read counts  $\mathbf{Y} = [\mathbf{y}_v]_{v \in L(T)}$ , copy-number tree  $T$ , copy-number profile vector  $\mathbf{c} = [c_v]_{v \in V(T)}$ , and supported losses  $\mathcal{L}$ , find the mutation matrix  $\mathbf{B}^* \in \mathcal{B}_{T, \mathbf{c}, \mathcal{L}}$  that maximizes  $\Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B})$ .*

We show the ML-LSR is NP-hard by reduction from the Minimum Flip Problem [23] (Section SC.4.3). Since current datasets have mutation matrices with hundreds–thousands of cells, we derive an algorithm in the next section that finds an approximate solution to the ML-LSR problem by subdividing the ML-LSR problem into  $k$  instances of the maximum likelihood Incomplete Directed Perfect Phylogeny problem.

#### 4.4.3 Scarlet Algorithm for Maximum-Likelihood Loss-Supported Refinement Problem

We introduce SCARLET (Single-Cell Algorithm for Reconstructing the Loss-supported Evolution of Tumors), an algorithm to find a loss-supported phylogeny  $T'$  from single-cell DNA sequencing data. SCARLET aims to solve the ML-LSR problem by finding the maximum likelihood mutation matrix  $\mathbf{B}^*$  such that there exists a solution  $T', \mathbf{c}', \mathbf{B}'$  to the LSR problem where  $\mathbf{B}'$  is an augmentation of  $\mathbf{B}^*$  (i.e.,  $\mathbf{B}^* = \mathbf{B}'[L(T)] = \mathbf{B}'[L(T')]$ ). We proceed here by finding  $\mathbf{B}'$ , which gives us  $\mathbf{B}^*$ . In Section 2.3, we showed that finding  $\mathbf{B}'$  for the LSR problem decomposes into a set of IDP instances if we know the mutation profiles  $\mathbf{R} = [\mathbf{b}'_{r(v)}]_{v \in I(T)}$  of the roots of subtrees  $\mathcal{T}$ . Given  $\mathbf{B}$ , we inferred  $\mathbf{R}$  recursively, starting with the leaves  $L(T)$  whose mutation profiles are known. In the ML-LSR, however, we are not given  $\mathbf{B}$ . Therefore, SCARLET uses a heuristic to find  $\mathbf{B}^*$  where we first compute maximum likelihood mutation profiles  $\mathbf{R}^*$  of the roots, and then solve a set of instances of a maximum likelihood IDP (ML-IDP) problem given  $\mathbf{R}^*$ . We compute  $\mathbf{R}^*$  by marginalizing over possible mutation profiles for each mutation (Section SC.3.2). As ML-IDP input, we define a ternary matrix  $\bar{\mathbf{B}}'_v = [\bar{\mathbf{b}}'_w]_{w \in L(T'_v)}$  for each vertex  $v \in V(T)$ . For  $v \in I(T)$ ,

we define  $\bar{\mathbf{b}}'_{p(r(v))}$  as in Section 4.4.1 given the mutation profile  $\mathbf{b}'_{r(v)}$  of the root  $r(v)$  of  $T'_v$ . For  $v \in L(T)$ , we have that  $\bar{\mathbf{b}}'_{p(r(v))} = \mathbf{b}_v$  but unlike in the LSR problem, we are not given the mutation profile  $\mathbf{b}_v$  in the ML-LSR problem. Instead, we are able to compute the likelihood of  $\mathbf{b}_v$  as in Equation C.1. As such, finding  $\mathbf{B}^*$  is equivalent to find the maximum likelihood submatrices  $\{\mathbf{B}^*[\{v : c_w = c_v\}] : v \in I(T)\}$  such that  $\bar{\mathbf{B}}'_v$  admits an incomplete directed perfect phylogeny. We show in Section SC.3.3 how to compute these maximum likelihood submatrices using an integer-linear programming (ILP) formulation.

This heuristic is not guaranteed to find the overall maximum likelihood solution  $\mathbf{B}^*$  as there may be cases where  $\mathbf{B}^*$  does not admit a solution with the maximum likelihood set of roots  $\mathbf{R}^*$ . However, we showed in the Section 4.2.2 that SCARLET is both accurate and fast in practice. SCARLET can be used with any likelihood model that assumes conditional independence between variant read counts given the mutation matrix and total read counts as described in the previous section. In this work, we used a beta-binomial model for variant read counts, similar to the one used by SciPhi [138], which accounts for overdispersion due to whole-genome amplification as well as sequencing error. Additional details are in Section SC.3.1.



## Chapter 5

# Haplotype Phasing in Single-Cell DNA Sequencing Data

### 5.1 Introduction

In recent years, single-cell DNA sequencing technologies have enabled the measurement of the genomic changes in individual cells [56]. This technology has been used to measure somatic mutations in normal tissue [86, 98], to quantify somatic evolution in cancer [110, 153], to investigate the genomes of unculturable microorganisms [97], and for other applications.

Unfortunately, it is not yet possible to directly sequence the DNA molecule(s) present in a single cell. Rather, current single-cell DNA sequencing technologies first perform whole-genome amplification (WGA), in order to obtain sufficient DNA to sequence. Several WGA methods have been introduced, with the three most common being degenerate oligonucleotide primed PCR (DOP-PCR), multiple displacement amplification (MDA), and multiple annealing and looping based amplification cycles (MALBAC) [56]. The lengths of the amplified genomic fragments, or amplicons, range from

200-300bp for DOP-PCR, to 1-5kb for MALBAC, and up to 10-100kb for MDA [136]. As WGA uses repeated cycles of amplification, any errors or non-uniformity in coverage obtained during early cycles of amplification are amplified in later cycles. Thus, WGA results in highly non-uniform coverage of the genome Figure 5.1(a), which is apparent in the observed strong correlations between read depth at genomic loci within the distance of an amplicon [169].

The amplification bias resulting from WGA leads to challenges in identifying genomic variants in single cells, and thus is a negative of single-cell sequencing technology. Considerable efforts have been made to develop analysis algorithms that overcome this bias [7, 55] and to design WGA methods with less bias [22, 120].

In this paper, we demonstrate a positive aspect of amplification bias: since neighboring genomic loci are often co-amplified, the correlation in sequence coverage between neighboring alleles on the same chromosome can be used to phase haplotypes in diploid genomes (Figure 5.1(b)). Diploid genomes, such as the human genome, consist of pairs of homologous chromosomes, distinguished by single-nucleotide polymorphisms (SNPs), and other small genomic differences. Current DNA sequencing technologies yield sequence reads that originate from a mixture of both homologous chromosomes, losing information about the chromosomal origin of each read. Thus, for any pair of heterozygous SNPs that are further apart than a read length, the alleles that are present on the same chromosome are unknown.

Haplotype assembly is the process of reconstructing the haplotypes of an individual – i.e. assigning the alleles of heterozygous SNPs to the corresponding chromosome of origin – from sequence reads obtained from an individual. Since each read generally derives from a single chromosome, if a read spans multiple SNPs, then the observed alleles are presumed from a single haplotype. Haplotype and SNP phase information has applications in population genetics [148] and clinical and medical genomics [58, 127, 152] as well as being used to improve other analyses such as SNP imputation and genotyping [16, 96] and somatic variant calling [11]. Obtaining long continuous haplotype blocks is a challenge as the distance between adjacent SNPs is longer than reads and read fragments in most sequencing technologies. Long reads and other technologies that provide long-range phase

information, such as linked-read and Hi-C data, have been used to improve the length of haplotype assemblies [35, 118, 121, 171], and there are microfluidic techniques designed to recover haplotypes of single-cell data [26, 46]. However, none of these techniques can be applied to existing short-read single-cell data, and may be prohibitively expensive for new experiments.

We describe an algorithm that exploits amplification bias across a collection of sequenced single cells to assemble haplotypes (Figure 5.1(c)). This algorithm is based on the observation in [169] that homologous chromosomes are amplified almost independently during WGA, and thus show similar rates of amplification bias. Thus, amplicon-scale correlations in read depth provide a signal to phase heterozygous SNPs across amplicon lengths. Specifically, our model leverages *allele dropout*, where one of the two alleles of a heterozygous SNP is not measured in a cell, a common feature of single-cell sequencing data [56]. Alleles of two nearby SNPs on the same chromosome are likely to drop out (not be covered by an amplicon) concurrently. We derive a statistical test of concurrent dropout of alleles of heterozygous SNPs. We validate our approach using both whole-genome and whole-exome DNA single-cell sequencing data, and show that our approach predicts haplotype phase with high accuracy, achieving  $> 90\%$  accuracy on top-ranked 22% of pairs of SNPs within amplicon-length distances. We use pairs of SNPs exhibiting high rates of concurrent dropout to define amplification fragments that we input into an existing haplotype assembler [35]. We obtain haplotype blocks that are 3-4 orders of magnitude larger (10.2kb vs. 312bp on whole-genome data, 9.2kb vs. 41bp on whole-exome data) than obtained using read information alone, with low increase in assembly errors.

## 5.2 Methods

Whole-genome amplification (WGA) methods used in single-cell sequencing result in datasets with high rates of *allele dropout*, where alleles that are present in the genome are not observed in a sequenced cell. The primary cause of allele dropout is the failure of amplification of a genomic region from one of the two homologous chromosomes during whole-genome amplification. Thus, one

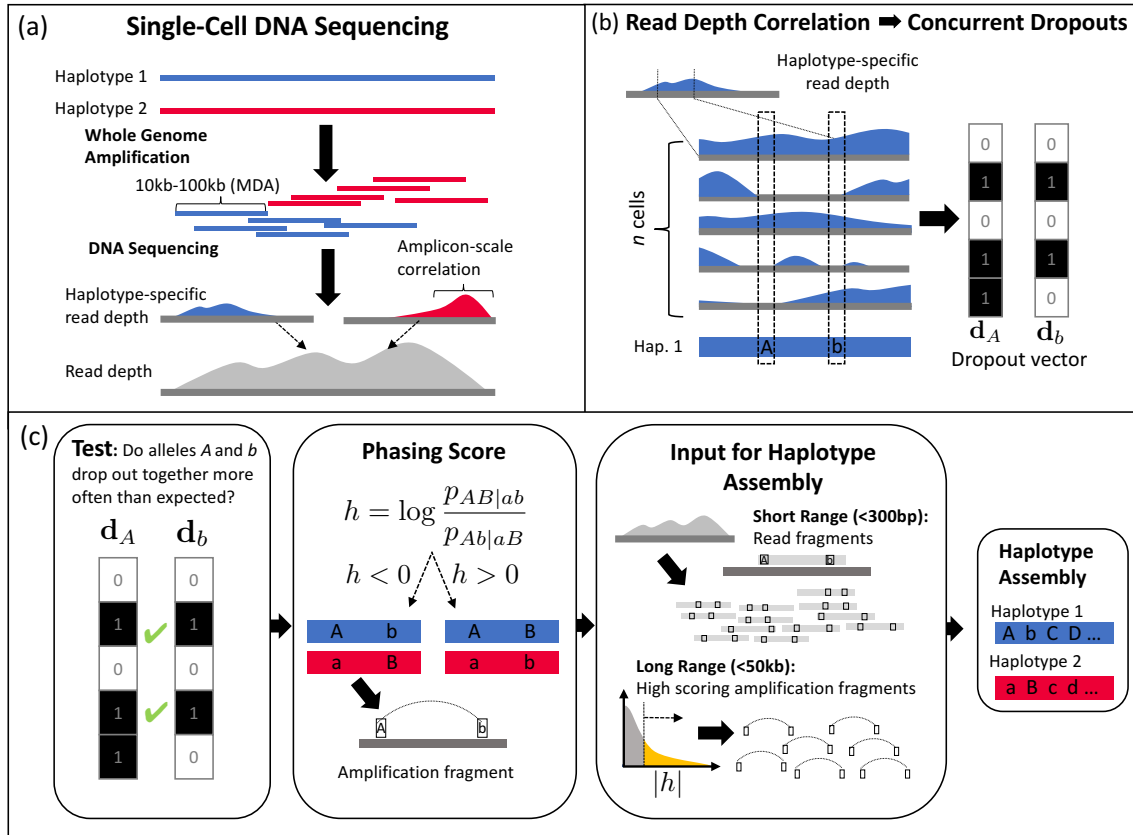


Figure 5.1: (a) Single-cell DNA sequencing typically requires whole-genome amplification (WGA) to obtain sufficient quantities of DNA, which results in non-uniform read depth with correlation at scale of amplicons. Since the two homologous chromosomes are amplified independently, read-depth correlations are strongest between sequence reads originating from the same chromosome/haplotype. (b) Amplicon-scale read-depth correlations, combined with high rates of allelic dropout result in increased rates of concurrent allelic dropout for pairs of alleles originating from the same haplotype, where entries of the dropout vectors  $\mathbf{d}_A$  and  $\mathbf{d}_b$  indicate whether alleles  $A$  and  $b$ , respectively, are measured in each cell. (c) We derive a phasing score for pairs of nearby SNPs based on the  $p$ -values of concurrent dropout for different phasings of alleles. High or low values of the phasing score correspond to amplification fragments containing pairs of alleles that are likely to be on the same haplotype. These amplification fragments are used as input to haplotype assembly algorithms, augmenting phasing information from read fragments containing alleles found on the same read.

expects to observe correlations in rates of allele dropout between alleles of SNPs whose distances are within the length of an amplicon (up to 10-100kb, depending on the WGA method). More specifically, since an amplicon contains DNA sequence from one homologous chromosome, one expects to observe a higher rate of concurrent dropout (dropout of both alleles in one cell) for a

pair of alleles on the same haplotype and within the length of an amplicon, than for two alleles on different haplotypes. We describe a statistical test to evaluate such concurrent dropout between alleles for a pair of SNPs. We then use pairs of alleles which show strong evidence of significant concurrent dropout as input to a haplotype assembly algorithm.

### 5.2.1 Quantifying Concurrent Dropout

We obtain DNA sequencing data from  $n$  single cells from the same individual, and assume that these cells share  $m$  heterozygous SNPs. Consider a pair of heterozygous SNPs with alleles  $A, a$  for the one SNP and alleles  $B, b$  for the other SNP. For an allele  $A$ , we define the *dropout vector*  $\mathbf{d}_A$  to be a binary vector of length  $n$ , where  $d_{A,s} = 1$  if we do not observe any reads containing allele  $A$  in cell  $s$ , and  $d_{A,s} = 0$  otherwise. Thus,  $d_A$  indicates the dropout for allele  $A$  across cells. For an allele  $A$ , let  $n_A = \sum_s d_{A,s}$  be the number of *dropouts*, or cells where  $A$  is not measured. Similarly, for a pair of alleles  $A$  and  $B$ , let  $n_{AB} = \mathbf{d}_A \cdot \mathbf{d}_B = \sum_{s=1}^n d_{A,s} \times d_{B,s}$  be the number of *concurrent dropouts* of alleles, i.e. the number of cells where both  $A$  and  $B$  are not observed. The key idea of our model is that if the distance between SNPs is less than the length of an amplicon, and if alleles  $A$  and  $B$  are on the same haplotype, then concurrent dropout of  $A$  and  $B$  is more frequent than expected by chance. Conversely, if SNPs are far apart or  $A$  and  $B$  are present on different haplotypes, then we expect that amplification of these alleles is independent, and concurrent dropouts are random events.

Let  $N_{AB}$  be a random variable indicating this number of concurrent dropouts between alleles  $A$  and  $B$  across  $n$  cells. Under the null model, dropouts between allele  $A$  and allele  $B$  are independent. To compute the distribution of  $N_{AB}$  under the null, we need to compute the probability  $w_{X,s} = \Pr(d_{X,s} = 1)$  that allele  $X$  drops out in cell  $s$  for each allele and cell. However,  $w_{X,s}$  varies by locus, allele  $X$ , and cell  $s$ . Locus-specific and allele-specific variability in dropout rates results from context-specific amplification, sequencing, or alignment biases. Cell-specific variability in dropout rates results from differences in sequencing depth and uniformity of coverage across cells. Since it is difficult to model each of these effects directly, we instead compute a weighted exact distribution

$P(N_{AB} = n_{AB} | n_A, n_B, \mathbf{w})$ , where  $\mathbf{w}$  are cell-specific weights obtained from the observed number of dropouts across all loci in each cell. See [81] for details of similar weighted tests used in other biological applications.

Specifically, let  $\mathbf{D}$  be the  $2m \times n$  matrix whose rows correspond to the  $2m$  dropout vectors for the set of alleles of  $m$  heterozygous SNPs. We compute the  $p$ -value  $\Pr(N_{AB} \geq n_{AB} | \mathbf{D}_{1\cdot}, \dots, \mathbf{D}_{2m\cdot}, \mathbf{D}_{\cdot 1} \cdots \mathbf{D}_{\cdot n})$  of observing  $n_{ab}$  or more concurrent dropouts, conditioned on the observed row sums and column sums of the matrix  $\mathbf{D}$ . Computing this  $p$ -value is non-trivial. Leiserson et al. [81] introduced the WExT algorithm to compute a saddlepoint approximation of a  $p$ -value for the related problem of mutually exclusive events. We use a recent release of the WExT software that computes a saddlepoint approximation for the co-occurrence test statistic  $N_{AB}$ .

### 5.2.2 Augmenting Haplotype Assembly with Amplification Fragments

Haplotype assembly is the reconstruction of haplotypes from local information about groups of alleles that are present on the same chromosome. The input for haplotype assembly is a set of fragments  $\mathcal{F}$ , where a fragment  $f \in \mathcal{F}$  defines a phasing over a set of SNPs – e.g.,  $f = ABc|abC$ , where  $|$  delineates the two chromosomes, indicates alleles  $A, B, c$  are on one chromosome and alleles  $a, B, C$  are on the other. Haplotype assembly algorithms aim to find the most likely haplotypes from the set  $\mathcal{F}$ . Typically, the set  $\mathcal{F}$  of fragments is equal to  $\mathcal{F}_r$ , the set of sequenced reads (or paired-end reads in the case of mate pair libraries). This is because alleles of two SNPs measured on the same read (or paired-read) are highly likely to reside on the same haplotype.

We extend the fragment set  $\mathcal{F}$  using a set of *amplification fragments* defined from pairs of alleles from neighboring SNPs that demonstrate concurrent dropout, using the statistical test defined in the previous section. A pair of SNPs, having alleles  $A, a$  for one SNP and alleles  $B, b$  for the second SNP, can be phased in two ways:  $AB|ab$  or  $Ab|aB$ . As noted by Zhang et al. [169], during WGA, homologous chromosomes are amplified independently. Thus, we assume the random variables  $N_{AB}$  and  $N_{ab}$  are independent under the null hypothesis, and combine the  $p$ -values  $p_{AB}$  and  $p_{ab}$  using

Fisher’s method to obtain a single  $p$ -value  $p_{AB|ab} = \Pr(T \geq t)$  where  $t = -2(\log p_{AB} + \log p_{ab})$  and  $T$  follows the  $\chi_4^2$  distribution, since two  $p$ -values are combined. For each pair of SNPs, this procedure yields two  $p$ -values,  $p_{AB|ab}$  and  $p_{Ab|aB}$ , corresponding to the strength of evidence against the null model of independence for each phasing. Under phasing  $AB|ab$ , we expect high dropout concurrence for allele pair  $\{A, B\}$  and allele pair  $\{a, b\}$  and independent dropout for allele pairs  $\{A, b\}$  and  $\{a, B\}$ . Thus, we expect to see a low  $p$ -value  $p_{AB|ab}$  and a high  $p$ -value  $p_{Ab|aB}$ .

We summarize the evidence in support of each phasing using a *phasing score* defined as

$$h = -\log \left( \frac{p_{AB|ab}}{p_{Ab|aB}} \right). \quad (5.1)$$

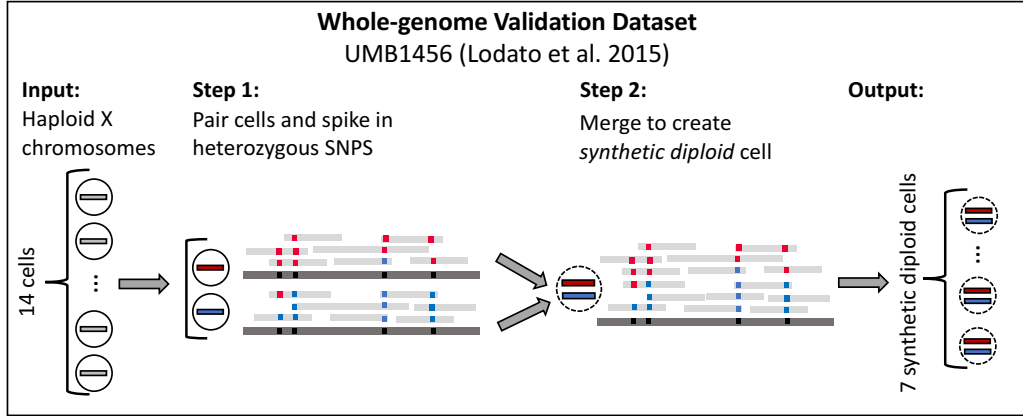
Here,  $h > 0$  indicates stronger evidence for phasing  $AB|ab$ , while  $h < 0$  indicates stronger evidence for phasing  $Ab|aB$ . Large values of  $|h|$  suggest that the allele pairs for one phasing have more concurrent dropout than expected by chance compared to the allele pairs for the other phasing, and thus are more likely to have come from same haplotypes. We define a set of amplification fragments  $\mathcal{F}_c$  containing allele pairs whose absolute values of the phasing scoring  $|h| \geq c$ , for a non-negative threshold  $c$ . We then use a haplotype assembly algorithm, HapCut2 [35] to assemble haplotypes using the combined set of sequence and amplification fragments  $\mathcal{F} = \mathcal{F}_r \cup \mathcal{F}_c$  as input.

## 5.3 Results

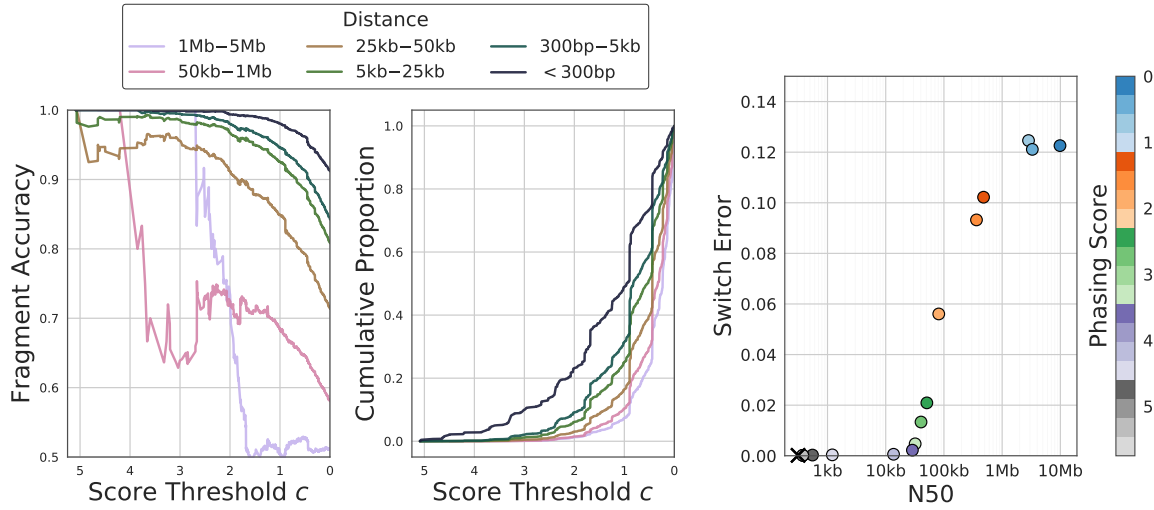
We tested our model on two single-cell DNA sequencing datasets: whole-genome sequencing of neurons [86] and whole-exome sequencing of breast cancer cells [153].

### 5.3.1 Whole-genome Single Cell Sequencing of Neurons

We first evaluated the performance of our model on synthetic diploid cells constructed from sequenced X chromosomes in single-cell DNA sequencing of neurons from a male individual, UMB1465



(a) Experimental Setup



**Figure 5.2: Haplotype Assembly on Whole-Genome DNA Sequencing Data.** (a) We form a validation dataset of seven synthetic diploid cells with known haplotypes from X chromosomes in whole-genome DNA sequencing data of single neuron cells from a male [86]. (b) (Left) The accuracy of the predicted phase for the set  $\mathcal{F}_c$  of amplification fragments with the absolute value of the phasing score  $|h| > c$ . We observe highly accurate prediction of phase for pairs of SNPs whose distance is less than the length of amplicons (here 95th percentile of amplicon length is 103kb). (Right) The proportion of SNP pairs included in the set of amplification fragments  $\mathcal{F}$ . (c) The  $N50$  and switch error for haplotype assembly as we vary the phasing score threshold  $c$ . The  $N50$  and switch error for the haplotype assembly with no amplification fragments is marked with an ‘x’.

[86]. This dataset includes whole-genome sequencing from  $n = 16$  single cells and two bulk samples.

The single cells were amplified using MDA, and a recent analysis [136] estimates that these samples



had MDA amplicons lengths up to 200kb, with a median of 19kb and 95th percentile of 103kb. To create synthetic diploid cells, we extracted the haploid X chromosomes from each sequenced cell (Figure 5.2a), excluding the pseudo-autosomal regions PAR1 and PAR2 on chromosome X as defined in human reference genome assembly GRCh37.p13. To account for variability in dropout rates between cells, we downsampled all cells to a fixed number of reads, and formed pairs of cells based on the total number of covered positions in the samples. We generated simulated haplotypes using population-level allele frequencies acquired from dbSNP, and spiked these alleles into the sequencing data. We introduced sequencing error into the spiked-in alleles based on the Phred quality scores of individual reads. Further details on the simulation can be found in Supplementary Section A1.

To infer haplotypes, we applied the statistical test described in Section 5.2 to all pairs of SNPs within 50kb of each other. Next, we constructed a set  $\mathcal{F}_c$  of amplification fragments for a phasing score value  $|h| = c$  and input the amplification fragments  $\mathcal{F}_c$  and read fragments  $\mathcal{F}_r$  into HapCut2 to assemble haplotypes.

We evaluated the ability of the phasing score to accurately phase pairs of SNPs at varying distances and over a range of values  $|h|$  (Figure 5.2b). We found that the proportion of fragments whose phase was predicted correctly, increases with the absolute value  $|h|$  of the phasing score, indicating that  $|h|$  is correlated with the accuracy of phasing. With larger amplicon lengths (50kb-1Mb and 1-5Mb) the fragment accuracy decreases rapidly as the phasing score decreases. In addition, relatively few fragments have high phasing scores  $|h|$ , demonstrating that few SNPs are accurately phased when the distance between SNPs exceed the length of WGA amplicons. For haplotype assembly, we use pairs of SNPs whose distance is 50kb or less. Over this set of SNP pairs, we are able to correctly phase fragments with 77% accuracy overall. However, if we restrict to the set of fragments with scores  $|h| \geq 1$ , the top 22% of fragments, we obtain an accuracy of 91%.

We assemble haplotypes using amplification fragments  $\mathcal{F}_c$  and read fragments  $\mathcal{F}_r$  as input to HapCut2 [35] as described in Section 5.2. We ran HapCut2 with a range of thresholds  $c$  for the phasing

scores, where for each  $c$ , we supply HapCut2 with  $\mathcal{F}_r \cup \mathcal{F}_c$ , the set of sequence fragments and amplification fragments. We also run HapCut2 using only sequence reads  $\mathcal{F}_r$  to get a baseline measure of results using only short reads. In each case, we ran HapCut2 with default parameters.

We evaluate the resulting haplotype assemblies using the following metrics (Figure 5.2c).

1. The *N50* is the length of a haplotype block such that half of all phased variants are in a block at least as long.
2. *Switch error* is the proportion of phase connections between adjacent SNPs that are incorrect.

Not surprisingly, we observe a trade-off between switch error rate and length of resulting haplotype blocks. Without any amplification fragments ( $\mathcal{F}_c = \emptyset$ ), HapCut2 obtains haplotype blocks with a median (N50) length of 312bp. As the phasing score threshold  $c$  decreases, the block lengths increase by several orders of magnitude with only relatively small corresponding increases in switch error rate. With phasing score threshold  $c = 2.25$ , we obtain a block length of  $N50 = 10.2\text{kb}$ , with a switch error rate of 0.02. At lower values of phasing score threshold, we see larger increases in error, which corresponds to the observed decrease fragment accuracy at the same values (Figure 5.2b). Depending on the downstream analysis that utilizes the resulting haplotypes, these low thresholds may prove useful. For example, at a threshold of  $c = 0$ , we acquire blocks with an  $N50 = 9.96\text{Mb}$ , with 92.6% of blocks containing no switch errors.

### 5.3.2 Whole-exome Single-cell Sequencing of Breast Cancer

Whole-exome sequencing comprises a significant proportion of available single-cell sequencing data sets [110], due to both the lower sequencing costs compared to whole-genome sequencing and because of interest in measuring variation in coding regions. Assembling haplotypes from short-read whole-exome sequencing data is challenging as reads (or paired reads) are shorter than the lengths of most introns. Introns are estimated to have a median length of approximately 1334bp in the human genome [69], while the median exon size is approximately 122bp [28]. Thus haplotype phase

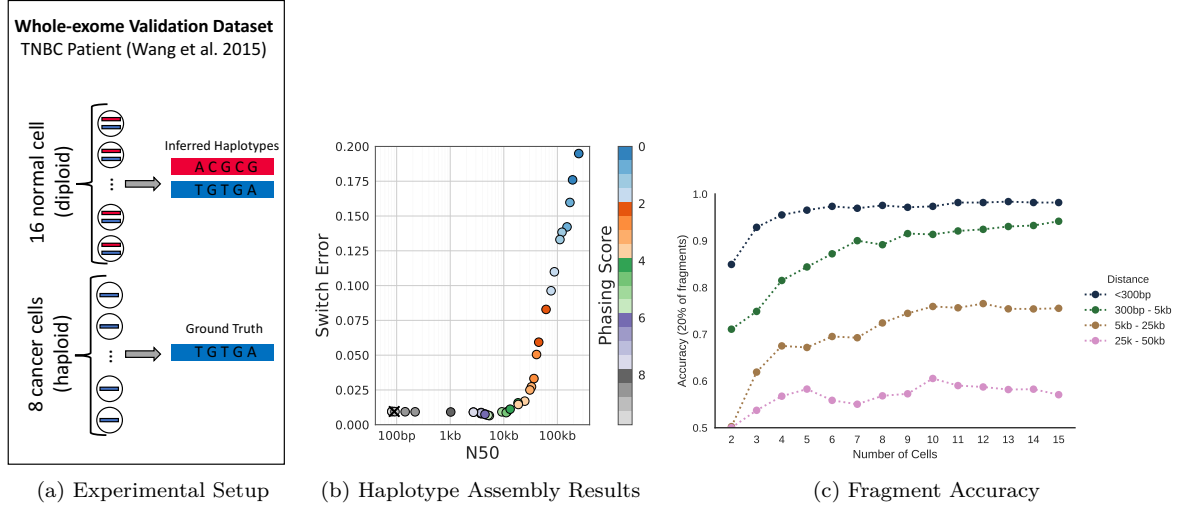


Figure 5.3: **Haplotype Assembly on Whole-Exome DNA Sequencing Data.** (a) We validate haplotype assemblies on whole-exome DNA sequencing data from an individual breast cancer patient [153], by comparing to the haplotype of chromosome 17, whose haplotype we can determine from the 8 cancer cells that have lost one homolog of this chromosome. (b) Haplotype block length ( $N50$ ) as a function of haplotype switch error for varying threshold of phasing score. The  $N50$  and switch error for the assembly with no amplification fragments is marked with an ‘ $\times$ ’. Amplification fragments increase the length of haplotype assemblies by orders of magnitude with small increase in switch error. (c) The accuracy of the phasing for the highest scoring 20% of amplification fragments for varying numbers of cells.

is difficult to determine across exons from short read data.

Since WGA amplicons are typically longer than an intron, we hypothesized that we could use our model to obtain haplotype blocks from single-cell whole-exome data that were substantially longer than blocks obtained with short-read sequencing data. To test this hypothesis, we evaluated our model on single-cell whole-exome data from a triple-negative breast cancer patient [153]. We observed that one copy of chromosome 17 was lost in 8 cancer cells in this datasets that were indicated to be hypodiploid. Using these 8 cells we obtained the true haplotypes for chromosome 17 (Figure 5.3a). We then applied our model to 14 normal (non-cancerous) diploid cells from the same individual. We ran the model and constructed amplification fragments as in previous sections. Figure 5.3b shows the features of the resulting haplotype assemblies.

Using only read fragments  $\mathcal{F}_r$ , the haplotype assemblies have a median block length of  $N50 = 41\text{bp}$ ,

which, as expected, is shorter than the length of a single exon. Using amplification fragments  $\mathcal{F}_c$  we are able to increase the block length by several orders of magnitude, with only small increases in switch error (Figure 5.3b). For example, when the phasing score threshold  $c = 2.75$ , we obtain a median block length  $N50 = 9.3\text{kb}$ , with a corresponding switch error rate of 0.04. This indicates that we are able to phase across multiple exons. Indeed this haplotype block length is of the same order of magnitude of the typical gene length.

While read-based phasing is limited by the length of fragments, reference-based phasing algorithms are generally able to phase over longer genomic distances. We compared the haplotype assemblies we obtain to a reference-based phasing algorithm, EAGLE2 [87] (Supplementary Figure 1). We show that while EAGLE2 was able to provide a phasing over the whole chromosome, our amplification-fragment based phasing provides lower error rates within the blocks that we obtain.

To investigate the number of single-cells required to obtain accurate amplification fragments, we ran the model on subsets of cells of size  $2 - 15$  (Figure 5.3c). We find that on this data set, the fragment accuracy levels off after 8-10 cells. Thus, we can obtain accurate amplification fragments even with relatively few cells.

## 5.4 Discussion

Single-cell DNA sequencing is increasingly being used to explore the genomic content of individual cells, but requires analysis algorithms that are robust to the errors and biases in this data. In this paper, we exploit one bias in single-cell sequencing data, amplification bias, and show how we can leverage this local information to assemble haplotypes.

Our results demonstrate that concurrent dropout between nearby alleles can provide amplicon-scale correlations that lead to better haplotype assemblies than using only correlations between alleles on the same sequence read. However, there are several limitations and avenues for further improvement of our approach. First, many recent haplotype assembly algorithms, including the HapCut2 [35]

used here, employ more sophisticated probabilistic models for error in fragments. Extending our model to estimate error rates for amplification fragments could provide better integration with the error models in haplotype assemblers, and yield more accurate haplotype predictions. While our current model does not calculate a likelihood for each phasing, we may be able to estimate the error rates based on the empirical distribution of phasing scores. Alternatively, an improved probability model that accounts for several of the features of sequencing data – including sequencing error and observed read depth – could be developed and would likely outperform the straightforward model of concurrent dropout introduced here. Additionally, extending the model to consider groups of SNPs instead of just pairs of SNPs may be useful for identifying phase when pairwise relationships are weak. This can be particularly useful with the lower sequence coverages that are common in single-cell sequencing datasets.

If one’s goal is to obtain a high-quality phased diploid genomes, there are a number of good approaches. These include: the application of high-quality reference-based phasing algorithms [16, 30, 87, 143] that exploit large populations of genotyped individuals; long-read [58] or linked-read [171] sequencing; or specialized techniques such as Strand-Seq [124]. Phasing algorithms also exist for both bulk [21] and single-cell RNA-seq data [21]. We do not expect that researchers will perform single-cell sequencing if their only goal is to obtain a phased, diploid genome.

Rather, we anticipate that the approach described here will be a useful complement for specific analyses of single-cell sequencing data. For example, we expect that this model can be broadly useful in variant calling in single cells, which is typically significantly confounded by amplification bias. While we validated the model on diploid genomes, the model is readily adaptable to copy-number aberrations (CNAs), and thus can be applied to cancer genomes which often demonstrate high levels of aneuploidy. The information derived from our model may be useful for allele-specific copy number calling in single cells, which to our knowledge is not currently done in any existing single-cell copy-number caller. Information on haplotype phase will also be useful for calling retrotransposon insertions [45] or single nucleotide variants. Recently, [11] showed how they were able to reduce false positive rates in SNV calling in single-cell DNA sequencing by phasing SNVs

to nearby SNPs. However, many SNVs cannot be phased to SNPs with short reads, especially in whole-exome data. Our method is able to phase across much larger distances and across exons. An additional application is phasing of structural variants in single-cells. SNPs on either side of the breakpoints of a structural variant will also show more dropout concurrence than expected by chance. An extension of our model might allow for improved phasing of structural variants, which could be useful in reconstructing highly rearranged cancer genomes. Finally, additional extensions of dropout concurrence might be applied to single-cell RNA-seq data, e.g. by exploiting correlations in allele-specific expression or allele-specific alternative splicing.

## Chapter 6

# Conclusions

Cancer is the second leading cause of death in the United States. The American Cancer Society projects that in 2019, there will be 1.7 million new cases of cancer and 600,000 cancer deaths in the United States [137]. In recent years, doctors and researchers have made meaningful progress to prevent the incidence and reduce the mortality of cancer. Cancer death rates have dropped by 27% since 1991 [137]. However, cancer remains difficult to treat, largely due to its heterogeneity – each cancer case is unique with different sets of mutations and responses to treatment. Cancer treatment is further complicated by intratumor heterogeneity, where tumors are composed of genetically distinct subpopulations of cells. This heterogeneity has been associated with resistance to treatment and with relapse and metastasis [29]. Characterizing tumor heterogeneity and understanding the evolutionary process that produced it is a critical step towards effective treatment of cancer.

The broadening availability of DNA sequencing from tumors over the past decade, including multi-sample, multi-site, and longitudinal sequencing, has given researchers a much more extensive understanding of heterogeneity and the process of evolution in tumors. Tens of thousands of tumors across many cancer types have now been sequenced by individual centers and hospitals as well as large-scale projects, like the International Cancer Genome Consortium (ICGC) [167] and The

Cancer Genome Atlas (TCGA) [155]. Due to the vast quantity of cancer sequencing data, there is a need for specialized algorithms to perform accurate analyses.

The work in this thesis describes a set of algorithms for the reconstruction of tumor evolutionary histories from DNA sequencing data. The focus of this work is on probabilistic modeling of sequencing data (Chapters 2, 4 and 5), accounting for the complex interactions between different mutation types (Chapters 3 and 4), and tackling the challenges put forth by different types of sequencing data (Chapters 2, 3 – bulk; Chapters 4, 5 – single cell).

## 6.1 Future Work

Throughout this thesis, we observed uncertainty in phylogenetic reconstruction, where sequencing data admits multiple possible evolutionary histories. Tackling this uncertainty is a challenge moving forward. First, efforts must be made to reduce possible ambiguity phylogenetic reconstruction. We already have taken steps towards achieving this. Chapter 3 quantifies how changing the number of samples and sequencing depth can affect the size of the solution space, and Chapter 4 discusses how introducing copy-number data can constrain the set of mutation losses. In recent work, we also demonstrated how integrating information about the context of the tumor samples – for metastatic [42] and longitudinal [106] data – can inform phylogenetic reconstruction. Future work remains to identify how additional sequencing data or external information can be used to further reduce ambiguity. Beyond reducing uncertainty, there need to be efforts towards quantifying and communicating the uncertainty. Not doing so may result in incorrect or contradictory results in downstream applications; for example, we demonstrated recently that not accounting for possible alternate phylogenies may result in overly complex metastatic seeding histories [42]. We tackled the problem of quantifying uncertainty in Chapter 2 and 3, where in Chapter 2 we provide a posterior probability over all possible trees, and in Chapter 3 we provide an algorithm to enumerate all possible solutions. However, while presenting the full solution space is important, it can also be difficult to interpret. The field of tumor phylogenetics needs new ways to meaningfully



summarize the uncertainty in solutions. In classic phylogenetics, there is a body of work towards reconstructing *consensus trees* [17], and recent efforts have begun to apply these concepts to tumor phylogenetics [3, 59]. These methods have shown promising results but are not yet broadly applicable to probabilistic models for phylogenetic trees and evolutionary models beyond infinite sites. Moreover, consideration needs to be made into how downstream analyses should deal with uncertainty in phylogenetic reconstruction and consensus trees.

One of the focuses of this thesis was integrating copy-number data into phylogenetic reconstruction. The approach described in Chapters 3 and 4 primarily focused on phylogenetic reconstruction using SNVs as a marker. This work modeled how copy-number aberrations affect the observed signal from SNVs. As they do not use a true evolutionary model for CNAs, they do not provide a true integrated model of SNV and CNA evolution. The choice to focus on SNVs as a marker was made partially due to limitations of the data. In bulk data (Chapter 3), there is a greater ability to resolve small subclones distinguished by SNVs than by CNAs. In single-cell data (Chapter 4), current technologies have a trade-off between quality of SNV data and quality of CNA data [56]. Therefore, it is presently challenging to measure both SNVs and CNAs from the same set of cells. This limitation may be overcome in the near future thanks to improving sequencing technology. However, a second motivation to focus on SNVs as a marker was the complexity of copy-number models. While most models make the assumption that SNVs evolve independently, CNAs often overlap each other and affect the structure of the genome. Existing algorithms that reconstruct the evolutionary history of CNAs generally make simplifying assumptions about copy-number events in order to make the problem computationally tractable. More work must be done to accurately model phylogenetic reconstruction of CNAs. Moreover, true simultaneous inference of CNA and SNV evolution remains an open problem. As future work, an integrated approach may help overcome some of the challenges of reconstructing CNA evolution if SNVs are used to constrain possible solutions.

In the coming years, emerging sequencing technologies will deepen our understanding of tumor heterogeneity and evolution. The broadening availability of single-cell DNA sequencing data will

allow researchers to probe the heterogeneity of tumors at the highest resolution. Moreover, less invasive non-surgical technologies like blood assays for circulating tumor cells and cell-free DNA will enable more complete tracking of tumor evolution over time. While these technologies are promising, they also introduce new computational challenges. As described in this thesis, new technologies typically require specialized models and algorithms to interpret. In addition to new DNA sequencing technologies that will allow us to better measure tumor evolution, there are a wide variety of technologies that measure different properties of tumors, including the transcriptome and epigenome of cancer cells, the immunological composition of the tumor microenvironment, the spatial organization of tumor DNA, and interactions between DNA and regulatory proteins. In general, interpretation of this data presently focuses on quantifying properties of the tumor at the time of measurement. But tumors are dynamic ecosystems, in which the composition of the tumor and microenvironment fluctuates, and new mutations disrupt different cellular processes. Applying an evolutionary framework to these data, or integrating new data with matched DNA sequencing data, will allow us to reconstruct how these different processes change over the course of the evolutionary history of a tumor.

Finally, while reconstructing the past gives insight into the development of tumors and may lead to different preventative approaches, prediction of future evolution is a key open problem for the utility of phylogenetic approaches in clinical settings. There are already inroads in this direction with work in RNA velocity [80], measurement of relative fitness of individual clones [139, 157], and quantifying the relationship between survival and intratumor heterogeneity [72]. While these are promising beginnings, there is still a long road towards a predictive model of future tumor evolution, especially with regards to predicting response to treatment.

# Bibliography

- [1] 10X Genomics. Assessing tumor heterogeneity with single cell cnv. <https://www.10xgenomics.com/solutions/single-cell-cnv>. Accessed: 2019-Nov-05.
- [2] R. Agarwala and D. Fernández-Baca. A Polynomial-Time Algorithm For the Perfect Phylogeny Problem When the Number of Character States is Fixed. *SIAM Journal on Computing*, 23(6):1216–1224, Dec. 1994.
- [3] N. Aguse, Y. Qi, and M. El-Kebir. Summarizing the solution space in tumor phylogeny inference by multiple consensus trees. *Bioinformatics*, 35(14):i408–i416, 2019.
- [4] J. M. Alves, T. Prieto, and D. Posada. Multiregional tumor trees are not phylogenies. *Trends in cancer*, 3(8):546–550, 2017.
- [5] N. Andor, T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji, and C. C. Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113, Nov. 2015.
- [6] O. Bagci and S. Kurtgöz. Amplification of cellular oncogenes in solid tumors. *North American journal of medical sciences*, 7(8):341, 2015.
- [7] B. Bakker, A. Taudt, M. E. Belderbos, D. Porubsky, D. C. Spierings, T. V. de Jong, N. Halsema, H. G. Kazemier, K. Hoekstra-Wakker, A. Bradley, et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome biology*, 17(1):115, 2016.

- [8] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.
- [9] C. M. Bielski, A. Zehir, A. V. Penson, M. T. Donoghue, W. Chatila, J. Armenia, M. T. Chang, A. M. Schram, P. Jonsson, C. Bandlamudi, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature genetics*, 50(8):1189, 2018.
- [10] H. L. Bodlaender, M. R. Fellows, and T. Warnow. Two strikes against perfect phylogeny. In Kuich [78], pages 273–283. ISBN 3-540-55719-9. doi: 10.1007/3-540-55719-9\_80. URL [http://dx.doi.org/10.1007/3-540-55719-9\\_80](http://dx.doi.org/10.1007/3-540-55719-9_80).
- [11] C. L. Bohrsen, A. R. Barton, M. A. Lodato, R. E. Rodin, V. Viswanadham, D. Gulhan, I. Cortes, M. A. Sherman, L. J. Luquette, M. Kwon, et al. Linked-read analysis identifies mutations in single-cell dna sequencing data. *bioRxiv*, page 211169, 2017.
- [12] N. Bolli, H. Avet-Loiseau, D. C. Wedge, P. Van Loo, L. B. Alexandrov, I. Martincorena, K. J. Dawson, F. Iorio, S. Nik-Zainal, G. R. Bignell, J. W. Hinton, Y. Li, J. M. C. Tubio, S. McLaren, S. O’ Meara, A. P. Butler, J. W. Teague, L. Mudie, E. Anderson, N. Rashid, Y.-T. Tai, M. A. Shammash, A. S. Sperling, M. Fulciniti, P. G. Richardson, G. Parmigiani, F. Magrangeas, S. Minvielle, P. Moreau, M. Attal, T. Facon, P. A. Futreal, K. C. Anderson, P. J. Campbell, and N. C. Munshi. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature communications*, 5, Jan. 2014.
- [13] P. K. Brastianos. Genomic Characterization of CNS Metastases: Implications for Precision Medicine. *The Oncologist*, 2016.
- [14] P. K. Brastianos, S. L. Carter, S. Santagata, D. P. Cahill, A. Taylor-Weiner, R. T. Jones, E. M. Van Allen, M. S. Lawrence, P. M. Horowitz, K. Cibulskis, K. L. Ligon, J. Tabernero, J. Seoane, E. Martinez-Saez, W. T. Curry, I. F. Dunn, S. H. Paek, S.-H. Park, A. McKenna, A. Chevalier, M. Rosenberg, F. G. Barker, C. M. Gill, P. Van Hummelen, A. R. Thorner, B. E. Johnson, M. P. Hoang, T. K. Choueiri, S. Signoretti, C. Sougnez, M. S. Rabin, N. U. Lin, E. P. Winer, A. Stemmer-Rachamimov, M. Meyerson, L. Garraway, S. Gabriel, E. S. Lander,

- R. Beroukhim, T. T. Batchelor, J. Baselga, D. N. Louis, G. Getz, and W. C. Hahn. Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discovery*, 5(11):1164–1177, Nov. 2015.
- [15] C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, Sept. 1973.
- [16] B. L. Browning and Z. Yu. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics*, 85(6):847–861, 2009.
- [17] D. Bryant. A classification of consensus methods for phylogenetics. *DIMACS series in discrete mathematics and theoretical computer science*, 61:163–184, 2003.
- [18] P. Buneman. A Note on the Metric Properties of Trees. *J. Combinatorial Theory (B)*, 17: 48–50, 1974.
- [19] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.
- [20] Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–20, Oct 2013. doi: 10.1038/ng.2764.
- [21] S. E. Castel, P. Mohammadi, W. K. Chung, Y. Shen, and T. Lappalainen. Rare variant phasing and haplotypic expression from rna sequencing with phaser. *Nature communications*, 7:12817, 2016.
- [22] C. Chen, D. Xing, L. Tan, H. Li, G. Zhou, L. Huang, and X. S. Xie. Single-cell whole-genome analyses by linear amplification via transposon insertion (lianti). *Science*, 356(6334):189–194, 2017.

- [23] D. Chen, O. Eulenstein, D. Fernandez-Baca, and M. Sanderson. Minimum-flip supertrees: complexity and algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(2):165–173, 2006.
- [24] S. A. Chowdhury, E. M. Gertz, D. Wangsa, K. Heselmeyer-Haddad, T. Ried, A. A. Schäffer, and R. Schwartz. Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics*, 31(12):i258–67, June 2015.
- [25] S. A. Chowdhury, E. M. Gertz, D. Wangsa, K. Heselmeyer-Haddad, T. Ried, A. A. Schäffer, and R. Schwartz. Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics*, 31(12):i258–i267, 2015.
- [26] W. K. Chu, P. Edge, H. S. Lee, V. Bansal, V. Bafna, X. Huang, and K. Zhang. Ultraaccurate genome sequencing and haplotyping of single human cells. *Proceedings of the National Academy of Sciences*, 114(47):12512–12517, 2017.
- [27] S. Ciccolella, M. S. Gomez, M. Patterson, G. Della Vedova, I. Hajirasouliha, and P. Bonizzoni. Inferring cancer progression from single cell sequencing while allowing loss of mutations. *bioRxiv*, page 268243, 2018.
- [28] I. H. G. S. Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.
- [29] I. Dagogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2):81, 2018.
- [30] O. Delaneau, J.-F. Zagury, and J. Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1):5, 2013.
- [31] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, Feb. 2015.

- [32] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):35, 2015.
- [33] L. Dollo. The laws of evolution. *Bull. Soc. Bel. Geol. Paleontol*, 7:164–166, 1893.
- [34] N. Donmez, S. Malikic, A. W. Wyatt, M. E. Gleave, C. C. Collins, and S. C. Sahinalp. Clonality inference from single tumor samples using low coverage sequence data. In *International Conference on Research in Computational Molecular Biology*, pages 83–94. Springer, 2016.
- [35] P. Edge, V. Bafna, and V. Bansal. Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome research*, 27(5):801–812, 2017.
- [36] P. Eirew, A. Steif, J. Khattra, G. Ha, D. Yap, H. Farahani, K. Gelmon, S. Chia, C. Mar, A. Wan, E. Laks, J. Biele, K. Shumansky, J. Rosner, A. McPherson, C. Nielsen, A. J. L. Roth, C. Lefebvre, A. Bashashati, C. de Souza, C. Siu, R. Aniba, J. Brimhall, A. Oloumi, T. Osako, A. Bruna, J. L. Sandoval, T. Algara, W. Greenwood, K. Leung, H. Cheng, H. Xue, Y. Wang, D. Lin, A. J. Mungall, R. Moore, Y. Zhao, J. Lorette, L. Nguyen, D. Huntsman, C. J. Eaves, C. Hansen, M. a. Marra, C. Caldas, S. P. Shah, and S. Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, Feb. 2015.
- [37] M. El-Kebir. Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.
- [38] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, June 2015.
- [39] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

- [40] M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Systems*, 3(1):43–53, 2016.
- [41] M. El-Kebir, B. J. Raphael, R. Shamir, R. Sharan, S. Zaccaria, M. Zehavi, and R. Zeira. Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology*, 12(1):13, 2017.
- [42] M. El-Kebir, G. Satas, and B. J. Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature genetics*, 50(5):718, 2018.
- [43] D. Eppstein, M. Löffler, and D. Strash. Listing all maximal cliques in sparse graphs in near-optimal time. *CoRR*, abs/1006.5440, 2010. URL <http://arxiv.org/abs/1006.5440>.
- [44] G. F. Estabrook, C. S. Johnson Jr, and F. R. Mc Morris. An idealized concept of the true cladistic character. *Mathematical Biosciences*, 23(3-4):263–272, Apr. 1975.
- [45] G. D. Evrony, X. Cai, E. Lee, L. B. Hills, P. C. Elhosary, H. S. Lehmann, J. Parker, K. D. Atabay, E. C. Gilmore, A. Poduri, et al. Single-neuron sequencing analysis of l1 retrotransposition and somatic mutation in the human brain. *Cell*, 151(3):483–496, 2012.
- [46] H. C. Fan, J. Wang, A. Potanina, and S. R. Quake. Whole-genome molecular haplotyping of single cells. *Nature biotechnology*, 29(1):51–57, 2011.
- [47] D. Fernández-Baca. The perfect phylogeny problem. In *Steiner Trees in Industry*, pages 203–234. Springer, 2001.
- [48] A. Fischer, I. Vázquez-García, C. J. Illingworth, and V. Mustonen. High-definition reconstruction of clonal composition in cancer. *Cell Reports*, 7(5):1740–1752, 2014.
- [49] R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.
- [50] R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479, 2013.



- [51] W. M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.
- [52] W. M. Fitch. On the problem of discovering the most parsimonious tree. *American Naturalist*, pages 223–257, 1977.
- [53] H. N. Gabow and E. W. Myers. Finding All Spanning Trees of Directed and Undirected Graphs. *SIAM Journal on Computing*, 7(3):280–287, 1978.
- [54] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979. ISBN 0716710447.
- [55] T. Garvin, R. Aboukhalil, J. Kendall, T. Baslan, G. S. Atwal, J. Hicks, M. Wigler, and M. C. Schatz. Interactive analysis and assessment of single-cell copy-number variations. *Nature methods*, 12(11):1058, 2015.
- [56] C. Gawad, W. Koh, and S. R. Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175, 2016.
- [57] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, and C. Swanton. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892, Mar. 2012.
- [58] G. Glusman, H. C. Cox, and J. C. Roach. Whole-genome haplotyping approaches and genomic medicine. *Genome medicine*, 6(9):73, 2014.
- [59] K. Govek, C. Sikes, and L. Oesper. A consensus approach to infer tumor evolutionary histories. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 63–72. ACM, 2018.

- [60] M. Griffith, C. A. Miller, O. L. Griffith, K. Krysiak, and Z. L. Skidmore. Optimizing cancer genome sequencing and analysis. *Cell Systems*, 1(3):210–223, 2015.
- [61] G. Gundem, P. Van Loo, B. Kremeyer, L. B. Alexandrov, J. M. C. Tubio, E. Papaemmanuil, D. S. Brewer, H. M. L. Kallio, G. Högnäs, M. Annala, K. Kivinummi, V. Goody, C. Latimer, S. O’Meara, K. J. Dawson, W. Isaacs, M. R. Emmert-Buck, M. Nykter, C. Foster, Z. Kote-Jarai, D. Easton, H. C. Whitaker, ICGC Prostate UK Group, D. E. Neal, C. S. Cooper, R. A. Eeles, T. Visakorpi, P. J. Campbell, U. McDermott, D. C. Wedge, and G. S. Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, Apr. 2015.
- [62] D. Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991. ISSN 1097-0037. doi: 10.1002/net.3230210104. URL <http://dx.doi.org/10.1002/net.3230210104>.
- [63] D. Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.
- [64] D. Gusfield. The multi-state perfect phylogeny problem with missing and removable data: Solutions via integer-programming and chordal graph theory. *Journal of Computational Biology*, 17(3):383–399, 2010.
- [65] D. Gusfield. *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT Press, 2014. ISBN 9780262027526. URL <http://books.google.com/books?id=genMAwAAQBAJ>.
- [66] R. Gysel and D. Gusfield. Extensions and improvements to the chordal graph approach to the multistate perfect phylogeny problem. *IEEE/ACM Trans Comput Biol Bioinform*, 8(4):912–7, 2011. doi: 10.1109/TCBB.2011.27.
- [67] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, L. M. Prentice, N. Melnyk, A. McPherson, A. Bashashati, E. Laks, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, 2014.

- [68] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–86, Jun 2014. doi: 10.1093/bioinformatics/btu284.
- [69] X. Hong, D. G. Scofield, and M. Lynch. Intron size, abundance, and distribution within untranslated regions of genes. *Molecular biology and evolution*, 23(12):2392–2404, 2006.
- [70] L. Huang, F. Ma, A. Chapman, S. Lu, and X. S. Xie. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annual review of genomics and human genetics*, 16:79–102, 2015.
- [71] K. Jahn, J. Kuipers, and N. Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.
- [72] M. Jamal-Hanjani, G. A. Wilson, N. McGranahan, N. J. Birkbak, T. B. Watkins, S. Veeriah, S. Shafi, D. H. Johnson, R. Mitter, R. Rosenthal, et al. Tracking the evolution of non-small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.
- [73] Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37):E5528–E5537, 2016.
- [74] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15:35, 2014. doi: 10.1186/1471-2105-15-35.
- [75] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):35, 2014.
- [76] S. Kannan and T. Warnow. A fast algorithm for the computation and enumeration of perfect phylogenies. *SIAM Journal on Computing*, 26:1749–1763, 1996.
- [77] S. W. Kembel, J. A. Eisen, K. S. Pollard, and J. L. Green. The phylogenetic diversity of metagenomes. *PLoS One*, 6(8):e23214, 2011.

- [78] W. Kuich, editor. *Automata, Languages and Programming, 19th International Colloquium, ICALP92, Vienna, Austria, July 13-17, 1992, Proceedings*, volume 623 of *Lecture Notes in Computer Science*, 1992. Springer. ISBN 3-540-55719-9.
- [79] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome research*, 27(11):1885–1894, 2017.
- [80] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastrioti, P. Lönnerberg, A. Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494, 2018.
- [81] M. D. Leiserson, M. A. Reyna, and B. J. Raphael. A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics*, 32(17):i736–i745, 2016.
- [82] M. L. Leung, Y. Wang, C. Kim, R. Gao, J. Jiang, E. Sei, and N. E. Navin. Highly multiplexed targeted dna sequencing from single nuclei. *Nature protocols*, 11(2):214, 2016.
- [83] M. L. Leung, A. Davis, R. Gao, A. Casasent, Y. Wang, E. Sei, E. Sanchez, D. Maru, S. Kopetz, and N. E. Navin. Single cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome research*, pages gr–209973, 2017.
- [84] B. Li and J. Z. Li. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol*, 15(9):473, 2014.
- [85] S. Ling, Z. Hu, Z. Yang, F. Yang, Y. Li, P. Lin, K. Chen, L. Dong, L. Cao, Y. Tao, L. Hao, Q. Chen, Q. Gong, D. Wu, W. Li, W. Zhao, X. Tian, C. Hao, E. A. Hungate, D. V. T. Catenacci, R. R. Hudson, W.-H. Li, X. Lu, and C.-I. Wu. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):E6496–505, Nov. 2015.

- [86] M. A. Lodato, M. B. Woodworth, S. Lee, G. D. Evrony, B. K. Mehta, A. Karger, S. Lee, T. W. Chittenden, A. M. D’Gama, X. Cai, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–98, 2015.
- [87] P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A. Reshef, H. K. Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443, 2016.
- [88] E. Loo, P. Khalili, K. Beuhler, I. Siddiqi, and M. A. Vasef. Braf v600e mutation across multiple tumor types: correlation between dna-based sequencing and mutation-specific immunohistochemistry. *Applied immunohistochemistry & molecular morphology*, 26(10):709–713, 2018.
- [89] M. Łuksza, N. Riaz, V. Makarov, V. P. Balachandran, M. D. Hellmann, A. Solovyov, N. A. Rizvi, T. Merghoub, A. J. Levine, T. A. Chan, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*, 551(7681):517, 2017.
- [90] J. Ma, A. Ratan, B. J. Raney, B. B. Suh, W. Miller, and D. Haussler. The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, 105(38):14254–14261, 2008.
- [91] S. Malikic. Clonality inference in multiple tumor samples using phylogeny. Master’s thesis, Simon Fraser University, 2014.
- [92] S. Malikic, A. W. McPherson, N. Donmez, and C. S. Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, Jan 2015. doi: 10.1093/bioinformatics/btv003.
- [93] S. Malikic, A. W. McPherson, N. Donmez, and C. S. Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 2015.
- [94] S. Malikic, K. Jahn, J. Kuipers, S. C. Sahinalp, and N. Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):2750, 2019.

- [95] S. Malikic, F. R. Mehrabadi, S. Ciccolella, M. K. Rahman, C. Ricketts, E. Haghshenas, D. Seidman, F. Hach, I. Hajirasouliha, and S. C. Sahinalp. Phiscs: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research*, 2019.
- [96] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906, 2007.
- [97] Y. Marcy, C. Ouverney, E. M. Bik, T. Lösekann, N. Ivanova, H. G. Martin, E. Szeto, D. Platt, P. Hugenholtz, D. A. Relman, et al. Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated tm7 microbes from the human mouth. *Proceedings of the National Academy of Sciences*, 104(29):11889–11894, 2007.
- [98] M. J. McConnell, M. R. Lindberg, K. J. Brennand, J. C. Piper, T. Voet, C. Cowing-Zitron, S. Shumilina, R. S. Lasken, J. R. Vermeesch, I. M. Hall, et al. Mosaic copy number variation in human neurons. *Science*, 342(6158):632–637, 2013.
- [99] N. McGranahan, F. Favero, E. C. de Bruin, N. J. Birkbak, Z. Szallasi, and C. Swanton. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, 7(283):283ra54–283ra54, Apr. 2015.
- [100] N. McGranahan, F. Favero, E. C. de Bruin, N. J. Birkbak, Z. Szallasi, and C. Swanton. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science translational medicine*, 7(283):283ra54–283ra54, 2015.
- [101] N. McGranahan, A. J. S. Furness, R. Rosenthal, S. Ramskov, R. Lyngaa, S. K. Saini, M. Jamal-Hanjani, G. A. Wilson, N. J. Birkbak, C. T. Hiley, T. B. K. Watkins, S. Shafi, N. Murugaesu, R. Mitter, A. U. Akarca, J. Linares, T. Marafioti, J. Y. Henry, E. M. Van Allen, D. Miao, B. Schilling, D. Schadendorf, L. A. Garraway, V. Makarov, N. A. Rizvi, A. Snyder, M. D. Hellmann, T. Merghoub, J. D. Wolchok, S. A. Shukla, C. J. Wu, K. S. Peggs, T. A.

- Chan, S. R. Hadrup, S. A. Quezada, and C. Swanton. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, page aaf1490, Mar. 2016.
- [102] A. McPherson, A. Roth, E. Laks, T. Masud, A. Bashashati, A. W. Zhang, G. Ha, J. Biele, D. Yap, A. Wan, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature genetics*, 48(7):758, 2016.
- [103] C. A. Miller et al. Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665, Aug 2014. doi: 10.1371/journal.pcbi.1003665.
- [104] Mission Bio. Copy number variants and single nucleotide variants simultaneously detected in single cells. [https://missionbio.com/cnv\\_application\\_note](https://missionbio.com/cnv_application_note). Accessed: 2019-Nov-05.
- [105] S. Miura, T. Vu, J. Deng, T. Buturla, J. Choi, and S. Kumar. Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. *bioRxiv*, page 697318, 2019.
- [106] M. A. Myers, G. Satas, and B. J. Raphael. Calder: Inferring phylogenetic trees from longitudinal tumor samples. *Cell systems*, 2019.
- [107] N. E. Navin. Cancer genomics: one cell at a time. *Genome Biol*, 15(8):452, 2014. doi: 10.1186/s13059-014-0452-9.
- [108] N. E. Navin. Delineating cancer evolution with single-cell sequencing. *Sci Transl Med*, 7(296):296fs29, Jul 2015. doi: 10.1126/scitranslmed.aac8319.
- [109] N. E. Navin. The first five years of single-cell cancer genomics and beyond. *Genome Research*, 25(10):1499–1507, Oct. 2015.
- [110] N. E. Navin. The first five years of single-cell cancer genomics and beyond. *Genome research*, 25(10):1499–1507, 2015.

- [111] D. E. Newburger, D. Kashef-Haghighi, Z. Weng, R. Salari, R. T. Sweeney, A. L. Brunner, S. X. Zhu, X. Guo, S. Varma, M. L. Troxell, R. B. West, S. Batzoglou, and A. Sidow. Genome evolution during progression to breast cancer. *Genome Research*, 23(7):1097–1108, July 2013.
- [112] D. E. Newburger et al. Genome evolution during progression to breast cancer. *Genome Res*, 23(7):1097–108, Jul 2013. doi: 10.1101/gr.151670.112.
- [113] S. Nik-Zainal et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, May 2012. doi: 10.1016/j.cell.2012.04.023.
- [114] P. C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–8, Oct 1976.
- [115] L. Oesper, A. Mahmoody, and B. J. Raphael. Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol*, 14(7):R80, 2013. doi: 10.1186/gb-2013-14-7-r80.
- [116] L. Oesper, A. Mahmoody, and B. J. Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biology*, 14(7):R80, 2013.
- [117] L. Oesper, G. Satas, and B. J. Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–40, Dec 2014. doi: 10.1093/bioinformatics/btu651.
- [118] M. Patterson, T. Marschall, N. Pisanti, L. Van Iersel, L. Stougie, G. W. Klau, and A. Schönhuth. Whatshap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, 22(6):498–509, 2015.
- [119] I. Pe’er, T. Pupko, R. Shamir, and R. Sharan. Incomplete directed perfect phylogeny. *SIAM Journal on Computing*, 33(3):590–607, 2004.
- [120] A. J. Picher, B. Budeus, O. Wafzig, C. Krüger, S. García-Gómez, M. I. Martínez-Jiménez, A. Díaz-Talavera, D. Weber, L. Blanco, and A. Schneider. Trueprime is a novel method for



- whole-genome amplification from single cells based on tthprimpol. *Nature communications*, 7:13296, 2016.
- [121] Y. Pirola, S. Zaccaria, R. Dondi, G. W. Klau, N. Pisanti, and P. Bonizzoni. Hapcol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics*, 32(11):1610–1617, 2015.
  - [122] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1):91, 2015.
  - [123] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1):91, 2015.
  - [124] D. Porubský, A. D. Sanders, N. van Wietmarschen, E. Falconer, M. Hills, D. C. Spierings, M. R. Bevoa, V. Guryev, and P. M. Lansdorp. Direct chromosome-length haplotyping by single-cell sequencing. *Genome research*, 26(11):1565–1574, 2016.
  - [125] D. Pradhan and M. El-Kebir. On the non-uniqueness of solutions to the perfect phylogeny mixture problem. In *RECOMB International conference on Comparative Genomics*, pages 277–293. Springer, 2018.
  - [126] B. J. Raphael, J. R. Dobson, L. Oesper, and F. Vandin. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med*, 6(1):5, 2014. doi: 10.1186/gm524.
  - [127] J. C. Roach, G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.
  - [128] M. Rose-Zerilli, J. Gibson, J. Wang, W. Tapper, Z. Davis, H. Parker, M. Larrayoz, H. McCarthy, R. Walewska, J. Forster, et al. Longitudinal copy number, whole exome and targeted deep sequencing of good risk ighv-mutated cll patients with progressive disease. *Leukemia*, 2016.

- [129] E. M. Ross and F. Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):69, 2016.
- [130] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nat Methods*, 11(4):396–8, Apr 2014. doi: 10.1038/nmeth.2883.
- [131] J. Z. Sanborn, J. Chung, E. Purdom, N. J. Wang, H. Kakavand, J. S. Wilmott, T. Butler, J. F. Thompson, G. J. Mann, L. E. Haydu, et al. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proceedings of the National Academy of Sciences*, 112(35):10995–11000, 2015.
- [132] G. Satas and B. J. Raphael. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, 33(14):i152–i160, 2017.
- [133] A. Schuh et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–6, Nov 2012. doi: 10.1182/blood-2012-05-433540.
- [134] R. F. Schwarz, A. Trinh, B. Sipos, J. D. Brenton, N. Goldman, and F. Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS computational biology*, 10(4):e1003535, 2014.
- [135] N. Segata, D. Börnigen, X. C. Morgan, and C. Huttenhower. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4, 2013.
- [136] M. A. Sherman, A. R. Barton, M. A. Lodato, C. Vitzthum, M. E. Coulter, C. A. Walsh, and P. J. Park. Pasd-qc: Quality control for single cell whole-genome sequencing data using power spectral density estimation. *Nucleic acids research*, 2017.
- [137] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

- [138] J. Singer, J. Kuipers, K. Jahn, and N. Beerenwinkel. Single-cell mutation identification via phylogenetic inference. *Nature communications*, 9(1):5144, 2018.
- [139] P. Skums, V. Tsyvina, and A. Zelikovsky. Inference of clonal selection in cancer populations using single-cell sequencing data. *Bioinformatics*, page i398–i407, 2019.
- [140] D. I. Smith, Y. Zhu, S. McAvoy, and R. Kuhn. Common fragile sites, extremely large genes, neural development and cancer. *Cancer letters*, 232(1):48–57, 2006.
- [141] A. Sottoriva, H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press, D. Shibata, and C. Curtis. A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209–216, Mar. 2015.
- [142] A. Stengel, W. Kern, T. Haferlach, M. Meggendorfer, A. Fasan, and C. Haferlach. The impact of tp53 mutations and tp53 deletions on survival varies between aml, all, mds and cll: an analysis of 3307 cases. *Leukemia*, 31(3):705, 2017.
- [143] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.
- [144] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):e165, 2013. doi: 10.1093/nar/gkt641. URL <http://nar.oxfordjournals.org/content/41/17/e165.abstract>.
- [145] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res*, 41(17):e165, Sep 2013. doi: 10.1093/nar/gkt641.
- [146] D. P. Tabassum and K. Polyak. Tumorigenesis: it takes a village. *Nature Reviews Cancer*, 15(8):473, 2015.
- [147] A. M. Taylor, J. Shih, G. Ha, G. F. Gao, X. Zhang, A. C. Berger, S. E. Schumacher, C. Wang,

- H. Hu, J. Liu, et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer cell*, 33(4):676–689, 2018.
- [148] R. Tewhey, V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215, 2011.
- [149] S. T. Tokdar and R. E. Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.
- [150] S. Turajlic, H. Xu, K. Litchfield, A. Rowan, T. Chambers, J. I. Lopez, D. Nicol, T. O’Brien, J. Larkin, S. Horswell, et al. Tracking cancer evolution reveals constrained routes to metastases: Tracerx renal. *Cell*, 173(3):581–594, 2018.
- [151] S. Turajlic, H. Xu, K. Litchfield, A. Rowan, S. Horswell, T. Chambers, T. O’Brien, J. I. Lopez, T. B. Watkins, D. Nicol, et al. Deterministic evolutionary trajectories influence primary tumor growth: Tracerx renal. *Cell*, 173(3):595–610, 2018.
- [152] M. van de Ven, J.-O. Andressoo, G. T. van der Horst, J. H. Hoeijmakers, and J. R. Mitchell. Effects of compound heterozygosity at the xpd locus on cancer and ageing in mouse models. *DNA repair*, 11(11):874–883, 2012.
- [153] Y. Wang, J. Waters, M. L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H. Liang, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155, 2014.
- [154] Y. Wang et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–60, Aug 2014. doi: 10.1038/nature13600.
- [155] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.
- [156] E. O. Wiley and B. S. Lieberman. *Phylogenetics: theory and practice of phylogenetic systematics*. John Wiley & Sons, 2011.

- [157] M. J. Williams, B. Werner, T. Heide, C. Curtis, C. P. Barnes, A. Sottoriva, and T. A. Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature genetics*, 50(6):895, 2018.
- [158] T. Wu, V. Moulton, and M. Steel. Refining phylogenetic trees given additional data: An algorithm based on parsimony. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(1):118–125, 2009.
- [159] X. Xu, Y. Hou, X. Yin, L. Bao, A. Tang, L. Song, F. Li, S. Tsang, K. Wu, H. Wu, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–895, 2012.
- [160] K. Yuan, T. Sakoparnig, F. Markowetz, and N. Beerenwinkel. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1):36, 2015.
- [161] S. Zaccaria, M. El-Kebir, G. W. Klau, and B. J. Raphael. Phylogenetic copy-number factorization of multiple tumor samples. *Journal of Computational Biology*, 25(7):689–708, 2018.
- [162] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhong, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, 45(10):1134–40, Oct 2013. doi: 10.1038/ng.2760.
- [163] H. Zafar, A. Tzen, N. Navin, K. Chen, and L. Nakhleh. Sifit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome biology*, 18(1):178, 2017.
- [164] H. Zafar, N. Navin, K. Chen, and L. Nakhleh. Siclonfit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Research*, 2019.
- [165] H. Zahn, A. Steif, E. Laks, P. Eirew, M. VanInsberghe, S. P. Shah, S. Aparicio, and C. L. Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature methods*, 14(2):167, 2017.

- [166] H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau, and W. S. Noble. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol*, 10(7):e1003703, Jul 2014. doi: 10.1371/journal.pcbi.1003703.
- [167] A. W. Zhang, A. McPherson, K. Milne, D. R. Kroeger, P. T. Hamilton, A. Miranda, T. Funnell, N. Little, C. P. de Souza, S. Laan, et al. Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. *Cell*, 173(7):1755–1769, 2018.
- [168] C.-Z. Zhang, V. A. Adalsteinsson, J. Francis, H. Cornils, J. Jung, C. Maire, K. L. Ligon, M. Meyerson, and J. C. Love. Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat Commun*, 6:6822, 2015. doi: 10.1038/ncomms7822.
- [169] C.-Z. Zhang, V. A. Adalsteinsson, J. Francis, H. Cornils, J. Jung, C. Maire, K. L. Ligon, M. Meyerson, and J. C. Love. Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nature communications*, 6, 2015.
- [170] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, and A. Kasprzyk. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database (Oxford)*, 2011:bar026, 2011. doi: 10.1093/database/bar026.
- [171] G. X. Zheng, B. T. Lau, M. Schnall-Levin, M. Jarosz, J. M. Bell, C. M. Hindson, S. Kyriazopoulou-Panagiotopoulou, D. A. Masquelier, L. Merrill, J. M. Terry, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*, 34(3):303, 2016.

# Appendix A

## Supplementary Methods and Results for Chapter 2

### A.1 Supplementary Methods

#### A.1.1 Sequential Importance Sampling

Importance sampling functions well when the number of vertices in the tree  $k$  and the number of samples  $m$  are low. However, the number of dimensions we are sampling from for  $\mathbf{F}$  increases as  $k \cdot m$ . Thus, in the presence of noise, sometimes, for all samples  $\bar{F}$ ,  $\mathcal{T}_{\bar{F}}$  is empty, and thus the calculated posterior probability of all trees is 0. This occurs when the largest tree that  $\bar{F}$  admits,  $T_{\bar{F}}$  has less than  $k$  nodes,  $|V(T_{\bar{F}})| < k$ . However, some of these samples may in fact be near another value  $\bar{F}' \approx \bar{F}$  such that there exists a  $(T, \pi)$  for which  $\bar{F}'$  respects the Sum Condition. As the number of nodes in the largest tree admitted by samples increases, we would expect we are nearing such a value.

We generalize importance sampling to use multiple proposal distributions,  $Q_1, \dots, Q_N$ , where  $Q_i$  is the distribution used at step  $i$ . As the estimate obtained from each distribution  $Q_i$  is an unbiased estimator of  $\int P(\mathbf{X})d\mathbf{X}$ , the mean of these values is also an unbiased estimator.

$$\int P(\mathbf{X})d\mathbf{X} \approx \frac{1}{N} \sum_{i=1}^N \frac{P(\bar{X}_i)}{Q_i(\bar{X}_i)}. \quad (\text{A.1})$$

As such, we propose an MCMC inspired multiple-importance sampling approach.  $Q_0$  is the proposal distribution supplied by SciClone as described above. Let  $\text{Beta}(\bar{F})$  be the beta distribution with the values  $\bar{F}$  as the means, and matrix  $\Sigma$  as the variances. At step  $i + 1$ ,

1. let  $Q_{i+1} = \text{Beta}(\bar{F}_i)$  if  $|V(T_{\bar{F}_i})| > |V(T_{\bar{F}_{i-1}})|$  or with probability

$$p = \min \left( \frac{\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \bar{F}^{(i)})}{\Pr(\mathbf{A} = A \mid \mathbf{D} = D, \mathbf{F} = \bar{F}^{(i)})}, 1 \right).$$

2. Let  $Q_{i+1} = Q_0$  with probability  $\epsilon$ .

3. Else  $Q_{i+1} = Q_i$

This approach uses information from previous samples to find samples which may have finite probability. Unlike MCMC, this approach is able to use information outside of model likelihood, i.e. the unconstrained clustering information and the tree size, to guide sampling.

### A.1.2 Comparison to Other Models



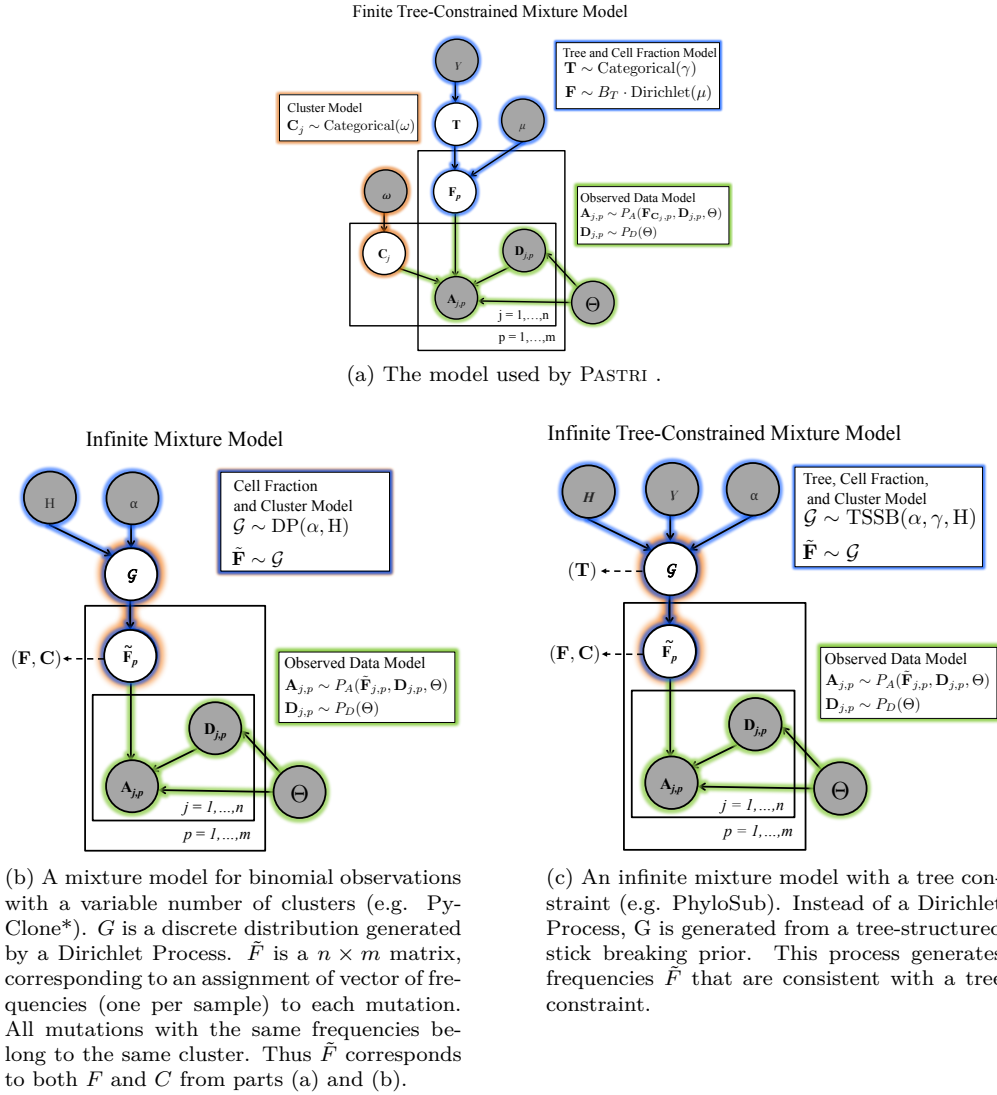


Figure A.1: Comparison of models for mixtures with binomial observations.

## A.2 Supplementary Results

### A.2.1 Additional Simulation Results

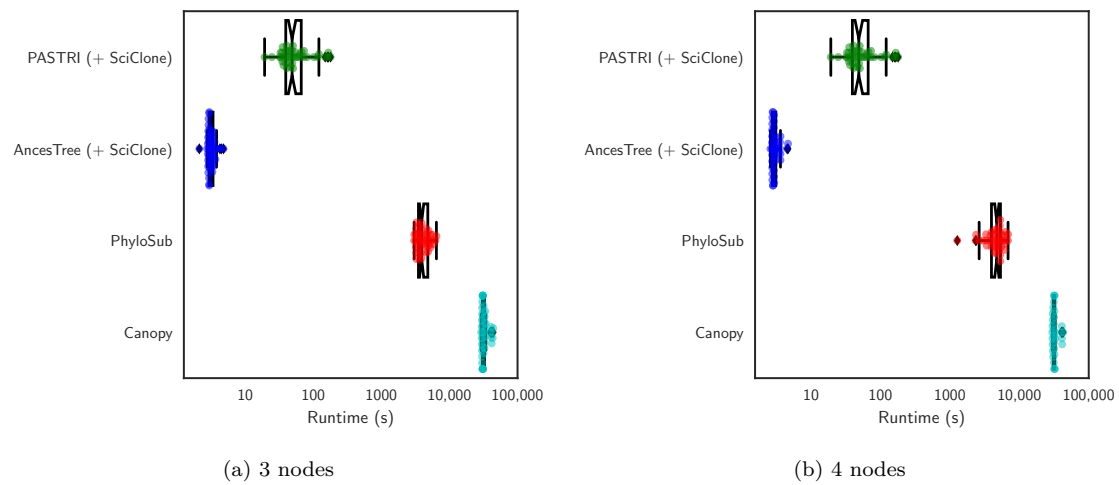


Figure A.2: **Runtime.**

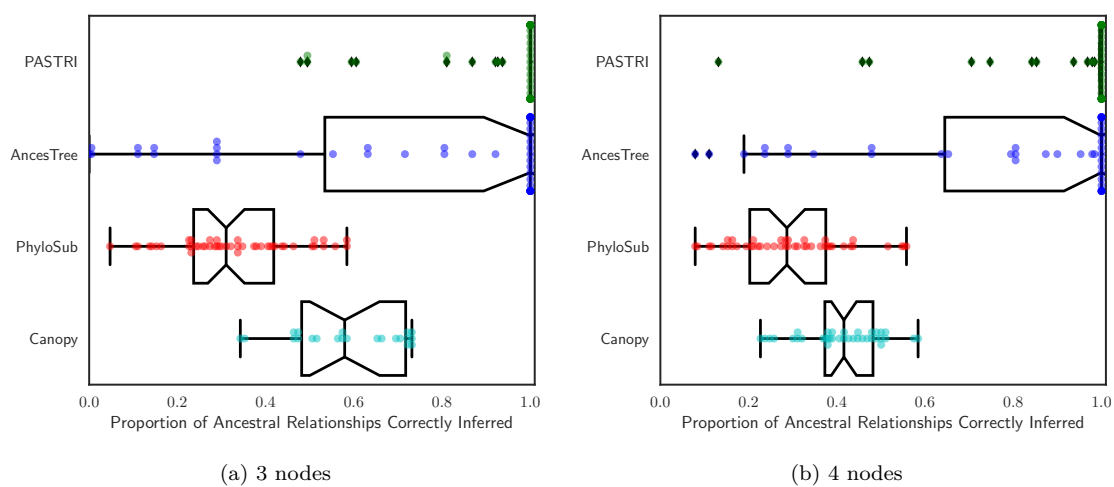
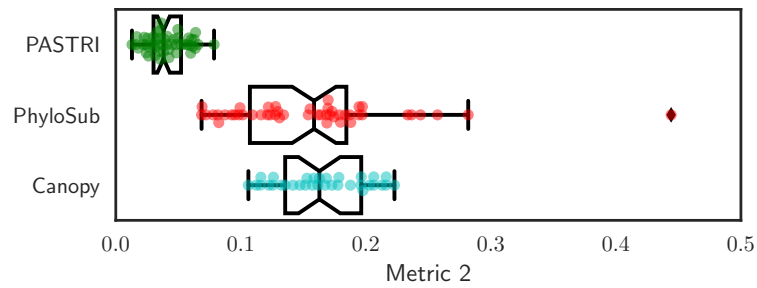
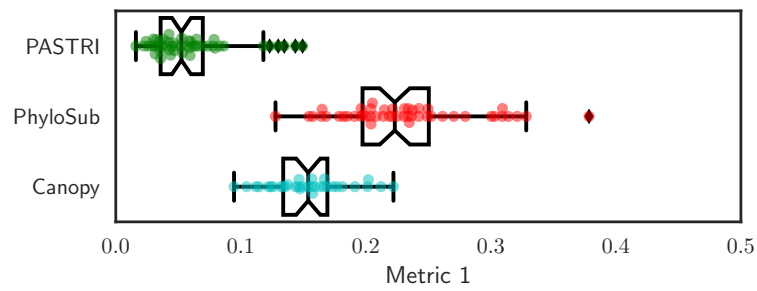
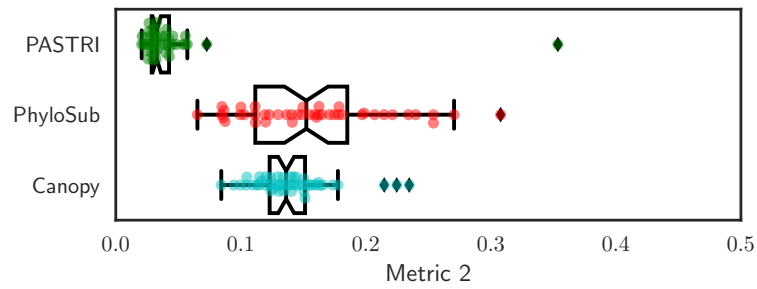
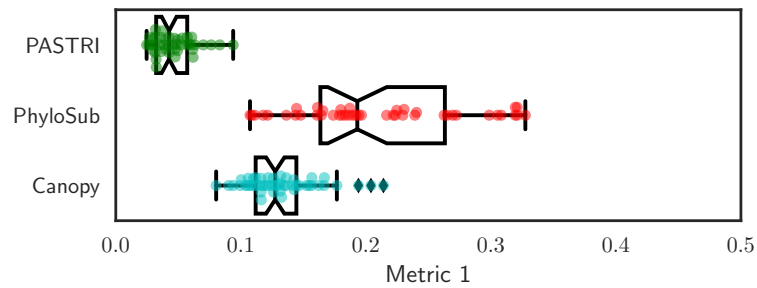


Figure A.3: **Accuracy in recovering ancestral relationships.**



(a) 3 nodes



(b) 4 nodes

Figure A.4: Accuracy in recovering cluster frequencies.

## Appendix B

# Supplementary Methods and Results for Chapter 3

### B.1 Supplementary Methods

In this section we present additional definitions and results. We start by introducing the Perfect Phylogeny Mixture Deconvolution Problem (PPMDP) in Section B.1.1. We show in Section B.1.2 that there exists a unique matrix  $U$  relating a given frequency tensor  $\mathcal{F}$  and complete perfect phylogeny tree  $T$ . Next, we derive in Section B.1.3 that solutions to the PPMDP correspond to constrained spanning trees in a directed, edge-labeled multi-graph. In Section B.1.4 we prove that the problem is NP-complete. We then proceed to consider a restricted version that consider cladistic characters in Section B.1.5. For this version, we present an enumeration algorithm in Section B.1.6. Finally, we tie things back to our cancer application in Section B.1.7, where we introduce a multi-state model that captures the somatic mutational process in cancer.

### B.1.1 The Perfect Phylogeny Mixture Deconvolution Problem

Recall that  $m$  is the number of samples,  $n$  is the number of characters and  $k$  is the number of states of each character. Our input measurements are given by the  $k \times m \times n$  frequency tensor  $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$  where  $f_{p,(c,i)}$  is the proportion of taxa of sample  $p$  that have state  $i$  for character  $c$ . We denote by  $F_i$  the slice of  $\mathcal{F}$  where the state of each character is  $i$ . Formally,  $\mathcal{F}$  is defined as follows.

**Definition 3.** An  $k \times m \times n$  tensor  $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$  is a frequency tensor provided  $f_{p,(c,i)} \geq 0$  and  $\sum_{i=0}^{k-1} f_{p,(c,i)} = 1$  for all characters  $c$  and samples  $p$ .

As mentioned in Section 3.2.1, the goal is to explain the observed frequencies  $\mathcal{F}$  as  $m$  mixtures of the leaves of a perfect phylogeny tree  $T$ , where each mixture corresponds to one sample. We recall the definition of a perfect phylogeny[47, 65].

**Definition 4.** A rooted tree  $T$  is a perfect phylogeny tree provided that (1) each vertex is labeled by a state vector in  $\{0, \dots, k-1\}^n$ , which denotes the state for each character; (2) the root vertex of  $T$  has state 0 for each character; (3) vertices labeled with state  $i$  for character  $c$  form a connected subtree  $T_{(c,i)}$  of  $T$ .

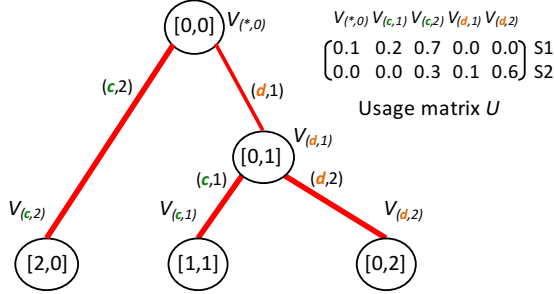
Rather than explaining  $\mathcal{F}$  as mixtures of the leaves of a perfect phylogeny tree, we aim to explain  $\mathcal{F}$  as  $m$  mixtures of all vertices of an  $n, k$ -complete perfect phylogeny tree, which is defined as follows.

**Definition 5.** An edge-labeled rooted tree  $T$  on  $n(k-1) + 1$  vertices is a  $n, k$ -complete perfect phylogeny tree provided each of the  $n(k-1)$  edges is labeled with exactly one character-state pair from  $[n] \times [k-1]$  and no character-state pair appears more than once in  $T$ . Let  $\mathcal{T}_{n,k}$  be the set of all  $n, k$ -complete perfect phylogeny trees.

We may do this without loss of generality, as each  $n, k$ -complete perfect phylogeny  $T$  can be mapped to a perfect phylogeny tree  $T'$  by extending inner vertices of  $T$  that have non-zero mixing proportions to leaves of  $T'$ . See Figure B.1 for an example. In the following we denote by  $T$  an  $n, k$ -complete perfect phylogeny tree  $T$ , by  $v_{(c,i)}$  the vertex of  $T$  whose incoming edge is labeled by  $(c, i)$ , and by

$v_{(*,i)}$  the root of  $T$ . Alternatively,  $v_{(1,0)}, \dots, v_{(n,0)}$  all refer to the root vertex  $v_{(*,0)}$ .

(A) Complete Perfect Phylogeny Tree  $T$



(B) Perfect Phylogeny Tree  $T'$

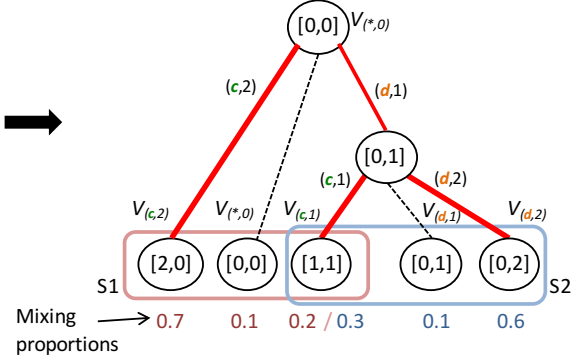


Figure B.1: Each complete perfect phylogeny tree (A) corresponds to a perfect phylogeny tree (B)

The observed frequencies  $\mathcal{F}$  are related to the vertices of  $T$  by an  $m \times (n(k-1) + 1)$  usage matrix  $U = [u_{p,(c,i)}]$  whose rows correspond to samples and columns to vertices of  $T$  such that each entry  $u_{p,(c,i)}$  indicates the mixing proportion of the vertex  $v_{(c,i)}$  of  $T$  in sample  $p$ . More specifically, each frequency  $f_{p,(c,i)}$  is the sum of mixing proportions of all vertices of  $T$  that possess state  $i$  for character  $c$ , i.e.  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$  (Figure B.2a). Formally, we define a usage matrix  $U$  as follows.

**Definition 6.** An  $m \times (n(k-1) + 1)$  matrix  $U = [u_{p,(c,i)}]$  is an  $m, n, k$ -usage matrix provided  $u_{p,(c,i)} \geq 0$ , and  $u_{p,(*,0)} + \sum_{c=1}^n \sum_{i=1}^{k-1} u_{p,(c,i)} = 1$  for all samples  $p$ . Let  $\mathcal{U}_{m,n,k}$  be the set of all  $m, n, k$ -usage matrices  $U$ .

Given  $\mathcal{F}$ , the goal is to infer a  $n, k$ -complete perfect phylogeny  $T$  and a usage matrix  $U$  such that mixing the vertices of  $T$  according to  $U$  results in  $\mathcal{F}$ , which was stated as Problem 1 in the main text.

**(Main Text) Problem 1** (Perfect Phylogeny Mixture Deconvolution Problem (PPMDP)). Given an  $k \times m \times n$  frequency tensor  $\mathcal{F} = [[f_{p,(c,i)}]]$ , find a  $n, k$ -complete perfect phylogeny tree  $T$  and a  $m, n, k$ -usage matrix  $U = [u_{p,(c,i)}]$  such that  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$  for all character-state pairs  $(c, i)$  and all samples  $p$ .

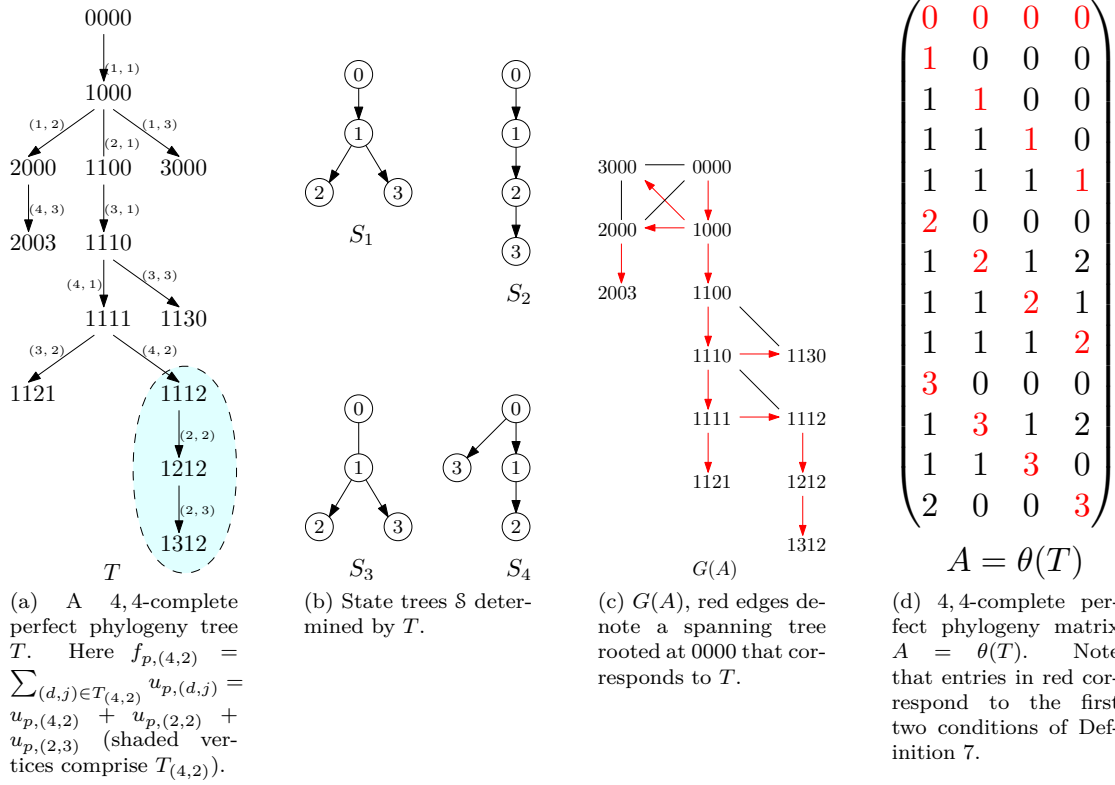


Figure B.2: Concepts of the Perfect Phylogeny Mixture Deconvolution Problem

### Two-State Perfect Phylogeny Mixtures

We now review and recast the main results for mixtures of a two-state ( $k = 2$ ) perfect phylogeny with all-zero root. Here, a character changes state from 0 to 1 at most once. The key insight is that the relative frequencies of the mutated ( $= 1$ ) states for a subset of characters constrain their potential ancestral relationships. This is because the mutated state persists in the tree. In particular, if  $(c, 1) \prec_T (d, 1)$  then all vertices in  $T$  that have state 1 for character  $d$  must also have state 1 for character  $c$ . A consequence is the following condition, called the *Ancestry Condition* (AC) in [38]:

$$f_{p,(c,1)} \geq f_{p,(d,1)} \text{ for all samples } p \text{ and characters } (c, 1) \prec_T (d, 1). \quad (\text{AC})$$



In fact, a stronger condition than the ancestry condition can be derived by considering the relationships between subtrees of  $T$ . Specifically, for each character  $c$ , the subtree  $T_{(c,1)}$ , consisting of all vertices with state 1 for character  $c$ , is *identical to* the subtree  $\bar{T}_{(c,1)}$  rooted at a vertex  $v_{(c,1)}$ . Moreover,  $\bar{T}_{(c,1)}$  is the disjoint union of  $v_{(c,1)}$  and the subtrees rooted at its children:  $\bar{T}_{(c,1)} = v_{(c,1)} \cup \left( \bigcup_{(d,1) \in \delta(c,1)} \bar{T}_{(d,1)} \right)$  (Figure B.15(a)). Combining this fact together with the equation  $f_{p,(c,1)} = \sum_{(d,1) \in T_{(c,1)}} u_{p,(d,1)}$  yields the key equation  $f_{p,(c,1)} = u_{p,(c,1)} + \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)}$ . Recalling that  $u_{p,(c,1)} \geq 0$ , we can relax this equation to the following inequality, referred to as the *Sum Condition* (SC) in [38],

$$f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \quad \text{for all samples } p \text{ and characters } c. \quad (\text{SC})$$

The sum condition is both necessary and sufficient for  $\mathcal{F}$  to be a mixture of  $T$ . The sum and ancestry conditions provide a combinatorial characterization of solutions as constrained spanning trees of a directed acyclic graph, which was called the *ancestry graph* in [38]. This derivation of the sum condition from the fact that  $\bar{T}_{(c,1)} = T_{(c,1)}$  in the two-state case is not explicitly stated in previous work, but this turns out to be the key ingredient in the generalization to the multi-state case.

### Reformulating Problem 1

In the remainder of this section we show how Problem 1 can be restated as a linear algebra problem. We start by observing that each vertex  $v_{(c,i)}$  of  $T$  defines a *state vector*  $\mathbf{a}_{(c,i)} \in \{0, \dots, k-1\}^n$  indicating the state of each character at that vertex. The root vertex  $v_{(*,0)}$  has state vector  $\mathbf{a}_{(*,0)} = (0, \dots, 0)$ , i.e.  $a_{(*,0),d} = 0$  for each character  $d \in [n]$ . The state vector  $\mathbf{a}_{(c,i)}$  of the remaining vertices  $v_{(c,i)} \neq v_{(*,0)}$  is the same as the state vector  $\mathbf{a}_{\pi(c,i)}$  of the parent vertex  $v_{\pi(c,i)}$  except at character  $c$  where the state is  $i$ . The state vectors of all the vertices of an  $n, k$ -complete perfect phylogeny tree  $T$  correspond to an  $(n(k-1) + 1) \times n$  matrix  $A$  (Figure B.2d).

We now define a subset of matrices  $A \in \{0, \dots, k-1\}^{(n(k-1)+1) \times n}$  that we call  *$n, k$ -complete perfect*

*phylogeny matrices* whose rows encode the state vectors of the vertices of an  $n, k$ -complete perfect phylogeny tree  $T$ . Let  $A \in \{0, \dots, k-1\}^{(n(k-1)+1) \times n}$ . We define  $G(A)$  as the undirected graph whose vertices correspond to the rows of  $A$ , and whose edges set consists of all pairs of vertices whose corresponding state vectors differ at exactly one position, i.e. have Hamming distance 1. We require that  $G(A)$  is connected (Figure B.2c), that every row  $\mathbf{a}_{(c,i)}$  of  $A$  (where  $i \in [k]$ ) introduces the character-state pair  $(c, i)$  and that there is a row  $\mathbf{a}_{(*,0)}$  that contains only 0-s (Figure B.2d). Formally, we say that  $A$  is an  $n, k$ -complete perfect phylogeny matrix if the following holds.

**Definition 7.** *Matrix  $A = [a_{(c,i),d}] \in \{0, \dots, k-1\}^{(n(k-1)+1) \times n}$  is a  $n, k$ -complete perfect phylogeny matrix provided  $a_{(*,0),d} = 0$  for all characters  $d$ ,  $a_{(c,i),c} = i$  for all character-state pairs  $(c, i)$  and  $G(A)$  is connected. Let  $\mathcal{A}_{n,k}$  be the set of all  $n, k$ -complete perfect phylogeny matrices.*

Unlike the general multi-state perfect phylogeny problem [78], we can recognize complete perfect phylogeny matrices in polynomial time, as these matrices form a restricted subset of multi-state perfect phylogeny matrices whose rows unambiguously encode all the vertices of a corresponding tree. We relate complete perfect phylogeny trees to complete perfect phylogeny matrices by defining the following function.

**Definition 8.** *The function  $\theta : \mathcal{T}_{n,k} \rightarrow \mathcal{A}_{n,k}$  maps a complete perfect phylogeny tree  $T \in \mathcal{T}_{n,k}$  to the complete perfect phylogeny matrix  $\theta(T) = A = [a_{(c,i),d}]$  where*

$$a_{(c,i),d} = \begin{cases} 0, & \text{if } (c, i) = (*, 0), \\ i, & \text{if } d = c, \\ a_{\pi(c,i),d}, & \text{if } d \neq c. \end{cases} \quad (\text{B.1})$$

**Lemma 2.** *The function  $\theta : \mathcal{T}_{n,k} \rightarrow \mathcal{A}_{n,k}$  is a surjection.*

*Proof.* The set of complete perfect phylogeny trees corresponding to a matrix  $A \in \mathcal{A}_{n,k}$  is exactly the set of spanning trees of  $G(A)$  rooted at  $v_{(*,0)}$ . This set is nonempty as by Definition 7,  $G(A)$  is connected and thus has at least one spanning tree for any  $A \in \mathcal{A}_{n,k}$ .  $\square$   $\square$

We now have the following convenient parameterization of the problem. We define matrix  $A_i = [a_{(d,j),c}^i]$  as

$$a_{(d,j),c}^i = \begin{cases} 1, & \text{if } a_{(d,j),c} = i, \\ 0, & \text{otherwise.} \end{cases}$$

Note that  $\sum_{i=0}^{k-1} iA_i = A$ . Since each sample is a mixture of the vertices of  $T$ , captured by the complete perfect phylogeny matrix  $A$ , with proportions defined in the usage matrix  $U$ , the observed frequency tensor  $\mathcal{F} = [F_i]$  satisfies

$$F_i = UA_i \tag{B.2}$$

for all states  $i \in \{0, \dots, k-1\}$ . Assuming no errors in  $\mathcal{F}$ , our goal is thus to find  $U \in \mathcal{U}_{m,n,k}$  and  $A \in \mathcal{A}_{n,k}$  satisfying (B.2). We thus may restate Problem 1 as follows.

**Problem 1** (Perfect Phylogeny Mixture Deconvolution Problem (PPMDP)). *Given an  $k \times m \times n$  frequency tensor  $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$ , find a  $n, k$ -complete perfect phylogeny tree  $T$  and a  $m, n, k$ -usage matrix  $U = [u_{p,(c,i)}]$  such that  $F_i = UA_i$  for all  $i \in \{0, \dots, k-1\}$  where  $A = \theta(T)$ .*

### B.1.2 Uniqueness of $U$ given $F$ and $T$

Remarkably,  $\mathcal{F} = [F_i]$  and  $A \in \mathcal{A}_{n,k}$  *uniquely* define the matrix  $U$  such that  $F_i = UA_i$  for all states  $i$  as we prove in the following.

We start by defining a set of  $(n(k-1)+1) \times nk$  binary matrices  $B_{n,k}$  that are in 1-1 correspondence to  $\mathcal{A}_{n,k}$ . We do so by defining the undirected graph  $H(B)$  for a matrix  $B \in \{0,1\}^{(n(k-1)+1) \times nk}$ . The vertices of  $H(B)$  correspond to the rows of  $B$  and there is an edge in  $H(B)$  if and only if the two corresponding state vectors differ at exactly two positions, i.e. have Hamming distance 2. Formally, we define a *binary  $n, k$ -complete perfect phylogeny matrix* as follows.

**Definition 9.** *A matrix  $B = [b_{(d,j),(c,i)}] \in \{0,1\}^{(n(k-1)+1) \times nk}$  matrix is a binary  $n, k$ -complete perfect phylogeny matrix provided*

- $\sum_{c=1}^n \sum_{i=0}^{k-1} b_{(d,j),(c,i)} = n$  where  $(d,j) \in [n] \times [k-1]$ ,

- $\sum_{c=1}^n b_{(d,j),(c,i)} = 1$  where  $(d,j) \in [n] \times [k-1]$  and  $i \in \{0, \dots, k-1\}$ ,
- $b_{(*,0),(c,0)} = 1$  where  $c \in [n]$ ,
- $b_{(c,i),(c,i)} = 1$  for all  $(c,i) \in [k] \times [n-1]$  and
- $H(B)$  is connected.

Let  $\mathcal{B}_{n,k}$  be the set of all binary  $n, k$ -complete perfect phylogeny matrices.

We now define the following function  $\psi$  that maps a complete perfect phylogeny matrix  $A$  to a binary matrix.

**Definition 10.** The function  $\psi$  maps a complete perfect phylogeny matrix  $A \in \mathcal{A}_{n,k}$  to the binary matrix  $\psi(A) = B = [A_0 \dots A_{k-1}]$ .

We show that  $\psi(A)$  is a binary  $n, k$ -complete perfect phylogeny matrix for each  $A \in \mathcal{A}_{n,k}$ , and that  $\psi$  is in fact a bijection.

**Lemma 3.** The function  $\psi$  is a bijection between  $\mathcal{A}_{n,k} \rightarrow \mathcal{B}_{n,k}$ .

*Proof.* Let  $A \in \mathcal{A}_{n,k}$ . We claim that  $B = \psi(A) = [A_0 \dots A_{k-1}] \in \mathcal{B}_{n,k}$ . Recall that  $A_i = [a_{(d,j),c}^i]$  where

$$a_{(d,j),c}^i = \begin{cases} 1, & \text{if } a_{(d,j),c} = i, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore  $a_{(d,j),c}^i = b_{(d,j),(c,i)}$ . We thus have that  $\sum_{c=1}^n \sum_{i=0}^{k-1} b_{(d,j),(c,i)} = n$  and  $\sum_{c=1}^n b_{(d,j),(c,i)} = 1$  where  $i \in \{0, \dots, k-1\}$ . Moreover, because  $a_{(*,0),d} = 0$  for all  $d \in [n]$ , we have that  $b_{(*,0),(d,0)} = 1$ . Also, as  $a_{(c,i),c} = i$ , we have  $b_{(c,i),(c,i)} = 1$ . Furthermore,  $G(A)$  and  $H(B)$  are isomorphic with  $u_{(c,i)} \leftrightarrow v_{(c,i)}$  where  $u_{(c,i)} \in V(G)$ ,  $v_{(c,i)} \in V(H)$ . Thus,  $H(B)$  is connected as  $G(A)$  is connected. Hence,  $A \in \mathcal{A}_{n,k}$ .

Let  $B = [b_{(d,j),(c,i)}] \in \mathcal{B}_{n,k}$ . We claim that  $A = [a_{(d,j),c}] \in \mathcal{A}_{n,k}$  where  $a_{(d,j),c} = i$  such that  $b_{(d,j),(c,i)} = 1$ . Since  $b_{(*,0),(c,0)} = 1$  where  $c \in [n]$ , we have that  $a_{(*,0),c} = 0$ . Moreover, as  $b_{(c,i),(c,i)} = 1$  for all  $(c,i) \in [k] \times [n-1]$ , we have that  $a_{(c,i),c} = i$ . Furthermore,  $H(B)$  and  $G(A)$  are

isomorphic with  $u_{(c,i)} \leftrightarrow v_{(c,i)}$  for  $u_{(c,i)} \in V(G)$ ,  $v_{(c,i)} \in V(H)$ . Thus,  $G(A)$  is connected as  $H(B)$  is connected. Hence,  $B \in \mathcal{B}_{n,k}$ .  $\square$   $\square$

We flatten frequency tensor  $\mathcal{F} = [F_i]$  into the  $m \times nk$  matrix  $F = [F_0 \dots F_{k-1}]$  and prove that the problem is equivalent to factorizing  $F$  into  $U$  and  $B$ .

**Lemma 4.** *Let  $\mathcal{F} = [F_i]$  be frequency tensor and let  $U \in \mathcal{U}_{m,n,k}$  be a usage matrix. There exists a matrix  $A \in \mathcal{A}_{n,k}$  such that  $F_i = UA_i$  for all states  $i \in \{0, \dots, k-1\}$  if and only if there exists a matrix  $B \in \mathcal{B}_{n,k}$  such that*

$$F = [F_0 \dots F_{k-1}] = UB. \quad (\text{B.3})$$

*Proof.* By Lemma 3, let  $A \in \mathcal{A}_{n,k}$  and  $B \in \mathcal{B}_{n,k}$  be corresponding matrices. Note that  $B = \psi(A) = [A_0 \dots A_{k-1}]$  where  $A_i = [a_{(d,j),c}^i]$  such that

$$a_{(d,j),c}^i = \begin{cases} 1, & \text{if } a_{(d,j),c} = i, \\ 0, & \text{otherwise.} \end{cases}$$

Since  $a_{(d,j),c}^i = b_{(d,j),(c,i)}$ , we have

$$f_{p,(c,i)} = u_{p,(*,0)} a_{(*,0),c}^i + \sum_{d=1}^n \sum_{j=1}^{k-1} u_{p,(d,j)} a_{(d,j),c}^i = u_{p,(*,0)} b_{(*,0),(c,i)} + \sum_{d=1}^n \sum_{j=1}^{k-1} u_{p,(d,j)} b_{(d,j),(c,i)}$$

for all  $p \in [m]$ ,  $c \in [n]$  and  $i \in \{0, \dots, k-1\}$ .  $\square$   $\square$

Hence, we may restate Problem 1 as a matrix factorization problem.

**Problem 3** (Perfect Phylogeny Mixture Deconvolution Problem (PPMDP)). *Given an  $k \times m \times n$  frequency tensor  $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$ , find a  $n, k$ -complete perfect phylogeny tree  $T$  and a  $m, n, k$ -usage matrix  $U = [u_{p,(c,i)}]$  such that  $F = UB$  where  $F = [F_0 \dots F_{k-1}]$  and  $B = \psi(\theta(T))$ .*

We now have all the ingredients to show that  $\mathcal{F}$  and  $T$  uniquely define a matrix  $U$ . We first show that any matrix  $B \in \mathcal{B}_{n,k}$  has full row rank.

**Lemma 5.** *Any matrix  $B \in \mathcal{B}_{n,k}$  has row rank  $n(k-1) + 1$ .*

*Proof.* By Definition 9, we have that  $B = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ C & D \end{pmatrix}$  where  $C$  has dimensions  $n(k-1) \times n$ ,  $D$  has dimensions  $n(k-1) \times n(k-1)$ ,  $\mathbf{1}$  is the  $1 \times n$  matrix of all 1-s and  $\mathbf{0}$  is the  $1 \times n(k-1)$  matrix of all 0-s. We show that the square submatrix  $D = [b_{(d,j),(c,i)}]$ , where  $(c,i), (d,j) \in [n] \times [k-1]$ , has full rank by performing Gaussian elimination according to a breadth-first search on  $H(B)$ , starting from the all-zero ancestor  $v_{(*,0)}$ . Let  $\ell(v)$  denote the breadth-first search (BFS) level of vertex  $v \in V(H(B))$ . Note that  $\ell(v_{(*,0)}) = 0$ . We claim that this process results in the  $n(k-1) \times n(k-1)$  identity matrix  $I$  where row  $\mathbf{i}_{(c,i)}$  corresponds to vertex  $v_{(c,i)} \in V(H(B)) \setminus \{v_{(*,0)}\}$ .

We show this constructively by induction on the BFS level  $l$ . The claim is that at BFS level  $l$  all rows  $\mathbf{i}_{(c,i)}$  where  $\ell(v_{(c,i)}) \leq l$  have been generated from  $D$  using elementary row operations. Initially, at  $l = 1$ , for each vertex  $v_{(d,j)}$  with BFS level  $\ell(v_{(d,j)}) = l = 1$  it holds that  $b_{(d,j),(d,j)} = 1$  and  $b_{(d,j),(c,i)} = 0$  for all  $(c,i) \in [n] \times [k-1] \setminus \{(d,j)\}$ . Therefore the vertices  $v_{(d,j)}$  at BFS level 1 correspond directly to rows  $\mathbf{i}_{(d,j)}$  of  $I$ . At every iteration  $l > 1$ , we generate, using elementary row operations, the rows  $\mathbf{i}_{(d,j)}$  of  $I$  that correspond to vertices  $v_{(d,j)}$  at BFS level  $\ell(v_{(d,j)}) = l$ . In order to obtain row  $\mathbf{i}_{(d,j)}$  of  $I$ , we subtract from  $\mathbf{b}_{(d,j)}$  the rows  $\mathbf{i}_{(c,i)}$  where  $b_{(d,j),(c,i)} = 1$  and  $(c,i) \neq (d,j)$ . Observe that by Definition 9,  $H(B)$  is connected and that every character-state pair  $(c,i)$  must have been introduced by vertex  $v_{(c,i)}$ , which must on a path to the root  $v_{(*,0)}$  from  $v_{(d,j)}$ . Hence,  $\ell(v_{(c,i)}) < l$  for every  $(c,i) \in [n] \times [k-1] \setminus \{(d,j)\}$  where  $b_{(d,j),(c,i)} = 1$ . Since the corresponding vertices  $v_{(c,i)}$  are at a BFS level strictly smaller than  $l$ , the corresponding rows  $\mathbf{i}_{(c,i)}$  have already been generated by the induction hypothesis. Therefore at the final iteration, we obtain the identity matrix  $I$  from  $D$  using elementary row operations. It thus follows that  $D$  is full rank.

Since  $D$  is full rank, the row rank of  $\begin{pmatrix} C & D \end{pmatrix}$  equals the rank of  $D$ , which is  $n(k-1)$ . Furthermore, the first row  $\begin{pmatrix} \mathbf{1} & \mathbf{0} \end{pmatrix}$  of  $B$  cannot be expressed as a linear combination of  $\begin{pmatrix} C & D \end{pmatrix}$ . This implies that the row rank of  $B$  is  $n(k-1) + 1$ . □ □

This means that given the  $m \times nk$  frequency matrix  $F$  and the  $(n(k-1) + 1) \times nk$  binary complete perfect phylogeny matrix  $B$ , there exists a *unique*  $m \times (n(k-1) + 1)$  matrix  $U$  such that (B.3) holds,

i.e.  $U = FB^{-1}$  where the  $nk \times (n(k-1)+1)$  matrix  $B^{-1}$  is the right inverse of  $B$  such that  $BB^{-1} = I$  where  $I$  is the  $(n(k-1)+1) \times (n(k-1)+1)$  identity matrix. Given  $F$  and  $T$ , we now define the unique matrix  $U = [u_{p,(c,i)}]$  without explicitly computing the right inverse of  $B$ . We do this using the notion of a *descendant set* of character-state pairs  $(c, i)$ . The descendant set  $D_{(c,i)}$  is the set of states for character  $c$  that are descendants of  $v_{(c,i)}$  in  $T$ . Formally,  $D_{(c,i)} = \{j \mid (c, i) \prec_T (c, j)\}$ . Recall that  $T_{(c,i)}$  is the subtree of  $T$  consisting of all vertices that have state  $i$  for character  $c$ , and that  $\bar{T}_{(c,i)}$  the subtree rooted at vertex  $v_{(c,i)}$ . The descendant set of a character precisely determines the relationship between  $\bar{T}_{(c,i)}$  and  $T_{(c,i)}$ ; namely we have  $\bar{T}_{(c,i)} = \bigcup_{l \in D_{(c,i)}} T_{(c,l)}$ . In the two-state ( $k = 2$ ) case, we have that  $T_{(c,1)} = \bar{T}_{(c,1)}$  (see Figure B.15). Recall that  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ . We define the *cumulative frequency*  $f_p^+(D_{(c,i)}) = \sum_{l \in D_{(c,i)}} f_{p,(c,l)}$ . In the following lemma we show that given  $T \in \mathcal{T}_{n,k}$ , the cumulative frequencies  $f_p^+(D_{(c,i)})$  for the descendant sets defined by  $T$  *uniquely* determine the usage matrix  $U$ . For an intuitive explanation of the usage equation (B.4), see Figure B.3.

**Lemma 6.** *Let  $T \in \mathcal{T}_{n,k}$  and  $\mathcal{F} = [[f_{p,(c,i)}]]$  be a frequency tensor. For a character-state pair  $(c, i)$  and sample  $p$ , let*

$$u_{p,(c,i)} = f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}). \quad (\text{B.4})$$

*Then  $U = [u_{p,(c,i)}]$  is the unique matrix whose entries satisfy  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ .*

*Proof.* Let  $(c, i)$  be a character-state pair and  $p$  be a sample. Let  $A = \theta(T)$  and  $B = \psi(A)$ . Recall that  $T_{(c,i)}$  is the set of vertices  $\{v_{(d,j)} \mid b_{(d,j),(c,i)} = 1\}$ . Note that by definition, the vertices of  $T_{(c,i)}$  induce a connected subtree in  $T$ . We thus need to show that

$$f_{p,(c,i)} = u_{p,(*,0)} b_{(d,j),(c,i)} + \sum_{d=1}^n \sum_{j=1}^{k-1} u_{p,(d,j)} b_{(d,j),(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}.$$

We introduce the following shorthand  $\Delta(c, i) = \bigcup_{(d,j) \in T_{(c,i)}} \delta(d, j) \setminus T_{(c,i)}$ , which is the set of vertices

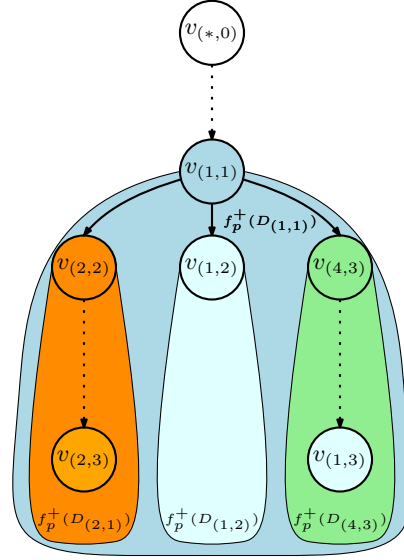


Figure B.3: **Example of the usage equation (B.4).** Recall that  $f_{p,(c,i)}$  is the sum of the usages of the vertices that have state  $i$  for character  $c$ , i.e.  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ . The observation here is that the cumulative frequency  $f_p^+(D_{(c,i)})$  equals the sum of the usages of all vertices in the subtree of  $T$  rooted at vertex  $v_{(c,i)}$ , i.e.  $f_p^+(D_{(c,i)}) = \sum_{l \in D_{(c,i)}} f_{p,(c,l)} = \sum_{l \in D_{(c,i)}} \sum_{(d,j) \in T_{(c,l)}} u_{p,(d,j)}$ . Hence,  $u_{p,(1,1)} = f_p^+(D_{(1,1)}) - (f_p^+(D_{(2,2)}) + f_p^+(D_{(1,2)}) + f_p^+(D_{(4,3)})) = (f_{p,(1,1)} + f_{p,(1,2)} + f_{p,(1,3)}) - (f_{p,(2,2)} + f_{p,(2,3)} + f_{p,(1,2)} + f_{p,(4,3)})$ .

$\{v_{(d,j)}\}$  that are not in  $T_{(c,i)}$  but whose parent  $v_{\pi(d,j)}$  is in  $T_{(c,i)}$ . Thus,

$$\begin{aligned} f_{p,(c,i)} &= \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)} \\ &= \sum_{(d,j) \in T_{(c,i)}} \left( f_p^+(D_{(d,j)}) - \sum_{(e,l) \in \delta(d,j)} f_p^+(D_{(e,l)}) \right). \end{aligned}$$

Observe that in the equation above, for every  $(d,j) \in T_{(c,i)} \setminus \{(c,i)\}$ , there are two terms  $f_p^+(D_{(d,j)})$  and  $-f_p^+(D_{(d,j)})$ , which consequently cancel out. The remaining terms are  $f_p^+(D_{(c,i)})$ , and  $-f_p^+(D_{(c,l)})$



for each  $(c, l) \in \Delta(c, i)$ . In addition, we have that the state trees  $\{D_{(c,l)}\}$  corresponding to the elements of  $(c, l) \in \Delta(c, i)$  are pairwise disjoint. Moreover,  $D_{(c,i)} \setminus \{i\} = \bigcup_{(c,l) \in \Delta(c,i)} D_{(c,l)}$ . Thus,

$$\begin{aligned} f_{p,(c,i)} &= f_p^+(D_{(c,i)}) - \sum_{(c,l) \in \Delta(c,i)} f_p^+(D_{(c,l)}) \\ &= f_{p,(c,i)} + \sum_{(c,l) \in \Delta(c,i)} f_p^+(D_{(c,l)}) - \sum_{(c,l) \in \Delta(c,i)} f_p^+(D_{(c,l)}) \\ &= f_{p,(c,i)}. \end{aligned}$$

By Lemma 5, the equation  $F = UB$  has only one solution given  $F$  and  $B$ . Thus  $U = [u_{p,(c,i)}]$  is the unique matrix such that  $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$  for all samples  $p$  and character-state pairs  $(c, i)$ .  $\square$   $\square$

### B.1.3 Combinatorial Characterization of the PPMDP

We say that  $T$  *generates*  $\mathcal{F}$  if the corresponding matrix  $U$ , defined by (B.4), is a usage matrix. It turns out that positivity of the values  $u_{p,(c,i)}$  is a necessary and sufficient condition for  $T$  to generate  $\mathcal{F}$ . We show this in the following theorem.

**(Main Text) Theorem 1.** *A complete perfect phylogeny tree  $T$  generates  $\mathcal{F}$  if and only if*

$$f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) \geq 0 \quad (\text{MSSC})$$

for all character-state pairs  $(c, i)$  and samples  $p$ .

*Proof.* ( $\Rightarrow$ ) We start by proving the forward direction. Let  $F = [F_0 \dots F_{k-1}]$ ,  $T$  be a tree that generates  $F$ ,  $B = \psi(\theta(T))$  be the corresponding binary matrix of  $T$  and let  $U = [u_{p,(c,i)}]$  be the corresponding usage matrix. Since  $T$  generates  $F$ , we have that  $u_{p,(c,i)} \geq 0$  for all character-state pairs  $(c, i)$  and samples  $p$ . By Lemma 6, (MSSC) thus holds.

( $\Leftarrow$ ) As for the reverse direction, we need to show that if (MSSC) is met, the matrix  $U$  as defined

in Lemma 6 is a usage matrix. Let  $p \in [m]$ . We prove this direction by showing that (i)  $u_{p,(c,i)} \geq 0$  for all character-state pairs  $(c,i)$ , and (ii)  $u_{p,(*,0)} + \sum_{c=1}^n \sum_{i=1}^{k-1} u_{p,(c,i)} = 1$ .

(i) By Lemma 6 and (MSSC), we have that  $u_{p,(c,i)} \geq 0$ .

(ii) We now have

$$u_{p,(*,0)} + \sum_{c=1}^n \sum_{i=1}^{k-1} u_{p,(c,i)} = f_p^+(D_{(*,0)}) - \sum_{(d,j) \in \delta(*,0)} f_p^+(D_{(d,j)}) + \sum_{c=1}^n \sum_{i=1}^{k-1} \left( f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) \right).$$

Observe that for each  $(c,i) \in [n] \times [k-1]$  there are exactly two terms in the above equation:  $+f_p^+(D_{(c,i)})$  when  $v_{(c,i)}$  is considered as a parent and  $-f_p^+(D_{(c,i)})$  when  $v_{(c,i)}$  is considered as a child. Hence, all these terms cancel out. Since  $D_{(*,0)} = \{0, \dots, k-1\}$  and  $\sum_{i=0}^{k-1} f_{p,(c,i)} = 1$ , we have that  $f_p^+(D_{(*,0)}) = 1$ . Thus,  $u_{p,(*,0)} + \sum_{c=1}^n \sum_{i=1}^{k-1} u_{p,(c,i)} = 1$ .

□

□

Although the sets  $D_{(c,i)}$  are *a priori* unknown, we show that the following condition (MSAC) must hold for any tree  $T$  that generates  $\mathcal{F}$  where  $(c,i) \prec_T (d,j)$ .

**Definition 11.** Let  $(c,i)$  and  $(d,j)$  be distinct character-state pairs and let  $D_{(c,i)}, D_{(d,j)} \subseteq \{0, \dots, k-1\}$ . A pair  $(D_{(c,i)}, D_{(d,j)})$  is a valid descendant set pair provided

$$f_p^+(D_{(c,i)}) - f_p^+(D_{(d,j)}) \geq 0 \tag{MSAC}$$

for all samples  $p$ ; and additionally if  $c = d$  then  $D_{(c,j)} \subsetneq D_{(c,i)}$ .

It turns out that there are potentially many valid descendant set pairs as shown by the following lemma.

**Lemma 7.** Let  $(D_{(c,i)}, D_{(d,j)})$  be a valid descendant set pair. If  $D_{(c,i)} \subseteq D'_{(c,i)}$  and  $D'_{(d,j)} \subseteq D_{(d,j)}$  then  $(D'_{(c,i)}, D'_{(d,j)})$  is a valid descendant set pair.

*Proof.* Let  $D_{(c,i)} \subseteq D'_{(c,i)}$  and  $D'_{(d,j)} \subseteq D_{(d,j)}$ . Let  $p \in [m]$ . Since  $f_{p,(c,i)} \geq 0$  (by Definition 3),  $D_{(c,i)} \subseteq D'_{(c,i)}$  and  $D'_{(d,j)} \subseteq D_{(d,j)}$ , we have that  $f_p^+(D'_{(c,i)}) \geq f_p^+(D_{(c,i)})$  and  $f_p^+(D_{(d,j)}) \geq f_p^+(D'_{(d,j)})$ . Moreover, if  $c = d$  then  $D_{(d,j)} \subsetneq D_{(c,i)}$  and thus  $D'_{(d,j)} \subsetneq D'_{(c,i)}$ . Hence,  $(D'_{(c,i)}, D'_{(d,j)})$  is a valid descendant set pair.  $\square$   $\square$

Since  $v_{(*,0)}$  is the all-zero ancestor, we have the following corollary that describes the *extreme* valid descendant set pair.

**Corollary 1.** *Let  $T$  be a complete perfect phylogeny tree that generates  $\mathcal{F}$ . If  $(c,i) \prec_T (d,j)$  and  $(c,i) \neq (c,0)$  then  $D_{(c,i)} = [k-1]$  and  $D_{(d,j)} = \{j\}$  form a valid descendant set pair  $(D_{(c,i)}, D_{(d,j)})$ .*

The following proposition shows that (MSAC) is a necessary condition to solutions of the PP-MDP.

**(Main Text) Proposition 1.** *Let  $T$  be a complete perfect phylogeny tree that generates  $\mathcal{F}$ . If  $(c,i) \prec_T (d,j)$  then there exist a valid descendant set pair  $(D_{(c,i)}, D_{(d,j)})$ .*

*Proof.* Let  $(c,i) \prec_T (d,j)$ . Let  $P = v_{(c_1,i_1)}, \dots, v_{(c_t,i_t)}$  where  $v_{(c_1,i_1)} = v_{(c,i)}$  and  $v_{(c_t,i_t)} = v_{(d,j)}$  be the unique path from  $v_{(c,i)}$  to  $v_{(d,j)}$  in  $T$ . Assume for a contradiction that there exists no valid descendant set pair  $(D_{(c,i)}, D_{(d,j)})$ . By Corollary 1, we thus have that  $D_{(c,i)} = [k-1]$  and  $D_{(d,j)} = \{j\}$  do not form a valid descendant set pair  $(D_{(c,i)}, D_{(d,j)})$ . Now, if  $c = d$ , we would have a contradiction as  $\{j\} \subseteq [k-1]$ . Therefore,  $c \neq d$  and  $f_p^+(D_{(c,i)}) - f_p^+(D_{(d,j)}) < 0$ . By Theorem 1, we have that  $f_p^+(D_{(c_l,i_l)}) - f_p^+(D_{(c_{l+1},i_{l+1})}) \geq 0$  for all  $1 \leq l < t$ . Therefore,  $f_p^+(D_{(c_1,i_1)}) - f_p^+(D_{(c_t,i_t)}) \geq 0$ , which leads to a contradiction.  $\square$   $\square$

We now define the multi-state ancestry graph  $G_{\mathcal{F}}$  whose vertices correspond to character-state pairs and whose edges correspond to valid descendant state pairs.

**Definition 12.** *The multi-state ancestry graph  $G_{\mathcal{F}}$  of the frequency tensor  $\mathcal{F}$  is an edge-labeled, directed multi-graph  $G_{\mathcal{F}} = (V, E)$  whose vertices  $v_{(c,i)}$  correspond to character-state pairs  $(c,i)$  and whose multi-edges are  $(v_{(c,i)}, v_{(d,j)})$  for all valid descendant set pairs  $(D_{(c,i)}, D_{(d,j)})$ .*

Note that  $v_{(1,0)}, \dots, v_{(n,0)}$  all refer to the same root vertex  $v_{(*,0)}$ . See Figure B.4 for an example

multi-state ancestry graph. We use the labels of the multi-edges to restrict the set of allowed spanning trees by defining a threading as follows.

**Definition 13.** A rooted subtree  $T$  of  $G_{\mathcal{F}} = (V, E)$  is a threaded tree provided (1) for every pair of adjacent edges  $(v_{(c,i)}, v_{(d,j)}), (v_{(d,j)}, v_{(e,l)}) \in E(T)$  with corresponding labels  $(D_{(c,i)}, D_{(d,j)})$  and  $(D'_{(d,j)}, D'_{(e,l)})$ , it holds that  $D_{(d,j)} = D'_{(d,j)}$ , and (2) for every pair of vertices  $v_{(c,i)}, v_{(c,j)} \in V(T)$  it holds that  $D_{(c,j)} \subseteq D_{(c,i)}$  if and only if  $(c, i) \prec_T (c, j)$ .

We now prove that solutions of an PPM DP instance  $\mathcal{F}$  correspond to threaded spanning trees of the ancestry graph  $G_{\mathcal{F}}$ .

**(Main Text) Theorem 2.** A complete perfect phylogeny tree  $T$  generates  $\mathcal{F}$  if and only if  $T$  is a threaded spanning tree of the ancestry graph  $G_{\mathcal{F}}$  such that (MSSC) holds.

*Proof.* ( $\Rightarrow$ ) Let  $T$  be a complete perfect phylogeny tree generating  $\mathcal{F}$ . We claim that  $T$  is a threaded spanning tree of the ancestry graph  $G_{\mathcal{F}}$ . We start by showing that every edge  $(v_{(c,i)}, v_{(d,j)}) \in E(T)$  is an edge of  $G_{\mathcal{F}}$  labeled by  $(D_{(c,i)}, D_{(d,j)})$ . By Theorem 1, we have that  $f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) \geq 0$  for all character-state pairs  $(c, i)$  and samples  $p$ . Let  $(c, i)$  be a character-state pair. By definition, we have that  $D_{(c,j)} \subsetneq D_{(c,i)}$  for all  $(c, j) \in \delta(c, i)$ . Moreover,  $i \in D_{(c,i)}$  and  $j \in D_{(d,j)}$  for all character-state pairs  $(d, j) \in \delta(c, i)$ . Thus,  $(D_{(c,i)}, D_{(d,j)})$  is a valid descendant set pair for all character-state pairs  $(d, j) \in \delta(c, i)$ . Hence, every edge  $(v_{(c,i)}, v_{(d,j)})$  of  $T$  is an edge of  $G_{\mathcal{F}}$  labeled by the valid descendant set pair  $(D_{(c,i)}, D_{(d,j)})$ . Thus,  $T$  is a tree of  $G$ . Next, we show that  $T$  is a threaded spanning tree.

1. By definition of  $D$ , we have that every pair of adjacent edges  $(v_{(c,i)}, v_{(d,j)}), (v_{(d,j)}, v_{(e,l)}) \in E(T)$  is labeled by  $(D_{(c,i)}, D_{(d,j)})$  and  $(D_{(d,j)}, D_{(e,l)})$ , respectively.
2. By definition of  $D$ , we have that for every edge  $(v_{(c,i)}, v_{(d,j)}) \in E(T)$  labeled by  $(D_{(c,i)}, D_{(d,j)})$ , it holds that  $D_{(c,i)} = \{l \mid (c, i) \prec_T (c, l)\}$  and  $D_{(d,j)} = \{l \mid (d, j) \prec_T (d, l)\}$ . Hence,  $(c, i) \prec_T (c, j)$  if and only if  $D_{(c,j)} \subseteq D_{(c,i)}$ .

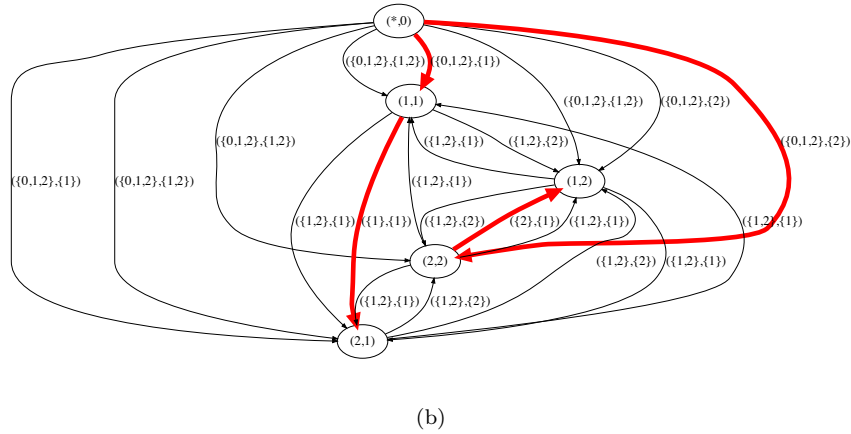
The conditions of Definition 11 are thus met. Therefore,  $T$  is a threaded spanning tree of  $G_{\mathcal{F}}$ .

$$F_0 = \begin{pmatrix} 0.3 & 0.2 \\ 0.4 & 0.5 \end{pmatrix}$$

$$F_1 = \begin{pmatrix} 0.6 & 0.5 \\ 0.3 & 0.1 \end{pmatrix}$$

$$F_2 = \begin{pmatrix} 0.1 & 0.3 \\ 0.3 & 0.4 \end{pmatrix}$$

(a)



(b)

Figure B.4: **Example PPMDP instance and solution.** (A) Frequency tensor  $\mathcal{F} = [F_0 F_1 F_2]$  where  $m = 2$ ,  $n = 2$  and  $k = 3$ . (B) Multi-state ancestry graph  $G_{\mathcal{F}}$ . Red edges denote an  $n, k$ -complete perfect phylogeny tree  $T$  that generates  $\mathcal{F}$ .

( $\Leftarrow$ ) Let  $T$  be a threaded spanning tree of the ancestry graph  $G_{\mathcal{F}}$  such that  $f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) \geq 0$  for all character-state pairs  $(c, i)$  and samples  $p$ . Observe that  $T$  is a complete perfect phylogeny tree. By condition (1) of Definition 11, we have that for all character-state pairs  $(c, 0) \neq (*, 0)$  and  $(d, j) \in \delta(c, i)$ , adjacent edges  $(v_{\pi(c,i)}, v_{(c,i)})$ ,  $(v_{(c,i)}, v_{(d,j)})$  are labeled by  $(D_{\pi(c,i)}, D_{(c,i)})$  and  $(D_{(c,i)}, D_{(d,j)})$ —where  $\pi(c, i)$  is the parent character-state pair of  $(c, i)$ . Hence, we may use  $D_{(c,i)}$  to unambiguously denote the descendant state set of the character-state pair  $(c, i)$ . Moreover, by condition (2) of Definition 11 and the fact that  $T$  is a spanning tree of  $G_{\mathcal{F}}$ , we have that  $D$  is defined in the same way as in Theorem 1. By Theorem 1, we thus have that  $T$  generates  $\mathcal{F}$ .  $\square$   $\square$

Figure B.4 shows an example instance  $\mathcal{F}$ , where  $k = 3$ ,  $m = 2$  and  $n = 2$ , the multi-state ancestry graph  $G_{\mathcal{F}}$  and a solution  $T$  that generates  $\mathcal{F}$ .

### B.1.4 Problem Complexity

In previous work, we have shown that the problem is NP-complete for general  $m$  [38]. An open question was the hardness for constant  $m$ , which we resolve with the following lemma.

**Theorem 4.** *The PPMDP is NP-complete even for  $m = 2$  and  $k = 2$ .*

*Proof.* Clearly, the problem is in NP—given matrices  $U$  and  $A$  we can check in polynomial time whether  $F_i = UA_i$  for all  $i \in \{0, \dots, k-1\}$ . We show NP-hardness by reduction from SUBSET SUM, which, given non-negative integers  $B = \{b_1, \dots, b_t\}$  and  $d$ , asks whether there exists a subset  $B' \subseteq B$  whose sum equals  $d$ . This problem is NP-complete [54].

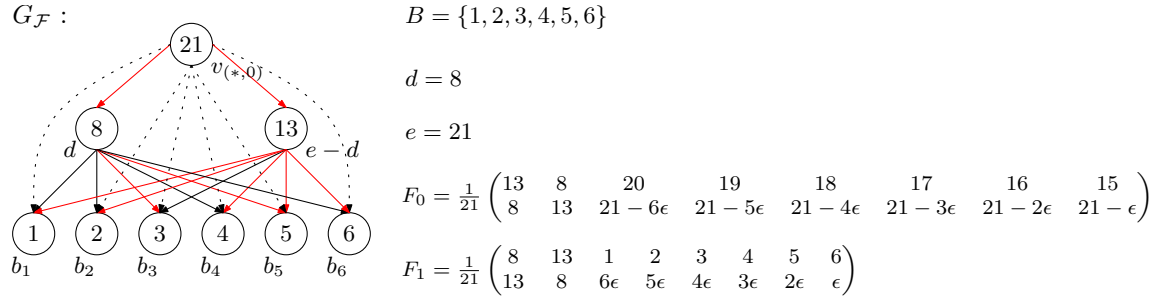
Let  $e = \sum_{\ell=1}^t c_\ell$ . Without loss of generality assume that  $e > 0$ ,  $b_\ell < d$  for all  $\ell \in [t]$  and that  $b_\ell \leq b_{\ell+1}$  for all  $\ell \in [t-1]$ . The corresponding frequency tensor  $\mathcal{F} = [F_0 F_1]$  is then defined as follows.

$$F_0 = \frac{1}{e} \begin{pmatrix} e-d & d & e-b_1 & e-b_2 & \dots & e-b_t \\ d & e-d & e-t\epsilon & e-(t-1)\epsilon & \dots & e-\epsilon \end{pmatrix}$$

$$F_1 = \frac{1}{e} \begin{pmatrix} d & e-d & b_1 & b_2 & \dots & b_t \\ e-d & d & t\epsilon & (t-1)\epsilon & \dots & \epsilon \end{pmatrix}.$$

Note that  $\mathcal{F}$  is a  $k \times m \times n$  tensor where  $k = 2$ ,  $m = 2$  and  $n = t+2$ . Also note that the normalization factor  $\frac{1}{e}$  ensures that each  $f_{p,(c,i)} \in [0, 1]$  and that  $f_{p,(c,0)} + f_{p,(c,1)} = 1$  for all  $p \in [m]$  and  $c \in [n]$ . Clearly  $\mathcal{F}$  can be obtained in polynomial time from a SUBSET SUM instance. By construction, the ancestry graph  $G_{\mathcal{F}}$  consists of a root vertex  $v_{(*,0)}$  with outgoing edges to all other vertices  $\{d, e-d, b_1, \dots, b_t\}$ . In addition, these vertices induce a complete bipartite graph with vertex sets  $\{d, e-d\}$  and  $\{b_1, \dots, b_t\}$ . See Figure B.5 for an illustration.

We claim that there exist  $U \in \mathcal{U}_{2,t+2,2}$  and  $A \in \mathcal{A}_{t+2,2}$  such that  $F_0 = UA_0$  and  $F_1 = UA_1$  if and only if there exists a subset  $B' \subseteq B$  whose sum is  $d$ . Equivalently, by Theorem 2, we claim that  $G_{\mathcal{F}}$  admits a threaded spanning tree  $T$  satisfying (MSSC) if and only if  $B$  has a subset  $B'$  whose

Figure B.5: **Reduction from SUBSET SUM**

sum is  $d$ .

We start by proving the forward direction. Since  $e = \sum_{\ell=1}^t b_{\ell}$  and  $T$  is a threaded spanning tree, the sum of the children  $\delta(d)$  equals  $d$  and the sum of the children  $\delta(e - d)$  equals  $e - d$ . Therefore  $B' = \delta(d)$  is a subset of  $B$  whose sum is  $d$ . As for the reverse direction, we have that the sum of  $B'$  equals  $d$  and thus that the sum of  $B \setminus B'$  equals  $e - d$ . The corresponding threaded spanning tree  $T$  where  $\delta(*, 0) = \{d, e - d\}$ ,  $\delta(d) = B'$  and  $\delta(e - d) = B \setminus B'$  is therefore a threaded spanning tree of  $G_{\mathcal{F}}$  that satisfies (MSSC).  $\square$

The result above settles the outstanding question posed in [38] of fixed-parameter tractability in  $m$ .

**Corollary 2.** *PPMDP is not fixed-parameter tractable in  $m$ .*

### B.1.5 Cladistic Perfect Phylogeny Mixture Deconvolution Problem

In the case of cladistic characters we are given a set  $\mathcal{S} = \{S_c \mid c \in [n]\}$  of state trees for each character. The vertex set of a state tree  $S_c$  is  $\{0, \dots, k-1\}$ , and the edges describe the relationships of the states of character  $c$ . A perfect phylogeny  $T$  is *consistent* with  $\mathcal{S}$  provided  $i \prec_{S_c} j$  if and only if  $(c, i) \prec_T (c, j)$  for all characters  $c$  and states  $i, j$ . In the cladistic PPMDP we seek to find a complete perfect phylogeny tree  $T$  that generates  $\mathcal{F}$  and is consistent with  $\mathcal{S}$ . The cladistic multi-state perfect phylogeny problem reduces to the binary case and is polynomial-time decidable [47]. Thus, it is not surprising that the PPMDP also simplifies in the cladistic case. In particular, the

set  $\mathcal{S}$  of states trees determines the set of descendant sets as  $D_{(c,i)} = \{j \mid i \prec_{S_c} j\}$ . Therefore the cladistic multi-state ancestry graph  $G_{(\mathcal{F},\mathcal{S})}$  is a simple graph with edges  $(v_{(c,i)}, v_{(d,j)})$  labeled by  $(D_{(c,i)}, D_{(d,j)})$  provided  $c \neq d$  and (MSAC) holds. Moreover, for each character  $c$  there is an edge  $(v_{(c,i)}, v_{(c,j)})$  provided state  $i$  is the parent of state  $j$  in  $S_c$ . Solutions of the cladistic PPMDP correspond to threaded spanning trees in  $G_{(\mathcal{F},\mathcal{S})}$  as shown by the following proposition.

**Proposition 2.** *A complete perfect phylogeny tree  $T$  generates  $\mathcal{F}$  and is consistent with  $\mathcal{S}$  if and only if  $T$  is a threaded spanning tree of the cladistic ancestry graph  $G_{(\mathcal{F},\mathcal{S})}$  such that (MSSC) holds.*

*Proof.* Let  $T$  be a threaded spanning tree of the ancestry graph  $G_{(\mathcal{F},\mathcal{S})}$ . We claim that  $T$  is consistent with  $\mathcal{S}$ , i.e.  $i \prec_{S_c} j$  if and only if  $(c,i) \prec_T (c,j)$  for all characters  $c$  and states  $i, j$ . By condition (1) of Definition 11, we have that for all character-state pairs  $(c,i) \neq (*,0)$  and  $(d,j) \in \delta(c,i)$ , adjacent edges  $(v_{\pi(c,i)}, v_{(c,i)})$ ,  $(v_{(c,i)}, v_{(d,j)})$  are labeled by  $(D_{\pi(c,i)}, D_{(c,i)})$  and  $(D_{(c,i)}, D_{(d,j)})$ , respectively—where  $\pi(c,i)$  is the parent character-state pair of  $(c,i)$ . By definition, we have  $D_{(c,i)} = \{l \mid i \prec_{S_c} l\}$  for each character-state pair  $(c,i)$ . By condition (2) of Definition 11 and the fact that  $T$  is a spanning tree of  $G_{(\mathcal{F},\mathcal{S})}$ , we have  $D_{(c,i)} = \{l \mid (c,i) \prec_T (c,l)\}$ . The lemma now follows.  $\square \quad \square$

Figure B.6 shows an example instance  $(\mathcal{F}, \mathcal{S})$ , where  $k = 3$ ,  $m = 2$  and  $n = 2$ , the cladistic multi-state ancestry graph  $G_{(\mathcal{F},\mathcal{S})}$  and a solution  $T$  that generates  $\mathcal{F}$  and is consistent with  $\mathcal{S}$ .

### B.1.6 Algorithm for the Cladistic-PPMDP

We now describe an algorithm for enumerating all trees  $T$  that are consistent with the given state trees  $\mathcal{S}$  and generate the given frequencies  $\mathcal{F}$ . The crucial observation is that any subtree of a consistent, threaded spanning tree  $T$  that satisfies (MSSC) must itself satisfy (MSSC) and be consistent and threaded. A subtree  $T'$  is *consistent* if it is rooted at  $v_{(*,0)}$ , and for each character  $c$  the set of states  $\{i \mid v_{(c,i)} \in V(T')\}$  induces a connected subtree in  $S_c$ . We can thus constructively grow consistent, threaded trees that satisfy (MSSC) by maintaining the following invariant.

**Invariant 1.** *Let tree  $T$  be the partially constructed tree. It holds that (1) for each  $v_{(c,i)} \in V(T) \setminus \{v_{(*,0)}\}$  and parent  $\pi(i)$  of  $i$  in  $S_c$ , the vertex  $v_{(c,\pi(i))}$  is the first vertex labeled by character  $c$  on the*



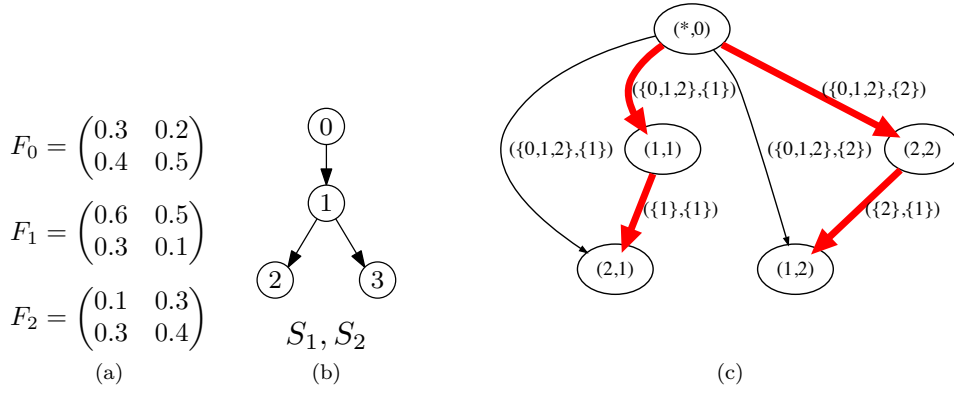


Figure B.6: **Example cladistic PPMDP instance and solution.** (A) Frequency tensor  $\mathcal{F} = [F_0 F_1 F_2]$  where  $m = 2$ ,  $n = 2$  and  $k = 3$ . (B) State trees  $\mathcal{S}$ . (C) Cladistic multi-state ancestry graph  $G_{(\mathcal{F}, \mathcal{S})}$ , red edges denote  $n, k$ -complete perfect phylogeny tree  $T$  that generates  $\mathcal{F}$  and is consistent with  $\mathcal{S}$ .

unique path from  $v_{(*,0)}$  to  $v_{(c,i)}$ ; and (2) for each vertex  $v_{(c,i)} \in V(T)$ , (MSSC) holds for  $T$  and  $\mathcal{F}$ .

We maintain a subset of edges  $H \subseteq E(G_{(\mathcal{F}, \mathcal{S})})$  called the *frontier* that can be used to extend a partial tree  $T$  such that the following invariant holds.

**Invariant 2.** Let tree  $T$  be the partially constructed tree. For every edge  $(v_{(c,i)}, v_{(d,j)}) \in H$ , (1)  $v_{(c,i)} \in V(T)$ , (2)  $v_{(d,j)} \notin V(T)$ , and (3) Invariant 1 holds for  $T'$  where  $E(T') = E(T) \cup \{(v_{(c,i)}, v_{(d,j)})\}$ .

Algorithm 1 gives the pseudo code of ENUMERATE described in the main text. The initial call is  $\text{ENUMERATE}(G, \{v_{(*,0)}\}, \delta(*, 0))$ . The partial tree containing just the vertex  $v_{(*,0)}$  satisfies Invariant 1. The set  $\delta(*, 0)$  corresponds to the set of outgoing edges from vertex  $v_{(*,0)}$  of  $G_{(\mathcal{F}, \mathcal{S})}$ , which by definition satisfies Invariant 2. Upon the addition of an edge  $(v_{(c,i)}, v_{(d,j)}) \in H$  (line 5), Invariant 2 is restored by adding all outgoing edges from  $v_{(d,j)}$  whose addition results in a consistent partial tree that satisfies (MSSC) (lines 8-9) and by removing all edges from  $H$  that introduce a cycle (lines 11-12) or violate (MSSC) (lines 13-14).

In order to extend this algorithm to the general PPMDP, we need to update the descendant sets  $D_{(c,i)}$  as we grow the tree. This in turn has implications for how we maintain the frontier: for each

**Algorithm 1:** ENUMERATE( $G, T, H$ )**Input:** Ancestry graph  $G_{(\mathcal{F}, \mathcal{S})}$ , perfect phylogeny tree  $T$ , frontier  $H$ **Output:** All complete perfect phylogeny trees that generate  $\mathcal{F}$  and are consistent with  $\mathcal{S}$ **if**  $H = \emptyset$  **and**  $|V(T)| = |V(G)|$  **then**    Return  $T$ **else**    **while**  $H \neq \emptyset$  **do**         $(v_{(c,i)}, v_{(d,j)}) \leftarrow \text{POP}(H)$          $E(T) \leftarrow E(T) \cup \{(v_{(c,i)}, v_{(d,j)})\}$         **foreach**  $(v_{(d,j)}, v_{(e,l)}) \in E(G)$  **do**

**if**  $v_{(e,l)} \notin V(T)$  **and**  $v_{(e,\pi(l))}$  *is the first vertex with character  $e$  on path from  $v_{(*,0)}$  to  $v_{(d,j)}$*  **and**  $f_p^+(D_{(d,j)}) \geq f_p^+(D_{(e,l)}) + \sum_{(f,s) \in \delta(d,j)} f_p^+(D_{(f,s)})$  **then**

                PUSH( $H, (v_{(d,j)}, v_{(e,l)})$ )

**foreach**  $(v_{(e,l)}, v_{(f,s)}) \in H$  **do**

**if**  $v_{(f,s)} = v_{(d,j)}$  **then**

                Remove  $(v_{(e,l)}, v_{(f,s)})$  from  $H$

**else if**

$v_{(e,l)} = v_{(c,i)}$  **and**  $\exists p \in [m]$  *such that*  $f_p^+(D_{(c,i)}) < f_p^+(D_{(f,s)}) + \sum_{(h,t) \in \delta(c,i)} f_p^+(D_{(h,t)})$

**then**

                    Remove  $(v_{(e,l)}, v_{(f,s)})$  from  $H$

        ENUMERATE( $G, T, H$ )         $E(T) \leftarrow E(T) \setminus \{(v_{(c,i)}, v_{(d,j)})\}$ 

potential frontier edge, we need to consider how its addition to  $T$  affects the descendant sets of the existing vertices of  $T$  and thereby (MSSC). We leave this extension as future work.

We account for errors by taking as input nonempty intervals  $[\underline{f}_{p,(c,i)}, \bar{f}_{p,(c,i)}]$  that contain the true frequency  $f_{p,(c,i)}$  for character-state pairs  $(c,i)$  in samples  $p$ . A tree  $T$  is *valid* if there exists a frequency tensor  $\mathcal{F}' = [[f'_{p,(c,i)}]]$  such that  $\underline{f}_{p,(c,i)} \leq f'_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$  and  $\mathcal{F}'$  generates  $T$ —i.e. (MSSC) holds for  $T$  and  $\mathcal{F}'$ . A valid tree  $T$  is *maximal* if there exists no valid supertree  $T'$  of  $T$ , i.e.  $E(T) \subsetneq E(T')$ . The task now becomes to find the set of all maximal valid trees. We recursively define  $\hat{\mathcal{F}} = [[\hat{f}_{p,(c,i)}]]$  where

$$\hat{f}_{p,(c,i)} = \max \left\{ \underline{f}_{p,(c,i)}, \sum_{(d,j) \in \delta(c,i)} \hat{f}_p^+(D_{(d,j)}) - \sum_{j \in D(c,i) \setminus \{i\}} \hat{f}_{p,(c,j)} \right\}. \quad (\text{B.5})$$

The intuition here is to satisfy (MSSC) by assigning the smallest possible values to the children. We

do this bottom-up from the leaves and set  $\hat{f}_{p,(c,i)} = \underline{f}_{p,(c,i)}$  for each leaf vertex  $v_{(c,i)}$ . It turns out that  $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$  is a necessary condition for  $T$  to be valid as shown in the following lemma.

**Lemma 8.** *If tree  $T$  is valid then  $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$  for all  $p$  and  $(c,i)$ .*

*Proof.* Let  $T$  be a valid tree and let  $\mathcal{F} = [[f_{p,(c,i)}]]$  be a frequency tensor that generates  $T$  such that  $\underline{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ . We claim that  $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ . We proof this by structural induction on  $T$  by working our way up to the root.

For the base case, let  $v_{(c,i)}$  be a leaf of  $T$ . That is,  $\delta(c,i) = \emptyset$ . Thus by definition,  $\hat{f}_{p,(c,i)} = \underline{f}_{p,(c,i)}$ . Therefore,  $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ . For the step, let  $v_{(c,i)}$  be an inner vertex of  $T$ . Thus,  $\delta(c,i) \neq \emptyset$ . The induction hypothesis (IH) states that  $\underline{f}_{p,(d,j)} \leq \hat{f}_{p,(d,j)} \leq f_{p,(d,i)} \leq \bar{f}_{p,(d,j)}$  for all descendants  $(d,j)$  of  $(c,i)$  in  $T$ . We distinguish two cases:

1. In case  $\hat{f}_{p,(c,i)} = \underline{f}_{p,(c,i)}$ , we have  $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ .
2. In case  $\hat{f}_{p,(c,i)} = \sum_{(d,j) \in \delta(c,i)} \hat{f}_p^+(D_{(d,j)}) - \sum_{j \in D(c,i) \setminus \{i\}} \hat{f}_{p,(c,i)}$ , we have  $\hat{f}_{p,(c,i)} \geq \underline{f}_{p,(c,i)}$  by definition. By (MSSC), we have that  $f_{p,(c,i)} \geq \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) - \sum_{j \in D(c,i) \setminus \{i\}} f_{p,(c,i)}$ . By the IH, we have  $\underline{f}_{p,(d,j)} \leq \hat{f}_{p,(d,j)} \leq f_{p,(d,i)} \leq \bar{f}_{p,(d,j)}$  for all  $(d,j) \in \delta(c,i)$  and  $\underline{f}_{p,(c,l)} \leq \hat{f}_{p,(c,l)} \leq f_{p,(c,l)} \leq \bar{f}_{p,(c,l)}$  for all  $l \in D(c,i) \setminus \{i\}$ . Therefore,  $\hat{f}_{p,(c,i)} = \sum_{(d,j) \in \delta(c,i)} \hat{f}_p^+(D_{(d,j)}) - \sum_{j \in D(c,i) \setminus \{i\}} \hat{f}_{p,(c,i)} \leq \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) - \sum_{j \in D(c,i) \setminus \{i\}} f_{p,(c,i)} \leq f_{p,(c,i)}$ . Hence,  $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ . □

□

It may be the case that the frequencies  $\hat{f}_{p,(c,i)}$  of a character  $c$  in a sample  $p$  do not sum to 1. Therefore,  $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$  is not a sufficient condition for  $T$  to be valid. However, if  $1 - \sum_{i=1}^{k-1} \hat{f}_{p,(c,i)} \leq \bar{f}_{p,(c,0)}$  holds,  $T$  is valid as it is generated by frequency tensor  $\mathcal{F}' = [[f'_{p,(c,i)}]]$  where

$$f'_{p,(c,i)} = \begin{cases} 1 - \sum_{i=1}^{k-1} \hat{f}_{p,(c,i)}, & \text{if } i = 0, \\ \hat{f}_{p,(c,i)}, & \text{otherwise.} \end{cases}$$

This is the case if  $\bar{f}_{p,(c,0)} = 1$ . By assuming the latter, we are able to enumerate all maximal valid trees by updating condition (2) of Invariant 1 so that (MSSC) holds for  $T$  and  $\hat{\mathcal{F}}$ .

Algorithm 2 gives the pseudo code of an enumeration procedure of all maximal valid trees given intervals  $[l_{p,(c,i)}, u_{p,(c,i)}]$  for each character-state pair  $(c, i)$  in each sample  $p$ . The initial call is  $\text{NOISYENUMERATE}(G, \{v_{(*,0)}\}, \delta(*, 0))$ . The partial tree containing just the vertex  $v_{(*,0)}$  satisfies Invariant 1. The set  $\delta(*, 0)$  corresponds to the set of outgoing edges from vertex  $v_{(*,0)}$  of  $G_{(\mathcal{F}, \mathcal{S})}$ , which by definition satisfies Invariant 2. Upon the addition of an edge  $(v_{(c,i)}, v_{(d,j)}) \in H$  (line 5), Invariant 2 is restored by adding all outgoing edges from  $v_{(d,j)}$  whose addition results in a consistent partial tree  $T'$  that satisfies (MSSC) for  $\hat{\mathcal{F}}$  (lines 9-10) and by removing all edges from  $H$  that introduce a cycle (lines 12-13) or violate (MSSC) for  $\hat{\mathcal{F}}$  (lines 14-15). Note that in line 13 we dropped the condition  $v_{(e,l)} = v_{(c,i)}$  as the newly added edge  $(v_{(c,i)}, v_{(d,j)})$  may affect the frequencies  $\hat{\mathcal{F}}$  of the vertices of the current partial tree  $T$ .

Since a maximal valid tree  $T$  does not necessarily span all the vertices, it may happen that for a character  $c$  not all states in  $S_c$  are present. We say that a maximal valid tree  $T$  is *state complete* if for each vertex  $v_{(c,i)}$  of  $T$ , all vertices  $v_{(c,j)}$  where  $j \in V(S_c)$  are also in  $V(T)$ . Our goal is to report all maximal valid and state-complete trees. Therefore, we post-process each maximal valid tree  $T$  and remove all vertices  $v_{(c,i)}$  where there is a  $j \in V(S_c)$  such that  $v_{(c,j)} \notin V(T)$ . The tree that we report corresponds to the connected component rooted at  $v_{(*,0)}$ . Since each maximal valid and state-complete tree is a partial valid tree rooted at  $v_{(*,0)}$ , our enumeration procedure reports all maximal valid and state-complete trees.

### B.1.7 Multi-State Model for the Somatic Mutational Process in Cancer

The characters in our model are positions in the genome whose states we model with a 4-tuple  $(x, y, \bar{x}, \bar{y})$  where  $x$  and  $y$  are the number of maternal and paternal copies of the position, respectively, and  $\bar{x}$  and  $\bar{y}$  are the number of mutated maternal and paternal copies, respectively. We define

---

**Algorithm 2:** NOISYENUMERATE( $G, T, H$ )

---

**Input:** Ancestry graph  $G_{(\mathcal{F}, \mathcal{S})}$ , perfect phylogeny tree  $T$ , frontier  $H$ 
**Output:** All maximal valid perfect phylogenies that are consistent with  $\mathcal{S}$ 
**if**  $H = \emptyset$  **then**

    Let  $T'$  be the subtree of  $T$  that only contains state-complete characters

    Return  $T'$ 
**else**

    **while**  $H \neq \emptyset$  **do**

         $(v_{(c,i)}, v_{(d,j)}) \leftarrow \text{POP}(H)$ 

         $E(T) \leftarrow E(T) \cup \{(v_{(c,i)}, v_{(d,j)})\}$ 

        **foreach**  $(v_{(d,j)}, v_{(e,l)}) \in E(G)$  **do**

            **if**  $v_{(e,l)} \notin V(T)$  **and**  $v_{(e,\pi(l))}$  is the first vertex with character  $e$  on path from  $v_{(*,0)}$  to

                 $v_{(d,j)}$  **and**  $\hat{f}_p^+(D_{(d,j)}) \geq \hat{f}_p^+(D_{(e,l)}) + \sum_{(f,s) \in \delta(d,j)} \hat{f}_p^+(D_{(f,s)})$  **then**

                    PUSH( $H, (v_{(d,j)}, v_{(e,l)})$ )

        **foreach**  $(v_{(e,l)}, v_{(f,s)}) \in H$  **do**

            **if**  $v_{(f,s)} = v_{(d,j)}$  **then**

                Remove  $(v_{(e,l)}, v_{(f,s)})$  from  $H$ 

            **else if**  $\exists p \in [m], \hat{f}_p^+(D_{(c,i)}) < \hat{f}_p^+(D_{(f,s)}) + \sum_{(h,t) \in \delta(c,i)} \hat{f}_p^+(D_{(h,t)})$  **then**

                Remove  $(v_{(e,l)}, v_{(f,s)})$  from  $H$ 

        NOISYENUMERATE( $G, T, H$ )

         $E(T) \leftarrow E(T) \setminus \{(v_{(c,i)}, v_{(d,j)})\}$ 

$z := \max\{\bar{x}, \bar{y}\}$ . As input we are given  $H = [h_{p,c}]$  where  $h_{p,c}$  is the VAF of character  $c$  in sample  $p$ . In addition, we are given  $\mathcal{M} = [[\mu_{p,c,(x,y)}]]$  where  $\mu_{p,c,(x,y)}$  is the *mixing proportion* of the copy-number state  $(x, y)$  of character  $c$  in sample  $p$ . The following linear system relates mixing proportions  $\mathcal{M}$  and variant allele frequencies  $H$  to state frequencies  $\mathcal{F} = [[f_{p,(c,(x,y,z))}]]$ .

$$\sum_z f_{p,(c,(x,y,z))} = \mu_{p,c,(x,y)} \quad \text{for all copy-number states } (x, y) \quad (\text{B.6})$$

$$\sum_{(x,y,z)} z \cdot f_{p,(c,(x,y,z))} = h_{p,c} \cdot \sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)} \quad (\text{B.7})$$

Given  $\mathcal{M}$  and  $H$ , we want to arrive at instances  $(\mathcal{F}, \mathcal{S})$  of the Cladistic-PPMDP. We do this by defining a *state graph*  $G$  whose vertices correspond to states  $(x, y, \bar{x}, \bar{y})$  and edges are allowed transitions between states. There are three types of edges:

## 1. Mutation edges:

- If  $\bar{x} = \bar{y} = 0$  and  $x > 0$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x, y, \bar{x} + 1, \bar{y}) \in E(G)$ .
- If  $\bar{x} = \bar{y} = 0$  and  $y > 0$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x, y, \bar{x}, \bar{y} + 1) \in E(G)$ .

## 2. Amplification edges:

- If  $x > 0$  and  $\bar{x} < x$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x + 1, y, \bar{x}, \bar{y}) \in E(G)$ .
- If  $x > 0$  and  $\bar{x} > 0$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x + 1, y, \bar{x} + 1, \bar{y}) \in E(G)$ .
- If  $x > y$ ,  $y > 0$  and  $\bar{y} < y$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x, y + 1, \bar{x}, \bar{y}) \in E(G)$ .
- If  $x > y$ ,  $y > 0$  and  $\bar{y} > 0$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x, y + 1, \bar{x}, \bar{y} + 1) \in E(G)$ .
- If  $x = y$ ,  $y > 0$  and  $\bar{y} < y$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (y + 1, x, \bar{y}, \bar{x}) \in E(G)$ .
- If  $x = y$ ,  $y > 0$  and  $\bar{y} > 0$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (y + 1, x, \bar{y} + 1, \bar{x}) \in E(G)$ .

## 3. Deletion edges:

- If  $x > \bar{x}$  and  $x > y$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x - 1, y, \bar{x}, \bar{y}) \in E(G)$ .
- If  $\bar{x} > 0$  and  $x > y$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x - 1, y, \bar{x} - 1, \bar{y}) \in E(G)$ .
- If  $y > \bar{y}$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x, y - 1, \bar{x}, \bar{y}) \in E(G)$ .
- If  $\bar{y} > 0$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (x, y - 1, \bar{x}, \bar{y} - 1) \in E(G)$ .
- If  $x = y$  and  $x > \bar{x}$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (y, x - 1, \bar{y}, \bar{x}) \in E(G)$ .
- If  $x = y$  and  $\bar{x} > 0$  then  $(x, y, \bar{x}, \bar{y}) \rightarrow (y, x - 1, \bar{y}, \bar{x} - 1) \in E(G)$ .

We define a *valid state tree* as follows.

**Definition 14.** A subtree  $S$  of  $G$  is a valid state tree provided that (1)  $S$  is rooted at  $(1, 1, 0, 0)$ , (2) there is at most one mutation edge in  $E(S)$ , and (3) vertices  $V_{(x,y)} \subseteq V(S)$  with identical copy-number states  $(x, y)$  form a connected subtree of  $S$ .

We remap each vertex  $v_{(x,y,\bar{x},\bar{y})} \in V(G)$  to  $v_{(x,y,z)}$  where  $z = \max\{\bar{x}, \bar{y}\}$ . We denote by  $S_{(x,y)}$  the subset of vertices that have copy-number state  $(x,y)$ . We now have the following lemma.

**Lemma 9.** *Let  $S$  be a valid state tree and let  $X = \{(x,y) \mid v_{(x,y,z)} \in V(S)\}$ . There exists at most one pair  $(x^*, y^*) \in X$  such that  $|V(S_{(x^*, y^*)})| = 2$  and  $|V(S_{(x,y)})| = 1$  for all other  $(x,y) \in X \setminus \{(x^*, y^*)\}$ .*

*Proof.* Since  $S$  is a valid state tree, there is at most one mutation edge in  $E(S)$ . If there is no mutation edge in  $S$  then, by Definition 14,  $S_{(x,y)} = \{v_{(x,y,0)}\}$  for all  $(x,y) \in X$ . Now consider the case where there is a mutation edge in  $E(S)$ . There must only be one such edge, which we denote by  $(x^*, y^*, 0) \rightarrow (x^*, y^*, 1)$ . By Definition 14, it holds that  $S_{(x^*, y^*)} = \{v_{(x^*, y^*, 0)}, v_{(x^*, y^*, 1)}\}$  and that  $|V(S_{(x,y)})| = 1$  for all  $(x,y) \in X \setminus \{(x^*, y^*)\}$ .  $\square$   $\square$

**Corollary 3.** *The system composed of (B.6) and (B.7) is fully determined.*

Let  $S$  be a valid state tree and let  $(x^*, y^*, 1)$  be a vertex of  $S$  such that  $|V(S_{(x^*, y^*)})| = 2$ . From (B.6) and (B.7) it follows that

$$f_{p,c,(x^*, y^*, 1)} = h_{p,c} \cdot \sum_{(x,y,z) \in V(S)} (x+y) \cdot \mu_{p,c,(x,y)} - \sum_{\substack{(x,y,z) \in V(S) \\ (x,y) \neq (x^*, y^*)}} z \cdot \mu_{p,c,(x,y)} \quad (\text{B.8})$$

$$= h_{p,c} \cdot \sum_{(x,y,z) \in V(S)} (x+y) \cdot \mu_{p,c,(x,y)} - \sum_{(x,y,z) \in V(S)} z \cdot \mu_{p,c,(x,y)} + \mu_{p,c,(x^*, y^*)}. \quad (\text{B.9})$$

If  $f_{p,c,(x^*, y^*, 1)} \geq 0$  for all samples  $p$  then  $S_c$  is *compatible* with character  $c$ . Given  $\mathcal{M}$  and  $H$ , the goal is to find all compatible valid state trees for each character  $c$ .

It may be the case that there exists two distinct compatible state trees  $S$  and  $S'$  such that for each distinct vertex  $(x,y,z) \notin V(S) \cap V(S')$  it holds that  $f_{p,c,(x,y,z)} = 0$ . We do not wish to distinguish between such state trees. Therefore, to deal with these degenerate cases, we remove all non-root vertices  $(x,y,z) \neq (1,1,0)$  where  $\mu_{p,c,(x,y)} = 0$  across all samples  $p$  through the following operation. We remove all incoming and outgoing edges of  $(x,y,z)$  and introduce edges  $\pi(x,y,z) \rightarrow (x',y',z')$  for every child  $(x',y',z') \in \delta(x,y,z)$ . The order in which we consider the vertices does

not matter.

## Noisy VAFs

We now show how to derive frequency intervals  $[\underline{f}_{p,(c,i)}, \bar{f}_{p,(c,i)}]$  from a valid state tree  $S_c$ , the VAF confidence interval  $[h_{p,c}, \bar{h}_{p,c}]$ , and the mixing proportions  $\mu_{p,c,(x,y)}$ . We define

$$[h'_{p,c}, \bar{h}'_{p,c}] = \left[ \frac{\sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)} - \mu_{p,c,(x^*,y^*)}}{\sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)}}, \frac{\sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)}}{\sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)}} \right] \cap [h_{p,c}, \bar{h}_{p,c}].$$

That is,  $[h'_{p,c}, \bar{h}'_{p,c}]$  is the VAF interval for which state tree  $S_c$  is compatible with character  $c$  in sample  $p$ . Character  $c$  is *compatible* with state tree  $S_c$  (in sample  $p$ ) only if  $[h'_{p,c}, \bar{h}'_{p,c}]$  is nonempty.

In that case, we have

$$\begin{aligned} [\underline{f}_{p,(c,(x^*,y^*,1))}, \bar{f}_{p,(c,(x^*,y^*,1))}] &= \left[ \underline{h}'_{p,c} \sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)} - \sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)} + \mu_{p,c,(x^*,y^*)}, \right. \\ &\quad \left. \bar{h}'_{p,c} \sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)} - \sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)} + \mu_{p,c,(x^*,y^*)} \right] \\ [\underline{f}_{p,(c,(x^*,y^*,0))}, \bar{f}_{p,(c,(x^*,y^*,0))}] &= [\mu_{p,c,(x^*,y^*)} - \bar{f}_{p,(c,(x^*,y^*,1))}, \mu_{p,c,(x^*,y^*)} - \underline{f}_{p,(c,(x^*,y^*,1))}] \end{aligned}$$

and

$$[\underline{f}_{p,(c,(x,y,z))}, \bar{f}_{p,(c,(x,y,z))}] = [\mu_{p,c,(x,y)}, \mu_{p,c,(x,y)}] \quad \text{for all } (x,y,z) \text{ where } (x,y) \neq (x^*,y^*).$$

## B.2 Supplemental Results

We include here additional results on both simulated data and real data not included in the main text.



- Figure B.7 shows simulated data results for  $n = 5$  instances.
- Figure B.8 shows simulated data results for instances with  $n = 15$ ,  $m = 10$  and a target coverage of 1,000x.
- Figure B.9 shows the recall of the representative tree in case of violations of the infinite alleles assumption on 20 simulated instances with  $n = 5$ ,  $m = 10$  and a target coverage of 1,000x.
- Figure B.10 shows the solution space of CLL077 [133].
- Figure B.11 shows the reported tree by Gundem et al.[61].
- Figure B.12 shows a summary of the thirteen characters of A22 that had no compatible state tree.
- Figure B.13 shows the distribution of maximal clique sizes in the pairwise compatibility graph of A22.
- Figure B.14 shows the trees referred to in Section 3.2.3.
- Figure B.15 shows the relationship between  $T_{(c,i)}$  and  $\overline{T}_{(c,i)}$ .

### B.2.1 Simulations

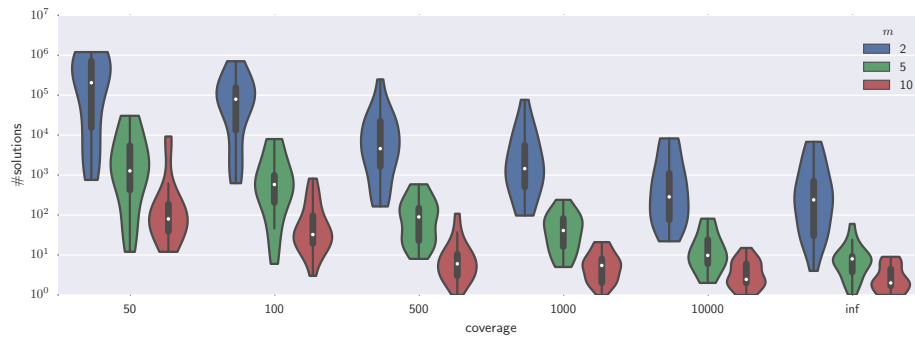
We generated simulated instances using the following procedure. We create 20 multi-state perfect phylogeny trees  $T^*$  containing  $n = 5$  characters, using randomly generated state trees that have at most two copy-number states per character with  $x \leq 3$  and  $y \leq 3$ . For each simulated tree  $T^*$ , we generate a frequency tensor  $\mathcal{F}$  containing  $m \in \{2, 5, 10\}$  samples, by mixing vertices from  $T^*$ . This results in 60 simulated mixtures. Next, we use the entries of the simulated  $\mathcal{F}$  to compute the input by calculating for each character  $c$  in sample  $p$ , the VAF  $h_{p,c}$  and the mixing proportions  $\mu_{p,c,(x,y)}$  as defined in Section 3.2.2. For each character  $c$  and sample  $p$ , we draw its total read count from a Poisson distribution parameterized by a target coverage. Next we draw the number of variant reads from a binomial parameterized by the previously drawn total read count and the true VAF.

We then consider the posterior distribution of observing the drawn total and variant read counts from which we compute 0.99 confidence intervals on the VAF [38].

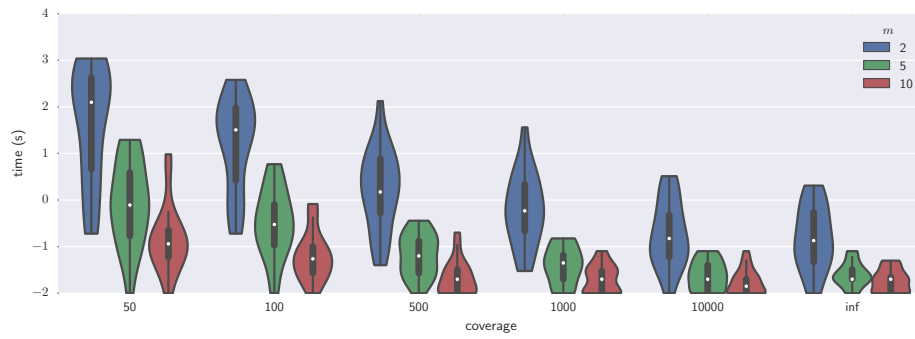
### B.2.2 Comparison against PhyloWGS

We run PhyloWGS with default parameters (2500 MCMC samples, 5000 Metropolis-Hastings iterations) on the simulated instances with  $n = 5$  characters and  $m \in \{2, 5, 10\}$  samples. As PhyloWGS does not have a “perfect data” mode, we provide correct VAF and copy number mixing proportions. We indicate that these are high-confidence by specifying a coverage of 1,000,000x. We also run PhyloWGS on larger instances with  $n = 15$  characters and noisy VAFs (1,000x coverage).

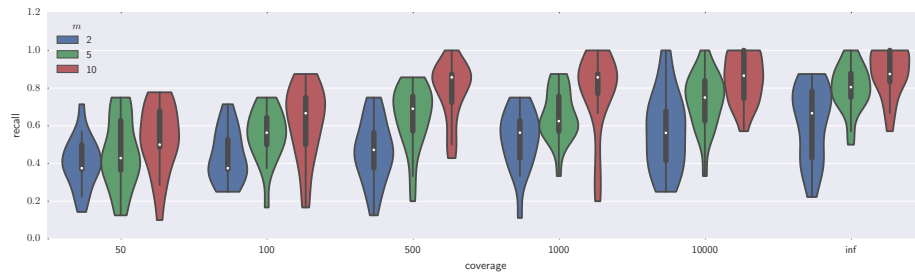
PhyloWGS has a two-state perfect phylogeny model where characters correspond to either CNA events or SNV events, and expected SNV frequencies are adjusted based on associated CNAs. As the method does not report the number of copies that contain each mutation, we could not directly map the results onto  $(x, y, z)$  states. Instead, for the comparison, we map the states in the true trees to an  $(x, y)$  copy number state, and an indicator  $w_c$  for the presence of a mutation in character  $c$ . We calculate recall as the fraction of edges (characterized by an ordered pair of states) from the simulated tree that are recovered in the solution. Some vertices from the output PhyloWGS trees contained clusters of mutations. In those instances, we include all combinations of pairs from the parent cluster to a child cluster.



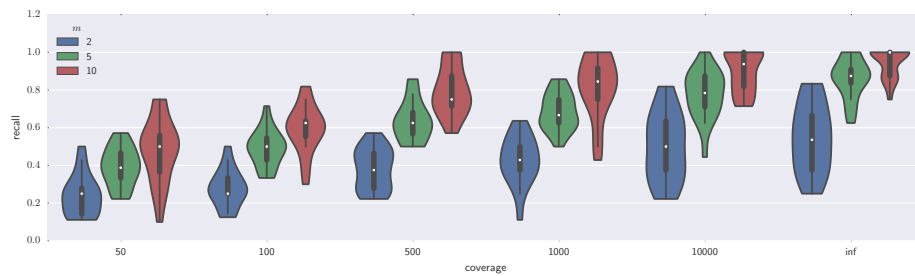
(a) Number of solutions (log-scale)



(b) Running time (seconds, log-scale)



(c) Recall of representative tree



(d) Median recall

Figure B.7: **Simulated data results for  $n = 5$  instances with noisy VAFs.** A coverage of ‘inf’ corresponds to error-free VAFs.

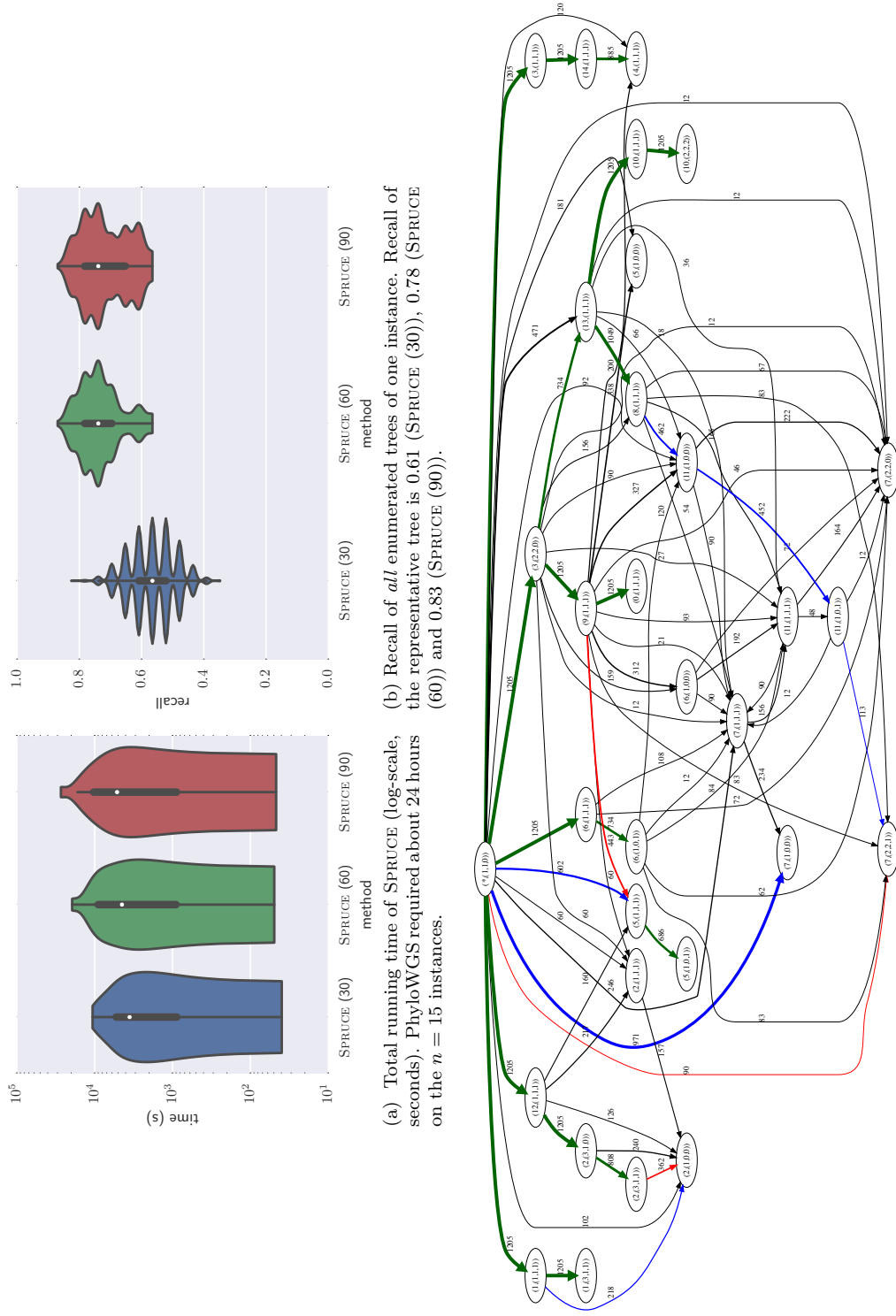


Figure B.8: Simulated data results for larger instances with  $n = 15$ ,  $m = 10$  and target coverage of 1,000x



Figure B.9: **Recall of the representative tree in case of violations of infinite alleles:** for 20 instances with  $n = 5$ ,  $m = 10$  and target coverage of 1,000x;  $x$ -axis denotes the number of violations.

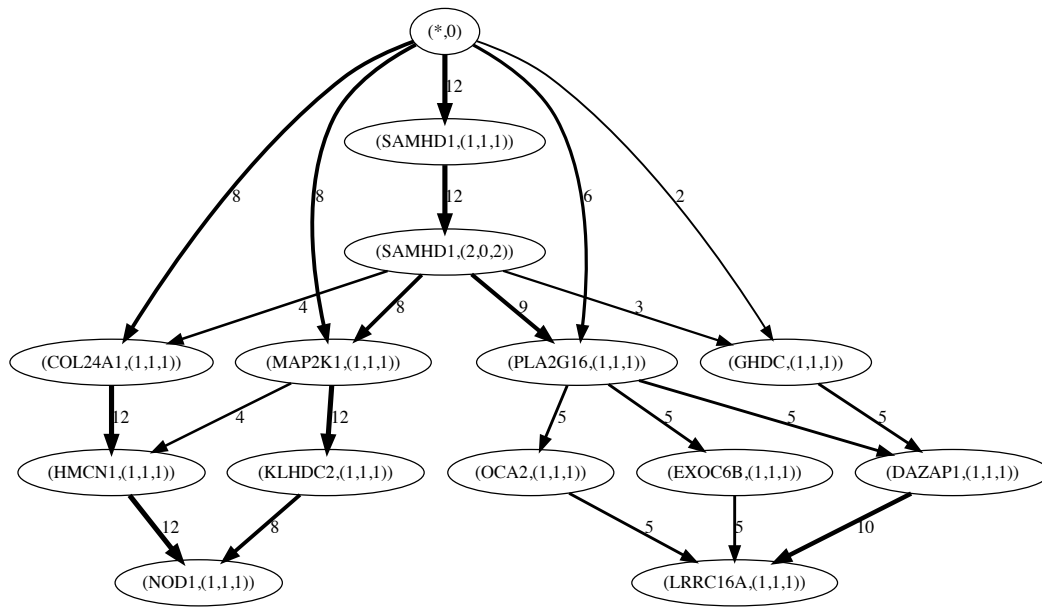


Figure B.10: **Enumerated solution space of tumor CLL077 (20 trees).** Vertices correspond to the vertices of the solution trees and each edge is labeled by the number of solutions in which it occurs. We note that GPR158 is not shown here as it was not contained in any of the 20 solutions.

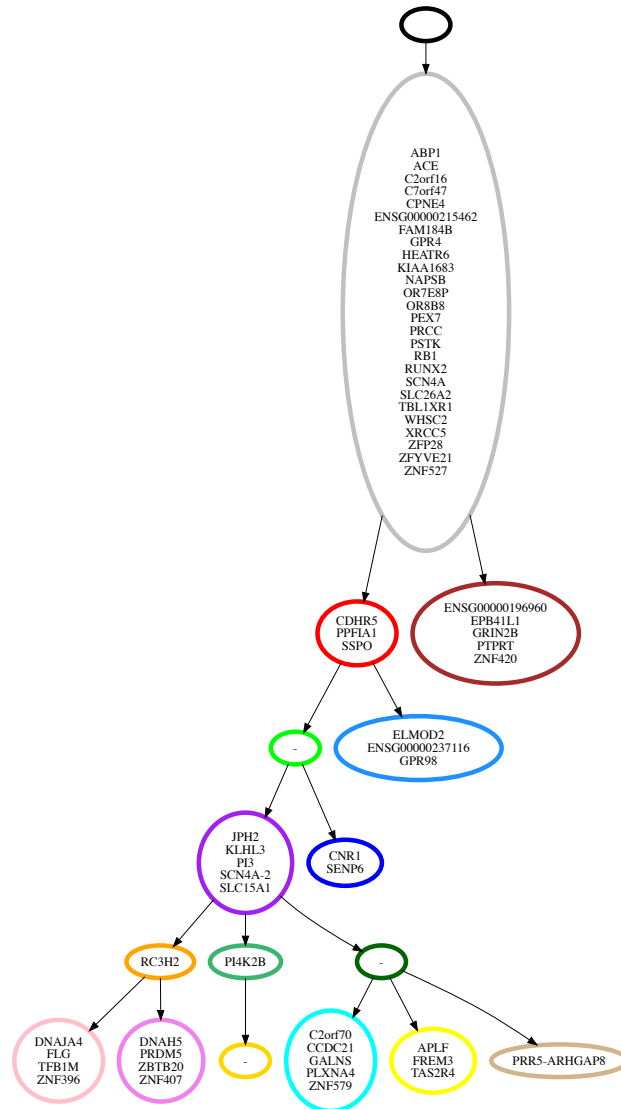


Figure B.11: **Tree on A22 reported by Gundem et al.[61]:** Vertices and coloring correspond to the reported clusters on SNVs. Note that we only show the subset of SNVs that pass the filtering criteria described in Section 3.2.3. As such, some of the reported clusters are empty.

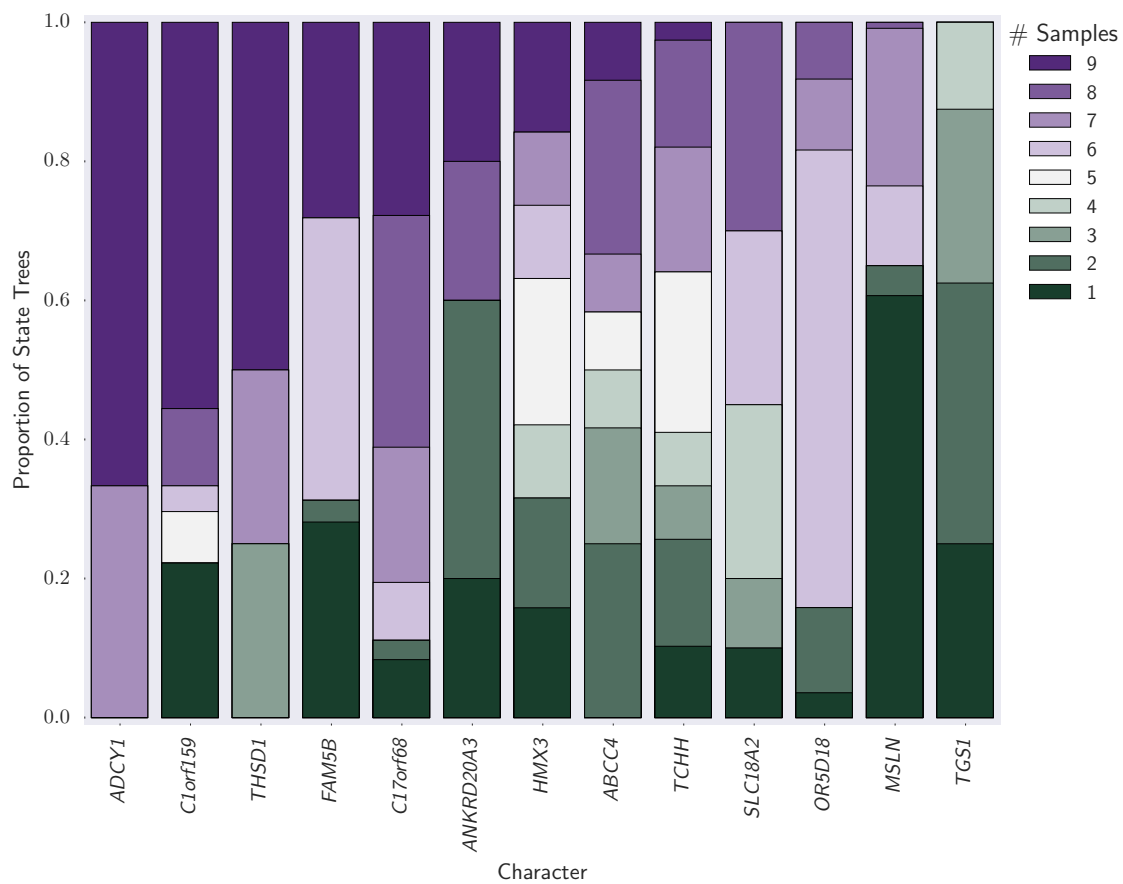


Figure B.12: **Thirteen characters had no state tree that was compatible across all samples.** Each column corresponds to a character, with stacked bars denoting the fraction of state trees compatible with a specified number of samples indicated by the color.

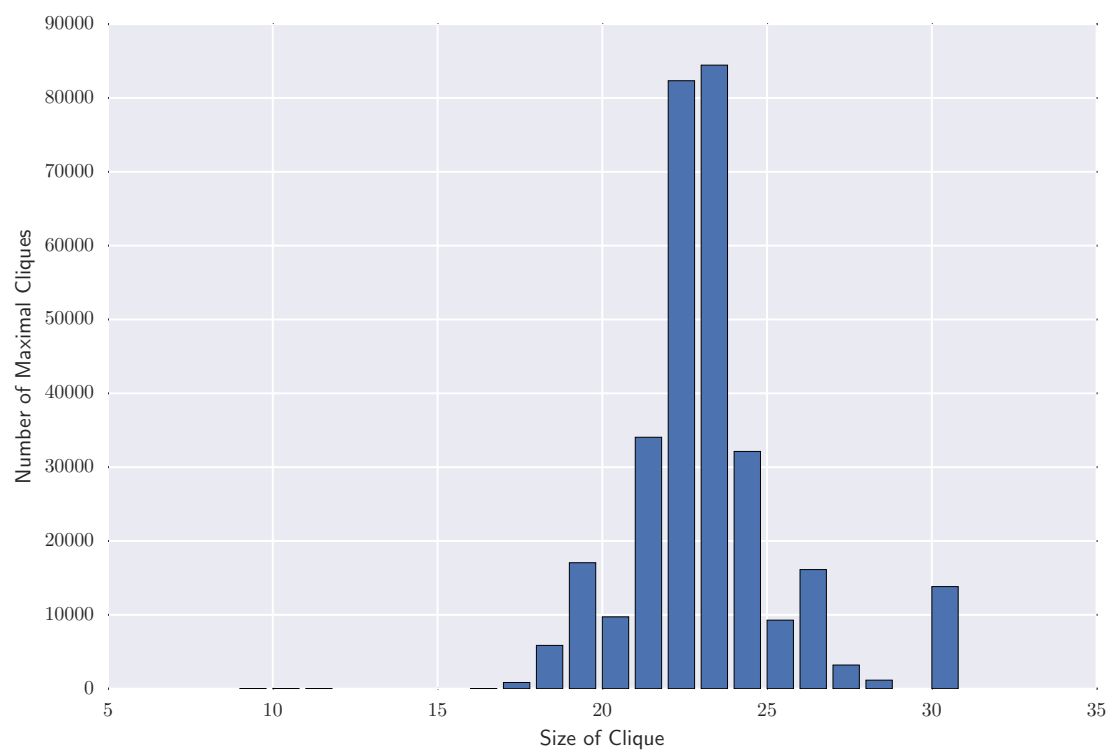


Figure B.13: Distribution of sizes of maximal cliques in the compatibility graph of A22.



Figure B.14: **A22 results.** (A) State tree  $S_1$  for  $PPFIA1$  results in a tree where  $PPFIA1$  is a parent of the blue cluster. This relationship is consistent with the tree reported in [61]. (B) State tree  $S_2$  for  $PPFIA1$  results in a tree in which the blue and the red clusters are siblings.

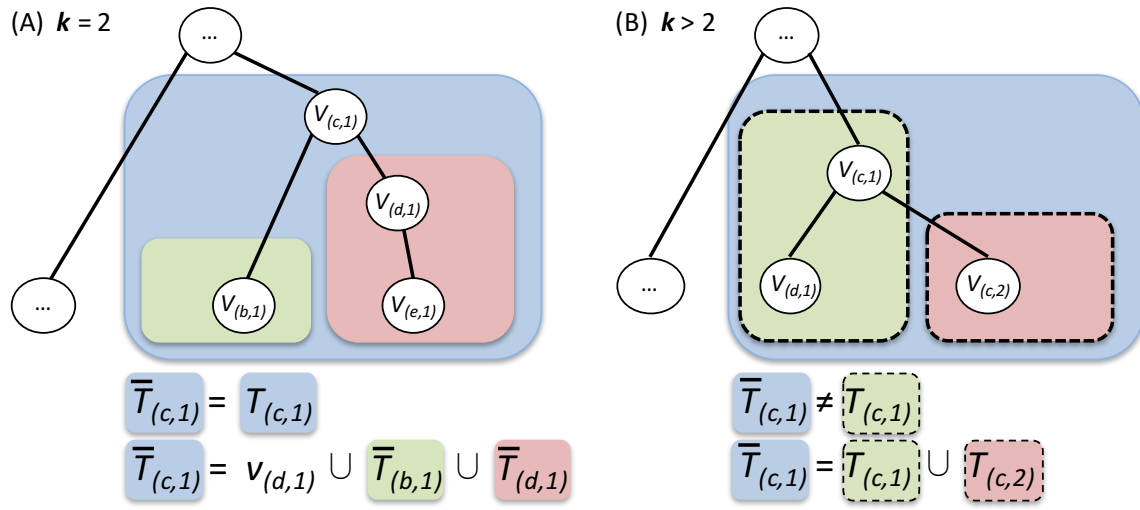


Figure B.15: **Relationship between  $T_{(c,i)}$  and  $\bar{T}_{(c,i)}$  in the case of  $k = 2$  and  $k > 2$  states.** (A) In the case of  $k = 2$  states, we have that  $T_{(c,1)} = \bar{T}_{(c,1)}$ . (B) In the case of  $k > 2$  states,  $T_{(c,i)} \neq \bar{T}_{(c,i)}$ . Instead,  $\bar{T}_{(c,i)} = \bigcup_{l \in D_{(c,i)}} T_{(c,l)}$  where  $D_{(c,i)}$  is the descendant set of  $(c,i)$ . Here,  $D_{(c,1)} = \{1, 2\}$ .

## Appendix C

# Supplementary Methods and Results for Chapter 4

### C.1 Supplementary Methods

### C.2 Supplementary Tables and Figures

### C.3 Supplementary Methods

#### C.3.1 Solving ML-LSR Problem: Likelihood model

We use a probabilistic model to evaluate mutation profiles using observed read counts. If mutation  $a$  is absent in cell  $v$  (i.e.,  $b_{v,a} = 0$ ), then the probability of observing a variant read corresponds to

Table C.1: Single-cell phylogenetic reconstruction algorithms

| Algorithm  | Publication | Evolutionary Model | Input Format      | Additional Data |
|------------|-------------|--------------------|-------------------|-----------------|
| OncoNem    | [129]       | Infinite Sites     | Binary            | Bulk Sequencing |
| SCITE      | [71]        | Infinite Sites     | Binary or Ternary |                 |
| B-SCITE    | [94]        | Infinite Sites     | Binary            |                 |
| SciPhi     | [138]       | Infinite Sites     | Read Counts       |                 |
| SASC       | [27]        | Dollo              | Binary            | Bulk Sequencing |
| SPhyR      | [37]        | Dollo              | Binary            |                 |
| SiFit      | [163]       | Finite Sites       | Binary or Ternary |                 |
| SiCloneFit | [164]       | Finite Sites       | Binary or Ternary |                 |
| PhiSCS     | [95]        | Finite Sites       | Binary            | Copy-number     |
| SCARLET    |             | Loss-Supported     | Read Counts       |                 |

the per-nucleotide rate of sequencing error  $\epsilon$ . For Illumina sequencing reads, we use  $\epsilon = 0.001$ . If mutation  $a$  is present in cell  $v$  (i.e.,  $b_{v,a} = 1$ ), then we model the variant counts at a locus using a beta-binomial distribution (similarly to the model previously used by SciPhi [138]). We estimate parameters  $\alpha$  and  $\beta$  empirically from the distribution of heterozygous germline single-nucleotide polymorphisms (SNPs) in the data. We thus define the data likelihood for observing  $x_{v,a}$  variant reads at locus  $a$  in cell  $v$  as follows,

$$\Pr(x_{v,a} \mid y_{v,a}, b_{v,a}) = \begin{cases} \text{Beta-Binomial}(x_{v,a} \mid n = y_{v,a}, \alpha, \beta) & \text{if } b_{v,a} = 1, \\ \text{Binomial}(x_{v,a} \mid n = y_{v,a}, p = \epsilon) & \text{if } b_{v,a} = 0. \end{cases} \quad (\text{C.1})$$

We assume variant read counts are independent across variants given  $Y$  and  $\mathbf{B}$ . This gives us the following total likelihood:

$$\Pr(X \mid Y, B) = \prod_{v=1}^n \prod_{a=1}^m \Pr(x_{v,a} \mid y_{v,a}, b_{v,a}). \quad (\text{C.2})$$

### C.3.2 Solving ML-LSR Problem: Finding max-likelihood subtree roots

As the first step of SCARLET, we find the maximum likelihood subtree roots  $R$  for all internal subtrees. We must select a set of subtree roots  $R$  such that there exists a loss-supported refinement

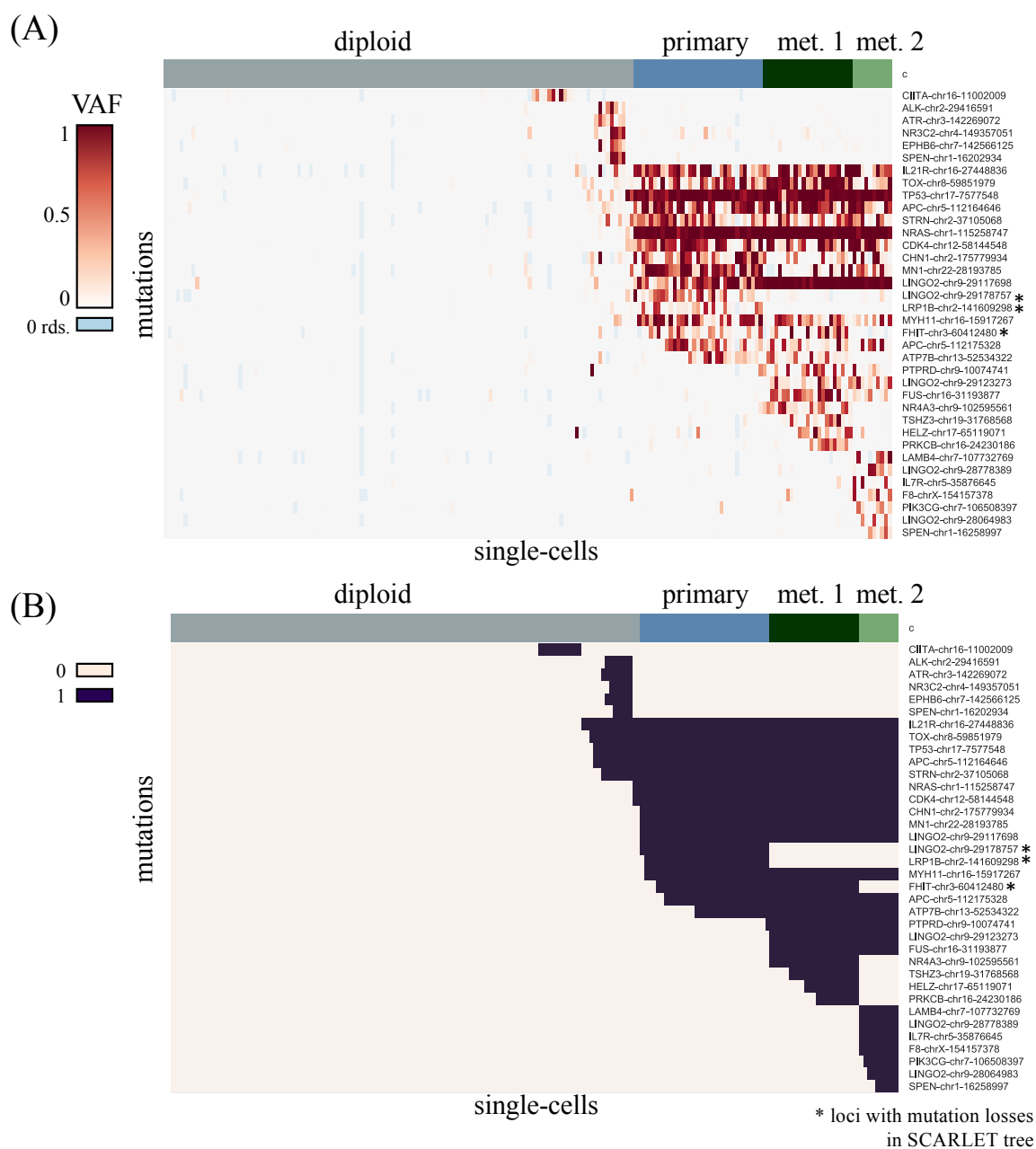


Figure C.1: **CRC2 Data and SCARLET Mutation Matrix.** (A) Variant allele frequencies for 36 single-nucleotide variants detected in CRC2. (B) Maximum likelihood mutation matrix identified by SCARLET. Loci in which SCARLET detected a mutation loss are marked with an \*.

$T', \mathbf{c}', \mathbf{B}'$  with subtree roots  $R$ . This constrains the possible mutation profiles of the roots. Specifically, by the Definition of a loss-supported phylogeny for each locus  $a$ , mutation  $a$  is gained at most once in  $T'$ . Let subtree  $T_a$  of  $T$  correspond to the set of vertices  $v$  for which mutation  $a$  is present in subtree  $T[\gamma(v)]$ .  $R$  is a valid mutation state assignment for roots  $R$  provided for each locus  $a$ , (1) that there exists a subtree  $S_a$  of  $T$  such that for all  $v \in V(T)$ ,  $r_{v,a} = 1$  if  $v \in V(S_a)$  and  $v$  is not the root of  $S_a$  and  $r_{v,a} = 0$  otherwise; and (2) for any edge  $(v, w) \in T$  such that  $v \in S_a$  and  $w \notin S_a$ ,  $a \in \mathcal{L}(c_v, c_w)$ . Any valid  $R$  uniquely defines a subtree  $S_a$  for each locus  $a$ . Roots  $R$  admit a mutation profile  $\mathbf{b}_a$  provided that  $\mathbf{b}_a$  satisfies the following.

1. If  $v \notin S_a$  then mutation  $a$  is absent in all cells  $v'$  such that  $c_{v'} = c_v$ .
2. If  $v \in S_a$  and  $v$  is not the root of  $S_a$  then mutation  $a$  is present in all cells  $v'$  such that  $c_{v'} = c_v$ .
3. If  $v \in S_a$  and  $v$  is the root of  $S_a$  then mutation  $a$  is either present or absent, as mutation  $a$  occurred in  $T'_v$

Let  $\beta_R = \{\mathbf{b}_a : R \text{ admits } \mathbf{b}_a\}$  be the set of mutation profile for mutation  $a$  admitted by roots  $R$ . The likelihood given a root assignment  $R$  is computed by marginalizing over admitted mutation profiles, as follows.

$$\Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{R}) = \prod_{a=1}^m \prod_{\mathbf{b}_a \in \beta_R} \Pr(\mathbf{x}_a \mid \mathbf{y}_a, \mathbf{b}_a) = \prod_{a=1}^m \prod_{v=1}^n \Pr(x_{v',a} \mid y_{v',a}, R) \quad (\text{C.3})$$

such that

$$\Pr(x_{v',a} \mid y_{v',a}, R) = \begin{cases} \Pr(x_{v',a} \mid y_{v',a}, b_{v',a} = 0) & \text{if } c_{v'} = c_v \text{ and } v \notin S_a \\ \Pr(x_{v',a} \mid y_{v',a}, b_{v',a} = 1) & \text{if } c_{v'} = c_v, v \in S_a \text{ and } v \text{ is not root of } S_a \\ \frac{1}{2} \Pr(x_{v',a} \mid y_{v',a}, b_{v',a} = 1) \\ \quad + \frac{1}{2} \Pr(x_{v',a} \mid y_{v',a}, b_{v',a} = 0) & \text{otherwise} \end{cases} \quad (\text{C.4})$$

We thus find  $R^*$  by enumerating valid mutation state assignments for roots for each mutation locus  $a$ , then computing the maximum likelihood as above.

### C.3.3 Solving ML-LSR Problem: Integer Linear Program (ILP) to solve maximum likelihood IDP

We define a cost-function

$$C_{p,a} = \begin{cases} \log \Pr(x_{v,a} \mid y_{v,a}, b_{v,a} = 1) - \log \Pr(x_{v,a} \mid y_{v,a}, b_{v,a} = 0) & \text{if } v \text{ is observed,} \\ M & \text{if } v \text{ is not observed and } \bar{b}_{v,a} = 0 \\ -M & \text{if } v \text{ is not observed and } \bar{b}_{v,a} = 0 \\ 0 & \text{if } v \text{ is not observed and } \bar{b}_{v,a} = ?. \end{cases} \quad (\text{C.5})$$

This gives us the following ILP. We introduce a set of auxiliary variables  $F, G, H$  to enforce the three gametes condition, where  $F_{a,b}$ ,  $G_{a,b}$  and  $H_{a,b}$  indicate that a pair of columns  $a, b$  show the  $(1, 1)$ ,  $(0, 1)$  and  $(1, 0)$  gametes respectively.

$$\text{maximize } \sum_{p,a} C_{p,a} \cdot B_{p,a}$$

subject to

$$\begin{aligned}
F_{a,b} + G_{a,b} + H_{a,b} &\leq 2 && \text{for all } a, b \\
F_{p,a,b} &\geq B_{p,a} + B_{p,b} - 1 && \text{for all } a, b, p \\
F_{p,a,b} &\leq B_{p,a} && \text{for all } a, b, p \\
F_{p,a,b} &\leq B_{p,a} && \text{for all } a, b, p \\
F_{a,b} &\leq \sum_i F_{p,a,b} && \text{for all } a, b \\
F_{a,b} &\geq F_{p,a,b} && \text{for all } a, b, p \\
G_{p,a,b} &\geq (1 - B_{p,a}) + B_{p,b} - 1 && \text{for all } a, b, p \\
G_{p,a,b} &\leq (1 - B_{p,a}) && \text{for all } a, b, p \\
G_{p,a,b} &\leq B_{p,b} && \text{for all } a, b, p \\
G_{a,b} &\leq \sum_i G_{p,a,b} && \text{for all } a, b \\
G_{a,b} &\geq G_{p,a,b} && \text{for all } a, b, p \\
H_{p,a,b} &\geq B_{p,a} + (1 - B_{p,b}) - 1 && \text{for all } a, b, p \\
H_{p,a,b} &\leq B_{p,a} && \text{for all } a, b, p \\
H_{p,a,b} &\leq (1 - B_{p,b}) && \text{for all } a, b, p \\
H_{a,b} &\leq \sum_i H_{p,a,b} && \text{for all } a, b \\
H_{a,b} &\geq H_{p,a,b} && \text{for all } a, b, p \\
B_{p,a} &\in \{0, 1\} && \text{for all } a, p \\
F_{p,a,b} &\in \{0, 1\} && \text{for all } a, b, p \\
G_{p,a,b} &\in \{0, 1\} && \text{for all } a, b, p \\
H_{p,a,b} &\in \{0, 1\} && \text{for all } a, b, p \\
F_{a,b} &\in \{0, 1\} && \text{for all } a, b \\
G_{a,b} &\in \{0, 1\} && \text{for all } a, b \\
H_{a,b} &\in \{0, 1\} && \text{for all } a, b
\end{aligned}$$



### C.3.4 Simulation Details

We simulate 50 trees with 20 mutations and 4 copy-states. We do not directly model copy-number evolution in our simulation. Instead, for each edge between cells  $p, q$  with different copy-number states  $i, j$ , we simulate the set of supported losses  $\mathcal{L}_{i,j}$  by selecting a subset of loci of size  $|\mathcal{L}_{i,j}| \sim \text{Poisson}(0.2*m)$ . A mutation loss  $-a$  is introduced on an edge with probability 0.5 provided that the source cell  $p$  contains mutation  $a$ ,  $a \in \mathcal{L}_{i,j}$ . The resulting trees have between 1-8 mutation losses. In no instances was the same mutation lost multiple times on the same tree. Thus, all 50 trees respect the  $k - \text{Dollo}$  model with  $k = 1$ . From each tree, we sample 100 cells and simulate sequencing error as follows. For each locus  $a$  in cell  $p$ , we generate a variant read count  $v_{p,a}$  and a total read count  $r_{p,a}$ . We simulate data with an allelic dropout rate of  $d = 0.15$ . With probability  $d^2$ , both alleles drop out, i.e.,  $v_{p,a} = r_{p,a} = 0$ . Otherwise, the total number of reads is  $r_{p,a} \sim \text{Poisson}(100)$ . If the variant is absent,  $v_{p,a} \sim \text{Binomial}(t_{p,a}, \epsilon)$  where  $\epsilon = 0.001$  models the sequencing error rate. If the variant is present we model the overdispersion in the variant read distribution resulting from whole-genome amplification using a Beta-Binomial model. Let  $f_{p,a} \sim \max\{\text{Beta}(\alpha, \alpha), \epsilon\}$  (with  $\alpha = 0.25$  in order to obtain a dropout rate of  $d \approx 0.15$ ) and  $v_{p,a} \sim \text{Binomial}(t_{p,a}, f_{p,a})$ .

SCITE and SPhyR and SiFit were run with default parameters. In the case where SCITE returned multiple solutions, we selected the first solution that was returned. SPhyR was run with a maximum number  $k = 1$  of mutation losses per locus. All algorithms were given the true allelic dropout and sequencing error rates. As SciPhi is designed for variant calling from a large set of putative variants, and thus has filters to remove false positives, we found that the default parameters removed a lot of the variants present in the tree. The parameters for these filters were (1) `cwm` the number of cells that show the mutation, (2) `nmc` the number of cells that must have a minimum alternate read count of (3) `ms`. As all variants including in the simulation are present in at least one cell, we set all three of these parameters to be 0, so as to not pre-filter out any variants.

### C.3.5 Copy-number analysis of CRC2

To assign cells to copy-number states for CRC2, we first binned reads at 10Mb. We next performed hierarchical clustering on cells. A cut-off for hierarchical clustering was selected to yield 4 clusters, the observed number of copy-number clones that was identified in the copy-number analysis performed by [83]. Supported losses were inferred by looking for significant shifts in read-depth in the bin containing the variant of interest using a Wilcoxon rank-sum test and a p-value threshold of  $p = 0.1$ .

## C.4 Proofs

### C.4.1 Proof of Theorem 1

**Theorem 5.** *Given copy-number tree  $T$ , copy-number state vector  $\mathbf{c}$ , mutation matrix  $B$  and supported losses  $\mathcal{L}$ , a phylogeny  $T'$ , copy-number state vector  $\mathbf{c}'$  and augmented mutation matrix  $B'$  are a solution to the LSR problem if and only if  $T'$  is a refinement of  $T$  and for all  $v \in V(T)$*

1. *,  $\mathbf{c}'_{v'} = \mathbf{c}_v$  for all  $v' \in \gamma(v)$ ;*
2.  *$T'[\gamma(v)]$  is a perfect phylogeny with respect to  $B'[\gamma(v)]$ ;*
3. *If  $b'_{r(v),a} = 1$  then  $b'_{p(r(v)),a} = 1$  for all  $a \notin \mathcal{L}(\mathbf{c}'_{p(r(v))}, \mathbf{c}'_{r(v)})$ ;*
4. *For all mutations  $a$ , there exists exactly one edge  $(v', w') \in E(T')$  with  $b'_{v',a} = 0$  and  $b'_{w',a} = 1$ .*

*Proof.* We first show that any solution that meets these constraints is a solution to the LSR problem. The constraints about  $T'$  being a refinement and copy-number consistency (constraint 1 in Problem 1) are explicitly enforced in this theorem. We must show that  $T'$  is also a loss-supported phylogeny. The requirements for a loss-supported phylogeny is that every mutation occurs only once, which is enforced by (4) in this theorem, and that every mutation loss is supported. By (2), there are

no mutation losses between cells that have the same copy-number state, and by (3) between cells with different copy-number states, mutation losses that are not supported are not allowed. Thus,  $(T', B', \mathbf{c}')$  is a solution to the LSR.

We next show that any solution to the LSR also meets these constraints. We will do this by showing that any  $(T', B', \mathbf{c}')$  that violates any one of these constraints cannot be a solution to the LSR.

1.  $T'$  being a refinement and copy-number consistency (item (1)) are explicitly required by the LSR problem statement, and thus cannot be violated by any solution.
2. If  $T'[\gamma(v)]$  is not a perfect phylogeny with respect to  $B'[\gamma(v)]$ , either there exists a mutation that occurs more than once in subtree  $T'[\gamma(v)]$ , in which case  $T'$  is not a Dollo phylogeny, or there exists a mutation that is lost in subtree  $T'[\gamma(v)]$ . As every vertex in  $T'[\gamma(v)]$  has the same copy-number state (by item (1)), this mutation loss is not supported. In either case it would mean  $T'$  is not a loss-supported phylogeny and thus not a solution to the LSR problem.
3. If this condition is violated, then there is a mutation loss that is not supported. Thus,  $T'$  is not a loss-supported phylogeny, and consequently not a solution to the LSR problem.
4. This condition must be true for  $T'$  to be a Dollo, and thus a loss-supported phylogeny, and thus cannot be violated by any solution.

□

### C.4.2 Proof of Lemma 1

**Lemma 10.** *If there exists a solution to the LSR problem for a given  $T$ ,  $\mathbf{c}$ ,  $\mathcal{L}$ ,  $\mathbf{B}$ , then there exists a solution  $T'$ ,  $\mathbf{c}'$ ,  $\mathbf{B}'$  that meets the following conditions.*

1. For all  $(v, w) \in E(T)$ ,  $p(r(w))$  is a leaf of subtree  $T'[\gamma(v)]$ .
2. For all  $v \in V(T)$  such that  $v$  is not the root of  $T$ ,  $b_{r(v)} = 1$  if  $b_{v'} = 1$  for all  $v' \in L(T'[\gamma(v)])$ .
3. For all  $v \in V(T)$  and all loci  $a$ ,  $b_{p(r(v)),a} \geq b_{r(v),a}$ .

*Proof.* We will show by construction that for any solution  $T', \mathbf{c}', \mathbf{B}'$  that violates these constraints, there exists another solution  $T'', \mathbf{c}'', \mathbf{B}''$  that meets these constraints.

**(1) For all  $(v, w) \in E(T)$ ,  $p(r(w))$  is a leaf of subtree  $T'[\gamma(v)]$ .** Consider an edge  $(p(r(w)), r(w)) \in E(T')$  such that  $p(r(w))$  is not a leaf of  $T'[\gamma(v)]$ . That is,  $p(r(w))$  has another child in  $T'[\gamma(v)]$ . We construct  $T''$  by splitting  $p(r(w))$  into two vertices  $u$  and  $u'$  such that there is an edge  $(u, u') \in E(T'')$ ,  $\mathbf{b}_u'' = \mathbf{b}_{u'}''$  and  $c_u'' = c_{u'}''$ , and the only outgoing edge from  $u'$  is  $(u', r(w))$ . Thus,  $u'$  is now the parent of  $r(w)$  and  $u$  is a leaf. This split preserves the rest of the tree and does not introduce violations of any of the conditions in Theorem 1 or any of the other assumptions in this Lemma. Thus  $T'', \mathbf{B}'', \mathbf{c}''$  is a solution to the LSR problem.

**(2) For all  $v \in V(T)$  such that  $v$  is not the root of  $T$ ,  $b_{r(v)} = 1$  if  $b_{v'} = 1$  for all  $v' \in L(T'[\gamma(v)])$ .** Assume that  $T', \mathbf{c}', \mathbf{B}'$  meet condition 1. If this constraint is violated, this means that there is some mutation  $a$  that is gained in a subtree  $T'[\gamma(v)]$  but there are no leaves of  $T'[\gamma(v)]$  that do not contain  $a$ . Let  $T'' = T'$ ,  $\mathbf{c}'' = \mathbf{c}'$ . Let  $b'_{v',a} = 1$  if  $v \in T'[\gamma(v)]$ . This change does not violate any of the conditions in Theorem 1. Specifically, this change does not introduce new mutation gains, and as this only alters the mutation profiles of internal vertices of  $T'[\gamma(v)]$  so this cannot introduce new mutation losses. As  $T''$  and  $\mathbf{c}''$  are preserved, refinement and copy-number consistency conditions are automatically met. This change may introduce violations to Assumption 3 in this Lemma that can subsequently be corrected as below.

**(3) For all  $v \in V(T)$  and all loci  $a$ ,  $b_{p(r(v)),a} \geq b_{r(v),a}$ .** This constraint states that there are no mutation gains on edges between subtrees. We construct  $T''$  by performing a similar split as we did for constraint (1). Suppose there's an edge  $(p(r(v)), r(v)) \in E(T')$  such that  $b_{p(r(v)),a} = 0$  and  $b_{r(v),a} = 1$ . Split  $p(r(v))$  into vertices  $u, u'$  such that  $b_{u,a} = 0$  and  $b_{u',a} = 1$ , and for all  $a' \neq a$ ,  $b_{u,a'} = b_{u',a'}$  and the only outgoing edge from  $u'$  is  $(u', r(v))$ . This split preserves the rest of the tree and does not introduce violations of any of the conditions in Theorem 1 or any of the other assumptions in this Lemma. Thus  $T'', \mathbf{B}'', \mathbf{c}''$  is a solution to the LSR problem.  $\square$

### C.4.3 Hardness of ML-LSR

**Lemma 11.** *The ML-LSR is NP-hard.*

*Proof.* We show this by reduction from the Flip problem which is known to be NP-Complete [23].

**Problem 4.** *Given a binary matrix  $\mathbf{B} \in \{0, 1\}^{m \times n}$  and integer  $\kappa \in \mathbb{N}$ , decide whether there exists a directed perfect phylogeny matrix  $\mathbf{B}' \in \{0, 1\}^{m \times n}$  such that no more than  $\kappa$  entries in  $\mathbf{B}'$  differ from  $\mathbf{B}$ .*

Let  $(\mathbf{B}, \kappa)$  be an instance of the Flip problem. For the corresponding instance of the ML-LSR problem, we let  $k = 1$  and define the inputs as follows:

1.  $T$  is the star phylogeny, where all leaves  $v \in V(T)$  are attached to a single internal vertex;
2.  $\mathbf{c} = \vec{1}$ ;
3. No mutation losses are supported in  $\mathcal{L}$ ;
4.  $\mathbf{X} = \mathbf{B}$ , and  $\mathbf{Y} = [1]^{m \times n}$ .

Define a likelihood function  $\Pr(X \mid Y, B^*)$  that is symmetric when  $x_{v,a} \in \{0, 1\}$  and  $y_{v,a} = 1$

$$\log(\Pr(x_{v,a} \mid y_{v,a}, b_{v,a})) = \begin{cases} \alpha & \text{if } x_{v,a} = b_{v,a} \\ \beta & \text{if } x_{v,a} \neq b_{v,a} \end{cases} \quad (\text{C.6})$$

such that  $\beta < \alpha$ . Thus the log-likelihood of a matrix  $B^*$  is

$$\log \Pr(X \mid Y, B^*) = \sum_{v=1}^n \sum_{a=1}^m \log(\Pr(x_{v,a} \mid y_{v,a}, b_{v,a})) \quad (\text{C.7})$$

$$= \lambda \cdot \beta + (mn - \lambda)\alpha \quad (\text{C.8})$$

$$(\text{C.9})$$

We claim that there exists a perfect phylogeny matrix  $\mathbf{B}'$  with at most  $\kappa$  changes if and only if there exists a solution  $\mathbf{B}^*$  to the ML-LSR

$$\log \Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B}^*) \geq \kappa \cdot \beta + (mn - \kappa) \cdot \alpha \quad (\text{C.10})$$

$$(\text{C.11})$$

We first show the forward direction. If there exists a perfect phylogeny matrix  $\mathbf{B}'$  with at most  $\kappa$  changes from  $\mathbf{B}' = \mathbf{X}$ , then the log-likelihood  $\Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B}') \geq \kappa \cdot \beta + (mn - \kappa) \cdot \alpha$ . Thus for the maximum likelihood solution,  $\mathbf{B}^*$ ,  $\Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B}^*) \geq \Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B}')$ .

We next show the reverse direction. If  $\Pr(\mathbf{X} \mid \mathbf{Y}, \mathbf{B}^*) \geq \kappa \cdot \beta + (mn - \kappa) \cdot \alpha$  then  $\mathbf{B}^*$  has at most  $\kappa$  changes from  $\mathbf{X} = \mathbf{B}$ . Thus, there exists a  $\mathbf{B}' = \mathbf{B}^*$ .

□

## Appendix D

# Supplementary Methods and Results for Chapter 5

### D.1 Supplementary Methods

#### D.1.1 Creating Synthetic Diploid Simulations

To create artificial haplotypes, we used variant positions from the set of common human variants from dbSNP build 150. We restrict to the set of single-nucleotide variants. dbSNP reports the population allele frequencies at each position,  $[p_A, p_C, p_G, p_T]$ . We generate two haplotypes  $H, \bar{H}$  for each individual such that  $H = [h_1, h_2, \dots, h_n]$ ,  $\bar{H} = [\bar{h}_1, \bar{h}_2, \dots, \bar{h}_n]$ , where  $h_i \in [A, C, G, T]$ , and  $\Pr(h_i = x) = p_x$ . Note that this approach does not maintain linkage disequilibrium between adjacent SNPs, but the algorithms tested on this data set do not rely on population level information.

Single cells were downloaded from SRA accession numbers SRP041470 and SRP061939. For the

simulation we used 16 cells from patient UMB1465. All cells were downsampled to a target coverage of 3x using samtools. Cells were separated in to pairs  $c, \bar{c}$  based on the total number of positions in chromosome X that were covered by at least one read, where cell  $c$  corresponds to haplotype  $H$  and cell  $\bar{c}$  corresponds to haplotype  $\bar{H}$ . Two cells were excluded due to irregular drop-out rates. This yields 7 pairs of cells. From a cell  $c$ , we extract the set of reads that cover at least one position  $i$  in the generated haplotypes  $H, \bar{H}$ . For each read  $\mathbf{r}$  covering position  $i$ , we alter the nucleotide  $r_i$  mapped to position  $i$  as

$$r_i = \begin{cases} h_i & \text{with probability } p_{r,i} \\ \{x \in A, C, G, T | x \neq h_i\} & \text{with probability } (1 - p_{r,i}/3) \end{cases}$$

with  $p_{r,i}$  is the probability of an incorrect base call implied by the reported Phred score  $q_{r,i} = -10\log_{10}p_{r,i}$ . These cells  $c, \bar{c}$  were merged into a single BAM file.

## D.2 Supplementary Results

### D.2.1 Comparison to Reference-Based Phasing

An alternate approach to determining the haplotype phasing of an individuals genome is statistical or reference-based phasing with a reference panel [16, 30, 87, 143]. These methods take advantage of linkage disequilibrium, the correlation between the occurrence of certain alleles at nearby variants across a population. They take as input genotype data from of one or more individuals as well as a panel of phased genomes from a larger population. Here we compare against the performance of EAGLE2 [87], the current state-of-the-art reference-based phasing approach on the single-cell breast cancer data from Wang et al. [153].

We used the Sanger Imputation Server which integrates the Eagle2 software, to perform phasing, using the 1000 Genomes Phase 3 reference panel. As the service was unable to run on the full set



of SNPs from chromosome 17, we created an input file which included all heterozygous variants identified in the sample, in addition to all homozygous variants in which the called allele had a population frequency of less than 0.95. This yielded a set of 119,990 SNPs.

On this dataset, we found that EAGLE2 had a total switch error of 17.3% over the set of heterozygous variants. As we do not obtain a phasing over the entire chromosome, we measured We evaluate the switch error of EAGLE2 on the same set of haplotype blocks we recover using amplification fragments for a range of phasing scores (see Figure 3 in the main text). Figure SD.1 shows the resulting switch errors. Note that the EAGLE2 switch error decreases with the cutoff used for the phasing score. While EAGLE2 does not have access to read information used to calculate the phasing score, we expect both the phasing score and the performance of EAGLE (based on linkage disequilibrium) to correlate with the distance between SNPs in the block. We see that for scores  $\geq 2$ , using amplification fragments with HAPCUT achieves lower switch-error rates than EAGLE2. These results suggest the potential utility of a hybrid approach, combining the accuracy of phasing from amplification fragments at short distances, with the ability of reference-based phasing to span pairs of SNPs beyond amplicon length.

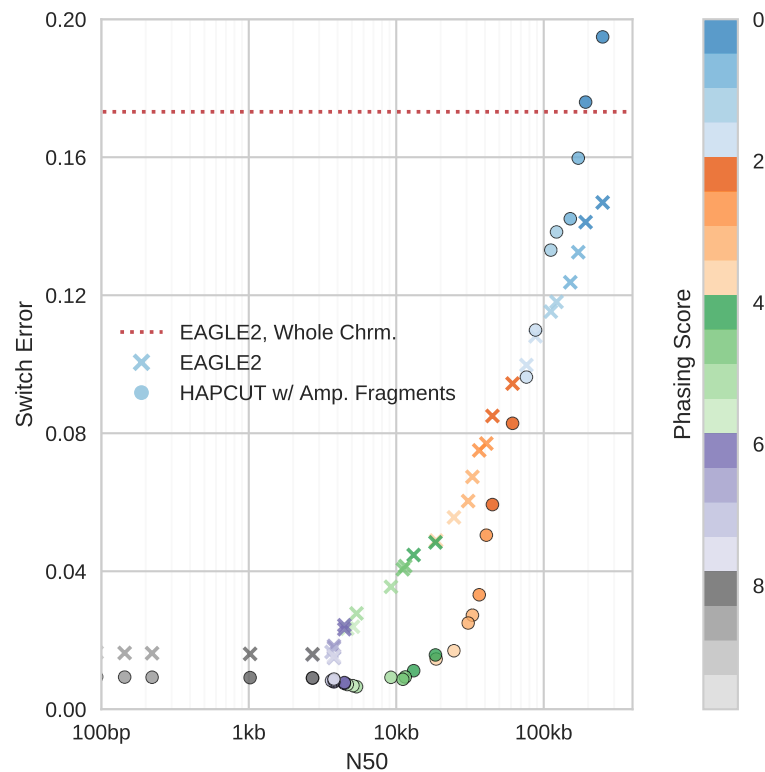


Figure D.1: **Comparison to EAGLE2 on whole-exome data.** We evaluate the switch error obtained by EAGLE2 over the set of haplotype blocks recovered using amplification fragments, on the single-cell breast cancer data from Wang et al. [153]. The total switch error rate across all the entire chromosome is shown as the red line.