Modeling Tumor Clonal Evolution through Longitudinal ctDNA Profiling

Siddharth Sabata, Russell Schwartz

Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University

Computational Biology Department

Abstract

Circulating tumor DNA (ctDNA) profiling has revolutionized the study of tumor evolution by enabling non-invasive, real-time tracking of clonal dynamics [3]. This work focuses on modeling cancer clonal evolution using longitudinal ctDNA samples from liquid biopsies. We developed an automated data pipeline for optimal mutation marker selection, integrating patient datasets from the Allegheny Health Network (AHN) and TRACERx consortia. Using PhyloWGS, we infer clonal phylogenies from variant allele frequencies (VAFs), generating time-series data on clonal abundances. This approach enables scalable and robust phylogenetic tracking of tumor evolution, with the long-term goal of predicting patient outcomes.

Introduction

Tumor evolution is a dynamic process driven by genetic heterogeneity, which shapes disease progression and treatment response. Liquid biopsy-based ctDNA profiling offers a real-time, minimally invasive window into tumor clonal dynamics [3]. However, integrating longitudinal ctDNA data into phylogenetic models remains a significant computational challenge, requiring new methods to reliably reconstruct tumor clone phylogenies over time [1]. We address this challenge by developing an automated pipeline for longitudinal phylogenetic tracking, leveraging high-performance computing (HPC) resources for data integration and analysis. In particular, this work builds on the Mase-phi framework, which introduced strategies to optimally select ctDNA mutation markers for refining tumor phylogeny models and tracking subclone frequencies across samples [1]. Our extended pipeline applies this methodology at scale to multiple cohorts, aiming to improve the prediction of tumor metastasis and relapse.

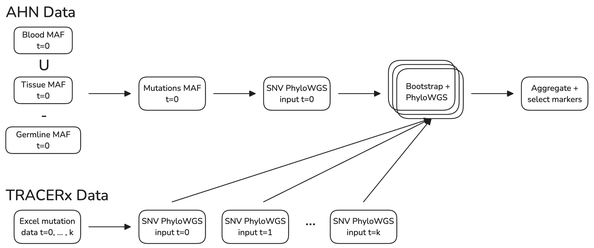

Figure 3: AHN and TRACERx data pipelines.

Methodology

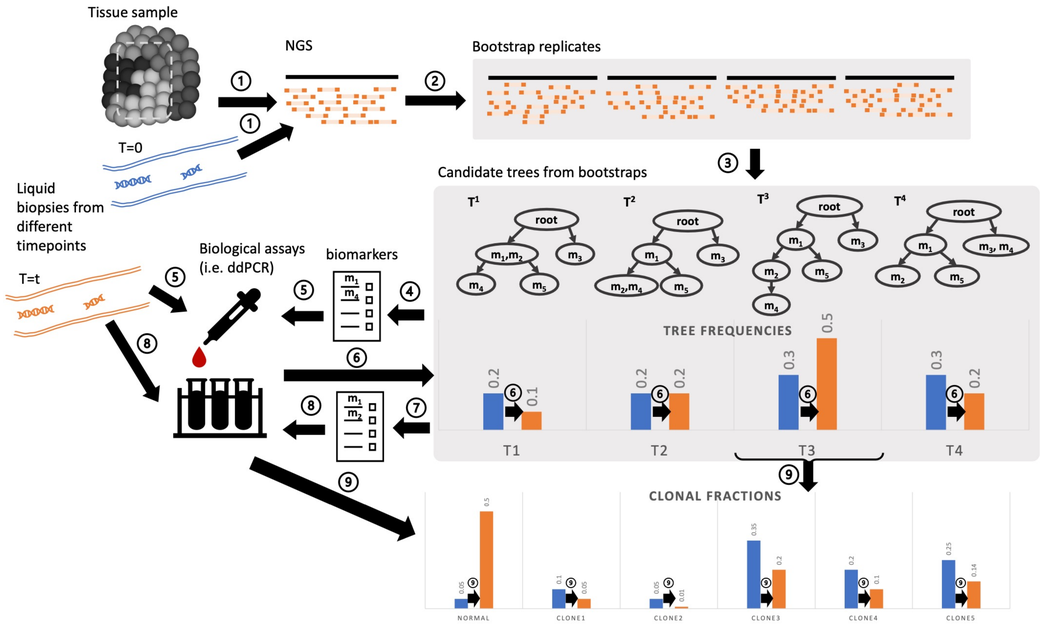

Figure 1: The overall inference pipeline. (1). We assume we have first sequenced tissue and liquid biopsy sample(s), obtaining reference germline and ctDNA. (2) We create bootstrapped samples over reads for each sequence set. (3) We infer a set of possible trees from the bootstrapped samples, serving as an estimated empirical tree distribution. (4) We then seek a set of optimal biomarkers of mutations to best reduce the tree uncertainty and (5) apply these in biological assays (e.g. ddPCR). (6) We then use the results of these assays to update the empirical tree distributions. (7) We further seek a set of optimal biomarkers to track subclone frequencies efficiently and (8) assay these biomarkers. (9) Finally, we then use the results of the assays to estimate clonal fractions at each sampled timepoint. [1]

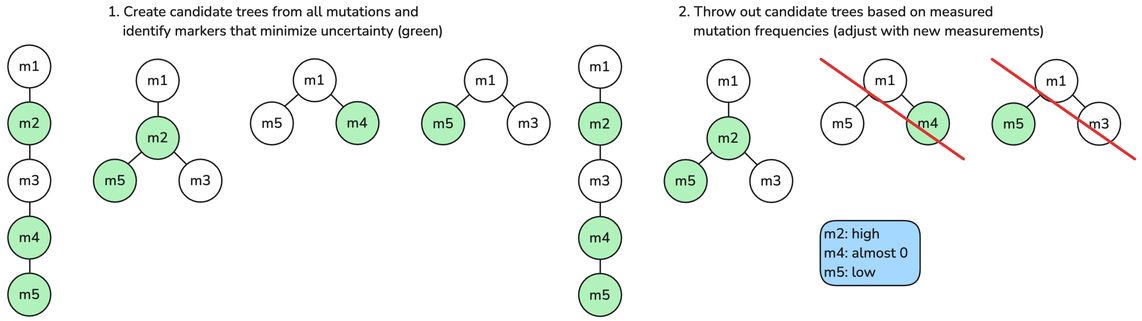

Figure 2: Phylogenetic Inference and Marker Selection

1. Apply PhyloWGS with bootstrapping to produce an ensemble of plausible phylogenies, capturing uncertainties in variant frequencies [2].
2. Examine the resulting ensemble to identify the major candidate trees and key differences among them.
3. For each mutation, measure how its placement or frequency varies across the candidate trees; a mutation that differs markedly between trees is highly informative.
4. Rank mutations by their expected reduction in uncertainty and select a top subset for targeted ctDNA assays.
5. Use targeted sequencing (e.g., ddPCR) to measure these markers in follow-up ctDNA samples and update the phylogeny as more ctDNA time points are collected.

Future Directions

We will use NLP to interpret each clone's functional significance by automatically summarizing gene functions, pathways, and clinical implications. This AI-assisted approach helps prioritize subclones most likely to drive disease—such as those enriched in proliferation or metastasis genes—and bridges the gap between raw genomic data and actionable insights, thus improving real-time decision-making in oncology.

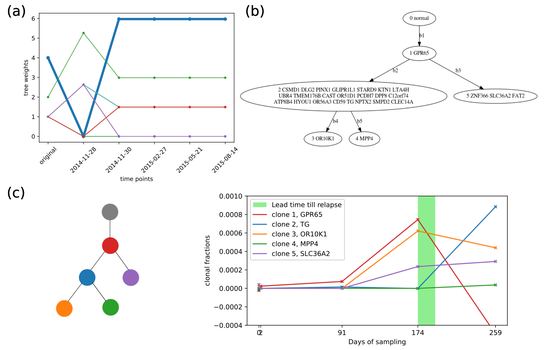

Figure 4: Results of applying our tree refinement and clonal tracking methods on the CRUK0044 sample from the TRACERx data. (a) Changes in tree weights for each topology identified in bootstrap sampling, after adjusting the tree distribution using the selected markers at each time point. (b) The inferred most likely tree after all serial samples, corresponding to the blue line in (a). (c) Inferred clonal frequencies as of each longitudinal sample derived from the selected marker set as of each day of sampling, with lines representing the clones color-coded as in the tree at left. [1]

Figure 5: Theoretical language model pipeline for clones described in Figure 4

Acknowledgements

We thank the Schwartz Lab, Thomas Rachman, Xuecong Fu, and Vivek Chelur

References

1. Fu, X. et al. Marker selection strategies for circulating tumor DNA guided by phylogenetic inference. Preprint at https://doi.org/10.1101/2024.03.21.585352 (2024).
2. Deshwar, A. G. et al. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol 16, 35 (2015).
3. Abbosh, C. et al. Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. Nature 616, 553–562 (2023).