Investigating longitudinal evolution of liquid cancers using computational and mathematical models

Nathan D. Lee

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Ivana Bozic, Chair

Hong Qian

Georg Luebeck

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Investigating longitudinal evolution of liquid cancers using computational and mathematical models

Nathan D. Lee

Chair of the Committee:
Associate Professor Ivana Bozic
Applied Mathematics

Cancer can result from a series of driver mutations, alterations in the genome sequence that confer a fitness advantage to cells containing them, resulting in their net proliferation. Mutations that have a neutral fitness effect ("passenger mutations") also accumulate in cancer cells, contributing to the significant heterogeneity observed in tumors. The genetically distinct subpopulations of cancer cells—or subclones—can have different levels of fitness and treatment sensitivity. Study of the individual subclones facilitates understanding the whole tumor's dynamics, treatment resistance, and potential for relapse. In this thesis I discuss several projects that employ mathematical and computational models to investigate subclonal evolution in longitudinal studies of leukemia. First, I discuss work that employs branching processes to reconstruct the evolutionary history of cancers, including when cancer was initiated and when subsequent driver mutations occurred. Next, I present a pipeline I developed in collaboration with a team of physician-scientists and clinical researchers to enable monitoring and interactive visualization of clonal evolution and cancer relapse in the clinic, by clustering mutations and inferring the inter-clonal architecture and relationships. I show several examples of this pipeline applied to data from a clinical trial of hematopoietic cell transplantation to treat acute myeloid leukemia patients. Last, I analyze the evolution of resistance in response to a new targeted therapy for chronic lymphocytic leukemia. Using the

high resolution afforded by ultra-deep sequencing data, we show that resistance mutations are present at very low frequencies pre-treatment and expand upon initiation of treatment, with key implications for cancer monitoring and resistance.

TABLE OF CONTENTS

Page

List of Figures iv

List of Tables xi

Chapter 1: Introduction 1

Chapter 2: Inferring parameters of cancer evolution in chronic lymphocytic leukemia 4
2.1 Abstract 4
2.2 Introduction 4
2.3 Results 6
2.4 Discussion 24
2.5 Materials and methods 27
2.6 Supplementary Materials 47

Chapter 3: Ultra-deep mutational landscape in chronic lymphocytic leukemia patients uncovers clonal dynamics of resistance to targeted therapies. . . 49
3.1 Introduction 49
3.2 Methods 51
3.3 Results 56
3.4 Discussion 63
3.5 Supplemental Materials 66

Chapter 4: Tracking clonal evolution in a cohort of post-transplant acute myeloid leukemia patients 69
4.1 Introduction 69
4.2 Methods 73
4.3 Results 78
4.4 Discussion and Future Work 86

Chapter 5: Using population information to predict chronic lymphocytic leukemia progression 88
5.1 Introduction 88
5.2 Results 90
5.3 Discussion 93

Bibliography 100

CITATIONS TO PREVIOUSLY PUBLISHED WORK

Chapter 2 has been published as [1]. Chapter 3 is in preparation to be submitted as "Ultra-deep mutational landscape in Chronic Lymphocytic Leukemia patients uncovers clonal dynamics of resistance to targeted therapies" by *David Woolston, *Elena Latorre-Esteves, *Nathan Lee, Mazyar Shadman, Xin Ray Tee, Jeanne Fredrickson, Brendan F. Kohrn, Olga Sala-Torra, Chaitra Ujjani, Ashley Eckel, Brian Till, Min Fang, Jerald Radich, **Ivana Bozic, **Rosa Ana Risques,**Cecilia CS Yeung (*co-first authors, **co-last authors).

LIST OF FIGURES

Figure Number Page

2.1 **Stochastic branching process model of tumor evolution.** (a) Stochastic branching process model for tumor expansion. Initiated tumor cells (blue) divide with birth rate b, die with death rate d, and accrue passenger mutations with mutation rate u. Type-1 cells, which carry the driver mutation, divide with birth rate b_1, die with death rate d_1, and accrue passenger mutations with mutation rate u. (b) The initiated tumor, or type-0, (blue) population growth is initiated from a single cell. A driver mutation occurs in a single type-0 cell at time t_1, starting the type-1 population (red). The tumor sample is collected and bulk sequenced at times $t_1 + t$ and $t_1 + t + \Delta$, where the driver fraction is α_1 and α_2, respectively. Tumor size (in number of cells) is M_1 and M_2 at first and second sample collection dates. (c) By the time the tumor is observed, it has a high level of genetic heterogeneity due to the mutations that have accrued in both type-0 (blue) and type-1 populations (red). Each yellow star represents a different passenger mutation. 8

2.2 **Accuracy of parameter inferences from simulated data**. We simulated tumor growth by performing a Monte Carlo simulation, which simulates the birth, death, and accumulation of mutations in the individual cells that make up a tumor, and generates the mutation frequency and tumor size data used by the estimates. Simulations are of fast-growing tumors with (a) single driver subclone and mutation rate $u = 1$, (b) single driver subclone and $u = 3$, (c) two nested driver subclones with $u = 1$, and (d) two sibling driver subclones with $u = 1$. Mean percent errors (MPEs) of estimates are shown in black above the plots, and mean absolute percent errors (MAPEs) are shown in gray. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error data points. Complete parameter values and number of runs are included in Table 2.3. 13

2.3 **Percent errors (PEs) for case with no death**. Accuracy of parameter inferences for Monte Carlo simulation of tumor with no cell death for (a) single driver subclone with mutation rate $u = 1$, (b) single driver subclone with $u = 10$, (c) two nested subclones with $u = 1$, and (d) two sibling subclones with $u = 1$. Mean percent error (MPEs) are the black numbers above the plots, and mean absolute percent errors (MAPEs) are the grey numbers below the MPEs. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error datapoints. Complete parameter values and number of runs are included in Table 2.3. 14

2.4 **Percent errors (PEs) for slow-growing tumor**. Accuracy of parameter inferences for surviving Monte Carlo simulation runs of slow-growing tumor for (a) single subclone with mutation rate $u = 1$, (b) single subclone with $u = 5$, (c) two nested subclones with $u = 1$, and (d) two sibling subclones with $u = 1$. Mean percent error (MPEs) are the black numbers above the plots, and mean absolute percent errors (MAPEs) are the grey numbers below the MPEs. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error data points. Complete parameter values and number of runs are included in Table 2.3. 15

2.5 **Accuracy for t estimate increases with tumor size.** A Monte Carlo simulation of a birth-death process was performed for (a) fast-growing, (b) slow-growing, and (c) no cell death parameter regimes. For each of the 100 surviving simulated tumors, the percent error of the t estimate (Eq. (2.3)) was calculated when the tumor first reached the specified tumor sizes. Means are indicated by red points and lines, $\pm$ one standard deviation is shown by the red region, and individual data points for each simulation run are shown as the grey points (with horizontal jitter for visibility). 16

2.6 **Model for tumor expansion with two driver mutations.** (a) Two nested driver subclones. Initiated tumor (type-0) cells in blue, cells with driver 1 (type-1) in red, and cells with both drivers (type-2) in orange. A driver mutation occurs in a type-0 cell at t_1. A second driver mutation occurs in a type-1 cell at $t_1 + t_2'$. Tumor is bulk sequenced at $t_1 + t_2' + t$ and $t_1 + t_2' + t + \Delta$. (b) Two sibling driver subclones. Type-0 cells (in blue). A driver mutation occurs in a type-0 cell at t_1. A second driver mutation occurs in a different type-0 cell at t_2. Tumor is bulk sequenced at $t_1 + \tau_1$ (or, equivalently $t_2 + \tau_2$) and $t_1 + \tau_1 + \Delta$ (equivalently $t_2 + \tau_2 + \Delta$). 17

2.7 **Corrections for observed mutation counts.** (a) If passenger mutations (circles with stars) that occur after the driver reach fixation in the driver population (red), then they are indistinguishable from the passengers that were present in the first cell with the driver, which accrued in the type-0 population (blue). The estimate of when the driver occurred needs to account for these mutations (circled). In (b), we compare percent errors of parameter estimates for time from tumor initiation until appearance of a driver subclone, t_1, with and without this correction (Eq. (2.6)). Errors for estimate with correction are shown in blue, and for estimate without correction (Eq. (2.5)) in orange. Errors are plotted as a kernel density estimate for Monte Carlo simulations of slow-growing tumor with mutation rate $u = 5$. Mean percent errors (MPEs) and mean absolute percent errors (MAPEs) are listed. (c) Mutations present on two or fewer variant reads (red) are filtered out in post-processing. Mutations with more than two variant reads (black) are included. The number of subclonal mutations between frequencies f_1 and f_2, γ, which is used in the mutation rate estimate, must be corrected for mutations that are filtered out. In (d), the percent errors for the observed (orange) and corrected (blue) γ (Eq. (2.7)) are plotted as kernel density estimates. Observed mutations are those that passed post-processing, i.e. those that have more than $L = 2$ mutant reads. True mutation frequencies were generated from 135 surviving runs of a Monte Carlo simulation of a fast-growing tumor with mutation rate $u = 1$, from which sequencing reads were simulated with 200x average coverage (see Materials and Methods). Percent errors are calculated relative to the true γ measured from the true mutation frequencies. 18

2.8 **Corrections for observed mutation counts.** (a) We compare percent errors of parameter estimates for time from tumor initiating until appearance of a driver subclone, t_1, with and without the correction for passengers that occur after the driver and reach fixation in the driver population (Eq. (2.6)). Errors for estimate with correction are shown in blue, and for estimate without correction ((2.5)) in orange. Errors are plotted as a kernel density estimate for Monte Carlo simulations of fast-growing tumor with mutation rate $u = 1$. Mean percent errors (MPEs) and mean absolute percent errors (MAPEs) are listed. (b) The percent errors for the observed (orange) and corrected (blue) number of subclonal mutations between frequencies f_1 and f_2, γ, (Eq. (2.7)) are plotted as kernel density estimates. Observed mutations are those that passed post-processing, i.e. those that have more than $L = 2$ mutant reads. True mutation frequencies were generated from 135 surviving runs of a Monte Carlo simulation of a fast-growing tumor with mutation rate $u = 1$, from which sequencing reads were simulated with 100x average coverage (see Materials and Methods). Percent errors are calculated relative to the true γ measured from the true mutation frequencies. 19

2.9 **Reconstructing the timeline of CLL evolution in patients.** We applied our methodology to estimate subclonal growth rates, mutation rates and evolutionary timelines in CLL tumors from Ref. [2]. Vertical height of a clone represents its $\log_{10}$-scaled size. Mutations were clustered into clones and phylogenetic trees were inferred using PhylogicNDT [3]. Tree edges are colored by clone number and are labeled with driver mutations, if any. For each patient, we show estimates for patient age at CLL initiation and times of appearance of CLL subclones. Dashed white line indicates when the patient was diagnosed. Solid black arrows indicate times of bulk sequencing measurements. 25

3.1 **Patient R001 CLL mutation evolution and treatment history.** Colored regions correspond to different treatment regimens. The WBC count (blue dotted line) can be used as an indicator of tumor burden and progression. Key resistance mutations are in bold in the legend. 57

3.2 **Patient R002 CLL mutation evolution and treatment history.** Colored regions correspond to different treatment regimens. The WBC count (blue dotted line) can be used as an indicator of tumor burden and progression. Key resistance mutations are in bold in the legend. 59

3.3 **Comparison of Duplex and NGS VAFs.** Variants detected with standard NGS are detected at similar frequencies with Duplex sequencing. 63

3.4 **Characterization of mutations identified pre- and post-pirtobrutinib treatment.** (A) Description of samples collected over time for both patients R001 (A,B,C,D,E,F) and R002 (A,B,C,D,E). Sample types include bone marrow and peripheral blood. Values for percent disease, days pre/post pirtobrutinib treatment, and mean coding depth for duplex sequencing are shown for each sample. (B) Mutation Frequency for coding and non-coding mutations for each sample collected over time. (C) Number of coding mutations found in genes associated with drug resistance in CLL for each sample. Genes associated with CLL resistance and covered by duplex sequencing probes include BAX, BCL2, BTK, PLCG2, and TP53. Each sample is represented by a single column; mutated genes are color-coded and represented as a fraction within the column. (D) Number of coding mutations found in genes associated with drug resistance in CLL for each sample collected over time. 67

3.5 **Reproducibility of duplex sequencing data.** Two bone marrow samples collected the same day from patient R001 were independently processed for duplex sequencing. Variant allele frequencies (VAF) were calculated by dividing the number of mutant duplex reads (alternative counts) by the duplex depth at the mutated position. Black dots represent the mutations found in 2 separate samples collected the same day for the same patient (R001). Spearman's rank correlation coefficient is used to measure the degree of association between two variables ($\rho = 0.89$), showing high reproducibility of measurements. 68

4.1 **Software pipeline for interactive visualization of clonal evolution in the clinic**. Variants called from Archer analysis are compiled as a variant call format (VCF) file. Variants are filtered based on clinical relevance and sequencing quality. PyClone-VI is used to cluster the filtered variants into clones. Pairtree is used to infer the phylogenetic trees describing the relationships between clones. The inferred subclonal frequencies are used to build fishplots visualizing clonal evolution. Figure created in part with `BioRender.com`. . . 80

4.2 **Patient 1 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. The right side of the clinical data provides a general category, and the left label provides the specific category. The top level shows HCT date, second level shows when the patient had grade 2 acute graft-vs-host disease, level 3 shows detection of relapse, level 4 shows treatment received, and the bottom level shows when the patient was deceased. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel. 81

4.3 **Patient 9 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. The top level of the clinical data show HCT date, second level shows the date of complete remission (CR) and measurable residual disease (MRD), the third level show treatment given, and the bottom level shows when the patient had moderate chronic GvHD. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel. 82

4.4 **Patient 17 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. In top to bottom order, the clinical panel shows transplant date, GvHD, staging results (CR = complete remission), G-CLAM treatment, and when the patient was deceased. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel. 84

4.5 **Patient 22 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. In top to bottom order, the clinical panel shows HCT date, GvHD and its grade, and date of morphologic relapse. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel. 85

5.1 **All patients fit to a logistic growth curve.** As observed in ref [2], the patients fall into two classes: finite carrying capacity and exponential growth. (A-C) Log-log plots of the fitted logistic parameter values. (D-F) Distributions of the parameters. K is bimodal, r is unimodal with some long tails, and y0 is more concentrated. 95

5.2 **Comparison of growth curve functions.** (A) Comparison of exponential, logistic, and Gompertz growth curves. (B) Correlation and linear regression of the α and β parameters in the Gompertz fits (C) Correlation and linear regression of the α and β parameters in the Gompertz model with a fitted intercept parameter. (D-F) For 3 patients, comparison of the Gompertz, logistic, and reduced Gompertz models. Dashed and solid lines indicate a model with and without an additional fitted parameter for the intercept, respectively. 96

5.3 **Metrics for goodness of fit of different models for the growth curve.** Mean absolute percentage error (A), mean percentage error (B), and Akaike information criterion (C) of the fit for each patients are evaluated, where the fit for each patient is represented by a single point. The following growth curves functions were fit by nonlinear least squares, with parameters fit in parentheses: Gompertz (α, β), Gompertz with a y-intercept parameter (α, β, y_0), logistic (r, K), logistic with an intercept parameter (r, K, y_0), reduced Gompertz model (β), and reduced Gompertz model with intercept parameter (β, y_0). The mean is displayed at the top and indicated by the red line. . . . 97

5.4 **Benefits of using population data for prediction with few measurements in simple simulated example.** Exponential phase of WBC growth curve was simulated for 20 patients. All panels shows the fitted curves for a single patient. Left to right, top to bottom, more points are added to the training set used to fit the model (black points). Blue points are not used to train the model. Training set is black points for patient under consideration, and the rest of the populations' time series. A curve fit is performed using nonlinear mixed effects (solid green line) and nonlinear least squares (dashed red line). 98

5.5 **Nonlinear mixed-effects model of Gompertz growth curve.** (A) Fixed effects curve (black solid line) for all the patients fitted with a 3 parameter Gompertz model (α, β, and intercept). All other lines are the growth curves of all the patient WBC counts. (B-D) Random effects for all the patients fitted with a 3 parameter Gompertz model. (E-H) Examples from 4 patients where the first 12 points of WBC count time series (black points) for patient under consideration are fit to 3 parameter Gompertz model using nonlinear least-squares (blue line) and nonlinear mixed effects (gray line). The mixed effects model makes use of WBC time series from the other patients. Red triangular point indicates first measurement > 2 years after last measurement used to fit the model. 99

LIST OF TABLES

Table Number Page

2.1 Inferred parameters for CLL patients with exponential growth patterns, for which there are at least two longitudinal bulk sequencing measurements before treatment. Estimates are computed from tumor size measurements and mutation frequencies from whole exome sequencing. Mutation rates are for the exome only. The time estimates are in terms of the patient's age in years. . . 21

2.2 Confidence intervals for inferred parameters for CLL patients with exponential growth patterns, for which there are at least two longitudinal bulk sequencing measurements before treatment. Estimates are computed from tumor size measurements and mutation frequencies from whole exome sequencing. Mutation rates are for the exome only. The time estimates are in terms of the patient's age in years. 22

2.3 **Parameter values.** Parameter values and number of surviving runs for Monte Carlo simulations. For all simulations $f_1 = 0.01$, $f_2 = 0.20$, $L = 2$. Table (`.xlsx` file) can be downloaded from: `https://doi.org/10.1371/journal.pcbi.1010677.s006` 47

ACKNOWLEDGMENTS

I'd like to thank Ivana for her guidance and mentorship over the course of my PhD. We both came to UW at similar times to start new phases in our academic careers—you as a new faculty hire and myself as an incoming graduate student—and it has been rewarding to see us both grow into these roles as we worked together over the past several years. I learned a lot from you during this time that contributed to my growth as a researcher. Your dedication to responsiveness and good communication gave me many opportunities to ask questions and test out ideas. I feel good about how things have come together in the past year, making a good conclusion to my time at UW.

I'm also grateful for the privilege of collaborating with a variety of highly knowledgeable physicians and scientists, including Elizabeth, Jerry, Olga, Cecilia, Rosana, David, Isaac, and Lan. I appreciated the encouragement, positivity, and variety of expertise that I could reliably count on from all of you. Our work together on some very interesting projects significantly enriched my PhD experience. I am also inspired by Elizabeth, Jerry, and Cecilia's seemingly endless energy and motivation to make a difference in their patients' lives.

I'd like to thank my committee members for the time, effort, and stimulating questions they contributed. Hong, I am thoroughly tested and engaged by your unique perspective on mathematical biology and evolutionary processes, which has an impressive capacity to produce questions ranging from the historical to philosophical in scale. Sasha, I appreciate your enthusiasm and the enjoyable time early on in grad school working on optimization problems. Georg, your keen understanding of my work led to many productive questions, delivered with your characteristic friendliness. Ben, I am glad I chose you as my GSR, as you went above and beyond what was required.

I'm grateful for the friends I've made in my time at UW, especially Sheridan, Diya, Mo, Adam, and Jeremy, as well as the group of friends from undergrad. A particularly meaningful part of my time in Washington was the memories I made in the Cascade and Olympic mountains with many of these people. David, thanks for introducing me to the world of backpacking here, and somehow that disastrous first Enchanted Valley trip didn't turn me off of backpacking for life. Thanks, Jeremy, for so many special excursions into Washington's mountains for backpacking and skiing. Those trips kept me energized throughout the ups and downs of the PhD.

Thanks to my parents for their support and encouragement. They never pressured me to follow a certain career path or life trajectory, and instead always encouraged my curiosity and personal passions, whether music, reading, or science. I think this is one of the greatest things that parents can pass on to their children. I'm also grateful for the rest of my family and the sense of community and support structure they provide.

Finally, I'm thankful for Bridget's love and support over the past 9 years. Six years is a long time to have to spend apart on opposite sides of the country while each enduring the challenges of a PhD, but you were always there for me. It's been an adventure together with our visits back and forth from Seattle to New York, but it will be nice to be together in the same city again.

Chapter 1

INTRODUCTION

Cancer is the end result of a complex multi-stage process driven by a sequence of genetic alterrations [4, 5], as well as a variety of microenvironmental, epigenetic, and transcriptional factors [6]. The rise of genome sequencing technologies in the past several decades has revealed the complex genomic landscape shaped by carcinogenesis [7–9]. In particular, it has identified key genetic events occurring over the course of the cancer's lifetime, from a normal tissue to a premalignant state to a malignant tumor [10–12]. In this work we focus on modeling the different genetically distinct cancer cell subpopulations—or subclones—which are determined by the set of mutations—or alterations in the genome sequence—they possess.

Of the mutations detected in a given cancer, some will have a causative role in carcinogenesis, while others will have a neutral effect on fitness or cancer progression [13]. "Driver" mutations confer a fitness advantage relative to neighboring cells, which can result in the increased proliferation of cells containing the driver. Most tumors contain multiple driver mutations, ranging from a few to dozens, depending on the cancer type [11]. Neutral mutations, which don't have an effect on fitness, are often referred to as "passenger" mutations, due to their ability to reach a large enough frequency to be detected by sequencing in a tumor by "hitchhiking" along with a driver mutation in an expanding subclone [14, 15]. Alternatively, neutral mutations can also reach observable frequency due to random chance [16–18]. When individual cancers are sequenced, it reveals the presence of many passenger mutations, with anywhere from tens to thousands of passengers per tumor [9, 19].

The specific structure of the different populations, or clones, comprising a cancer has an effect on the individual clones, as well as the overall properties of the tumor [16]. The

specific genotype (the genetic composition) of a clone influences its phenotypic properties (the higher level observable features), such as an increased growth capacity, metastatic potential, or drug resistance [16]. In this thesis, I discuss several projects that use mathematical and computational models of clonal evolution to investigate different cancer behaviors. The work presented here makes use of longitudinal cancer studies of liquid cancers, where multiple samples are collected from the same patient over time. Longitudinal studies allow examination of the temporal dynamics of cancer, and how they change in response to anti-cancer therapies. Additionally, a necessary first step of any quantitative analysis of clonal evolution is inferring the clonal structure, often by clustering of mutations into clones and identifying the ancestral relationships between them [20–23]. These relationships are often visualized as a phylogenetic tree, where each clone is represented by a node, and edges connect each clone to its parental clone. The existing clustering and tree inference methods for cancer sequencing data struggle to build a consensus tree for a single time point, but can take advantage of serial samples to better determine the structure [23]. Studies of liquid cancers like leukemias often provide longitudinal samples, due to the relative ease of collecting peripheral blood samples, compared to more invasive biopsies of solid cancers.

In Chapter 2 I discuss work with Ivana Bozic from our published paper "Inferring parameters of cancer evolution in chronic lymphocytic leukemia" [1], where we use stochastic process models of carcinogenesis to reconstruct the evolutionary history of individual liquid cancers. Next, in Chapter 3 I discuss soon-to-be-submitted work from a collaboration with Ivana Bozic, the Risques Lab at UW Medicine, and several groups at Fred Hutchinson Cancer Center, where we use ultra-deep sequencing data to investigate the evolution of resistance in response to targeted leukemia therapies. Chapter 4 includes work from a collaboration with Ivana Bozic and physician-scientists and clinical researchers at Fred Hutchinson Cancer Center, where we develop a pipeline to build an interactive data visualization platform for the oncologists to use to explore impact that different treatments have on a cancer's clonal evolution. We apply this pipeline to study data from a clinical trial of stem cell transplantation to treat an aggressive form of leukemia. Lastly, I present work with Sasha Aravkin and Ivana

Bozic on making use of population-level information to predict cancer progression.

All these projects share the theme of using mathematical, statistical, and computational models to quantify the evolutionary and temporal dynamics of tumors. We make use of simulations of cancer growth and sequencing data to better understand the complexities of carcinogenesis. However, simulations are ultimately idealizations of this messy process. Each chapter focuses on applying our methods to clinical data and addressing challenges encountered when using this often noisy data.

Additionally, each of these projects allows one to examine the characteristics and driving forces behind an individual patient's cancer evolution. For example, in Chapter 2 we infer the age when a patient's cancer was initiated, in Chapter 3 we study the genetic mechanisms of resistance to chemotherapy, in Chapter 4 we investigate the evolutionary patterns of cancer recurrence, and in Chapter 5 we model the growth dynamics of cancer progression pre-treatment. Such insights are key for developing treatment personalized to the unique features of each patient's cancer.

Chapter 2

INFERRING PARAMETERS OF CANCER EVOLUTION IN CHRONIC LYMPHOCYTIC LEUKEMIA

2.1 Abstract

As a cancer develops, its cells accrue new mutations, resulting in a heterogeneous, complex genomic profile. We make use of this heterogeneity to derive simple, analytic estimates of parameters driving carcinogenesis and reconstruct the timeline of selective events following initiation of an individual cancer, where two longitudinal samples are available for sequencing. Using stochastic computer simulations of cancer growth, we show that we can accurately estimate mutation rate, time before and after a driver event occurred, and growth rates of both initiated cancer cells and subsequently appearing subclones. We demonstrate that in order to obtain accurate estimates of mutation rate and timing of events, observed mutation counts should be corrected to account for clonal mutations that occurred after the founding of the tumor, as well as sequencing coverage. Chronic lymphocytic leukemia (CLL), which often does not require treatment for years after diagnosis, presents an optimal system to study the untreated, natural evolution of cancer cell populations. When we apply our methodology to reconstruct the individual evolutionary histories of CLL patients, we find that the parental leukemic clone typically appears within the first fifteen years of life.

2.2 Introduction

When a cell accrues a sequence of driver mutations—genetic alterations that provide a proliferative advantage relative to surrounding cells—it can begin to divide uncontrollably and eventually develop the complex features of a cancer [5, 6, 24]. Thousands of specific driver mutations have been implicated in carcinogenesis, with individual tumors harboring from

few to dozens of drivers, depending on the cancer type [11]. Mutations that don't have a significant effect on cellular fitness also arise, both before and after tumor initiation [13]. These neutral mutations, or "passengers", can reach detectable frequencies by random genetic drift or the positive selection of a driver mutation in the same cell [14, 15, 17, 18]. Mutational burden detectable by bulk sequencing reveals tens to thousands of passengers per tumor [9, 19].

Genome sequencing technologies have revealed the heterogeneous, informative genetic profiles produced by the evolutionary process driving carcinogenesis [25, 26]. These genetic profiles have been used to obtain insight into specific features of the carcinogenic process operating in individual patients. For example, the molecular clock feature of passenger mutations has been employed to measure timing of early events in tumor formation, as well as identify stages of tumorigenesis and metastasis [27–35]. Other studies have estimated mutation rates [13, 36, 37], selective growth advantages of cancer subclones [2, 38–40], and the effect of spatial structure on cancer evolution [41–43]. We note that previous approaches typically only estimate one or a few parameters of cancer evolution. In addition, many state-of-the-art methods make use of computationally expensive approaches [37, 42, 44] or simplifying assumptions, such as approximating tumor expansion as deterministic or ignoring cell death [2, 44]. Our approach relies on analytic formulas and sampling, which for realistic numbers of subclones and time points is efficient, and does not require simulation of tumor growth or computationally expensive model fitting.

Mathematical models of cancer progression, especially when used in conjunction with experimental and clinical data, can provide important insights into the evolutionary history of cancer [15, 32, 45–49]. Branching processes—a type of a stochastic process—can be used to model how different populations of dividing, dying, and mutating cells in a tumor evolve over time [50]. Their theory and applications have been well developed to model the multistage nature of cancer development [38, 41, 47, 50–52]. Here we use a branching process model of carcinogenesis to derive a comprehensive reconstruction of an individual tumor's evolution.

Tumors can grow for many years, even decades, before they reach detectable size [29]. Typically, tumor samples used for sequencing would be obtained at the end of the tumor's

natural, untreated progression. More recently, longitudinal sequencing, where a tumor is sequenced at multiple times during its development, has provided better resolution of tumor growth dynamics and evolution in various cancer types [2, 3, 53–55]. Chronic lymphocytic leukemia (CLL) is an ideal system for studying cancer evolution because it can be monitored, via peripheral blood samples, without treatment until disease progression [56].

We establish that two longitudinal bulk sequencing and tumor size measurements are sufficient to reconstruct virtually all parameters (mutation rate, growth rates, times of appearance of driver mutations, and time since the driver mutation) of cancer evolution in individual patients. Our analytic approach yields simple formulas for the parameters; thus, estimation of the parameters governing cancer growth is not computationally intensive, regardless of tumor size. Our framework makes possible a personalized, high-resolution reconstruction of a cancer's timeline of selective events and quantitative characterization of the evolutionary dynamics of the subclones making up the cancer cell population.

2.3 Results

Model

We consider a multi-type branching process of tumor expansion (Fig 2.1A). Tumor growth is started with a single initiated cell at time 0. Initiated tumor cells divide with rate b and die with rate d. These cells already have the driver mutations necessary for expansion, so we assume $b > d$. The population of initiated cells can go extinct due to stochastic fluctuations, or survive stochastic drift and start growing (on average) exponentially with net growth rate $r = b - d$. We will focus only on those populations that survived stochastic drift.

At some time $t_1 > 0$ a new driver mutation occurs in a single initiated tumor cell, starting a new independent birth-death process, with birth rate b_1 and death rate d_1 (Fig 2.1B). Net growth rate of cells with the new driver is $r_1 = b_1 - d_1$. The new driver increases the rate of growth, i.e., $r_1 > r$. We define the driver's selective growth advantage by $g = (r_1/r - 1)$. In addition, both populations of cells (with and without the driver) accrue passenger mutations

with rate u (Fig 2.1C).

After the driver mutation occurs, an additional time t passes before the tumor is observed. Type-0 cells are original initiated tumor and type-1 cells contain the driver mutation. In Materials and Methods we also analyze the more general case of two nested or sibling driver mutations, as well as the fully generalized case of any clonal structure that might arise during tumor expansion.

Parameter estimates from two longitudinal measurements

We demonstrate that with two longitudinal bulk sequencing measurements, it is possible to accurately estimate net growth rates, time of appearance of a driver mutation, time between a driver mutation and observation, and mutation rate in the tumor. The tumor is first sequenced at time of observation, $t_1 + t$, where both time of driver mutation, t_1, and time from driver mutation to observation, t, are yet unknown (Fig 2.1B). A second bulk sequencing is performed at $t_1 + t + \Delta$, a known Δ time units after the tumor is first observed (Fig 2.1B). Later, we apply our method to the CLL data from Ref. [2], where the average size of Δ for all the pre-treatment samples sequenced is 1.8 years (0.6-4.9 years). In general, we expect that in the case of smaller Δ values measurement errors would have a larger effect on the estimated growth rates, due to an expected smaller change in cancer cell count and subclonal structure during a smaller time interval. From the bulk sequencing data, the fraction of cells carrying the driver mutation, α_1 and α_2, can be measured at the time points $t_1 + t$ and $t_1 + t + \Delta$, respectively. We denote total number of cells in the tumor at the two bulk sequencing time points as M_1 and M_2. For liquid cancers, cell counts of the relevant cancer cell population serve as indicators of cancer progression. In the case of CLL, white blood cell (WBC) count is useful as a measure of tumor burden in peripheral blood, as it is routinely taken and includes the cancerous cell population. More precise estimates of tumor burden would include absolute lymphocyte count (ALC) and number of B lymphocytes. Both ALC and WBC counts can suffer from inaccuracies due to the prevalence of smudge cells in CLL,

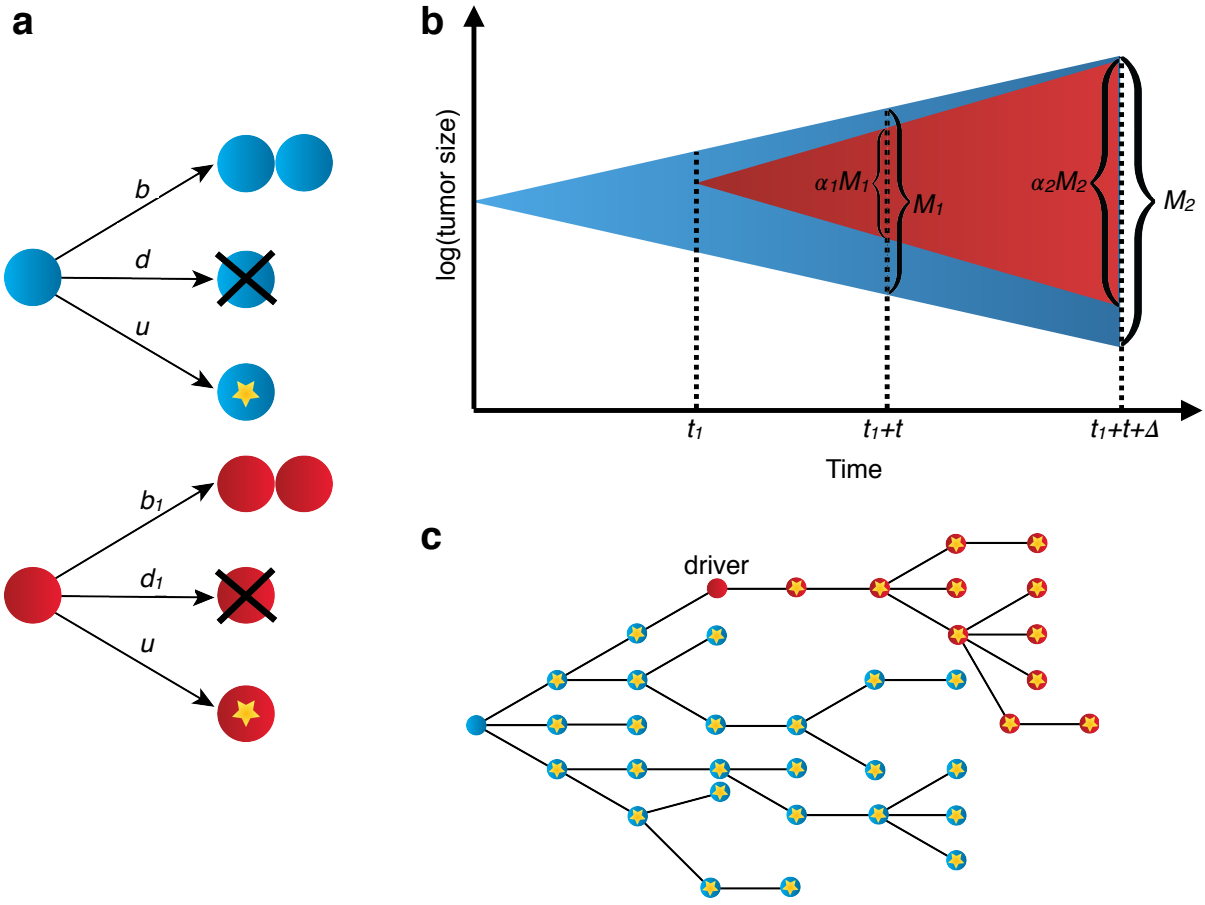

Figure 2.1: **Stochastic branching process model of tumor evolution.** (a) Stochastic branching process model for tumor expansion. Initiated tumor cells (blue) divide with birth rate b, die with death rate d, and accrue passenger mutations with mutation rate u. Type-1 cells, which carry the driver mutation, divide with birth rate b_1, die with death rate d_1, and accrue passenger mutations with mutation rate u. (b) The initiated tumor, or type-0, (blue) population growth is initiated from a single cell. A driver mutation occurs in a single type-0 cell at time t_1, starting the type-1 population (red). The tumor sample is collected and bulk sequenced at times $t_1 + t$ and $t_1 + t + \Delta$, where the driver fraction is α_1 and α_2, respectively. Tumor size (in number of cells) is M_1 and M_2 at first and second sample collection dates. (c) By the time the tumor is observed, it has a high level of genetic heterogeneity due to the mutations that have accrued in both type-0 (blue) and type-1 populations (red). Each yellow star represents a different passenger mutation.

often resulting in an underestimate of these counts [57].

Equating expected values of the sizes of the type-0 and type-1 population at the two bulk sequencing time points with the measured numbers of cells present in clones 0 and 1, we obtain estimates of the net growth rates of the two subclones:

$$r = \frac{1}{\Delta} \log\left(\frac{(1-\alpha_2)M_2}{(1-\alpha_1)M_1}\right) \tag{2.1}$$

$$r_1 = \frac{1}{\Delta} \log\left(\frac{\alpha_2 M_2}{\alpha_1 M_1}\right). \tag{2.2}$$

From the growth rate estimates and subclone sizes, we can approximate the expected value of the time a population in a branching process takes to reach an observed size [50]. This yields an estimate of the time t from the appearance of driver mutation until observation:

$$t = \frac{1}{r_1} \log(M_1 \alpha_1). \tag{2.3}$$

Using the bulk sequencing data from the second time point, γ, the number of subclonal passengers between the specified frequencies f_1 and f_2, can be measured. Using results from previous work [58], we derive the expected value of γ (Materials and Methods), which can be used to estimate the mutation rate u:

$$u = \frac{f_1 f_2 r r_1 \gamma}{(f_2 - f_1)(\alpha_2 r + r_1(1-\alpha_2))}. \tag{2.4}$$

The m passenger mutations that were present in the original type-1 cell when the driver mutation occurred (Fig 2.1C) are present in all type-1 cells. m can be estimated from bulk sequencing data and used to estimate time of appearance of the driver. We maximize the likelihood function $P(m|t_1)$ with respect to time of appearance of the driver, t_1, (see Materials and Methods) to obtain the maximum likelihood estimate

$$t_1 = \frac{m}{u}. \tag{2.5}$$

Using formulas (2.4) and (2.5), we can now estimate t_1.

Estimates verified in simulated tumors

To assess the accuracy of the parameter estimates for several modes of tumor evolution, we simulate tumor growth by performing a Monte Carlo simulation, which simulates the birth, death, and accumulation of mutations in the individual cells that make up a tumor. This simulation generates the mutation frequency and tumor size data used by the estimates (see Methods section for details of simulation). We simulate three different types of tumors (slow growing, fast growing, and no cell death), with a high and a low mutation rate for each (Table 2.3).

In a simulation of a fast-growing tumor with a single subclonal driver mutation that confers a strong selective growth advantage of 100%, we can accurately estimate growth rates, mutation rate, time of driver event, and time since driver event (Fig 2.2AB). Growth rates of both initiated tumor and driver subclones can be estimated with a high degree of accuracy, achieving mean percentage error (MPE) of 0.03% and -0.07% for the lower mutation rate ($u = 1$) scenario. The mutation rate u and estimates for time of driver appearance, t_1, and time since driver, t, can also be estimated accurately, with MPEs of -0.9%, 3.8%, and -0.4%, respectively. Estimates for u, t_1, and t have a somewhat greater degree of variation compared to the growth rate estimates, due to the inherent randomness of the number of mutations and time to reach the observed size that occur in each realization of the stochastic process.

For the parameter regime with no cell death and the regime for a slow-growing tumor, we again achieve high accuracies for the net growth rates (Fig. 2.3AB, Fig. 2.4AB). In the lower mutation rate ($u = 1$) scenario, parameter estimates for the mutation rate u and time of driver appearance t_1 can be accurately estimated for both regimes, with MPEs of -1.3% and 4.9% for the no cell death case, and MPEs of -3% and 3.7% for the slow-growing tumor.

We note that the estimator for t (time since driver event) is biased, with the extent depending on the ratio of birth rate to net growth rate, and the tumor size. The underlying cause of the bias is due to a simplifying assumption in the estimator's derivation (see Methods,

"Derivation of estimates of evolutionary parameters"), and this bias decreases as tumor size increases and as the ratio of growth and division rate gets closer to 1. For the three main modes of growth in our study, we performed additional Monte Carlo simulations to precisely quantify the effect of death:birth ratio and tumor size on the estimator's accuracy (Fig. 2.5). For all three modes of growth, we observe a monotonic decrease in error as tumor size increases to more clinically realistic sizes. For a tumor size of 10^9, all modes of growth have a MPE of less than 4%, so for a clinically realistic cancer size—10^{11} for the CLL dataset—we expect an even better accuracy.

We also perform Monte Carlo simulations for the more complex cases of two nested and two sibling driver subclones (see Methods for derivations of estimators) for the same three modes of cancer growth used for the single driver subclone case above: fast growth (Fig. 2.2CD), no cell death (Fig. 2.3CD), and slow growth (Fig. 2.4CD). For two nested driver subclones, the second driver subclone also carries its parental subclone's driver mutation (Fig. 2.6A). For two sibling driver subclones, the drivers occur in separate subclones (Fig. 2.6B). The growth rate estimates show good agreement with the ground truth values, with MPEs close to 0. The mutation rate estimates also have good accuracy, with the absolute values of their MPEs all $\leq 4\%$. As for the single subclone cases already discussed, the time estimates for the nested and sibling subclone simulations have a greater variance. The estimate for t—time between the last driver mutation and diagnosis—shows good accuracy for the fast-growing tumors, but larger errors for the no cell death and slow growth cases. For both the nested and sibling simulations, the estimates for the times of driver mutations 1 and 2 (t_1 and t_2, respectively) have MPEs less than 6%.

Correcting mutation counts observed from genome sequencing data

We note that in our estimate for the time of appearance of the driver, t_1 (see formula (2.5)), used for comparison to simulated data, we employed a correction to m, the number of mutations that were present in the founder type-1 cell at t_1. From sequencing data, these m mutations are indistinguishable (Fig 2.7A) from mutations that occurred after t_1 in type-1

cells and reached fixation in the type-1 population [58]. Thus, the value of m observed from sequencing data, m_{obs}, will overestimate the true m. In Materials and Methods we show that the expected value of the number of passengers that occurred after t_1 and reached fixation in the type-1 population is u/r_1. We subtract this correction factor from m_{obs}:

$$m = m_{obs} - u/r_1. \tag{2.6}$$

The correction for the m mutations present in the original type-1 cell (2.6) at time t_1 improves the accuracy of the estimate for time of appearance of driver mutation t_1. For the fast-growing tumor with mutation rate $u = 1$ (Fig. 2.8A), the correction lowers the mean percent error (MPE) of the t_1 estimate from 14.0% to 3.8%. For the slow-growing tumor with mutation rate $u = 5$ (Fig 2.7B), the correction lowers the MPE of the t_1 estimate from 22.0% to 5.7% (Fig 2.7B).

Another issue arises from obtaining mutation count γ, number of mutations with frequency between f_1 and f_2, from genome sequencing data. When sequencing data is post-processed by filtering out mutations with L or fewer variant reads, low-frequency mutations will be difficult to detect [47] (Fig 2.7C). For a sample with average sequencing coverage of R and tumor purity p, mutations with mutant allele frequency below $L/(pR)$ will typically not be observable. As a result, since mutations with frequencies between f_1 and f_2 count towards γ, if $f_1 \leq 2L/(pR)$, the observed number of subclonal mutations between frequencies f_1 and f_2, γ_{obs}, will underestimate the true value, γ. For cancers with low mutational burden, such as CLL, we set a relatively low f_1 (1%) to have sufficient resolution to infer mutation rate. Consequently, some mutations with frequency above f_1 will likely be filtered out, and we account for this by correcting for the expected number of such subclonal mutations present at cancer cell frequencies (CCFs) between f_1 and $2L/(pR)$ (see Materials and Methods):

$$\gamma = \gamma_{obs}\left(\frac{\frac{1}{f_1} - \frac{1}{f_2}}{\frac{pR}{2L} - \frac{1}{f_2}}\right). \tag{2.7}$$

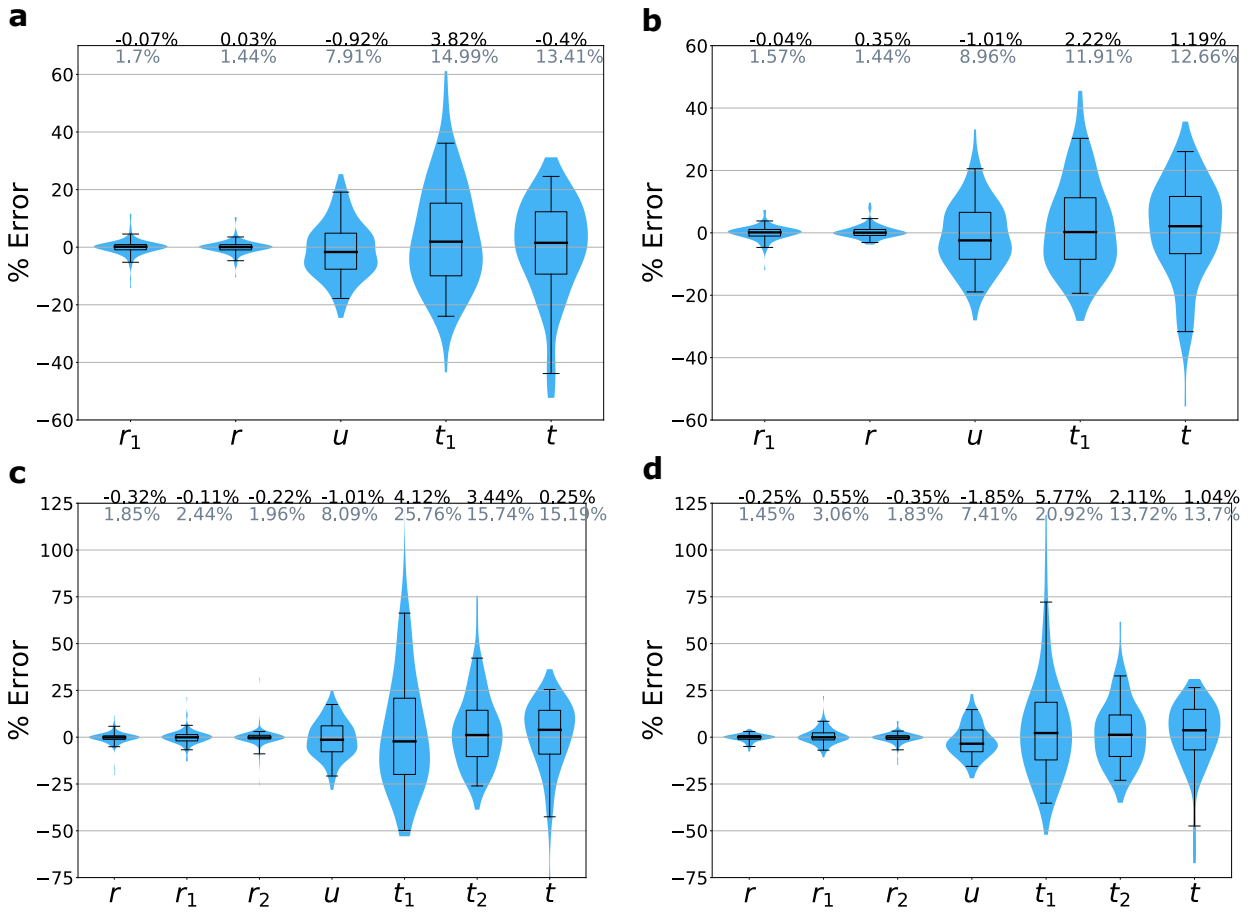

Figure 2.2: **Accuracy of parameter inferences from simulated data**. We simulated tumor growth by performing a Monte Carlo simulation, which simulates the birth, death, and accumulation of mutations in the individual cells that make up a tumor, and generates the mutation frequency and tumor size data used by the estimates. Simulations are of fast-growing tumors with (a) single driver subclone and mutation rate $u = 1$, (b) single driver subclone and $u = 3$, (c) two nested driver subclones with $u = 1$, and (d) two sibling driver subclones with $u = 1$. Mean percent errors (MPEs) of estimates are shown in black above the plots, and mean absolute percent errors (MAPEs) are shown in gray. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error data points. Complete parameter values and number of runs are included in Table 2.3.

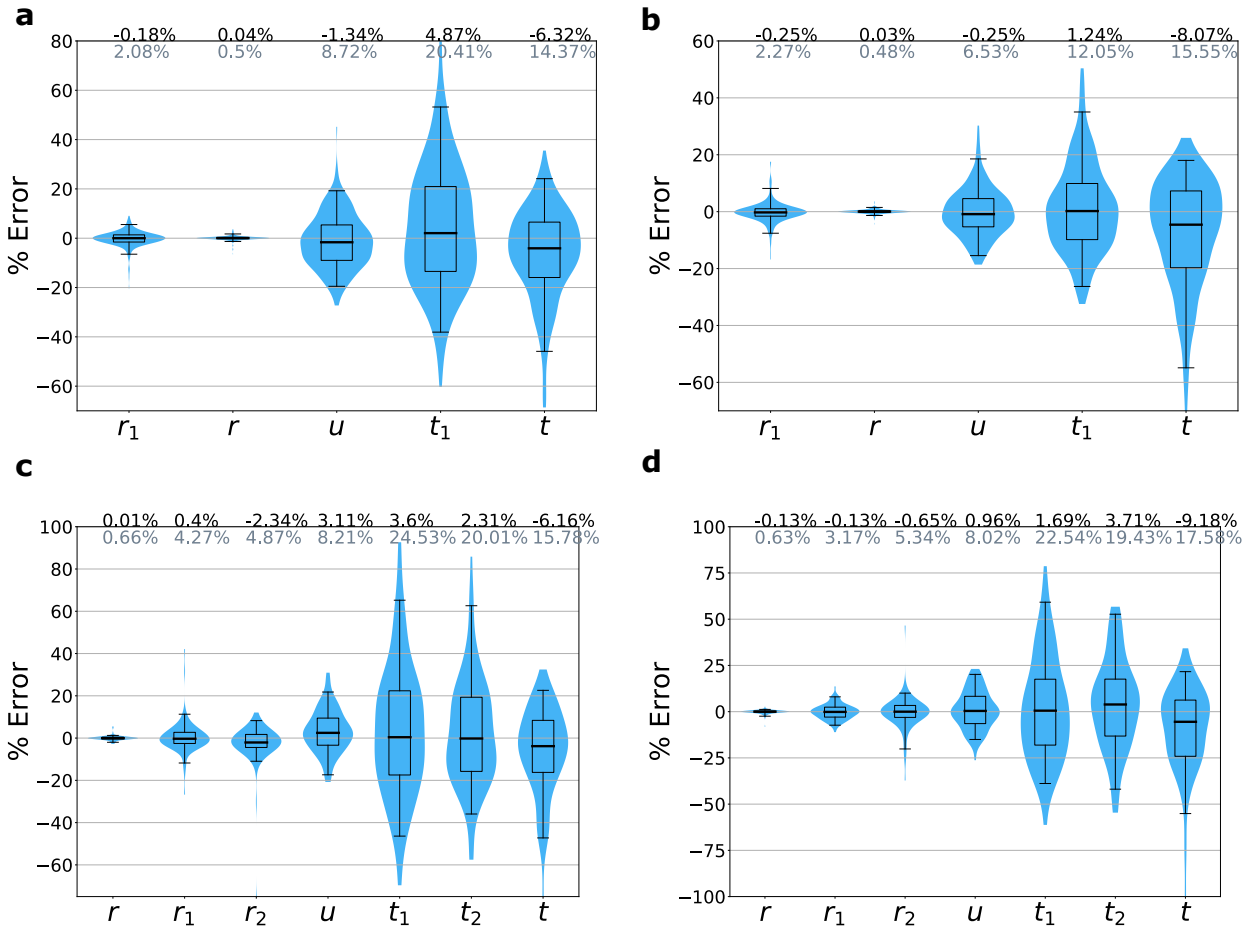

Figure 2.3: **Percent errors (PEs) for case with no death**. Accuracy of parameter inferences for Monte Carlo simulation of tumor with no cell death for (a) single driver subclone with mutation rate $u = 1$, (b) single driver subclone with $u = 10$, (c) two nested subclones with $u = 1$, and (d) two sibling subclones with $u = 1$. Mean percent error (MPEs) are the black numbers above the plots, and mean absolute percent errors (MAPEs) are the grey numbers below the MPEs. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error datapoints. Complete parameter values and number of runs are included in Table 2.3.

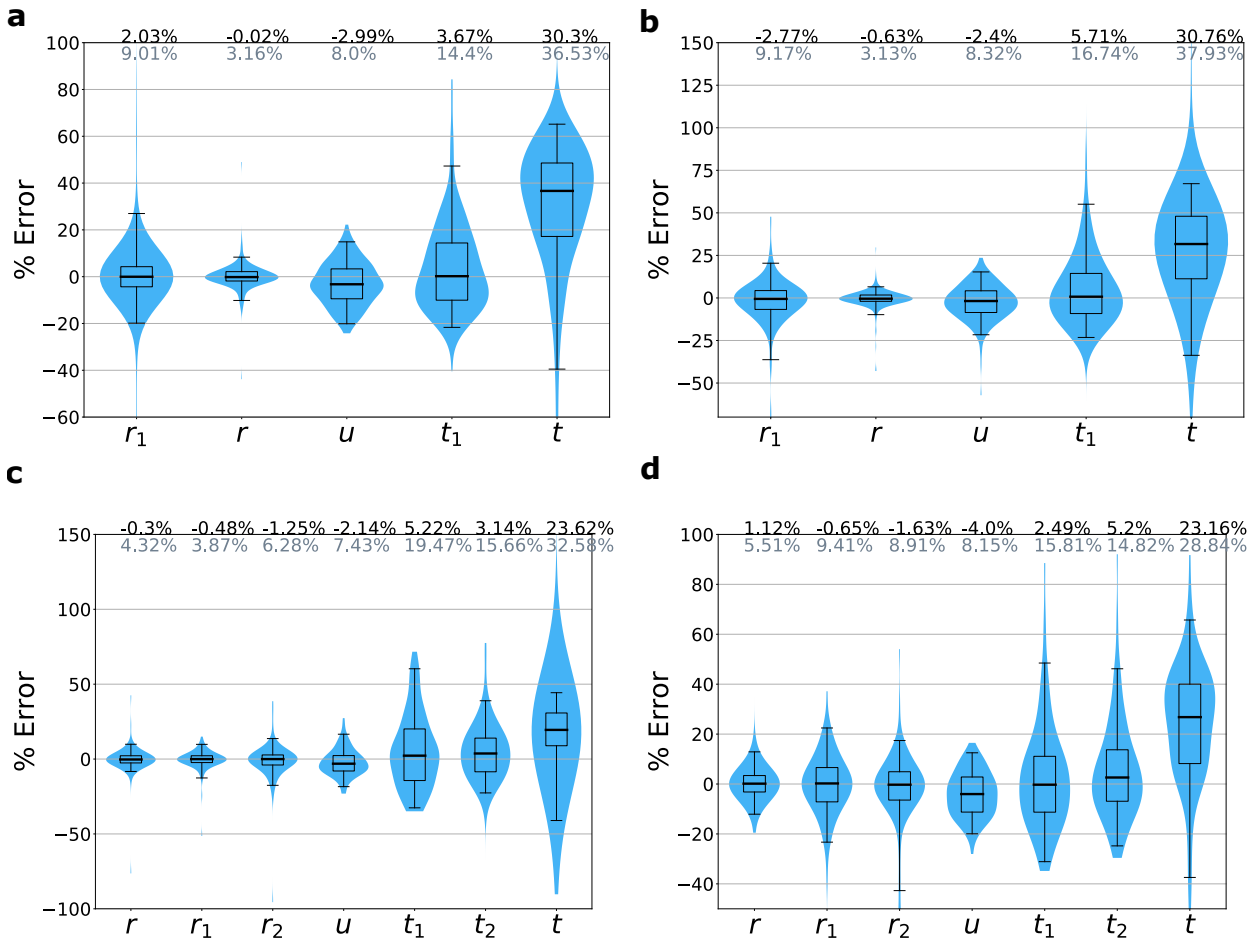

Figure 2.4: **Percent errors (PEs) for slow-growing tumor**. Accuracy of parameter inferences for surviving Monte Carlo simulation runs of slow-growing tumor for (a) single subclone with mutation rate $u = 1$, (b) single subclone with $u = 5$, (c) two nested subclones with $u = 1$, and (d) two sibling subclones with $u = 1$. Mean percent error (MPEs) are the black numbers above the plots, and mean absolute percent errors (MAPEs) are the grey numbers below the MPEs. Boxes contain 25th-75th quartiles, with median indicated by thick horizontal black line. Whiskers of boxplots indicate 2.5 and 97.5 percentiles. Violins are smoothed density estimates of the percent error data points. Complete parameter values and number of runs are included in Table 2.3.

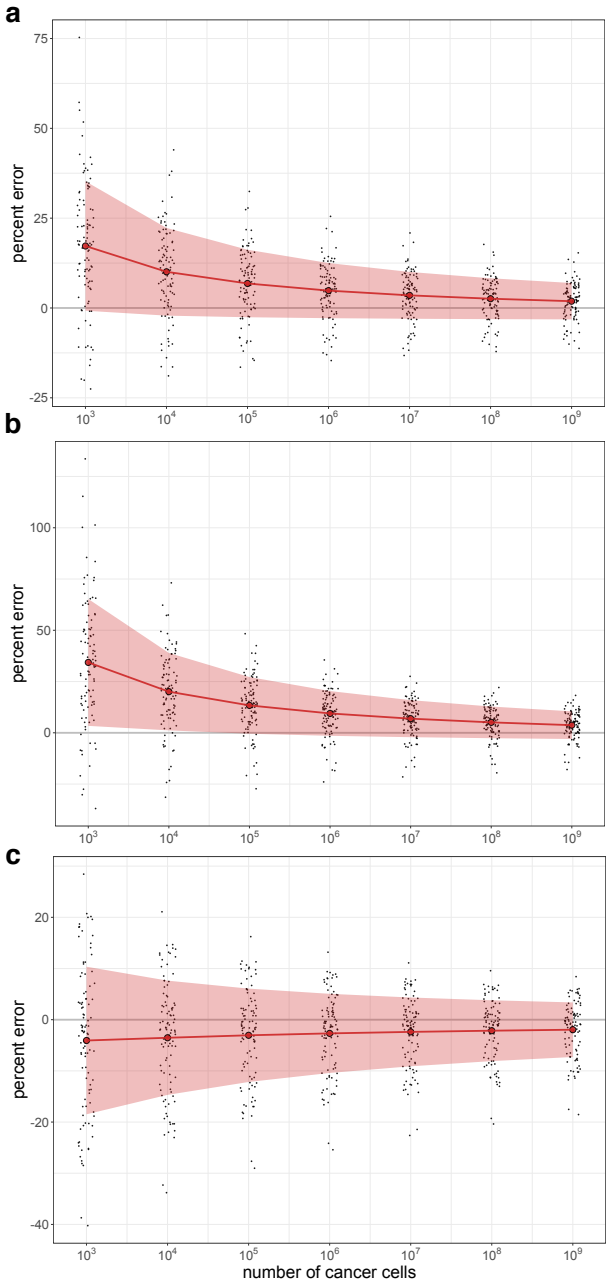

Figure 2.5: **Accuracy for t estimate increases with tumor size.** A Monte Carlo simulation of a birth-death process was performed for (a) fast-growing, (b) slow-growing, and (c) no cell death parameter regimes. For each of the 100 surviving simulated tumors, the percent error of the t estimate (Eq. (2.3)) was calculated when the tumor first reached the specified tumor sizes. Means are indicated by red points and lines, ± one standard deviation is shown by the red region, and individual data points for each simulation run are shown as the grey points (with horizontal jitter for visibility).

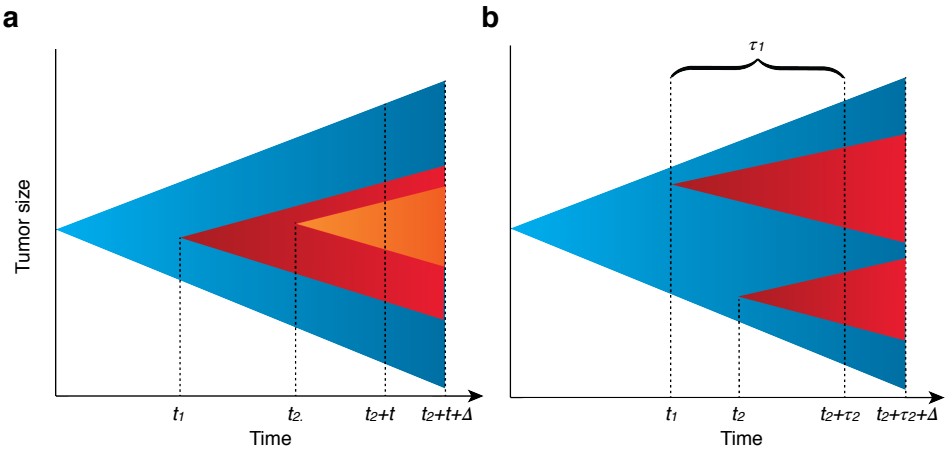

Figure 2.6: **Model for tumor expansion with two driver mutations.** (a) Two nested driver subclones. Initiated tumor (type-0) cells in blue, cells with driver 1 (type-1) in red, and cells with both drivers (type-2) in orange. A driver mutation occurs in a type-0 cell at t_1. A second driver mutation occurs in a type-1 cell at $t_1 + t_2'$. Tumor is bulk sequenced at $t_1 + t_2' + t$ and $t_1 + t_2' + t + \Delta$. (b) Two sibling driver subclones. Type-0 cells (in blue). A driver mutation occurs in a type-0 cell at t_1. A second driver mutation occurs in a different type-0 cell at t_2. Tumor is bulk sequenced at $t_1 + \tau_1$ (or, equivalently $t_2 + \tau_2$) and $t_1 + \tau_1 + \Delta$ (equivalently $t_2 + \tau_2 + \Delta$).

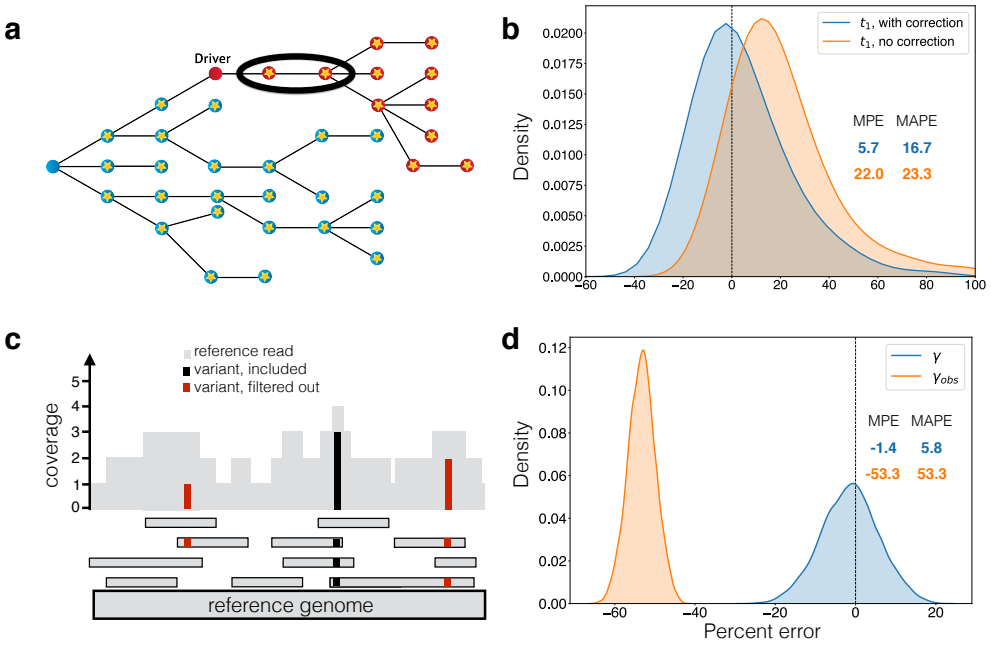

Figure 2.7: **Corrections for observed mutation counts.** (a) If passenger mutations (circles with stars) that occur after the driver reach fixation in the driver population (red), then they are indistinguishable from the passengers that were present in the first cell with the driver, which accrued in the type-0 population (blue). The estimate of when the driver occurred needs to account for these mutations (circled). In (b), we compare percent errors of parameter estimates for time from tumor initiation until appearance of a driver subclone, t_1, with and without this correction (Eq. (2.6)). Errors for estimate with correction are shown in blue, and for estimate without correction (Eq. (2.5)) in orange. Errors are plotted as a kernel density estimate for Monte Carlo simulations of slow-growing tumor with mutation rate $u = 5$. Mean percent errors (MPEs) and mean absolute percent errors (MAPEs) are listed. (c) Mutations present on two or fewer variant reads (red) are filtered out in post-processing. Mutations with more than two variant reads (black) are included. The number of subclonal mutations between frequencies f_1 and f_2, γ, which is used in the mutation rate estimate, must be corrected for mutations that are filtered out. In (d), the percent errors for the observed (orange) and corrected (blue) γ (Eq. (2.7)) are plotted as kernel density estimates. Observed mutations are those that passed post-processing, i.e. those that have more than $L = 2$ mutant reads. True mutation frequencies were generated from 135 surviving runs of a Monte Carlo simulation of a fast-growing tumor with mutation rate $u = 1$, from which sequencing reads were simulated with 200x average coverage (see Materials and Methods). Percent errors are calculated relative to the true γ measured from the true mutation frequencies.

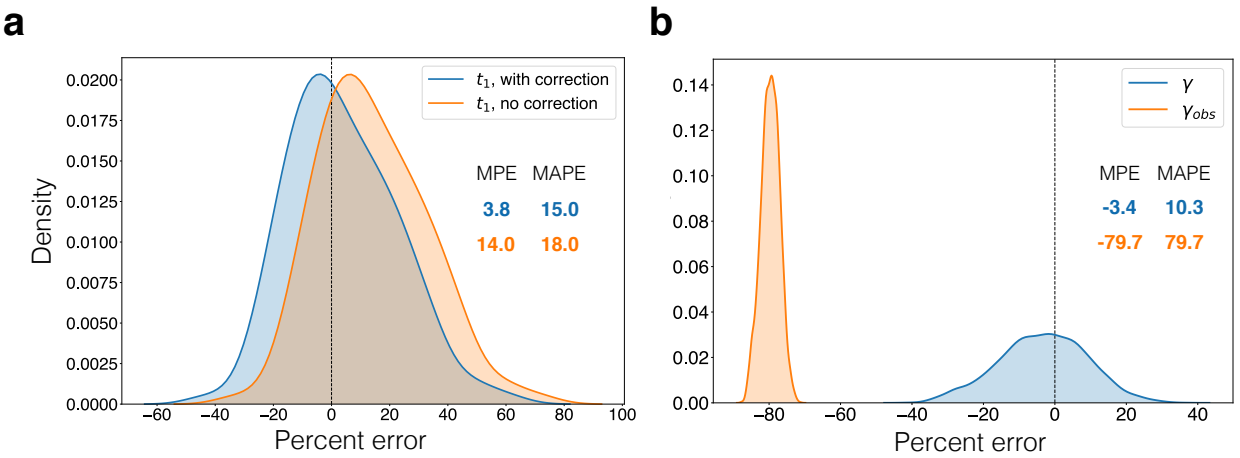

Figure 2.8: **Corrections for observed mutation counts.** (a) We compare percent errors of parameter estimates for time from tumor initiating until appearance of a driver subclone, t_1, with and without the correction for passengers that occur after the driver and reach fixation in the driver population (Eq. (2.6)). Errors for estimate with correction are shown in blue, and for estimate without correction ((2.5)) in orange. Errors are plotted as a kernel density estimate for Monte Carlo simulations of fast-growing tumor with mutation rate $u = 1$. Mean percent errors (MPEs) and mean absolute percent errors (MAPEs) are listed. (b) The percent errors for the observed (orange) and corrected (blue) number of subclonal mutations between frequencies f_1 and f_2, γ, (Eq. (2.7)) are plotted as kernel density estimates. Observed mutations are those that passed post-processing, i.e. those that have more than $L = 2$ mutant reads. True mutation frequencies were generated from 135 surviving runs of a Monte Carlo simulation of a fast-growing tumor with mutation rate $u = 1$, from which sequencing reads were simulated with 100x average coverage (see Materials and Methods). Percent errors are calculated relative to the true γ measured from the true mutation frequencies.

Before applying our methodology to patient sequencing data, we estimated the validity of the above correction applied to observed simulated mutation counts. When we simulate sequencing reads from simulated mutation frequencies (see Materials and Methods) and post-process by removing mutations with $L = 2$ or fewer variant reads, the adjustment we derived for mutation count γ (2.7) is critical, even for average sequencing coverage of 200x (Fig 2.7D). Without any correction, the observed γ has MPE of -53.3% compared to true γ, but with the correction, the computed γ has MPE of -1.4%. When average coverage is 100x, this correction becomes even more important, as many of the low-frequency mutations are discarded (Fig. 2.8B). Without any correction, the observed γ has MPE of -79.7%. With the correction the computed γ has MPE of -3.4%. The accuracy of the γ measurement affects our estimate of the mutation rate (2.4).

Estimating parameters for individual patients with CLL

We use our formulas to infer the patient-specific parameters of cancer evolution for four patients with CLL whose growth patterns and clonal dynamics were analyzed in [2]. These CLLs had peripheral WBC counts measured and whole exome sequencing (WES) performed at least twice before treatment. We consider patients whose WBC counts were classified as having an exponential-like growth pattern, with average $\gamma_{obs} > 2$, and with 3 or fewer macroscopic subclones (i.e. subclones with cancer cell fractions of 20% or greater for at least one pre-treatment time point). Our framework is designed specifically to study naturally evolving cancer dynamics, unperturbed by treatment, which will drastically alter the cancer's dynamics and size. For calculation of the γ_{obs} mutations between frequencies f_1 and f_2, we set $f_1 = 1\%$ due to the difficulty of detecting low frequency variants $< 1\%$ [59, 60]. We set f_2 to 20% to minimize overlap with potential driver mutations of the macroscopic subclones. The average γ_{obs} for the four analyzed patients ranges from 2.5 to 19.3, with a median of 5.2. As in Ref. [2], we perform subclonal reconstruction for each patient using PhylogicNDT [3]. To obtain confidence intervals for our parameter estimates, we utilize a sampling procedure to account for model and measurement uncertainties, including uncertainties in subclone

frequencies, fitted growth curves, and the Poisson process for mutation accumulation (see Materials and Methods). For each patient's tumor, we compute estimates of the growth rate of each clone, exome mutation rate, the times that each subclone arose, and how long each subclone expanded before the tumor was detected (Table 2.1). We also estimate what time the cancer was clinically detectable, by sampling from the distribution of fitted growth parameters and solving the resulting root-finding problem for time to reach detectable size under our growth model (see Materials and Methods). For CLL specifically, we compute time of leukocytosis—an abnormally high WBC count. We reconstruct these histories for tumors with various clonal structures.

Table 2.1: Inferred parameters for CLL patients with exponential growth patterns, for which there are at least two longitudinal bulk sequencing measurements before treatment. Estimates are computed from tumor size measurements and mutation frequencies from whole exome sequencing. Mutation rates are for the exome only. The time estimates are in terms of the patient's age in years.

Parameter	Pt. 3	Pt. 6	Pt. 9	Pt. 21
r (/yr)	0.51	0.68	0.28	0.79
r_1 (/yr)	0.85	0.41	-0.40	1.52
r_2 (/yr)		0.46	0.67	
r_3 (/yr)		1.09	0.63	
u (mut/yr)	0.48	0.15	0.36	0.20
MRCA (yr)	14.6	2.8	4.9	6.4
t_1 (yr)	33.5	35.4	18.8	19.6
t_2 (yr)		46.7	21.3	
t_3 (yr)		45.9	24.8	
age at diagnosis (yr)	63	58	54	35
age at leukocytosis (yr)	61.9	65.7	51.8	34.4

Table 2.2: Confidence intervals for inferred parameters for CLL patients with exponential growth patterns, for which there are at least two longitudinal bulk sequencing measurements before treatment. Estimates are computed from tumor size measurements and mutation frequencies from whole exome sequencing. Mutation rates are for the exome only. The time estimates are in terms of the patient's age in years.

Parameter	Pt. 3	Pt. 6	Pt. 9	Pt. 21
r (/yr)	[0.20, 0.85]	[0.15, 1.30]	[0.17, 0.42]	[0.30, 1.14]
r_1 (/yr)	[0.65, 1.04]	[0.08, 0.73]	[-0.45, -0.19]	[1.01, 2.04]
r_2 (/yr)		[0.08, 0.85]	[0.49, 0.94]	
r_3 (/yr)		[0.65, 1.78]	[0.39, 0.86]	
u (mut/yr)	[0.39, 0.59]	[0.12, 0.19]	[0.35, 0.37]	[0.19, 0.23]
MRCA (yr)	[1.4, 26.8]	[0.1, 13.2]	[1.2, 10.8]	[0.3, 16.7]
t_1 (yr)	[24.1, 39.2]	[21.7, 46.1]	[8.8, 35.1]	[10.8, 24.0]
t_2 (yr)		[25.6, 57.5]	[7.7, 31.7]	
t_3 (yr)		[31.3, 54.6]	[10.3, 37.6]	
age at leukocytosis (yr)	[60.3, 62.4]	[64.2, 67.1]	[51.6,51.9]	[32.8,34.6]

Patients 3 and 21 are examples of a CLL with a single subclone (Fig 2.9). For Patient 3, Clone 0, the most recent common ancestor (MRCA) of this patient's CLL, was initiated when the patient was 14.6 [1.4, 26.8] years old (median and [95% confidence interval] of estimate). Clone 0 grew with a net growth rate of 0.51 [0.20, 0.85] per year. Approximately two decades later, Clone 1 was initiated when the patient was 33.5 [24.1, 39.2] years old. Clone 1 expanded with a growth rate of 0.85 [0.65, 1.04] per year (corresponding to a selective growth advantage of 68.7% over Clone 0), and the patient was diagnosed approximately three decades later at age 63.

For patient 21, we estimate that the parental clone (MRCA, Clone 0) of this patient's

CLL was initiated when the patient was 6.4 [0.3, 16.7] years old, and grew with a net growth rate of 0.79 [0.30, 1.14] per year. Clone 1 appeared when the patient was 19.6 [10.8, 24.0] years old, and grew more quickly than Clone 0, with a selective growth advantage of ∼90% over Clone 0). Clone 1 contained a FGFR1 mutation, which might have been acting as a driver of the increased net proliferation. Clone 1 then grew for ∼15 years before the patient was diagnosed at age 35.

Patients 6 and 9 present more complex clonal structures (Fig. 2.9). Clone 0, the parental clone of the CLL of Patient 9, arose when the patient was 4.9 [1.2, 10.8] years old, and had a growth rate of 0.28 [0.17, 0.42] per year. Clone 1 arose when the patient was 18.8 [8.8, 35.1] years old. Interestingly, during clinical observation between diagnosis and treatment, Clone 1 was declining in size, with a growth rate of -0.40 [-0.45, -0.19] per year. In line with recent findings [61], we found that sometimes the estimated growth rate during the period of observation, such as the negative growth rate of Clone 1, is smaller than the minimal possible growth rate necessary to reach the observed clone size. In that case, for calculating mutation rate, time of the driver(s), time of detectability, and time between driver(s) and diagnosis we use the minimal growth rate. Clone 2, containing a KRAS mutation, had the largest net growth rate of the three clones (0.67 [0.49, 0.94] per year), corresponding to a selective growth advantage of 140.9% over the parental clone. Clone 2 arose when the patient was 21.3 [7.7, 31.7] years old.

We estimate that the CLL of Patient 6 was initiated when the patient was 2.8 [0.1, 13.2] years old. The leukemic parental clone, Clone 0, then grew at a rate of 0.68 [0.15, 1.30] per year. Approximately 33 years after the appearance of Clone 0, when the patient was 35.4 [21.7, 46.1] years old, the first subclone, Clone 1 appeared. Clone 3 arose from within Clone 1 when the patient was 45.9 [31.3, 54.6] years old. Clone 3 harbored a driver mutation in ASXL1 and had selective growth advantage of 60.8% over Clone 0. The patient was diagnosed at age 58, eventually needing treatment 12.0 years after diagnosis.

The average mutation rate in the four CLL patients we analyze is 0.30 mutations/year. This rate is over the exome, which accounts for $\sim 1\%$ of the human genome. Our average

estimated mutation rate in CLL exomes is similar to the measured rate of accumulation of mutations in human tissues of 40 mutations per year over the entire genome [62]. Other recent work has estimated a mutation rate of 17 mutations per year in human haematopoietic stem cell/multipotent progenitors [63]. Our estimated mutation rates during CLL progression are on par or higher than the recent estimates in healthy hematopoietic cells [63], in line with the expectation that mutation rates may be increased in cancer. The estimated times of appearance of CLL subclones are very long, on the order of 10 years or more. This finding is in agreement with results from Gruber et al. [2], who find few new CLL subclones over years to a decade of evolution. We observe that CLL initiation occurred early in most patients, within the first fifteen years of their lives, consistent with recent work in other cancer types [32, 48]. We find that CLL patients reach leukocytosis an average of 1.5 years before the first timepoint at which cancer genome sequencing was performed. For three of the patients, our estimated time of leukocytosis was before diagnosis, on average 1.3 years prior to diagnosis.

2.4 Discussion

We use a stochastic branching process model to reconstruct the timing of driver events and quantify the evolutionary dynamics of different subclonal populations of cancer cells. We estimate growth rates of tumor subclones, selective growth advantage of individual driver mutations, mutation rate in the tumor, time between tumor initiation and appearance of a subclonal driver mutation, and time between driver mutation and tumor observation. Together, this allows us to estimate the age of the patient at tumor initiation, as well as the age at appearance of a subclonal driver.

Previous work has computed relative order of driver events [31, 34, 64], while other studies have given estimates for scaled mutation rates and time of events [37, 44]. However, we present estimates for absolute, unscaled mutation rates and times, which are easily interpretable and don't implicitly depend on unknown parameters. We assume that mutations accrue with time, which simplifies derivations and is supported by recent experimental data that shows that non-dividing cells may accrue mutations at a similar rate as dividing cells [7].

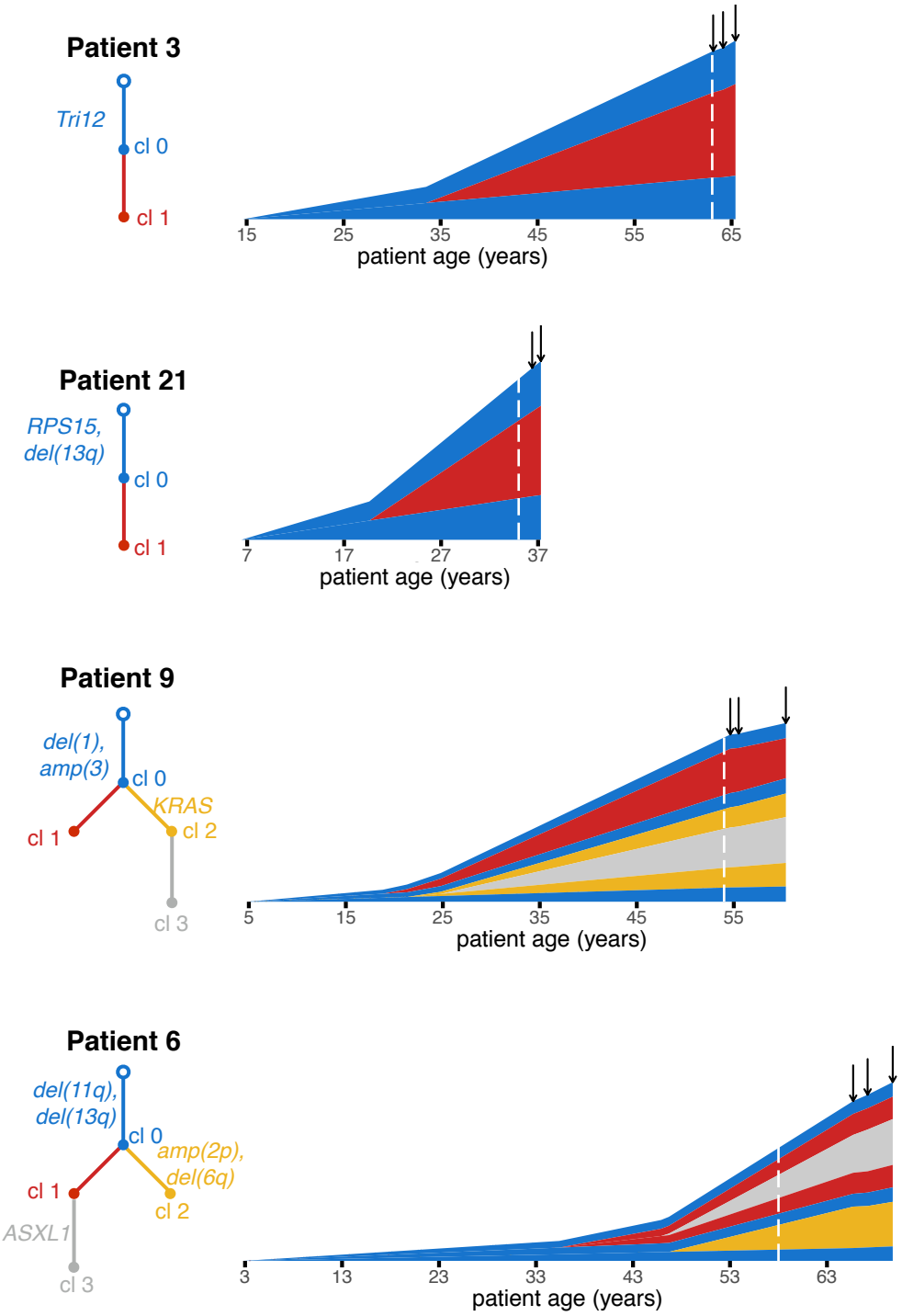

Figure 2.9: **Reconstructing the timeline of CLL evolution in patients.** We applied our methodology to estimate subclonal growth rates, mutation rates and evolutionary timelines in CLL tumors from Ref. [2]. Vertical height of a clone represents its $\log_{10}$-scaled size. Mutations were clustered into clones and phylogenetic trees were inferred using PhylogicNDT [3]. Tree edges are colored by clone number and are labeled with driver mutations, if any. For each patient, we show estimates for patient age at CLL initiation and times of appearance of CLL subclones. Dashed white line indicates when the patient was diagnosed. Solid black arrows indicate times of bulk sequencing measurements.

Other potential assumptions regarding mutation accumulation include mutations occurring at cell division [65] or assuming mutation rate is proportional to the copy number state [66]. For example, recent work reported that some mutational signatures in human cancers are generated during mitosis [65]. Other work has shown that the rate of accumulation of somatic single nucleotide variants is proportional to copy number [66]. We further assume that all cancer subpopulations have the same passenger mutation rate. In the case that mutations occur predominantly at cell division, assuming that the rate of cell division is comparable across all tumor subclones, our estimates would still be valid. In the case of a subclone that has an elevated mutation rate (e.g. due to a chromosomal amplification, mutation in a DNA repair pathway gene or an increased cell division rate), we would underestimate the mutation rate and overestimate the time of driver mutation(s) in that subclone. In the other subclones, the opposite would be true.

For individual CLLs that underwent bulk sequencing at two time points [2], we infer growth rates of individual subclones, mutation rate in the tumor, the times when cancer subclones began growing, the time between driver mutations and the patient's diagnosis, and time when the cancer is clinically observable. Our inferences are limited by the relatively low number of mutations present in CLL, as well as sequencing coverage [2], so we set a minimum passenger mutation count when selecting specific cases to analyze. The accuracy of estimates presented here is expected to be higher with whole genome sequencing available, with higher sequencing coverage, or in cancer types with more mutations, with some important limitations. Exponential growth—the mean behavior of our branching process model—has been well documented in vivo [2, 67–69], but tumors can also often exhibit sigmoidal growth (e.g. logistic, Gompertz models), where initial exponential growth is followed by a deceleration in growth [68, 70–73]. Our estimators should only be used for cancers exhibiting exponential growth; for other modes of growth, such as the logistic-growing class of CLL patients in Ref. [2], the parameter estimates would have to be derived specifically for the particular mode of growth observed. Exponential growth is the simplest common cancer growth pattern, and yet, estimating the exponential growth rates requires at least two longitudinal timepoints. To

fit all parameters for patients with more complex growth dynamics, additional longitudinal samples will be needed; this type of analysis would be further limited due to the scarcity of longitudinal pre-treatment samples in many cancer types. In the case of solid tumors, the number of cells can be estimated from measurements of tumor volume [74], however multiple biopsies would potentially be needed to fully account for the existing genetic heterogeneity. Furthermore, a solid tumor's spatial structure, mode of evolution, and biopsy collection influence how well selection and mutation spectra can be observed [42, 43, 75]. Recent modeling and computational work, in combination with careful multi-region sequencing and single cell sequencing, have begun to disentangle these confounding factors [39, 41, 42].

Our model and derivations assume a fixed mutation rate u after transformation and fixed growth rates of cancer subclones, similar to previous approaches [37, 42, 47]. Some individual cancer subclones (such as Clone 1 from Pt. 9) not only do not grow exponentially, they actually decline in absolute cell numbers, even if the overall tumor is undergoing expansion. This phenomenon has been previously observed [2, 76], and could be caused by the declining subclone getting outcompeted by more fit subclones. Sudden genomic instability events, or a change in cancer mutation and/or growth rate over time could also introduce errors into our parameter inferences. Recent sequencing data points to mutational processes that change over time during cancer evolution [33, 77]; incorporating possible changes in the mutation and/or growth rate into the model would require much higher density of sequencing and clinical data [49], as would employing a more complex growth model (e.g. boundary-driven or sigmoidal growth).

2.5 Materials and methods

Branching process model of tumor evolution

We employ a continuous, multi-type branching process model of cancer evolution. For the case of a single driver subclone, there are two cell types, type-0 and type-1. Tumor expansion is initiated by a single type-0, or initiated tumor cell. Type-0 cells divide with rate b and

die with rate d, yielding a net growth rate of $r = b - d$. At time t_1, a single driver mutation is introduced into a randomly selected cell in the type-0 population, founding a new type-1 population of cells. This type-1 population undergoes its own independent branching process. They divide with rate b_1, die with rate d_1, and have net growth rate $r_1 = b_1 - d_1$. If the driver mutation gives type-1 cells a selective growth advantage over the type-0 population, then $r_1 > r$. With the ratios of the growth rates denoted as $s = r_1/r$, the growth advantage can be quantified as $g = (s - 1) \cdot 100\%$. In the case of neutral evolution, $g = 0$. If there is a selective advantage, $g > 0$. Neutral mutations, or passengers, have no effect on the cell's fitness, and accrue according to a Poisson process with rate u. We assume an infinite alleles model such that there is no back mutation and an infinite sites model such that every new passenger mutation is unique. Only surviving populations are considered. All derivations below will condition on survival. The type-0 and type-1 populations at time t will be denoted as $X_0(t)$ and $X_1(t)$, respectively.

Measurements sufficient to determine evolutionary history

Here we derive estimates for parameters describing the carcinogenic process for a single driver subclone, using measurements taken from two time points late in the tumor's development. We require sequencing of the tumor at the two time points, when the tumor is first observed at the unknown time $t_1 + t$ and a specified Δ later, at $t_1 + t + \Delta$. From these two bulk sequencing measurements, we obtain measurements of α_1 and α_2, the fraction of cells carrying the driver mutation at $t_1 + t$ and $t_1 + t + \Delta$, respectively. In addition, from the bulk sequencing at $t_1 + t + \Delta$, we obtain measurements of m, the number of mutations present in the founder type-1 cell, as well as γ, the number of mutations with frequency between the specified f_1 and f_2. The total population size at these times, M_1 and M_2, is also measured.

Expected value of γ, number subclonal mutations

For a population consisting of a single clone with birth and death rates b and d, the expected number of subclonal mutations present at a frequency larger than f is shown to be [58]

$$\frac{\bar{u}(1-f)}{(1-\delta)f} \tag{2.8}$$

where $\delta = d/b$ and $\bar{u}$ is the probability that a daughter cell gains a new passenger mutation at cell division. In this paper, we allow mutations to occur at any point in time and consider the absolute mutation rate per cell, u, which is equal to $\bar{u}b$. Then the expected number of subclonal mutations between f_1 and f_2, $\mathbb{E}\gamma$, is

$$\mathbb{E}\gamma = \frac{u(1-f_1)}{b(1-\delta)f_1} - \frac{u(1-f_2)}{b(1-\delta)f_2} \tag{2.9}$$

$$= \frac{u}{r}(1/f_1 - 1/f_2) \tag{2.10}$$

where $r = b - d > 0$.

Now we derive $\mathbb{E}\gamma$ in the case of clones 0 through k, each clone with growth rate $r_i > 0$ and fraction α_i^c. Each clone i has $\alpha_i^c \frac{u}{r_i}(1/f_1 - 1/f_2)$ expected subclonal passengers between frequencies f_1 and f_2. Thus, the total expected number of passengers with frequencies between f_1 and f_2 is

$$\mathbb{E}\gamma = (1/f_1 - 1/f_2)\sum_{i=0}^{k} \frac{u\alpha_i^c}{r_i}. \tag{2.11}$$

For the simplest case we consider, a tumor with a single driver mutation occurring in the initiated tumor population, there is a type-0 population with growth rate r and a type-1 population with growth rate r_1. Equation (2.11) reduces to

$$\mathbb{E}\gamma = \left(\frac{u\alpha}{r_1} + \frac{u(1-\alpha)}{r}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right) \tag{2.12}$$

where α is the fraction of cells having the driver mutation.

Derivation of estimates of evolutionary parameters for single driver subclone

With the cancer bulk sequenced at the two time points $t_1 + t$ and $t_1 + t + \Delta$, we are able to derive estimates for t_1, t, r, r_1, and u. First we solve for r and r_1, based on the estimated cell counts at $t_1 + t$ and $t_1 + t + \Delta$. The observed type-i cell count is equated to the expected value of the type-i population size, conditioned on survival. For a birth-death process started with a single type-i cell at time 0, we have $\mathbb{E}[X_i(t)] = e^{r_i t}$. That process has extinction probability d_i/b_i [50]. Then,

$$\mathbb{E}[X_i(t)] = \mathbb{E}[\mathbb{E}[X_i(t)|I_{X_i(t)>0}]] \tag{2.13}$$

$$\approx \mathbb{E}[X_i(t)|X_i(t) = 0](d_i/b_i) + \mathbb{E}[X_i(t)|X_i(t) > 0](1 - d_i/b_i) \tag{2.14}$$

$$= \mathbb{E}[X_i(t)|X_i(t) > 0](1 - d_i/b_i) \tag{2.15}$$

where $I_{X_i(t)>0}$ is a random variable and indicator function defined as

$$I_{X_i(t)>0} = \begin{cases} 0 & \text{if } X_i(t) = 0 \\ 1 & \text{if } X_i(t) > 0 \end{cases}.$$

Thus, from (2.15), for large enough time t,

$$\mathbb{E}[X_i(t)|X_i(t) > 0] \approx \frac{1}{1 - d_i/b_i} e^{r_i t} = \frac{b_i}{r_i} e^{r_i t}. \tag{2.16}$$

It then follows that for the type-0 population,

$$\mathbb{E}[X_0(t_1 + t)|X_0(t_1 + t) > 0] = \frac{b}{r} e^{r(t_1+t)} = (1 - \alpha_1)M_1 \tag{2.17}$$

$$\mathbb{E}[X_0(t_1 + t + \Delta)|X_0(t_1 + t + \Delta) > 0] = \frac{b}{r} e^{r(t_1+t+\Delta)} = (1 - \alpha_2)M_2. \tag{2.18}$$

Proceeding similarly for the type-1 population, we obtain

$$r_1 = \frac{1}{\Delta} \log\left(\frac{\alpha_2 M_2}{\alpha_1 M_1}\right) \tag{2.19}$$

$$r = \frac{1}{\Delta} \log\left(\frac{(1 - \alpha_2)M_2}{(1 - \alpha_1)M_1}\right). \tag{2.20}$$

The expected value of the first time a population of type-1 cells in a branching process reaches the observed size $\alpha_1 M_1$ is [50]

$$\mathbb{E}t = \frac{1}{r_1}\log\Big(\frac{\alpha_1 M_1 r_1}{b_1}\Big) - \frac{1}{r_1}\int_0^\infty e^{-z}\log z dz \tag{2.21}$$

$$= \frac{1}{r_1}\log\Big(\frac{\alpha_1 M_1 r_1}{b_1}\Big) + \frac{0.5772}{r_1} \tag{2.22}$$

$$= \frac{1}{r_1}\Big(\log(\alpha_1 M_1) + \log(r_1/b_1) + 0.5772\Big) \tag{2.23}$$

$$\approx \frac{1}{r_1}\log(\alpha_1 M_1). \tag{2.24}$$

The last approximation is justified because for realistic cell counts, the first term in (2.23) dominates the other two, which is also evident in simulation studies (Fig. 2.5). For example, if $r_1 = \frac{1}{2}b_1$, then the second term $\log(r_1/b_1) = -0.69$, compared to the first term $\log(\alpha_1 M_1) = 19.11$. Even if r_1 is as low as $0.1b_1$, the second term is -2.30. In this case, the percent error of the approximation (2.24) is 7.3%. In general, the accuracy increases with increased tumor size.

With the measurement of γ, the number of subclonal passengers with frequency between f_1 and f_2, we can estimate the mutation rate u. In the previous section we derive the expected value of γ as

$$\mathbb{E}\gamma = \left(\frac{u\alpha}{r_1} + \frac{u(1-\alpha)}{r}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right). \tag{2.25}$$

Using the estimates of r and r_1 from (2.19) and (2.20), and the measured value of γ from the second bulk sequencing, equation (2.25) can be solved for the mutation rate u,

$$u = \frac{f_1 f_2 r r_1 \gamma}{(f_2 - f_1)(\alpha_2 r + r_1(1-\alpha_2))}. \tag{2.26}$$

When estimating mutation rate for the CLL patients from Ref. [2], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (2.26) is applied for each time point with the respective CCFs and observed γ values for each time point.

To derive the maximum likelihood estimates of t_1, we consider the likelihood function $P(m|t_1)$. The number of passenger mutations present in the founder type-1 cell that appeared at time t_1 is a Poisson process with rate u. Thus,

$$P\big(m|t_1\big) \propto \frac{(ut_1)^m e^{-ut_1}}{m!}. \tag{2.27}$$

Maximizing the logarithm of the likelihood function with respect to t_1 yields a MLE for t_1 in terms of estimated or measured quantities:

$$t_1 = m/u. \tag{2.28}$$

Estimating number of unobserved subclonal mutations from sequencing data

When sequencing data is post-processed by filtering out any mutations with L or fewer variant reads, the number of mutations between f_1 and f_2 will likely be underestimated if $2L/(Rp) > f_1$, where R is average sequencing coverage and p is tumor purity. Define γ_{obs} as the observed number of mutations between frequencies f_1 and f_2, after post-processing has been performed that filtered out any mutations with L or fewer variant reads. The expected number of subclonal mutations between frequencies f_1 and x is given by

$$\gamma(x) = c(1/f_1 - 1/x) \tag{2.29}$$

where c is a constant that will vary depending on the patient and sample. It can be fit on the sequencing data by noting

$$\gamma_{obs} = \gamma(f_2) - \gamma(2L/(Rp)) \tag{2.30}$$

$$= c(Rp/(2L) - 1/f_2). \tag{2.31}$$

Therefore, c can be estimated from the sequencing data as

$$c = \frac{\gamma_{obs}}{Rp/(2L) - 1/f_2}. \tag{2.32}$$

Then, we can estimate γ as

$$\gamma = \gamma_{obs}\left(\frac{\frac{1}{f_1} - \frac{1}{f_2}}{\frac{Rp}{2L} - \frac{1}{f_2}}\right). \tag{2.33}$$

Number of passengers reaching fixation after t_1

We estimate the number of passengers that occurred after t_1 and reached fixation in the type-1 population in order to adjust the m_{obs} mutation count. From [58], when mutations occur at cell division, the expected number of clonal passengers is $\delta\bar{u}/(1-\delta)$. $\bar{u}$ is the probability that a daughter cell gains a new passenger mutation at cell division, so the mutation rate is $u = \bar{u}b_1$. For the type-1 population, $\delta = d_1/b_1 < 1$. When mutations accrue over time, and not only at divisions, the expected number of clonal passengers is thus

$$\bar{u}/(1-\delta) = u/r_1. \tag{2.34}$$

Similarly, for a clone i, the expected number of passengers that occur after time t_i and reach fixation is

$$u/r_i \tag{2.35}$$

where $r_i = b_i - d_i > 0$.

Simulation of tumor evolution and sequencing data

To assess the accuracy of the analytic results, we perform a continuous time Monte Carlo simulation to model tumor evolution and collection of sequencing data with an implementation of the Gillespie algorithm [78]. Simulations are written in C/C++.

The type-j population has division rate b_j, death rate d_j, and mutation rate u. Mutations can occur at any point of the cell cycle, not just during division. z_n is the number of type-j cells with passenger n as their most recent passenger mutation. The type-0 population is initiated with a single cell at time 0, and the type-j population for $k \geq j > 0$ is initiated with a single cell at time t_j. Let a be the vector recording the ancestor of new mutations. Element a_i is the subclonal ancestor of the ith passenger mutation. For each $j \in 0, 1, \ldots, k$, repeat 1-4 while time is less than $t_k + t + \Delta$.

1) Set $\Gamma = N_j(b_j + d_j + u)$. Time increment to next event time is randomly sampled from Exp[Γ].

- If $j < k$, if time is greater than or equal to t_{j+1} for first time, randomly select type-j subclone i to have driver mutation, remove one cell from type-j population count, and set $N_{j+1} = 1$. Record the true value of m_{j+1}, the number of passenger mutations present in the founder type-$(j+1)$ cell.

2) Randomly select cell, with most recent passenger mutation i, to have the event.

3) Determine which type of event and update population and mutation frequencies. Sample Y from Uniform$[0, \Gamma]$ to determine event type:

   i) $y \in (0, b_j) \rightarrow$ birth. $N_j \mathrel{+}= 1$, $z_i \mathrel{+}= 1$.

   ii) $y \in (b_j, b_j + d_j) \rightarrow$ death. $N_j \mathrel{-}= 1$, $z_i \mathrel{-}= 1$.

   iii) $y \in (b_j + d_j, b_j + d_j + u) \rightarrow$ passenger mutation. Suppose it's the pth passenger, $z_i \mathrel{-}= 1$, $z_p = 1$. Update ancestor: $a_p = i$.

4) For $j = 0$, if time is less than t_1 and population goes extinct, restart simulation. For $j \geq 1$, if time is greater than t_j and population goes extinct, restart type-j simulation at t_j with a single cell.

5) Reindex to remove extinct passenger mutations, and traverse back through ancestor vector a to sum total number of cells with each passenger.

Measurements are taken at bulk sequencing times $t_k + t$ and $t_k + t + \Delta$. If time is greater than or equal to $t_k + t$, we measure $M_1 = \sum_{j=0}^{k} N_j$ and CCF of clone j as N_j/M_1. Then an additional bulk sequencing measurement is taken at the final time $t_k + t + \Delta$, where we measure $M_2 = \sum_{j=0}^{k} N_j$ and the CCF of clone j as N_j/M_2. At $t_k + t + \Delta$, we measure γ, the number of mutations with frequency between f_1 and f_2.

To measure $m_{j,obs}$, the observed number of passengers in the founder type-j cell, we count the number of passengers present in all type-j cells. We also save the true value of m_j.

For when we calculate a percent error of corrected and observed γ values in Fig. 2.7D and Fig. 2.8B, we simulate sequencing data by sampling from the mutation frequencies obtained in the Monte Carlo simulation, outlined above, using the approach of [47]. Define average sequencing coverage as R, number of cells at time of sequencing as M, Z_i as the number of cells with mutation i, R_i as read coverage, and χ_i as the true mutation frequency from Monte Carlo simulation. For each saved Monte Carlo simulation run, repeat the following 100 times:

1) Generate read coverage: $R_i \sim \text{Binomial}[M, R/M]$.

2) Generate number of cells carrying mutation i: $Z_i \sim \text{Binomial}[R_i, \chi_i/2]$.

3) Post-processing. If there are $L = 2$ or fewer variant reads, discard mutation.

4) Measure γ_{obs}, the observed number of subclonal mutations between frequencies f_1 and f_2: $\gamma_{obs} = \sum_i I(f_1 \leq 2Z_i/R \leq f_2, Z_i > L)$.

5) Calculate the truth, γ_{true}, from the true mutation frequencies: $\gamma_{true} = \sum_i I(f_1 \leq \chi_i \leq f_2)$.

Parameter values for simulations

For the simulations we consider three parameter sets corresponding to three modes of tumor evolution: a fast-growing tumor, slow-growing tumor, and tumor with no cell death, each with multiple mutation rates. We simulate three clonal structures: single driver subclone, two nested driver subclones, and two sibling driver subclones. All parameter values are listed in Table 2.3. Mutation rate parameter values lie within observed genome wide point mutation rates per day [79]. For simulation of parental clone and subclone, the fast-growing tumor dynamics are from [46]. The slower growing tumor parameter regime has a reduced net growth of $r = 0.025$, compared to the fast-growing tumor's net growth rate of $r = 0.07$.

Subclonal reconstruction of CLL sequencing data

The sequencing data from all CLLs analyzed is from Ref. [2], Supplementary Tables 2-4. As in that publication, we use PhylogicNDT [3] to perform subclonal reconstruction. We run the Cluster and BuildTree modules of PhylogicNDT on the longitudinal mutation data from Supplementary Table 3 of [2], using mutation alternate/reference counts, copy number, and tumor purity at all pre-treatment time points. Then for each patient, PhylogicNDT outputs a clonal reconstruction, which includes a phylogenetic tree of the subclones and posterior distributions of subclone CCFs. Additionally, it clusters mutations and assigns them to clones. We directly use subclone assignments and posteriors generated from PhylogicNDT. In our analysis we focus on estimating timing and growth rates of macroscopic subclones whose CCFs are greater than 20% for at least one pre-treatment time point.

Accounting for uncertainties in subclone frequencies and growth rates

Our estimates for parameters of cancer evolution require as input the information on the number of subclonal populations in the tumor, their CCFs and their phylogenetic relationships. In order to obtain this information, we use PhylogicNDT [3], which performs subclonal reconstruction of longitudinal cancer sequencing data. The uncertainty in subclone CCFs reported by PhylogicNDT affects our estimates for subclone growth rates, which in turn affect the estimates of mutation rate and time t between driver(s) and diagnosis. We account for this uncertainty by drawing from the CCF posterior distributions that are output by PhylogicNDT. Using these sampled CCF values, we then calculate growth rates, mutation rate u, and time t between driver(s) and diagnosis, thereby generating confidence intervals for these parameters due to CCF uncertainty.

To estimate subclonal growth rates, we fit an exponential growth curve to subclonal sizes measured at two or more time points. This regression yields fitted values for each clone's growth rate and age. To account for uncertainty in the curve fit (in the case of more than two longitudinal samples), we sample the growth rates and age of clone from a bivariate

normal distribution with mean equal to the fitted parameters and variance equal to the covariance matrix of the fitted parameters. When the estimated growth rate during the period of observation—including negative growth rates—is smaller than the minimal possible growth rate necessary to reach the observed clone size, we use the minimal growth rate for calculating mutation rate, time of the driver(s), time between driver(s) and diagnosis, and time of detectability.

Estimating time of cancer detectability

The time a cancer is detectable is the time at which the cancer exceeds the minimum observable size. For the CLL data, we estimate the time that the patients first exhibited an abnormally high WBC count, or leukocytosis, characterized by a WBC count of 11,500/µL [80], or approximately 5.75 x 10^{10} total WBCs, assuming a total blood volume of 5 L. In the previous section, we describe how we fit the growth dynamics for the CLL data and obtain a distribution of the fitted growth parameters. Here, we sample from the distribution of the fitted parameters 10,000 times (using the minimal growth rate in the case of a growth rate too low to give rise to the observed WBC count), and numerically solve for the time at which the total WBC count was equal to 5.75 x 10^{10}. i.e., we numerically find the root with respect to t_i of

$$f(\hat{\theta}_i, t_i) - 5.75 \text{ x } 10^{10} = 0 \tag{2.36}$$

where t_i is the ith estimated time out of 10,000 estimates, $f(\cdot)$ is the exponential function describing the mean cancer growth, and $\hat{\theta}_i$ is the ith random sample from the fitted growth parameters (intercept and growth rate).

Accounting for model uncertainty

The largest source of model uncertainty is the Poisson process for how mutations accumulate, which is used to estimate the time t_1 of the driver mutation. In the fast-growing tumor simulation experiments, the time t_1 had the largest error and variation (Fig 2.2). The estimate

for t_1 depends on the m mutations present in all cells in the driver subclone. The observed m is a single random sample from a Poisson distribution. To account for the uncertainty in t_1 arising from m in the CLLs analyzed, we sample t_1 from the posterior distributions $P(t_1|m)$. This source of model uncertainty due to the Poisson process will be most significant for cancers like CLL with a smaller number of mutations.

The time t between driver mutation and diagnosis is a random variable due to the stochasticity of cancer cell growth, and will naturally have a certain amount of variation. Time between driver event and diagnosis in a branching process follows a Gumbel distribution [50] and will have a constant variance. The mean, however, will increase with the logarithm of the cancer cell counts, which for the CLLs analyzed are $\sim 10^{11}$. The simulations of cancer evolution grow to smaller tumor sizes ($\sim 10^5$) and, as a result, the estimate for t has a significant amount of uncertainty (Fig 2.2). However, for time scales necessary to generate a tumor, the estimate for t will be quite accurate. For commonly observed tumor sizes, the stochastic fluctuations in the time for the cancer to reach that size will be smaller relative to the magnitude of the time. For a cancer with cell count $\sim 10^{11}$, the standard deviation of the time t will be less than 5% of its expected value.

Tumor with two nested driver subclones

Here we consider the case where there are two nested driver subclones (Fig. 2.6A). "Nested" means that all cells carrying the second driver mutation also carry the first. Type-0, or initiated tumor, cells have birth rate b_0, death rate d_0, and net growth rate $r_0 = b_0 - d_0$. Type-1 cells, which only have the first driver, have birth rate b_1, death rate d_1, and net growth rate $r_1 = b_1 - d_1$. Type-2 cells, which carry both drivers, have birth rate b_2, death rate d_2, and net growth rate $r_2 = b_2 - d_2$. The first driver occurred in a type-0 cell at time t_1. The second driver occurred in a type-1 cell at $t_2 = t_1 + t_2'$. The mutation rate u is the same for all subclones.

At times $t_1 + t_2' + t$ and $t_1 + t_2' + t + \Delta$, the tumor is bulk sequenced. The bulk sequencing allows the measurement of the fraction of cells with driver 1 at time $t_1 + t_2' + t$, α_1; the fraction

of cells with driver 2 at $t_1 + t_2' + t$, α_2; fraction of cells with driver 1 at time $t_1 + t_2' + t + \Delta$, β_1; the fraction of cells with driver 2 at $t_1 + t_2' + t + \Delta$, β_2; and the observed number of subclonal passenger mutations between frequencies f_1 and f_2, γ_{obs}. Note that the fraction of the population that is a type-1 cell at the two times is $\alpha_1 - \alpha_2$ and $\beta_1 - \beta_2$. The fraction of type-0 cells at the two bulk sequencing time points are $1 - \alpha_1$ and $1 - \beta_1$. The total number of cells at bulk sequencing time points are M_1 and M_2. We then equate the estimated cell counts to the expected value of the type-i population size X_i, conditioned on survival.

$$\mathbb{E}\Big[X_i\Big(t_1 + t_2' + t\Big)\Big|X_i\Big(t_1 + t_2' + t\Big) > 0] = \begin{cases} \frac{b_0}{r_0} e^{r_0(t_1+t_2'+t)} & i = 0 \\ \frac{b_1}{r_1} e^{r_1(t_2'+t)} & i = 1 \\ \frac{b_2}{r_2} e^{r_2 t} & i = 2 \end{cases} \tag{2.37}$$

$$= \begin{cases} (1-\alpha_1)M_1 & i = 0 \\ (\alpha_1 - \alpha_2)M_1 & i = 1 \\ \alpha_2 M_1 & i = 2 \end{cases} \tag{2.38}$$

$$\mathbb{E}\Big[X_i\Big(t_1 + t_2' + t + \Delta\Big)\Big|X_i\Big(t_1 + t_2' + t + \Delta\Big) > 0\Big] = \begin{cases} \frac{b_0}{r_0} e^{r_0(t_1+t_2'+t+\Delta)} & i = 0 \\ \frac{b_1}{r_1} e^{r_1(t_2'+t+\Delta)} & i = 1 \\ \frac{b_2}{r_2} e^{r_2(t+\Delta)} & i = 2 \end{cases} \tag{2.39}$$

$$= \begin{cases} (1-\beta_1)M_2 & i = 0 \\ (\beta_1 - \beta_2)M_2 & i = 1 \\ \beta_2 M_2 & i = 2 \end{cases} \tag{2.40}$$

Solving the above equations for r_i, we obtain the growth rate estimates:

$$r_0 = \frac{1}{\Delta}\log\Big(\frac{(1-\beta_1)M_2}{(1-\alpha_1)M_1}\Big) \tag{2.41}$$

$$r_1 = \frac{1}{\Delta}\log\Big(\frac{(\beta_1-\beta_2)M_2}{(\alpha_1-\alpha_2)M_1}\Big) \tag{2.42}$$

$$r_2 = \frac{1}{\Delta}\log\Big(\frac{\beta_2 M_2}{\alpha_2 M_1}\Big). \tag{2.43}$$

The expected value of the first time a population of type-2 cells in a branching process reaches the observed size $\alpha_2 M_1$ [50],

$$\mathbb{E}t = \frac{1}{r_2}\log(\frac{\alpha_2 M_1 r_2}{b_2}) - \frac{1}{r_2}\int_0^\infty e^{-z}\log z dz \tag{2.44}$$

$$= \frac{1}{r_2}\log(\frac{\alpha_2 M_1 r_2}{b_2}) + \frac{0.5772}{r_2} \tag{2.45}$$

$$\approx \frac{1}{r_2}\log(\alpha_2 M_1) \tag{2.46}$$

where the approximation in (2.46) is justified as for (2.24). By (2.11),

$$\mathbb{E}\gamma = u\bigg(\frac{1-\beta_1}{r_0} + \frac{\beta_1-\beta_2}{r_1} + \frac{\beta_2}{r_2}\bigg)\bigg(\frac{1}{f_1} - \frac{1}{f_2}\bigg). \tag{2.47}$$

Using the estimates for r_0, r_1, and r_2 from (2.41)-(2.43), and setting (2.47) equal to the value of γ obtained from (2.33) and the second bulk sequencing, u can be estimated:

$$u = \frac{f_1 f_2 \gamma}{(f_2-f_1)(\frac{1-\beta_1}{r_0} + \frac{\beta_1-\beta_2}{r_1} + \frac{\beta_2}{r_2})}. \tag{2.48}$$

When estimating mutation rate for the CLL patients from Ref. [2], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (2.48) is applied for each time point with the respective CCFs and observed γ values for each time point.

Every type-1 cell carries the m_1 passenger mutations that were present in the original type-1 cell when the first driver mutation occurred at t_1. Similarly, every type-2 cell carries the m_2 passengers that were present in the founder type-2 cell when the second driver mutation

occurred at t_2. Note, none of the m_1 mutations are counted towards m_2. Now we consider the likelihood function

$$P(m_1, m_2|t_1, t_2'). \tag{2.49}$$

$$P(m_1, m_2|t_1, t_2') \propto P(m_1|t_1)P(m_2|t_2') \tag{2.50}$$

$$\propto \frac{(ut_1)^{m_1}e^{-ut_1}}{m_1!}\frac{(ut_2')^{m_2}e^{-ut_2'}}{m_2!} \tag{2.51}$$

Now, maximizing the logarithm of (2.51) with respect to t_1 and t_2',

$$t_1 = \frac{m_1}{u} \tag{2.52}$$

$$t_2' = \frac{m_2}{u}. \tag{2.53}$$

The number of passengers present in the founder type-i cell cannot be directly observed, but we can measure $m_{i\,obs}$, the number of passengers present in all type-i cells. An expected u/r_1 passengers occurring after t_1 in type-1 cells and reaching fixation in the type-1 subclone will be incorrectly included in $m_{1\,obs}$, rather than in $m_{2\,obs}$ (see Methods). Similarly, an expected u/r_2 passengers occurring after t_2 in type-2 cells and reaching fixation in the type-2 subclone will be incorrectly included in $m_{2\,obs}$. Thus,

$$m_1 = m_{1\,obs} - u/r_1 \tag{2.54}$$

$$m_2 = m_{2\,obs} - u/r_2 + u/r_1. \tag{2.55}$$

Tumor with two sibling driver subclones

Here we consider a tumor with two "sibling" driver mutations (Fig. 2.6B). Sibling driver mutations are drivers that occur in separate subclones. In this case, cells are either initiated tumor cell (type-0), carry driver 1 (type-1), or carry driver 2 (type-2). No cells contain both drivers. Driver 1 occurred in a type-0 cell at time t_1. Driver 2 occurred in a type-0 cell at t_2. Type-0 cells have birth rate b_0, death rate d_0, and net growth rate $r_0 = b_0 - d_0$. Type-1

cells, which carry driver 1, have birth rate b_1, death rate d_1, and net growth rate $r_1 = b_1 - d_1$. Type-2 cells, which carry driver 2, have birth rate b_2, death rate d_2, and net growth rate $r_2 = b_2 - d_2$. The mutation rate u is the same for all subclones.

Suppose time τ_i elapses between driver mutation i and tumor observation. Bulk sequencing of the tumor is performed at $t_1 + \tau_1$ (or equivalently $t_2 + \tau_2$), and a known Δ later. Sequencing the tumor allows the measurement of the fraction of cells with driver 1 at the first sequencing, α_1; the fraction of cells with driver 2 at the first sequencing, α_2; fraction of cells with driver 1 at the second sequencing, β_1; the fraction of cells with driver 2 at the second sequencing, β_2; and the number of subclonal passenger mutations between frequencies f_1 and f_2, γ. The fraction of type-0 cells at the two bulk sequencing time points are $1 - \alpha_1 - \alpha_2$ and $1 - \beta_1 - \beta_2$. The total number of cells at the two sequencing time points are M_1 and M_2.

We then equate the estimated cell counts to the expected value of the type-i population size X_i, conditioned on survival.

$$\mathbb{E}\Big[X_i\Big(t_i + \tau_i\Big)\Big|X_i\Big(t_i + \tau_i\Big) > 0\Big] = \begin{cases} \frac{b_0}{r_0} e^{r_0(t_1+\tau_1)} & i = 0 \\ \frac{b_i}{r_i} e^{r_i(\tau_i)} & i = 1, 2 \end{cases} \tag{2.56}$$

$$= \begin{cases} (1 - \alpha_1 - \alpha_2) M_1 & i = 0 \\ \alpha_i M_1 & i = 1, 2 \end{cases} \tag{2.57}$$

$$\mathbb{E}\Big[X_i\Big(t_i + \tau_i + \Delta\Big)\Big|X_i\Big(t_i + \tau_i + \Delta\Big) > 0\Big] = \begin{cases} \frac{b_i}{r_i} e^{r_i(t_1+\tau_1+\Delta)} & i = 0 \\ \frac{b_i}{r_i} e^{r_i(\tau_i+\Delta)} & i = 1, 2 \end{cases} \tag{2.58}$$

$$= \begin{cases} (1 - \beta_1 - \beta_2) M_2 & i = 0 \\ \beta_i M_2 & i = 1, 2 \end{cases} \tag{2.59}$$

Solving the above equations for r_i, we obtain

$$r_0 = \frac{1}{\Delta} \log\Big(\frac{(1 - \beta_1 - \beta_2) M_2}{(1 - \alpha_1 - \alpha_2) M_1}\Big) \tag{2.60}$$

$$r_i = \frac{1}{\Delta} \log\Big(\frac{\beta_i M_2}{\alpha_i M_1}\Big) \qquad i = 1, 2 \tag{2.61}$$

The expected value of the first time a population of type-i cells in a branching process reaches the observed size $\alpha_i M_1$ is [50]

$$\mathbb{E}\tau_i = \frac{1}{r_i}\log\left(\frac{\alpha_i M_1 r_i}{b_i}\right) - \frac{1}{r_i}\int_0^\infty e^{-z}\log z dz \tag{2.62}$$

$$= \frac{1}{r_i}\log\left(\frac{\alpha_i M_1 r_i}{b_i}\right) + \frac{0.5772}{r_i} \tag{2.63}$$

$$\approx \frac{1}{r_i}\log(\alpha_i M_1) \qquad i = 1, 2 \tag{2.64}$$

where the approximation in (2.64) is justified as for (2.24). By (2.11),

$$\mathbb{E}\gamma = u\left(\frac{1-\beta_1-\beta_2}{r_0} + \frac{\beta_1}{r_1} + \frac{\beta_2}{r_2}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right) \tag{2.65}$$

Using the estimates for r_0, r_1, and r_2 from (2.60) and (2.61), and setting (2.65) equal to the value of γ obtained from (2.33) and the second bulk sequencing, u can be estimated.

$$u = \frac{f_1 f_2 \gamma}{(f_2 - f_1)(\frac{1-\beta_1-\beta_2}{r_0} + \frac{\beta_1}{r_1} + \frac{\beta_2}{r_2})} \tag{2.66}$$

When estimating mutation rate for the CLL patients from Ref. [2], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (2.66) is applied for each time point with the respective CCFs and observed γ values for each time point.

Every type-1 cell carries the m_1 passenger mutations that were present in the original type-1 cell when the first driver mutation occurred at t_1. Similarly, every type-2 cell carries the m_2 passengers that were present in the founder type-2 cell when the second driver mutation occurred at t_2. We assume that m_1 and m_2 don't contain any shared mutations. In the CLL dataset we use, this is true. We consider the likelihood function $P(m_1, m_2|t_1, t_2)$

$$P(m_1, m_2|t_1, t_2) \propto P(m_1|t_1)P(m_2|t_2) \tag{2.67}$$

$$\propto \frac{(ut_1)^{m_1}e^{-ut_1}}{m_1!}\frac{(ut_2)^{m_2}e^{-ut_2}}{m_2!}. \tag{2.68}$$

Maximizing the logarithm of (2.68) with respect to t_1 and t_2 yields the maximum likelihood

estimates:

$$t_1 = \frac{m_1}{u} \tag{2.69}$$

$$t_2 = \frac{m_2}{u}. \tag{2.70}$$

Using the same approach as in the case of a single driver, we obtain the corrections for the observed number of mutations present in all cells of each subclone:

$$m_1 = m_{1\,obs} - u/r_1 \tag{2.71}$$

$$m_2 = m_{2\,obs} - u/r_2. \tag{2.72}$$

Fully generalized estimates for any phylogeny of k drivers

Here we derive estimates for a completely general tumor phylogeny. Suppose a tumor has k driver mutations. In this general case, define a type-i cell as a cell where its most recent driver mutation was driver i. Note that a type-i cell can have between 0 and $k-1$ other driver mutations. A phylogenetic reconstruction of the k driver mutations is necessary for the completely general case. From this phylogenetic tree, the ancestor of each subclone can be obtained. Define the function $a(i)$ as the ancestor of the type-i population. That is, if all driver mutations contained in the type-i population are ordered, $a(i)$ gives the driver mutation that occurred prior to i. Define t_i as the time between when driver i occurred and when the type-i cells' previous driver mutation occurred. At time of observation, assume the type-i population has κ_i total driver mutations, where $1 \leq \kappa_i \leq k$ for all $1 \leq i \leq k$. Denote the time between the type-i's κ_i, or last, driver mutation and when the tumor is observed as τ_i. This is the time between the founder type-i cell's birth and tumor observation. Then the tumor is first observed and bulk sequenced at $T_1 \equiv (\sum_{j=0}^{\kappa_i - 1} t_{a^j(i)}) + \tau_i$ (equivalently τ_0 for $i = 0$), where we denote a^j as the jth iterate of the function a:

$$a^0(i) \equiv i \tag{2.73}$$

$$a^j(i) \equiv a(a^{j-1}(i)) \qquad \forall j \geq 1. \tag{2.74}$$

The tumor is also bulk sequenced at $T_2 \equiv (\sum_{j=0}^{\kappa_i - 1} t_{a^j(i)}) + \tau_i + \Delta$ (equivalently $\tau_0 + \Delta$ for $i = 0$). These assumptions allow for any subclone phylogeny, including combinations of the previously discussed sibling and nested subclone types.

The bulk sequencing allows the measurement of the fraction of cells with driver i at T_1, α_i; the fraction of cells with driver i at time T_2, β_i; and the number of subclonal passenger mutations between frequencies f_1 and f_2, γ. Again, the total number of cells at measurement times T_1 and T_2 are M_1 and M_2. To write the type-i frequencies, α_i^c and β_i^c, in terms of the driver frequencies, we subtract the fraction of cells descending from type-i cells but gaining additional driver mutation(s) after i, from the fraction of cells containing driver i:

$$\alpha_i^c = \begin{cases} \alpha_i - \sum_{j=1}^{k} \delta_{i,a(j)} \alpha_j & 1 \leq i \leq k \\ 1 - \sum_{j=1}^{k} \alpha_j^c & i = 0 \end{cases} \tag{2.75}$$

$$\beta_i^c = \begin{cases} \beta_i - \sum_{j=1}^{k} \delta_{i,a(j)} \beta_j & 1 \leq i \leq k \\ 1 - \sum_{j=1}^{k} \beta_j^c & i = 0 \end{cases} \tag{2.76}$$

where $\delta_{i,a(j)}$ is the Kronecker delta, defined as

$$\delta_{i,a(j)} = \begin{cases} 0 & \text{if } i \neq a(j) \\ 1 & \text{if } i = a(j) \end{cases} .$$

We equate the estimated cell counts at the first bulk sequencing time point to the expected value of the type-i population size X_i, conditioned on survival.

$$\begin{aligned} \mathbb{E}[X_i(T_1) | X_i(T_1) > 0] &= \frac{b_i}{r_i} e^{r_i \tau_i} \\ &= \alpha_i^c M_1 \end{aligned} \tag{2.77}$$

And similarly, at the second bulk sequencing time point,

$$\mathbb{E}[X_i(T_2) | X_i(T_2) > 0] = \frac{b_i}{r_i} e^{r_i(\tau_i + \Delta)} \tag{2.78}$$

$$= \beta_i^c M_2. \tag{2.79}$$

Solving the above equations for r_i, we obtain

$$r_i = \frac{1}{\Delta} \log\left(\frac{\beta_i^c M_2}{\alpha_i^c M_1}\right) \qquad \forall i = 0, 1, \ldots, k. \tag{2.80}$$

By (2.11)

$$\mathbb{E}\gamma = \left(u \sum_{i=0}^{k} \frac{\beta_i^c}{r_i}\right)\left(\frac{1}{f_1} - \frac{1}{f_2}\right). \tag{2.81}$$

Now, using the growth rate estimates r_i and the subclone sizes, we can estimate each τ_i. The expected value of the first time a population of type-i cells in a branching process reaches the observed size $\alpha_i^c M_1$ is [50]

$$\mathbb{E}\tau_i = \frac{1}{r_i} \log\left(\frac{\alpha_i^c M_1 r_i}{b_i}\right) - \frac{1}{r_i} \int_0^\infty e^{-z} \log z dz \tag{2.82}$$

$$= \frac{1}{r_i} \log\left(\frac{\alpha_i^c M_1 r_i}{b_i}\right) + \frac{0.5772}{r_i} \tag{2.83}$$

$$\approx \frac{1}{r_i} \log(\alpha_i^c M_1) \tag{2.84}$$

where the approximation in (2.84) is justified as for (2.24).

Using the $(k+1)$ r_i estimates from (2.80), and setting (2.81) equal to the value of γ obtained at the second bulk sequencing from (2.33), u can be estimated:

$$u = \frac{f_1 f_2 \gamma}{(f_2 - f_1)(\sum_{i=0}^{k} \frac{\beta_i^c}{r_i})}. \tag{2.85}$$

When estimating mutation rate for the CLL patients from Ref. [2], for which there is bulk sequencing at two or more time points, we average the mutation rate calculated at each of these time points. (2.85) is applied for each time point with the respective CCFs and observed γ values for each time point.

The number of passengers present in the original type i founder cell cannot be directly observed, but we can measure m_i, the number of clonal passengers present in the type i population, only including passengers not present in other clones. We will assume that the m_i don't contain any shared mutations, which is true for the CLL dataset we consider. The

likelihood function $P(m_1, \ldots, m_k | t_1, \ldots, t_k)$ is proportional to

$$\prod_{i=1}^{k} P(m_i|t_i) \propto \prod_{i=1}^{k} \frac{(ut_i)^{m_i} e^{-ut_i}}{m_i!}. \tag{2.86}$$

Then, maximizing the logarithm of (2.86) with respect to $t_1, t_2, \ldots, t_k$,

$$t_i = \frac{m_i}{u} \qquad \forall i = 1, \ldots, k. \tag{2.87}$$

The observed clonal passengers in the founder type-i cell will incorrectly include passengers that reached fixation in the type-i population after driver mutation i occurred, instead of correctly being counted toward the descendant of clone i. As a result, we again correct for the expected number of these passengers, u/r_i. That is,

$$m_i = m_{i,\,obs} - u/r_i + u/r_{a(i)} \qquad \forall i = 1, \ldots, k. \tag{2.88}$$

2.6 Supplementary Materials

Supplementary Tables

Table 2.3: **Parameter values.** Parameter values and number of surviving runs for Monte Carlo simulations. For all simulations $f_1 = 0.01$, $f_2 = 0.20$, $L = 2$. Table (`.xlsx` file) can be downloaded from: `https://doi.org/10.1371/journal.pcbi.1010677.s006`

Supplementary Methods. Unbiasedness of growth rate for Chapter 2.

Denote our estimator for growth rate of the type-i clone $\widehat{r}_i$, with true parameter value r_i. The type-i population has size $X_i(t)$ at time t. Then,

$$
\begin{aligned}
bias(\widehat{r}_i) &= \mathbb{E}(\widehat{r}_i) - r_i && (2.89)\\
&= \frac{1}{\Delta}\mathbb{E}\log\Big(\frac{X_i(t_{dx}+\Delta)}{X_i(t_{dx})}\Big) - r_i && (2.90)\\
&\approx \frac{1}{\Delta}\Big(r_i\Delta + \mathbb{E}\log\ U - \mathbb{E}\log V\Big) - r_i && (2.91)\\
&= \frac{1}{\Delta}(r_i\Delta + (\gamma + \log r_i) - (\gamma + \log r_i)) - r_i && (2.92)\\
&= 0 && (2.93)
\end{aligned}
$$

where U and V are i.i.d. $\mathrm{Exp}(r_i/b_i)$, and t_{dx} is the time of diagnosis.

Chapter 3

ULTRA-DEEP MUTATIONAL LANDSCAPE IN CHRONIC LYMPHOCYTIC LEUKEMIA PATIENTS UNCOVERS CLONAL DYNAMICS OF RESISTANCE TO TARGETED THERAPIES.

This chapter is from a paper which will soon be submitted, entitled "Ultra-deep mutational landscape in chronic lymphocytic leukemia patients uncovers clonal dynamics of resistance to targeted therapies" and on which I am a co-first author. The complete list of authors is: David Woolston[a*], Elena Latorre-Esteves[b*],Nathan Lee[b*], Mazyar Shadman[ab], Xin Ray Tee, Jeanne Fredrickson[b], Brendan F. Kohrn[b], Olga Sala-Torra[a], Chaitra Ujjani[ab c], Ashley Eckel[b], Brian Till[abc], Min Fang[abc], Jerald Radich[abc], Ivana Bozic[b†], Rosa Ana Risques[b†], Cecilia CS Yeung[abc†]

---

[a]Fred Hutchinson Cancer Center

[b]University of Washington

[c]Seattle Cancer Care Alliance
*Co-first authors
†Co-last authors

3.1 Introduction

Chronic lymphocytic leukemia/lymphoma (CLL) is one of the most common leukemias/lymphomas in adult patients with approximately 21,000 new CLL diagnoses reported in 2021 and its global incidence has risen over the past 30 years [81]. Its presentation ranges from indolent to aggressive, with aggressive CLL portending overall survival of 12 months after a Richter's transformation [82, 83]. Therapies for CLL now utilize Bruton's tyrosine kinase inhibitors (BTKi) such as the irreversible covalently binding ibrutinib with or without rituximab [84, 85]. First- and second-line BTK inhibitors irreversibly inhibit BTK activity by covalently binding

the cysteine residue C481 in BTK's ATP-binding site and demonstrated excellent patient outcomes (progression-free survival of 75% after 26 months of treatment) [85–87]. BTK is a key component of different B cell receptor pathways including P13K, MAPK, and NF-κB and regulates proliferation and differentiation of B cells [88,89]. Although 80-90% of CLL patients respond to targeted therapy, nearly half will relapse by 5 years [90] and most (70-90%) will eventually develop resistance and relapse [91]. Strikingly, 85% of the resistance seen in first line BTKi can be attributed to mutations in BTK or phospholipase C, γ 2 (PLCG2) [92, 93].

When patients with CLL who have been treated with BTKi relapse, 85% have been reported to develop resistance mutations either at the BTK drug binding site after 24 to 48 months of treatment [92, 94, 95] or activating mutations in PLCG2 either in isolation or in combination with mutations in BTK12. These mutations allow for BTK-independent activation pathways, thus enabling tumor cells to be less reliant on BTK [93,96]. Development of resistance is common under treatment with first and second generation BTKi; ibrutinib as the first in class reports 11-38% acquired primary resistance in CLL patients [97–99] whereas acalabrutinib reports 15% acquired primary resistance [100]. Resistance mutations altering the binding site on BTK or its immediate downstream effector PLCG2 have been discovered in a significant portion of cases demonstrating progression on BTKi [92, 98–100]. Reversible BTK inhibitors were developed to overcome the toxicity profiles of the earlier BTKi [101].

Pirtobrutinib (LOXO-305), a reversible selective BTK inhibitor which inhibits Y223 autophosphorylation of active BTK mutants at position 481 that cause resistance, has shown efficacy in studies in treating both ibrutinib naive and ibrutinib resistant CLL patients [101–103]. Early phase 1 and phase 2 studies reported pirtobrutinib as safe and effective for CLL treatment [104]. The success of BTK inhibitors and the availability of multiple second-, third-, and fourth-line therapies [85, 105] also changed the paradigm of disease monitoring and data now show that demonstration of measurable residual disease (MRD) is an important prognostic indicator in CLL [106,107]. Discovery of ultra-low levels of resistance mutations and ability to track the growth patterns of such clones could be an essential element of CLL management and care during treatment with BTK inhibitors. Laboratory

methods for detecting MRD in CLL are increasingly important and a routine part in guiding patient management decisions [108]. Demonstration of undetectable MRD at the end of treatment serves as an independent indicator of prognosis correlating with favorable outcomes for patients [109]. Most common laboratory methods used to date for MRD detection in CLL have been flow cytometry, although some studies have implemented allele specific oligonucleotide PCR [109].

Duplex sequencing is one of the most accurate sequencing methods to date and as such has been used for MRD detection in Ph+ALL [110], as well as for detection of resistance mutations in acute and chronic myeloid leukemias and acute lymphoblastic leukemia [110–112]. During library preparation, both strands of the targeted DNA region are tagged with a unique and complementary double-stranded nucleotide "barcode", which enables independent error correction in each DNA strand as well as duplex error correction by comparing both strands. This 3-layer correction method provides unprecedented resolution for the identification of mutant variants ($< 1/10,000$) [113–116] which could change the paradigm for the early identification of therapy resistance and MRD in CLL. However, this method has not been used in CLL.

In this study, we follow the complex clinical course of two patients who have developed resistance to CLL therapies and where we were able to study the molecular and cytogenetic clonal evolution of their CLL disease. We employed duplex sequencing to screen for mutations in TP53, BTK, BAX, PLCG2, and BCL2 at ultra-low allelic frequencies in serial marrow and blood samples from these two patients with CLL and investigated the mutational profiles that emerged when their disease progressed through pirtobrutinib therapy.

3.2 Methods

3.2.1 Patients and samples

This study features two patients who were originally consented and enrolled into a separate study for pirtobrutinib treatment of refractory CLL. When patients demonstrated resistance

to treatment with pirtobrutinib, they were consented under a Fred Hutch IRB approved CLL biorepository protocol to have their archival tissues and an additional sample obtained for this study for tumor banking, and resistance mutation profiling. Post-pirtobrutinib specimens were collected, and archival specimens were retrieved from as many timepoints as possible from the following clinical laboratories: flow cytometry, cytogenetic, molecular, and histology laboratories. For the first patient (R001), we analyzed 6 samples (A to F), which included 4 peripheral blood (PB) samples and 2 bone marrow (BM) samples. For the second patient (R002), we analyzed 5 samples (A to E), which included 1 PB sample and 4 BM samples. Archival DNA samples (n=8) required no further extraction. From fresh PB or BM samples (n=3), mononuclear cells were isolated using Ficoll-Paque media and density gradient centrifugation, then suspended in fetal bovine serum with 10% DMSO and cryopreserved in liquid nitrogen until DNA extraction. Cryopreserved cells were thawed, their DNA was immediately extracted using the Gentra Puregene Blood Kit (Qiagen) and DNA was quantified using the Qubit dsDNA HS Assay kit (ThermoFisher Scientific) per manufacture protocol. Clinical archival specimens were extracted according to CLIA/CAP clinical standard operating procedures (SOP)s per the specific originating laboratory.

3.2.2 DNA Duplex Sequencing library preparation

DNA from each sample (500ng) was prepared into libraries for Duplex sequencing according to published protocols [113, 115] and using commercially available kits (TwinStrand Biosciences, Seattle, WA). Library preparation consisted sequentially of sonication, end-repair, A-tailing, ligation to duplex adapters, fragment amplification, hybridization capture with biotinylated probes, and library amplification. The capture panel was designed to target TP53 mutations and other cancer mutations that have been identified in drug resistance and relapse in CLL [92–95]. The panel included 56 probes covering the coding region of TP53 as well as known hotspot areas for BCL2, BAX, BTK, and PLCG2, for a total size of 4843bp. Given the small size of the panel, two rounds of hybridization capture were performed to increase efficiency [117]. Proper library fragment size was confirmed by Agilent 4200 High Sensitivity

TapeStation. Indexed libraries were quantified using the Qubit dsDNA HS Assay kit, diluted, and pooled for sequencing. Libraries were sequenced using 150 PE reads on a NovaSeq Illumina platform on site or HiSeq at Genewiz (South Plainfield, NJ), allocating ∼11 million reads per sample.

3.2.3 Duplex Sequencing analysis

Sequencing reads were analyzed using pipeline v2.1.2 available at `https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline`. First, raw reads were demultiplexed and grouped using the double stranded molecular tag included in the duplex adapters. Reads sharing the same tag were used to produce consensus Single-Strand Consensus Sequence (SSCS) reads. Then, SSCS reads with complementary tags were compared to produce a single, highly accurate Duplex Consensus Sequence (DCS) or "duplex read". Duplex reads were aligned to the human genome reference hg38 (GRCh38), end-trimmed (15bp at 5', 5bp at 3'), and overlap-trimmed. Variants were called using VarDict Java, and output VCF files were converted to MAF files using the Vcf2Maf script (`https://github.com/mskcc/vcf2maf`) with VEP version 104 [118]. Masking was performed for TP53 areas prone to sequencing artifacts.

3.2.4 Mutational analysis

R scripts were used to process MAF files using R version 4.2.1 including the tidyverse library [119]. Variants were discarded if they had depth < 50 reads or were in masked areas. Variants were annotated based on their classification in the MAF file as missense, nonsense, silent, indel, splice, UTR or intronic. Intronic mutations, mutations in UTRs, and mutations in intronic splice regions were considered non-coding. All other mutations were considered coding.

Variant allele frequencies (VAF) were calculated by dividing the number of mutant duplex reads (alternative counts) by the duplex depth at the mutated position. Variants that were present in all the samples from a given patient at VAF> 0.9 (homozygous) or VAF between

0.4 and 0.6 (heterozygous) and had a dbSNP identifier were considered single nucleotide polymorphisms (SNP)s. Variants that had a dbSNP identifier, were present in all samples, and had a difference in VAF across samples (maximum minus minimum VAF) > 0.2, were considered SNP-loss of heterozygosity (SNP-LOH). All variants that were not SNP or SNP-LOH and had depth > 1000 were considered mutations. For each sample, mutation frequency (MF) was calculated independently for coding and non-coding regions as the number of mutant positions divided by the total number of duplex nucleotides sequenced in the corresponding regions.

3.2.5 Calculation of cancer cell fraction (CCF)

To study clonal evolution of CLL under therapy, we converted the VAF of each mutation to its CCF, the fraction of cancer cells containing the mutation. The calculation of the variant CCF incorporates tumor purity of the sample and the ploidy at the genomic location, as explained in the section "Variant CCF of heterozygous mutations in a diploid scenario". In Duplex sequencing, every Duplex read corresponds to an original DNA molecule. While mutations in a single molecule are reliably detected, they are subjected to higher sampling error. Thus, we only focused on mutations detected in two or more molecules in a given sample to increase precision in CCF calculations. In patient R001, we analyzed the four PB samples but not the two BM samples because they were collected on the same day and within days of the third PB sample. In patient R002, the CCF could not be calculated for the sample with 0% tumor purity, resulting in the analysis of 4 samples total: one PB and three BM samples.

3.2.6 Variant CCF of heterozygous mutations in a diploid scenario

Let r be the total number of reads for mutation i occurring in a diploid region of the genome, with v being the number of variant reads. Then $VAF_i = v/r$. The population of cells in the sample will be a mixture of normal cells and cancer cells. The tumor purity of a sample, p, is the fraction of cancer cells in the sample. Thus $r(1-p)$ reads are expected to correspond to

normal cells, and rp reads are expected to come from cancer cells. Furthermore, a somatic mutation in a diploid region is typically expected to be present in only one allele of a gene in question. In other words, the fraction of cancer cells containing mutation i is given by:

$$CCF_i = \frac{2v}{pr} = \frac{2VAF_i}{p} \tag{3.1}$$

3.2.7 SNP-LOH frequency and CCF

We use VAF of heterozygous SNPs to determine the frequency of SNP-LOH affecting a gene of interest. As described above, variants that had a dbSNP identifier, were present in all samples, and had a difference in VAF across samples (maximum minus minimum VAF) > 0.2 were considered SNP-loss of heterozygosity (SNP-LOH). The VAF of these SNPs was used to infer the frequency of cells with SNP-LOH in the sample. To describe the methodology in more detail, we consider the case of two SNPs on the opposite alleles of the same gene, SNP-A and SNP-B. There is a SNP-LOH in an unknown fraction of cells, x, in the sample, resulting in the loss of the allele containing SNPa. Let the measured VAFs of the two SNPs be denoted by VAF_a and VAF_b. Since SNPa is only present in cells without SNP-LOH, it follows:

$$VAF_a = \frac{1-x}{2(1-x)+x} = \frac{1-x}{2-x} \tag{3.2}$$

Note that $VAF_a < 0.5$ in this scenario, indicating that SNPa is present in the allele affected by SNP-LOH. Since SNPb is present in all cells, we have

$$VAF_b = \frac{1}{2-x} \tag{3.3}$$

Similarly, $VAF_b > 0.5$, indicating that SNPb is present in the allele not affected by LOH. The frequency of cells with SNP-LOH in the sample, x, can be calculated from any of the two equations above. To calculate the frequency of cancer cells (CCF) with SNP-LOH, we note that SNP-LOH should only be present in the cancer cell population. Thus, the SNP-LOH CCF is x/p, where p is the sample purity.

To validate the SNP-LOH CCF estimates, we cross referenced with clinical cytogenetic data from a combination of karyotype, FISH, and chromosomal genomic array testing (CGAT) on the same samples used for duplex sequencing or, when this was not available, at timepoints near duplex testing timepoint. Patient R001 had no cytogenetic alterations affecting BAX, PLCG2, and BTK confirming the presence of heterozygous SNPs at a VAF of $\sim 50\%$ in all samples. Patient R001 had two heterozygous SNPs (c.1101-375G>A and c.97-6C>T) that indicate SNP-LOH in TP53, at a CCF ranging from 0.51 to 0.91. This finding is corroborated with cytogenetic data including CGAT that confirms deletion of 17p in this CLL in two timepoints prior to treatment with pirtobrutinib. In patient R002, although other cytogenetic abnormalities were noted, no SNP-LOH was identified in the target gene region of our duplex sequencing assay to alter CCF calculations and thus a copy number of 2 was used.

3.2.8 Variant CCF in the presence of LOH

To calculate the CCF of a variant located in a gene affected by LOH, we first need to determine x, the fraction of cells with LOH in the sample, using the methodology described in the previous section. Let p be the tumor purity of the sample. If mutation j is present in a fraction f of cancer cells with LOH and in a fraction g of cancer cells without LOH, then

$$VAF_j = \frac{fx + g(p - x)}{2 - x} \tag{3.4}$$

$$CCF_j = \frac{fx + g(p - x)}{p} \tag{3.5}$$

Combining the two equations, it follows that

$$CCF_j = \frac{VAF_j(2 - x)}{p} \tag{3.6}$$

3.3 Results

3.3.1 Patients and specimens

Patient R001: A 67-year-old woman who had received a diagnosis of CLL with cytogenetic abnormalities including 11q and 17p rearrangements. She started treatment 2 years after

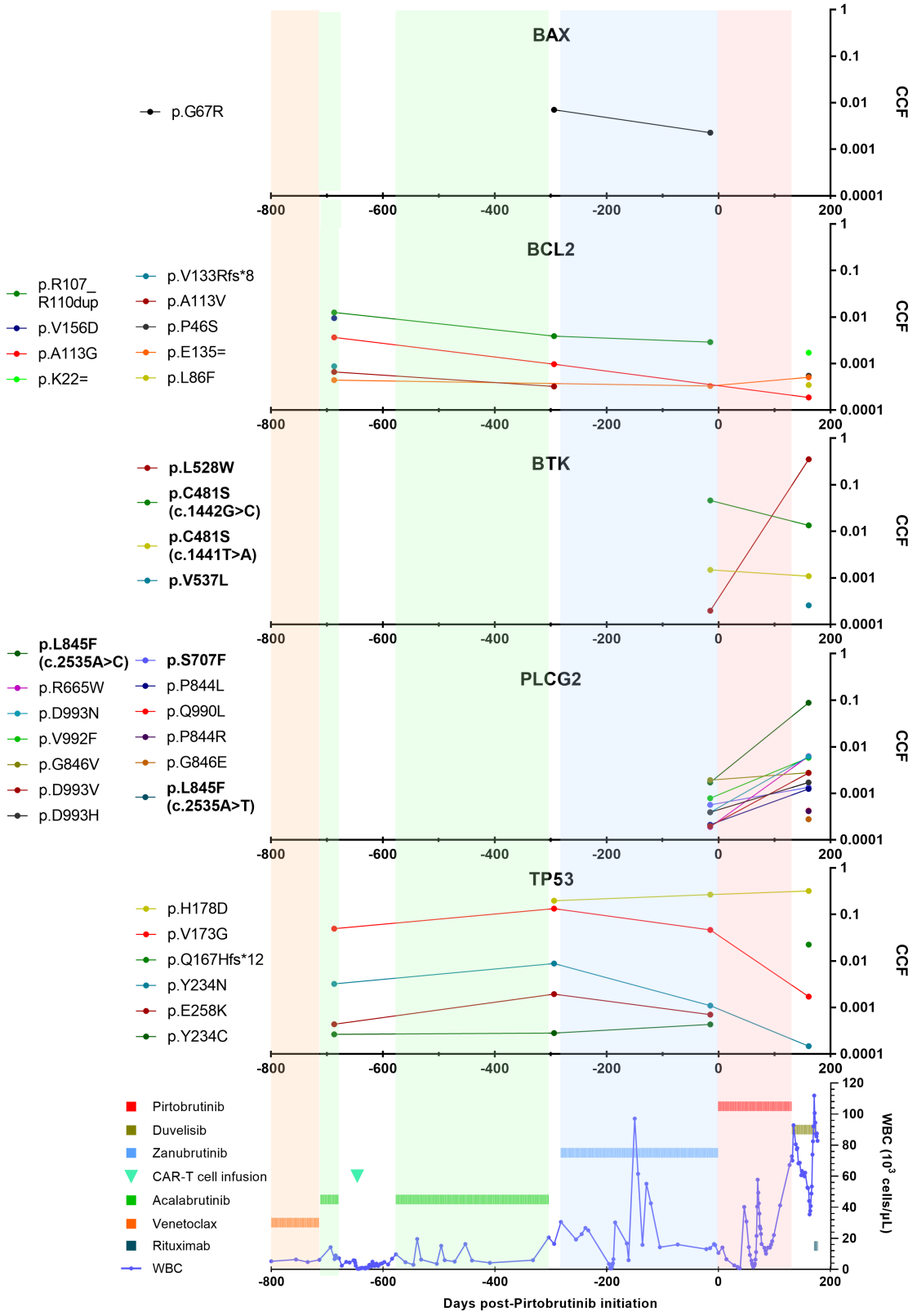

Figure 3.1: **Patient R001 CLL mutation evolution and treatment history.** Colored regions correspond to different treatment regimens. The WBC count (blue dotted line) can be used as an indicator of tumor burden and progression. Key resistance mutations are in bold in the legend.

diagnosis receiving bendamustine and rituximab, then ibrutinib for 7 months (stopped due to multiple soft tissue infections, arthralgia, and myalgia), and then switched to venetoclax for 17 months and stopped because of progressive disease (PD). During treatment with venetoclax, FISH and CGAT performed on a PD sample confirmed deletion 17p with loss of TP53, prompting several trials of combination therapies prior to an investigational CD19-targeted chimeric antigen receptor (CAR)-T cell therapy without response. She received acalabrutinib for 9 months (discontinued because of arthralgia and myalgia), and then switched to zanubrutinib for 9.25 months (discontinued due to PD). Patient was then treated with pirtobrutinib for 4 months until she was found to have PD. Subsequent treatment was duvelisib for 1.5 months without response. Patient died 2 months after stopping pirtobrutinib due to complications of CLL. Six total samples were available for study spanning over 3 years of follow up including both PB and BMs (Fig. 3.1, Fig. 3.4A).

Patient R002: A 73-year-old man with CLL for over 10 years prior to coming to our institution for treatment of relapse disease. This patient was initially treated with fludarabine, cyclophosphamide and rituximab, achieved a remission, and did not require treatment for 7 years. Upon arrival re-workup of his CLL showed several cytogenetic abnormalities such as gain of 1q41qter-, 3q26qter+, 7pterp21-, 8q13qter+, and deletion of 9p21 including copy-neutral LOH (cnLOH) of 9pterp13 but demonstrated no del 17p or alterations to TP53. Upon progression, he was treated with ibrutinib for 3 years until demonstrating PD, when he was switched to venetoclax for 2 years until further PD. He was placed on acalabrutinib when combination ibrutinib/venetoclax trial for 2 months caused atrial fibrillation. He did receive an experimental CD19-targeted CAR-T cell therapy but had a mixed response. He restarted acalabrutinib and had controlled disease for 9 months until his disease progressed again. Then he was treated with pirtobrutinib on a clinical trial for 2.5 months but showed PD. Subsequent treatments included: duvelisib (10 days, stopped due to gastrointestinal intolerance), an experimental bispecific anti CD20/CD3 antibody (PD after 1 dose), high-dose corticosteroids and palliative bendamustine. Patient died 2 months after stopping pirtobrutinib from complications of CLL. A single PB specimen was available for study from

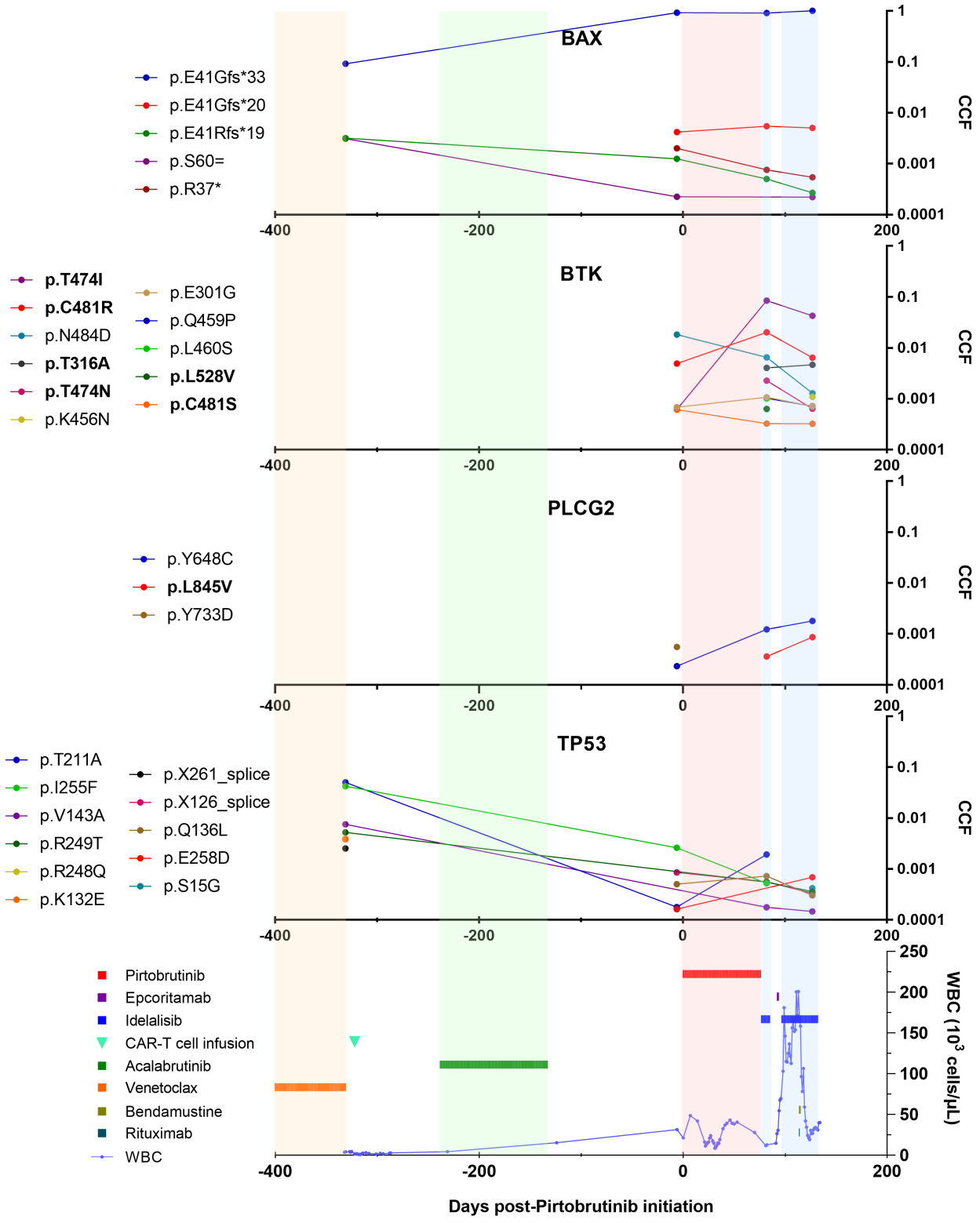

Figure 3.2: **Patient R002 CLL mutation evolution and treatment history.** Colored regions correspond to different treatment regimens. The WBC count (blue dotted line) can be used as an indicator of tumor burden and progression. Key resistance mutations are in bold in the legend.

this patient after demonstrating resistance to LOXO305, approximately 2 weeks before his death. 4 BM samples were available: 2 pre-pirtobrutinib treatment and 2 post-pirtobrutinib treatment (Fig. 3.2, Fig. 3.4A).

3.3.2 Targeted CLL drug resistance mutation testing by duplex sequencing

A total of 11 samples (6 from patient R001 and 5 from patient R002) were duplex sequenced with a panel including the most common CLL resistance genes (BAX, BCL2, BTK, PLCG2, and TP53). The average depth of sequencing across samples was 9,708x (min 7,812x, max 11,170x) (Fig. 3.4A). Both patients showed the highest number of mutations at the latest datapoint, with also corresponded to the highest percentage of disease (95%). In the latest sample from patient R001, tumor burden increased slightly from 82.7% to 95%, but coding mutation frequency (MF) nearly doubled from $1.2\text{x}10^6$ to $2.0\text{x}10^6$. In patient R002, after sequencing data was collected, electronic medical record (EMR) review revealed $< 1\%$ disease in sample R002-B (collected 294 days before initiating pirtobrutinib) by ClonoSeq. This indicates a mutational background in the absence of disease in this sample. Samples R002-D and R002-E, which were collected 82 and 127 days post-pirtobrutinib, more than doubled the MF of the sample without detectable disease. (Fig. 3.4B).

3.3.3 Using CCF to Study Clonal Dynamics During the Course of Disease

To determine clonal dynamics during the course of disease, for all mutations with 2 or more duplex reads in at least one of the samples analyzed, we transformed the VAF into CCF, taking into consideration the percentage of disease in the sample and the estimated LOH by SNP duplex sequencing data and cytogenetics (see Methods). The CCF values are plotted in Figure 3.1 (patient R001) and Figure 3.2 (patient R002).

For patient R001, CCF analysis revealed multiple mutations exhibiting putative resistance to pirtobrutinib, as noted by their significant increase in CCF in the resistant CLL (Fig. 3.1). Strikingly, even though pirtobrutinib treatment was able to suppress the pre-existing BTK C481S mutation, which was present at 5% CCF pre-pirtobrutinib and decreased to 1%

CCF post treatment, another BTK mutation, L528W, increased in CCF by three orders of magnitude following pirtobrutinib, from 0.0002 to 0.35. The BTK L528W mutation, being present in a third of CLL cells post-treatment, is thus potentially a significant driver of pirtobrutinib resistance in this patient. We detected this mutation at a very low VAF in a pre-pirtobrutinib sample using ultra-sensitive Duplex sequencing, demonstrating that it was pre-existing and did not arise de novo during pirtobrutinib treatment.

We also detected multiple PLCG2 mutations whose CCF increased significantly following pirtobrutinib. Most notably, the L845F mutation increased in CCF from 0.002 pre-pirtobrutinib to 0.09 post-pirtobrutinib. The majority of PLCG2 mutations detected in the resistant CLL were also detected at lower frequencies at the pre-treatment sample, indicating that they are pre-existing. We also found that most TP53 mutations decreased in frequency following pirtobrutinib. One notable exception is the H178D mutation, which was first detected at 0.2 CCF about a year prior to the start of pirtobrutinib treatment and remained fairly steady at a CCF of 0.32 post-treatment. The TP53 mutations occurred on the background of TP53 LOH/deletion, which was present at 80-100% CCF at all sequenced samples.

In both patients the first sequencing sample is close to cessation of BCL-2 inhibitor venetoclax. At this point, both patients had several TP53 mutations, and no mutations in either BTK or PLCG2. Patient R001 also had multiple BCL-2 mutations, whereas Patient R002 had multiple BAX mutations.

The second sequencing timepoint for both patients is collected after treatment with BTK inhibitor acalabrutinib. In Patient R002, multiple BTK mutations appeared following acalabrutinib, including known resistance mutations p.C481R, p.T474I and p.C481S, as well as two PLCG2 mutations. In parallel, TP53 mutant clones decreased in frequency.

In contrast, in Patient R001 no new mutations appeared following acalabrutinib, however multiple mutations in both BTK and PLCG2 appeared following subsequent treatment with BTK-inhibitor zanubrutinib, including known BTK resistance mutations p.L528W and p.C481S.

Pirtobrutinib was able to suppress p.C481S mutation in both patients, but other resistance mutations increased dramatically during pirtobrutinib treatment. Most notably, p.L528W in R001 increased in frequency more than 1000-fold (from 0.02% to 35%), and p.T474I in R002 increased in frequency from 0.06% to 8.5% (more than 100-fold). Mutations in PLCG2 also typically increased in frequency following pirtobrutinib treatment in both patients. In addition, for both patients, new BTK and PLCG2 mutations appeared after pirtobrutinib treatment. In patient 1, pirtobrutinib treatment suppressed minor clones with mutations in BAX, BCL2 and TP53 but enabled the expansion of the TP53 mutant clone p.H178D, which had appeared after acalabrutinib treatment. In patient 2, TP53 mutant clones were suppressed with acalabrutinib treatment and remained suppressed during pirtobrutinib treatment.

3.3.4 Biological and technical reproducibility of Duplex assay

Samples R001-D and R001-E were two technical replicates of the same BM sample collected 6 days prior to pirtobrutinib initiation. Despite sample R001-D being sequenced at slightly lower mean Duplex Depth than sample R001-E (7812x vs 10236x, Fig. 3.4A), both samples showed comparable VAF, which closely resembled those identified in the matching blood sample. All mutations with 3 or more reads per replicate were also detected in the second replicate with highly correlated VAFs (Spearman's rank correlation rho=0.89, p<2.2e-16, Fig. 3.5). Mutations identified in one or two reads are expected to be missed in replicate samples given their low abundance, random chance of selection, and lower depth in one of the replicates.

For a subset of samples, NGS data was available and was compared with duplex sequencing data (Fig. 3.3). NGS only identified one BTK mutation (p.C481S, c.1442G>C) and one TP53 mutation (p.H178D, c.532C>G) in Patient R001 and one BTK mutation (p.T474I, c.1421C>T) in Patient R002. All the mutations identified by NGS were also identified in the same samples by duplex-seq at a similar VAF. Importantly, however, for patient R002, Duplex sequencing identified the BTK p.T474I, C.1421C>T resistance mutation in an earlier sample that was missed by NGS (Sample R002-C). This resistance mutation clonally expanded after

pirtobrutinib treatment and at that point (sample R002-D) both methods could detect it. The fact that Duplex sequencing identified this preexisting resistance mutation at very low VAF (0.03%) prior to the treatment indicates the high sensitivity of the assay and its potential clinical utility to rule out resistance mutations prior to treatment.

PATIENT R001

gene	mutation	method VAF	R001-A	R001-B	R001-C	R001-D	R001-E	R001-F
BTK	p.C481S, c.1442G>C	NGS VAF	NP	NP	NP	10.0%	10.0%	NP
		Duplex VAF	0%	0%	1.6%	10.6%	10.5%	0.6%
TP53	p.H178D, c.532C>G	NGS VAF	NP	NP	NP	18.0%	18.0%	NP
		Duplex VAF	0%	7.7%	13.0%	17.0%	17.6%	27.9%

PATIENT R002

gene	mutation	method VAF	R002-A	R002-B	R002-C	R002-D	R002-E
BTK	p.T474I, c.1421C>T	NGS VAF	NP	NP	0%	6%	NP
		Duplex VAF	0%	0%	0.03%	3.9%	2.1%

NP = Not performed (clinical sequencing)

Figure 3.3: **Comparison of Duplex and NGS VAFs.** Variants detected with standard NGS are detected at similar frequencies with Duplex sequencing.

3.4 Discussion

We report mutation analysis of targeted DNA Duplex sequencing on serial PB and BM samples from two patients whose CLL demonstrated clinical resistance to pirtobrutinib. Our data highlighted the resistance mutations that complicate pirtobrutinib therapy and, more importantly, demonstrate the temporal alterations to the CCF and how the resistance mutations evolved under therapeutic pressure from different targeted therapies. We tracked both previously published resistance mutations seen in BTK, BAX, BCL2, PLCG2 [95, 100,

120, 121] in addition to known mutations in the coding regions as well as mutation in the non-coding regions. Our first patient has a more aggressive CLL disease with primary disease demonstrating 17p and 11q deletion. This patient's disease was treated within 2 years of initial diagnosis with different targeted agents including BTK inhibitors and CAR-T cells but progressed prior to enrollment into the pirtobrutinib clinical trial. Despite treatment with pirtobrutinib this patient had a BTK C481R resistance mutation which progressed from day 0 through 421 with a trend of increasing VAF/CCF. In contrast, our second patient had a long-standing history of CLL (over a decade) prior to starting a series of targeted therapies to which the CLL developed resistance, including BTK and BCL2 inhibitors, CAR-T cell therapy, and a novel bispecific antibody, before being enrolled into a clinical trial for treatment with pirtobrutinib. For both patients, we retrospectively obtained all available samples from our clinical laboratories and performed duplex sequencing for a panel of CLL resistance mutations along with TP53 mutations to track clonal trends over the course of their disease under treatment with different targeted agents, including pirtobrutinib, to assess how these treatments effect clonal evolution.

Duplex sequencing improved sensitivity for ultra-low levels of mutations which were likely undetectable by standard NGS. Orthogonal confirmation demonstrated in patient R002 sample C for variant BTK p.T474I, C.1421C>T where NGS showed 0% of this variant, and Duplex showed a very low level of mutation at 0.03%. Wang et al. describes the first 9 patients from the phase 1-2 BRUIN study, who showed resistance to pirtobrutinib and the types of mutations they noted by standard targeted NGS with a detection limit for variant allele frequency of $> 1\%$ [122]. In contrast to this study, our work utilizes duplex sequencing where all but one sample achieved over 9,000 duplex depth of coverage (range 7,812x-11,170x), and allows for a much more sensitive assessment of mutations in the samples. Duplex sequencing allows further error correction with Duplex tags that uniquely identify the DNA molecules. Because of this, we could detect ultra-low allelic frequencies of mutations down to a limit of detection of 1 in 10,000 duplexes or a VAF of 0.0001.

Pirtobrutinib, an oral, third generation non-covalent BTK inhibitor was developed to

target both wild type CLL and CLLs with BTK resistance mutations, such as C481 [104]. Initial reports from phase 1/2 trials show great promise, with high tolerability and overall response rates of 86% in patients who stay on the course of therapy [123]. Both of our patients with post-pirtobrutinib relapse demonstrated clonal expansion of resistant mutations and increased mutation in PLCG2 and BTK, with persistent mutations in TP53. Our first patient had additional expansion of BTK L528 variant while their BTK C481 variant clone diminished, along with a LOH in TP53. Although pirtobrutinib has been specifically developed to overcome mutations in the C481 residue, both patients carried persistent low level BTK C481 mutations during and after pirtobrutinib treatment suggesting alternative evasion mechanisms. These BTK C481 mutations coincided with a concomitant increase in PLCG2 mutations, which is observed in both patients with the addition of pirtobrutinib treatment. This observation is similar to those seen in the first 55 patients treated in Wang et al. who described 9 patients with pirtobrutinib-resistant disease, of whom 7 patients acquired mutations in the kinase domain of BTK and 9 developed PLCG2 mutations [122]. In contrast to the observations in the Wang et al. study, our two patients carry a greater variety of low-VAF PLCG2 mutations, with the highest VAF at 4%, while patient R001 showed a high VAF for resistance mutation BTK L528 at 16% at the last time point. More recently, Naeem et al. analyzed *in vitro* ibrutinib resistance models and primary CLL cells from initially pirtobrutinib-responding CLL patients and demonstrated that, at progression, CLL cells showed increasing resistance to pirtobrutinib [124]. The BAX frameshift mutation p.E41Gfs*33 was observed at high frequency in patient R002, and expanded over the course of acalabrutinib and pirtobrutinib treatment. Similar mutations were observed in half of the colorectal adenocarcinomas examined in Rampino et al. [125], and were recently implicated as a venetoclax resistance mechanism in CLL [126, 127] and acute myeloid leukemia [128].

The presence of TP53 mutations has important prognostic implications, even at low allelic burdens (10%). Patients with TP53 VAFs of greater 10% have lower overall survival [129]. Other groups also observed preexistent or emergent resistance mutations persisting at low allelic frequencies prior to relapse or progression [92, 130]. However, these studies are typically

a snapshot of a single timepoint or a comparison of the diagnosis sample to a relapse sample, without serial samples to follow the trends of a growing clone. In our study, most of the TP53 mutations were seen at a relatively low burden, although patient R001 did have one high burden TP53 H178 mutation demonstrating clonal expansion over the serial timepoints. R001 also demonstrated a more aggressive clinical course with significantly shorter survival than patient R002. Others have noted associations with shorter overall survivals and TP53 mutations [129]. We currently postulate that spurious TP53 mutations that are seen at one timepoint but show no increasing VAF (having a minimum of 2 clonal reads and in consecutive serial samples) in subsequent timepoints are not of clinical significance; however, additional patients and more serial studies are needed to understand the implications of temporal trends of TP53 mutations.

3.5 Supplemental Materials

3.5.1 Supplementary Figures

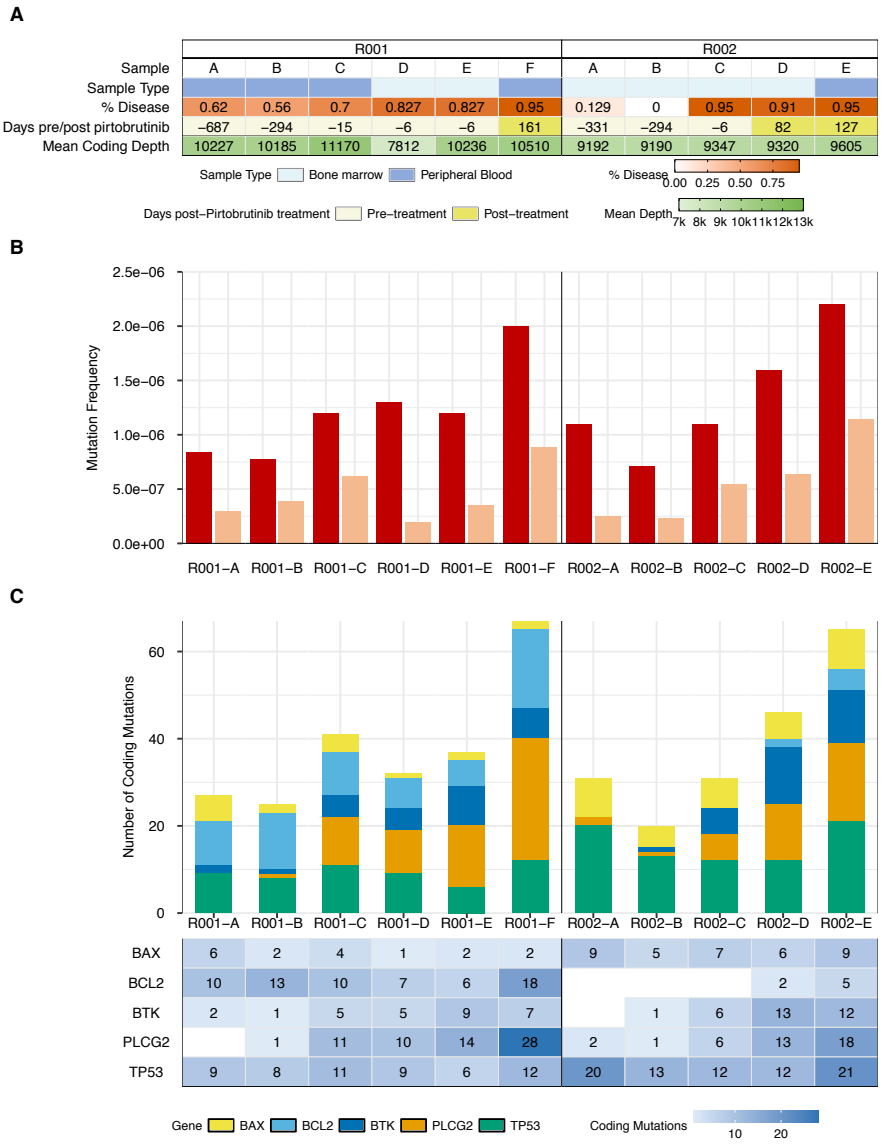

Figure 3.4: **Characterization of mutations identified pre- and post-pirtobrutinib treatment.** (A) Description of samples collected over time for both patients R001 (A,B,C,D,E,F) and R002 (A,B,C,D,E). Sample types include bone marrow and peripheral blood. Values for percent disease, days pre/post pirtobrutinib treatment, and mean coding depth for duplex sequencing are shown for each sample. (B) Mutation Frequency for coding and non-coding mutations for each sample collected over time. (C) Number of coding mutations found in genes associated with drug resistance in CLL for each sample. Genes associated with CLL resistance and covered by duplex sequencing probes include BAX, BCL2, BTK, PLCG2, and TP53. Each sample is represented by a single column; mutated genes are color-coded and represented as a fraction within the column. (D) Number of coding mutations found in genes associated with drug resistance in CLL for each sample collected over time.

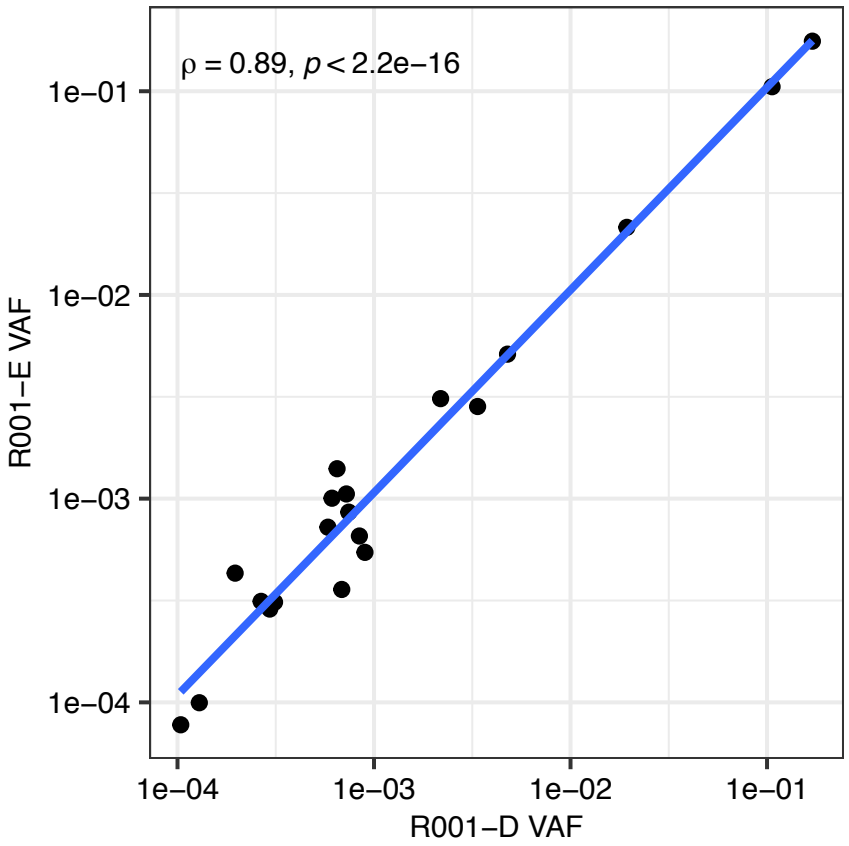

Figure 3.5: **Reproducibility of duplex sequencing data.** Two bone marrow samples collected the same day from patient R001 were independently processed for duplex sequencing. Variant allele frequencies (VAF) were calculated by dividing the number of mutant duplex reads (alternative counts) by the duplex depth at the mutated position. Black dots represent the mutations found in 2 separate samples collected the same day for the same patient (R001). Spearman's rank correlation coefficient is used to measure the degree of association between two variables ($\rho = 0.89$), showing high reproducibility of measurements.

Chapter 4

TRACKING CLONAL EVOLUTION IN A COHORT OF POST-TRANSPLANT ACUTE MYELOID LEUKEMIA PATIENTS

This chapter is from a collaboration with a team of physician-scientists and clinical researchers at Fred Hutch: Elizabeth Krakow[abc], Isaac Jenkins[a], Olga Sala-Torra[a], Lan Beppu[a], Jerald Radich[abc], Ivana Bozic[b], Cecilia CS Yeung[abc]

[a]Fred Hutchinson Cancer Center

[b]University of Washington

[c]Seattle Cancer Care Alliance

4.1 Introduction

4.1.1 Acute myeloid leukemia and myelodysplastic syndromes

Myelodysplastic syndromes (MDS) are a broad class of clonal hematopoietic disorders that are characterized by the bone marrow not producing enough healthy blood cells. Instead, an increased number of immature blood cells called blasts are produced. Clinical features of MDS include cytopenias—that is, low blood cell counts—and abnormal—or dysplastic—bone marrow [131]. Depending on the patient's risk level, treatment options for MDS can range from observation for low-risk patients, to chemotherapy regimens and hematopoietic stem cell transplants for high-risk patients [132]. Traditionally, when patients have more than 20% bone marrow or peripheral blood blasts, they are considered to have progressed to acute myeloid leukemia (AML), an aggressive neoplasm with annual incidence rate of 4.1 per 100,000 and 5 year survival of 13%[1]. Myeloid disorders exist on a continuum of clonal evolution, ranging

[1]The Haematological Malignancy Research Network (HMRN). Factsheets: Acute myeloid leukaemia 2022. www.hmrn.org/factsheets#acute_myeloid_leukaemia, (accessed 02 December 2022).

from clonal haematopoiesis of indeterminate potential (CHIP) to MDS and AML [133]. The boundary between MDS and AML has shifted over the years and remains ambiguous [131]. The 20% threshold is still widely used [131], including for clinical trial eligibility, originating from a 2001 World Health Organization classification [134]. Recently, alternative approaches that rely less on cutoffs for percent blasts and more on likelihood of leukemic progression have been proposed as more meaningful classifications of disease [131, 135, 136]. A recent study of more than 2000 AML and MDS patients performed genomic analyses to stratify patients into risk levels for leukemic progression and found that certain mutations correlated with survival time and disease progression [137]. Genomic and transcriptomic profiling of AML/MDS provides valuable insights into timing and type of treatment offered to patients, even before they are traditionally classified as having AML [133, 138]. Additionally, the somatic mutations driving AML and MDS have been extensively characterized [133, 139].

Allogeneic hematopoietic cell transplantation (HCT) is the only potentially curative treatment option for many AML/MDS patients [132]. Despite the overall decline in the risk of transplant-related mortality in the past several decades, the risk of relapse after transplant has increased [140], with relapse occurring in 30% to 60% of cases [140–143]. This highlights the need for further investigation of the clonal dynamics of post-transplant relapse and how monitoring clonal evolution in tandem with clinical indicators could improve treatment.

Single-cell genomic and transcriptomic profiling of blood and bone marrow provides one avenue for the construction of clonal phylogenies in AML. Studies have described the genetics of single cells in AML and characterized clonal architecture [144], as well as clonal evolution of primary cancer samples [145], allogeneic hematopoietic stem cell transplant recipients [146], leukemic hematopoiesis and transformation [147], and pre-treatment and paired relapse samples [148, 149]. Despite the strength of single-cell methods for recovering clonal lineages and genomic profiles, their widespread clinical use is limited by high cost and technical issues including noisy data and limited sensitivity [150–152].

Bulk sequencing remains the most common approach for performing clonal inference; though as single-cell methods improve in accurately detecting copy number alterations and

single nucleotide variants, the combination of single-cell methods with bulk sequencing will leverage the distinct strengths of each [23, 153, 154]. A wide array of sophisticated computational tools has been developed for the numerous steps involved in inferring clonal relationships from raw bulk sequencing data, including variant calling, clustering variants into clones, and building phylogenetic trees. Making use of such tools, in this work we develop a software pipeline and visualization platform to provide patient-specific information about the clonal evolution of their cancer. While our work is applicable to any cancer with serial samples, we tailor our approach to AML patients receiving HCT, based on the need for such tools in post-HCT settings, where donor chimerism—that is, the percent of hematopoietic cells that are derived from the donor—must be accounted for. We apply our methodology to a cohort of AML patients, from which pre-transplant and post-transplant bone marrow and peripheral blood samples were collected and sequenced. Our work has important implications for monitoring for relapse and informing clinicians' treatment decisions, such as indicating preemptive treatment for high-risk patients.

4.1.2 Insights into clonal evolution from individual variant allele frequencies

When cancer samples are sequenced, each site on a sequencing read can either have the reference allele or a variant allele. After sequencing post-processing and variant calling, one obtains a list of variants detected in the sample, as well as the number of reads the variant was detected on (alternate count) and the number of reads that the reference allele was detected on (reference count). From the read counts, it is possible to calculate the variant allele frequency (VAF), which is the fraction of total reads containing that variant. In the case of a copy number of two, heterozygous single nucleotide polymorphisms (SNPs) will have a VAF of approximately 0.5, and homozygous SNPs will have a VAF of approximately 1.0.

Tumors are highly heterogeneous, in part due to the accumulation of mutations in different genetically distinct subpopulations, or subclones, of the cancer cell population [15]. During bulk sequencing of the cancer, DNA is extracted from the collected cells, which are a mixture of cells from (at that point) unknown subclones. To further complicate matters, the sample

will also contain some normal cells. The percent of the sample consisting of cancer cells is referred to as "tumor purity." The single set of detected variants belongs to different subclones. Examining VAFs of individual variants provides limited insights into clonal evolution, except in the very unlikely scenario that each variant originates from its own unique subclone. Plotting VAFs over time shows how variant frequency changes; however, the dynamics of these variants are not independent, and the frequencies of variants originating from the same subclone will be correlated. The lineage of the clones is also obscured when observing individual VAFs. Analysis of individual variant frequencies does not reveal the order of occurrence of the subclones or their relationships.

All these issues motivate the use case for clustering variants into clones, which in turn allows the phylogenetic reconstruction for the inferred clones. In many cases it would be difficult to build a phylogenetic tree of all detected mutations. As I will discuss below, many tree inference methods struggle as the number of clones increase. As the upper bound on the number of clones is the number of mutations detected, it would not be feasible to use many of these methods to construct the mutational phylogeny. The clonal structure of cancers also has several key clinical implications. Clonal structure and heterogeneity have a fundamental connection to treatment resistance. Suppose a major and minor subclone in a tumor contain different oncogenic mutations. If the major subclone is eliminated via targeted therapy and the minor subclone survives, then it might expand and quickly become the dominant clone; whereas, if the two oncogenic mutations occurred in the same clone, they both would have been eliminated by the targeted therapy. Ideally treatment would be evaluated for its potential to select for certain high-risk pre-existing clones. Significant work has gone into identifying sets of mutations that are more likely to occur in mutually exclusive lineages or the same clone [155–162]. The co-occurrence of certain mutations can have a cooperative intrinsic fitness effect not possible with the individual mutations [163–166]. In the past decade, evidence has emerged that clones are not independent, non-interacting groups, and can in fact directly cooperate and compete in a non-autonomous manner [167–169]. As I discuss

in my review paper[2], clonal interactions can lead to cancer progression, tumor suppression, treatment resistance, and metastasis [168, 170, 171].

4.2 Methods

4.2.1 AML patient cohort

A cohort of 44 AML patients in remission were treated with HCT in a phase I/II trial. Conditioning consisting of clofarabine and low-dose (2 Gy) total body irradiation preceded HCT [172]. 43/44 patients received transplants from HLA-matched related or unrelated donors. 29 patients had serial bone marrow or blood samples collected for at least one pre-transplant and post-transplant timepoint. Variable numbers of post-transplant samples were sequenced for each patient, but for samples that were sequenced, collection generally occurred 28 days, 56 days, 100 days, 6 months, 1 year, and in one year intervals post-transplant, as well as at relapse. The 2-year relapse incidence was 17% and 49% for low and high-risk groups, respectively. Further details about the cohort of patients can be found in Krakow et al. [172].

4.2.2 DNA Sequencing

DNA extraction, sequencing, and sequencing data post-processing were performed by the Radich Lab at the Fred Hutch Cancer Center, by Lan Beppu and Dr. Olga Sala-Torra. Archer Analysis 6.0 was used to analyze single nucleotide variants and insertions/deletions. Variants and specific mutations of interest used for subsequent analysis were supplied to the Archer platform as Targeted Mutation Files so that reference read counts were obtained for all timepoints, even if the variant read wasn't detected at that time point. This is relevant for cases where a variant passes filtering because it is present at $> 5\%$ allele frequency at one timepoint, but not detected at other time points. This avoids the need to impute the reference read count for downstream analysis (specifically the Pairtree algorithm). The single

[2]Lee, Kaveh, Bozic, "Clonal interactions in cancer: integrating quantitative models with experimental and clinical data," *under review.*

nucleotide polymorphisms (SNPs), multiple nucleotide polymorphisms (MNPs), insertions, and deletions generated from the Archer analysis were compiled as Variant Call Format (VCF) files.

4.2.3 Variant filtering

Variants identified from the Archer analysis were then filtered to remove likely sequencing errors. Additionally, variants were selected based on their clinical relevance. The following 8 filtering criteria were used:

1. Variants must map uniquely to their genomic position. In other words, if multiple variants are detected at the same position, then these variants are removed. Such cases are often seen in hard-to-sequence regions, such as homopolymer regions (multiple repeated nucleotides in a row), which frequently produce sequencing errors.

2. At one or more timepoints, variant must meet allele frequency threshold,
$$\frac{\text{alternate observations}}{\text{alternate observations} + \text{reference observations}} \geq 5\%$$

3. Variant must have dbSNP ID and/or COSMIC ID. That is, the variant must be in the Catalogue Of Somatic Mutations In Cancer (COSMIC) and/or the Single Nucleotide Polymorphism Database (dbSNP). COSMIC is a comprehensive database of millions of somatic mutations in cancers, curated from tens of thousands of papers, as well as open-access databases (e.g., TCGA and ICGC) [173]. dbSNP is NCBI's large public database of small variants (e.g., SNPs, short insertions and deletions).

4. The allele fraction (AF) outlier p-value must be less than 1% at one or more timepoints. The allele fraction outlier p-value output by the Archer analysis is the probability that the mutation was due to background noise.

5. Remove any variants with sample strand bias. That is, remove variants with a sample strand bias p-value of ≤ 0.05.

6. Only variants on autosomes were used for subclonal reconstruction.

7. The PTEN mutation (dbSNP ID rs550122918) is present at similar subclonal frequencies pre- and post-transplant in 3/4 of the patients under consideration. It is unlikely that this variant, rare in the general population, would be present at a similar frequency both pre- and post-transplant when post-transplant chimerism is close to 100% in these patients. Thus we consider it a likely sequencing artifact and do not include it for subclonal reconstruction.

8. In the occasional case that the Archer analysis gives both missing reference and alternate read counts at some timepoints for variants that have been observed at other timepoints, we exclude these variants from the analysis.

4.2.4 Clustering variants into clones

There are a wide variety of options for inferring clonal structure from genomic data, including PhyloWGS [174], QuantumClone [175], EXPANDS [176], SciClone [177], FastClone [178], Ccube [179], cloneHD [180], LICHEe [181], Clomial [182], Canopy [183], PASTRI [184], and Sclust [185]. A recent alternative to these largely data-driven and statistical approaches is MOBSTER [186], which combines population genetics theoretical models with machine learning. We use PyClone-VI (`https://github.com/Roth-Lab/pyclone-vi`), a Bayesian statistical method for inferring the clonal structure of cancers [187], due to its accuracy, computational efficiency, and extensive benchmarking and usage. In benchmarking studies of different variant clustering methods (comparing PyClone, PyClone-VI, PhyloWGS, Sclust, Ccube, QuantumClone, and FastClone), in most cases PyClone-VI has an accuracy that is superior or similar (statistically insignificant differences) to the other methods [187, 188]. PyClone-VI has significant advantages in computational efficiency, especially runtime [187].

We run PyClone-VI with the beta-binomial distribution, 10 random restarts, and an initialization of 40 clusters. We input the variants that passed the previous filtering steps,

specifying their alternate and reference counts and copy number. The results presented here make use of several assumptions that will be addressed in future work: (1) assume the samples are diploid and (2) minor and major copy number of the segment overlapping the mutation are each 1, since only mutations on autosomes are being considered.

4.2.5 Phylogenetic reconstruction

With the filtered variants assigned to clones by PyClone-VI, the lineages of each clone and their relationships to one other are yet to be determined. The subclonal relationships can be represented by a phylogenetic tree, where each node represents a clone, and edges connect parental clones to their nested offspring subclones. For a single sample it is rare to be able to infer a single consensus tree, with multiple clones possible for the data [23]. With additional samples the space of possible trees becomes more constrained. SubMARine was developed to better quantify the space of possible clone trees for the given input data, and how these possible trees depend on the error in subclonal frequencies [34]. SubMARine finds the approximate Maximally-Constrained Ancestral Reconstruction (MAR), which represents all pairwise ancestral relationships possible given the input data.

A variety of methods have been developed in the past decade to build phylogenetic trees of the subclonal structure of tumors. The optimal tree inference method to use will depend on the modeling assumptions, the setting that the method was originally intended for, and the statistical and computational tools used in the algorithms. Some methods were optimized for multi-region sequencing (PICTograph [189], Canopy [183]), longitudinal cancer bulk sequencing data (CALDER [55], Canopy [183], Pairtree [190]), deep sequencing (CITUP [191, 192]), and joint inference of many samples from the same patient (PhylogicNDT [3]). Computational approaches used range from integer programming (CITUP [191], CALDER [55], AncesTree [193], parsimony methods [192]), Markov Chain Monte Carlo (MCMC) sampling (PhyloWGS [174], PhyloSub [194], PhylogicNDT [3], Canopy [183], Pairtree [190]), searching for spanning trees of directed graphs (PICTograph [189], LICHEe [181]), and importance sampling (PASTRI [184]).

Many of the early methods for subclonal reconstruction struggle to scale with increasing number of mutations and subclones. We use Pairtree (`https://github.com/morrislab/pairtree`) for the tree inference of clones identified first with PyClone. Pairtree reliably reconstructs complex trees from multiple serial samples. Briefly, Pairtree builds a "pairs tensor" that specifies the probability of all possible pairwise relationships between clones. An initial tree is constructed from this tensor. Trees are sampled using MCMC, and tree edges with low probability in the pairs tensor are rearranged. Acceptance or rejection of the tree is determined by the likelihood that it corresponds to the observed VAFs [190]. We use Pairtree, as it performs reliably and accurately for the metrics most important to the datasets we are studying. In particular, many of the above methods fail to produce trees, especially in cases of more than 10 subclones, whereas Pairtree reliably finds a tree. In their benchmarking of Pairtree, it generally had better accuracy than other state-of-the-art methods for 30 or fewer subclones, in terms of estimated subclonal VAFs and tree structure [190]. Unlike the other methods they tested, Pairtree was the only one that improved with increased number of samples [190]. This is a valuable feature for long-term longitudinal studies of liquid cancers, like the cohort of AML patients we apply our pipeline to. Pairtree also reports the uncertainty of the consensus tree and other possible trees.

The VCF files output by Archer analysis are converted into the simple somatic mutation (SSM) file required by Pairtree. It contains alternate and total read counts and variant read probabilities determined by the copy number. As noted for PyClone, currently we assume diploid loci. Cluster membership output by PyClone is specified by converting to the `.params.json` input file. As mentioned in the "DNA Sequencing" Methods section, with the parameters we use to run the Archer Analysis, we can avoid the need to impute missing reference read counts in most cases. However, in the occasional case that the Archer analysis gives both missing reference and alternate read counts at some timepoints for variants that have been observed at other timepoints, we exclude these variants from the analysis. The files for further analysis are saved using Pairtree's `plottree` module, specifying the `--tree-json` option.

4.2.6 Visualization of clonal evolution

Last, we visualize the clonal evolution using a fishplot. Fishplots show the prevalence of clones over time, like a stacked area plot, but preserve the phylogenetic structure of the clones. Thus, it shows the nested subclones arising from within their parental clones, and sibling (or branching) clones arise in separate regions. There are several existing fishplot visualization packages (timescape[3], EvoFreq[4] [195], fishplot[5] [196], ggmuller[6], pyfish[7]), but none exactly met our needs. Isaac Jenkins created a custom interactive fishplot software tailored to our specific needs: ease of aligning the fishplots with the clinical history and integrating their interactive features, making the time axis properly scaled, and reliable, robust automated visualization of many different fishplots.

4.3 Results

4.3.1 Pipeline to visualize longitudinal clonal evolution

We develop a pipeline to visualize clonal evolution of longitudinal cancer samples. After variants are called, we apply several post-processing steps to remove likely sequencing artifacts or errors (see Methods). Variants showing sequence strand bias, originating from background noise, non-uniquely mapping to the same positions as others, mapping to sex chromosomes, or missing both alternate and reference read counts are removed. Additionally, we focus on variants that are clinically relevant, by only including variants in the Catalogue Of Somatic Mutations In Cancer (COSMIC) and/or the Single Nucleotide Polymorphism Database (dbSNP), as well as requiring that variants are present at allele frequency of more than 5% at one or more timepoints. Variants remaining after filtering were clustered and assigned to subclones using PyClone-VI [187]. The ancestral relationships between the subclones were

[3] `https://github.com/shahcompbio/timescape`

[4] `https://github.com/MathOnco/EvoFreq`

[5] `https://github.com/chrisamiller/fishplot`

[6] `https://CRAN.R-project.org/package=ggmuller`

[7] `https://bitbucket.org/schwarzlab/pyfish`

inferred using Pairtree, which generates the most likely phylogenetic tree (as well as other possible trees) [190]. We then plot this data as a fishplot, which shows the clonal prevalence over time—like a stacked area plot—while portraying the clonal structure of nested clones arising within their parental clones and sibling or branched clones arising in a parallel manner. We align the fishplots with clinical indicators, biomarkers, and treatment regimens. This visualization includes interactive features, such as viewing mutations present in each clone and hovering over features to view tooltips.

4.3.2 The clonal evolution of relapse for four post-transplant AML patients

Here I present the preliminary analysis of four patients from the cohort of 29 AML patients who received hematopoietic stem cell transplants and had serial bone marrow and blood samples sequenced at pre- and post-transplant timepoints [172].

Patient 1

Patient 1 had bone marrow samples collected 21 days pre-transplant and 27 and 56 days post-transplant. The 3 samples had a total of 26 variants that passed filtering, which were clustered into 3 clones, consisting of two branched lineages (Fig. 4.2). The subclonal dynamics reflect the patient's transplant procedure. Clone 1, likely corresponding to variants from the patient, was the main clone at the pre-transplant timepoint 14 days before HCT. The patient received HCT at day 0, and by the first post-transplant timepoint, clone 1 significantly decreased to become a minor subclone. The other "sibling" lineage consisting of clone 2 likely corresponds to the donor variants, as it only emerges after transplant, and quickly becomes the dominant clone. The patient developed grade 2 graft versus host disease by day +27. They eventually relapsed (69 days post-transplant) and required two courses of azacitidine.

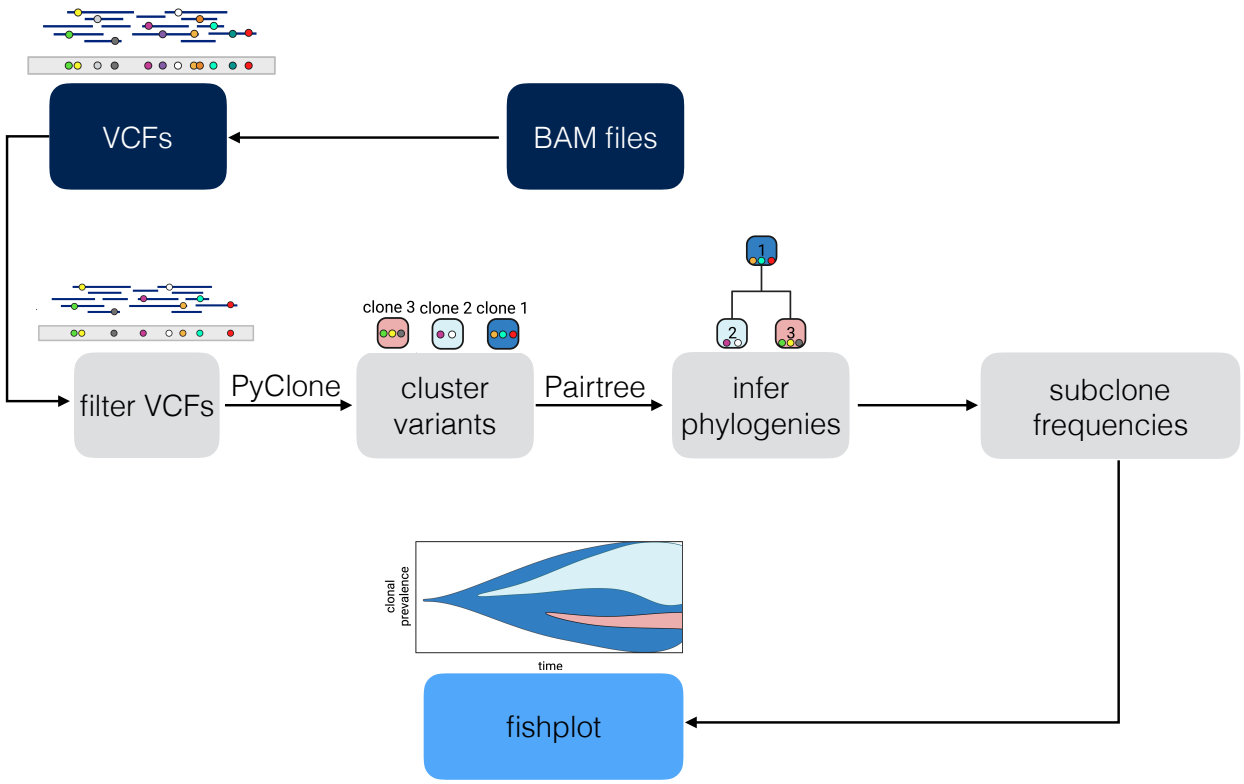

Figure 4.1: **Software pipeline for interactive visualization of clonal evolution in the clinic**. Variants called from Archer analysis are compiled as a variant call format (VCF) file. Variants are filtered based on clinical relevance and sequencing quality. PyClone-VI is used to cluster the filtered variants into clones. Pairtree is used to infer the phylogenetic trees describing the relationships between clones. The inferred subclonal frequencies are used to build fishplots visualizing clonal evolution. Figure created in part with `BioRender.com`.

Patient 9

Patient 9 had one pre-transplant peripheral blood sample collected 14 days pre-transplant, and 3 post-transplant bone marrow samples collected at 28, 61, and 84 days post-transplant and sequenced. After filtering, 10 variants remained. Patient 9 has the same phylogenetic tree as patient 1, and shows a similar change in clonal dynamics post-transplant (Fig. 4.3).

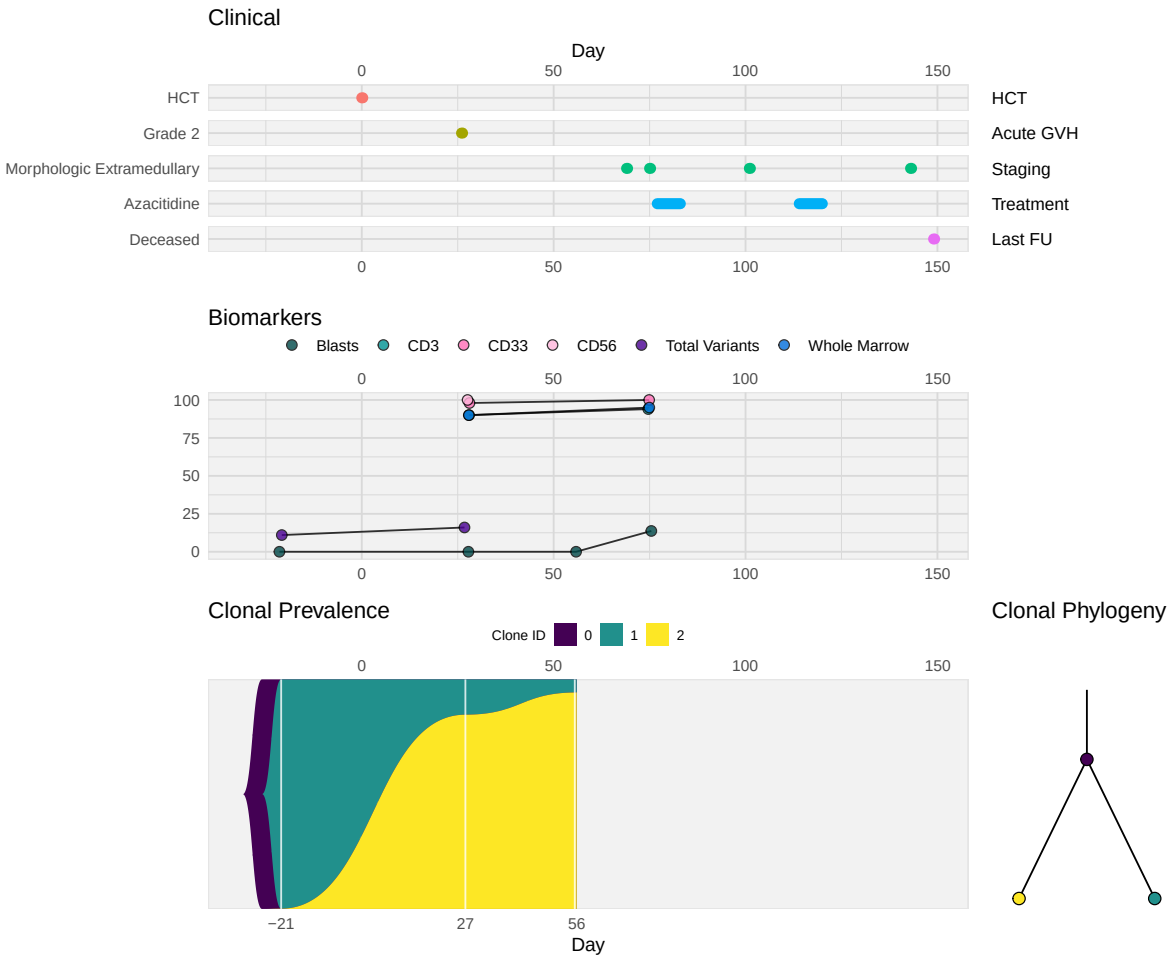

Figure 4.2: **Patient 1 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. The right side of the clinical data provides a general category, and the left label provides the specific category. The top level shows HCT date, second level shows when the patient had grade 2 acute graft-vs-host disease, level 3 shows detection of relapse, level 4 shows treatment received, and the bottom level shows when the patient was deceased. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel.

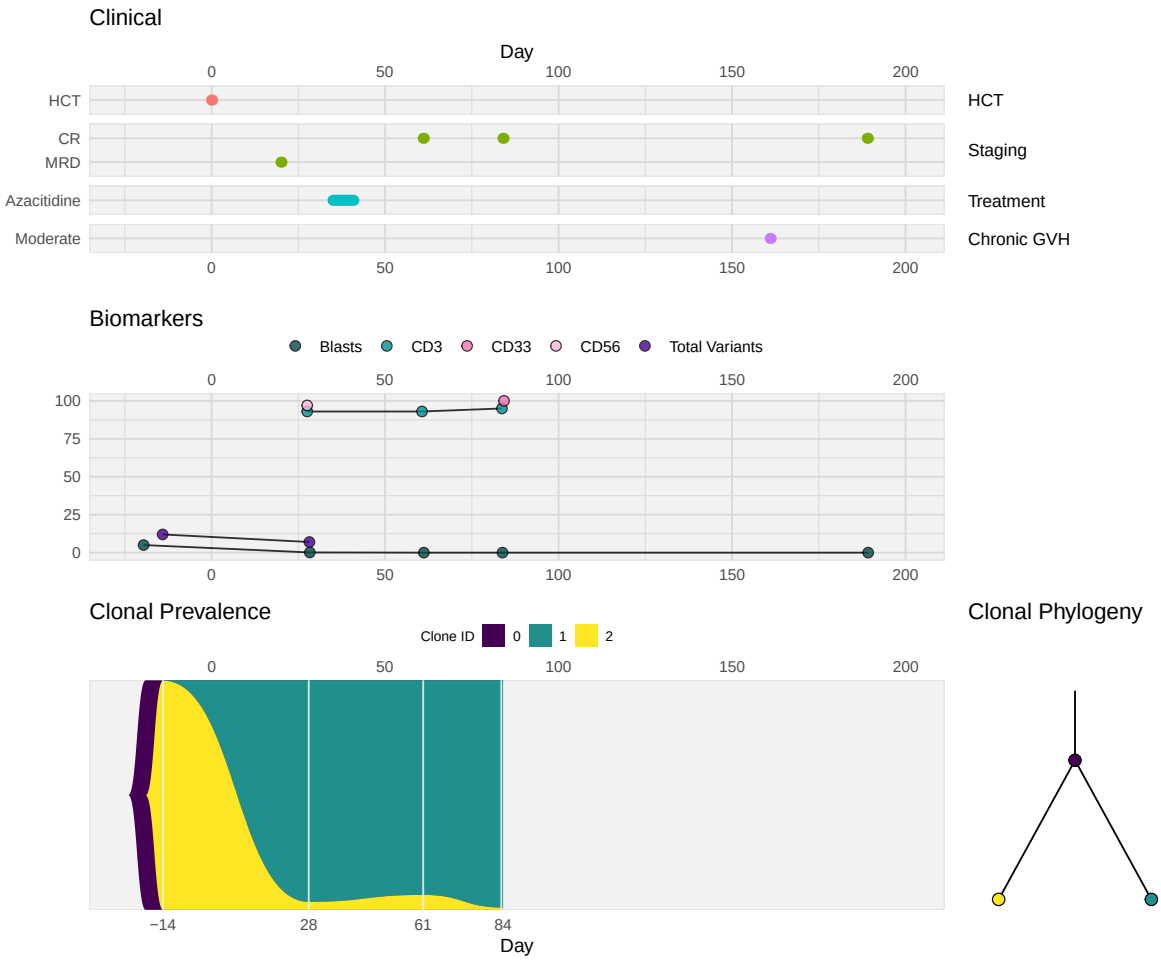

Figure 4.3: **Patient 9 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. The top level of the clinical data show HCT date, second level shows the date of complete remission (CR) and measurable residual disease (MRD), the third level show treatment given, and the bottom level shows when the patient had moderate chronic GvHD. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel.

The dominant clone pre-transplant (clone 2) corresponds to one of the branched lineages and becomes a minor clone post-transplant. The other lineage (clone 1) is observed by the first post-transplant timepoint, at which point it is the dominant subclone. Clone 2 does persist at low frequencies in the post-transplant timepoints. This patient had measurable residual disease MRD relapses at +20 and +504 days, as well as a morphologic relapse at +517. The patient received azacitidine after the first relapse and decitabine-MEC after the last. After both treatments the patient achieved complete remission.

Patient 17

Patient 17 offers a longer-term view of AML evolution, with post-transplant timepoints at +28 and +790 days. There were 36 variants remaining after filtering, which were clustered into 4 clones. Despite the greater number of clones, there are some key similarities with patients 1 and 9. The extra clone arises as an additional nested subclone. The two branching lineages likely correspond to donor and recipient variants. Again we observe the major pre-transplant subclone (clone 3) largely vanish after the HCT. The other lineage contains clones 1 and 2, which are only observed at significant frequencies post-transplant, by which point they have become the primary subclones. Morphological relapse was noted at days +481, +641, +711, +756, and +759.

Patient 22

Patient 22 again clearly demonstrates the binary pattern of clonal evolution, with two contrasting states existing pre- and post-transplant. This patient had 6 samples sequenced: one pre-transplant and 5 post-transplant samples (+29, +61, +82, +124, +181). All samples are bone marrow, except for the last, which is peripheral blood. There are 16 variants remaining after filtering. Clones 3 and 1 correspond to the patient's pre-transplant lineage, and clone 2 likely corresponds to the donor's variant. Clone 2 is the primary clone by day 29 and later.

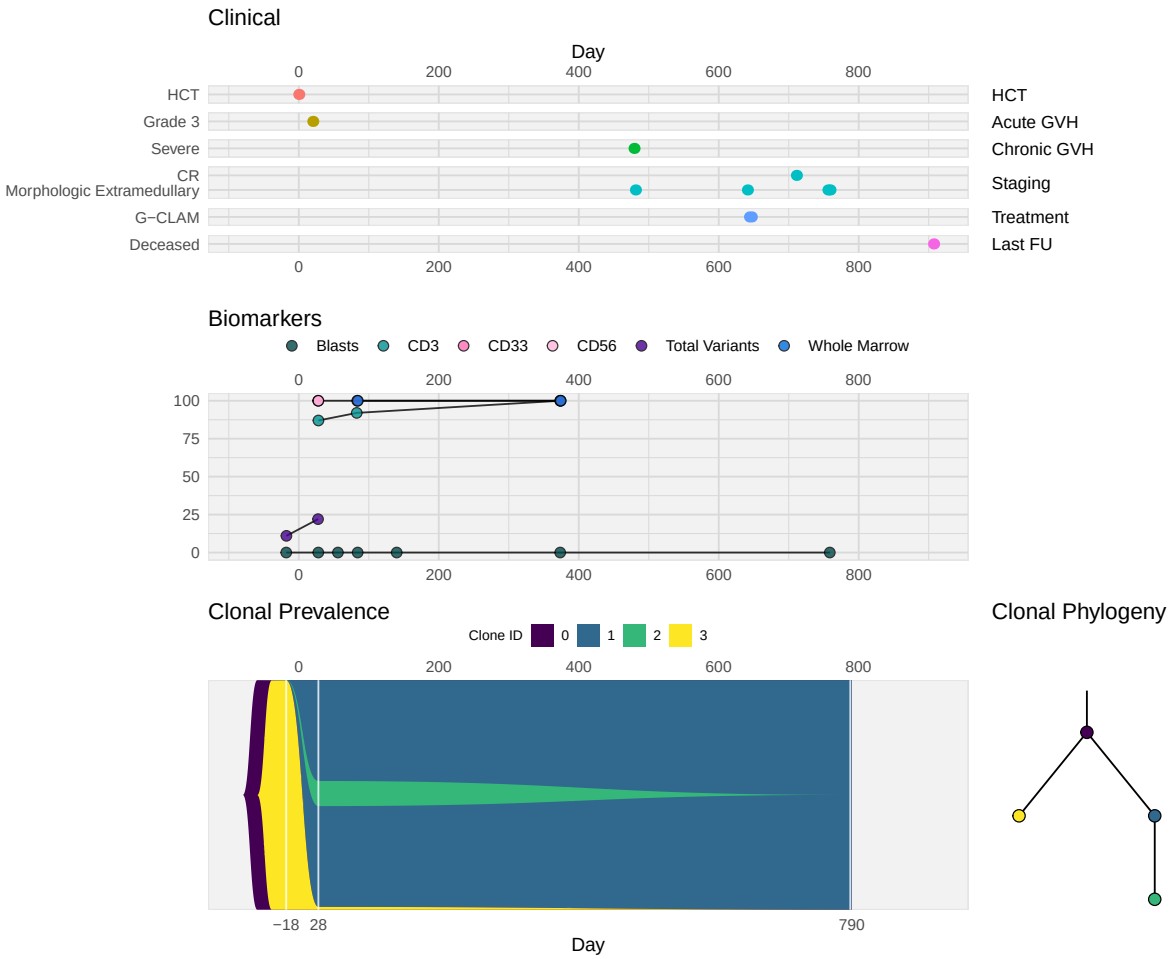

Figure 4.4: **Patient 17 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. In top to bottom order, the clinical panel shows transplant date, GvHD, staging results (CR = complete remission), G-CLAM treatment, and when the patient was deceased. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel.

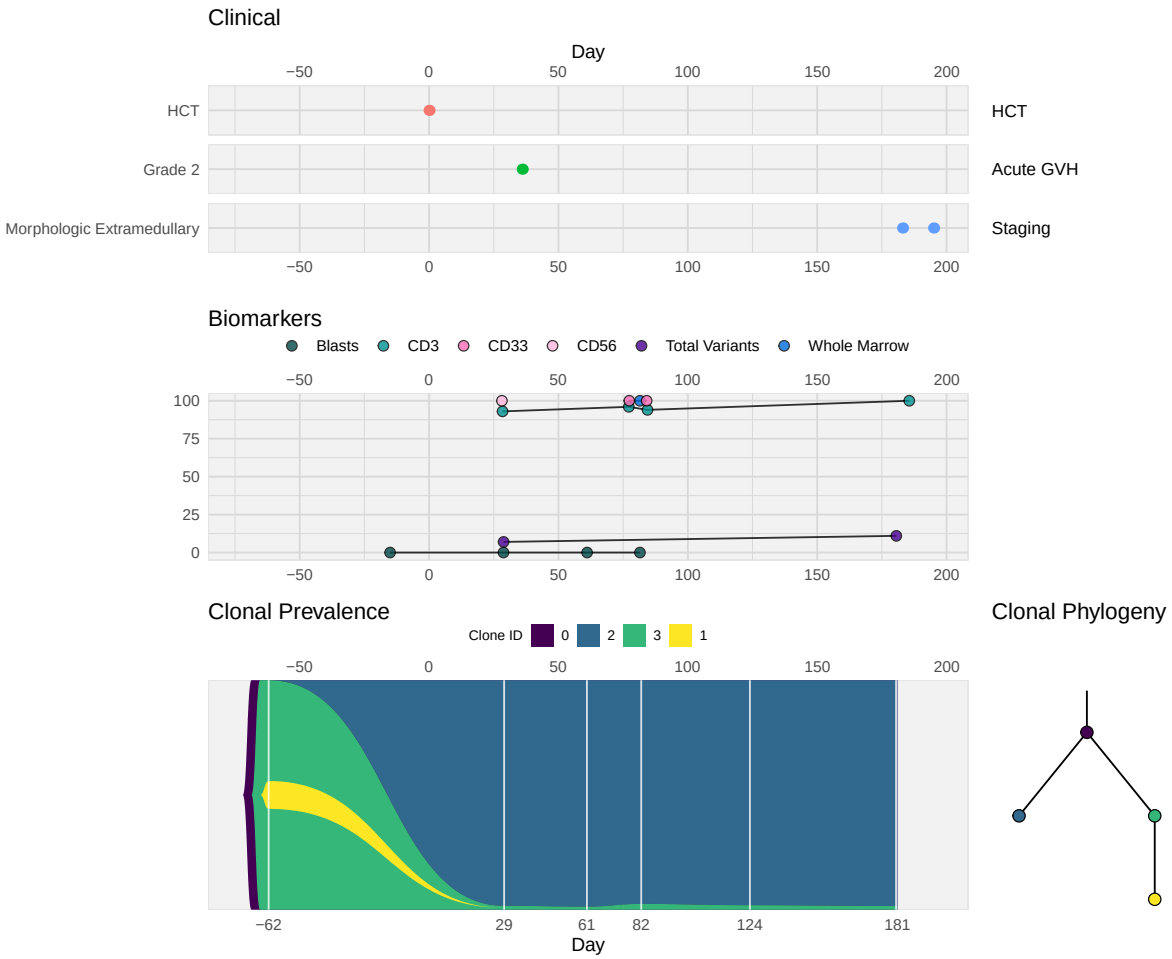

Figure 4.5: **Patient 22 clonal evolution**. Clinical and biomarker data are aligned with the fishplot showing inferred clonal structure and dynamics. In top to bottom order, the clinical panel shows HCT date, GvHD and its grade, and date of morphologic relapse. The biomarkers panel shows key information on transplant success and disease state, including percent blasts and chimerism. Colors in the clinical panel don't have any connection to the colors in the bottom clonal evolution panel.

4.4 Discussion and Future Work

We develop a pipeline to study the clinically relevant subclones detected in longitudinal studies of cancer. We use PyClone-VI [187] and Pairtree [190] for variant clustering and clonal phylogeny inference, respectively. Our approach is specifically created to apply to the setting of hematopoietic stem cell transplant patients, though our approach is applicable to any serially collected samples. In the HCT setting, at least one sample is collected before transplant alters the mutational landscape, as donor SNPs are more likely to be observed than the patient's once engraftment takes place. This likely accounts for the post-transplant clonal dynamics we observe in the four patients shown here, where the dominant clones pre-transplant decline post-transplant, at which point new clones arise and expand, likely corresponding to donor clones. Donor samples are going to be sequenced in the near future, after which it will be possible to identify which variants correspond to donor or patient. Then, the reemergence of any pre-transplant clones will be more easily observable. For cases where donor samples are not sequenced, additional care will have to be taken in interpreting the fishplots.

The infinite sites assumption (ISA)—that each mutation occurs only once—and the infinite alleles assumptions—that mutations do not revert back—is implicitly assumed for most variant clustering and tree inference methods, including PyClone and Pairtree [20, 23, 190, 197]. Pairtree does provide a module that can identify variants that are possible violations of the ISA, allowing them to be removed before tree building. Violations to the ISA have been shown to occur in cancer, with identical mutations arising more than once and some mutations being lost [198–200]. Mostly in single-cell sequencing settings, some phylogeny inference methods have relaxed or altogether removed these assumptions [201–207], and will be most important for studies of small numbers of mutations [23].

The similarity of the clonal evolution and phylogenies of these four patients' cancers is striking. All 4 have a truncal clone 0 that is clonal throughout all the samples. This clone 0 likely corresponds to common SNPs that are present in both patient and donor. The 4

patients also all have 2 branching lineages that arise from clone 0, which correspond to the dominant clones pre- and post-transplant. Further analysis of the remaining patients will examine whether these conserved patterns are seen more generally. Future work that will need to be carried out includes accounting for copy number alterations before performing the variant clustering and tree inference, identifying variants originating from the donor, and determining what factors predict relapse. There is particular interest in determining if relapse corresponds to a reemergence of pre-transplant clones, and if so, to what frequencies do the pre-transplant clones expand.

Chapter 5

USING POPULATION INFORMATION TO PREDICT CHRONIC LYMPHOCYTIC LEUKEMIA PROGRESSION

5.1 Introduction

A tumor's growth curve is the macromolecular result of all the individual cellular and microscopic processes occurring within a population of cancer cells. The simplest growth model - exponential growth - is characterized by rapid, uncontrolled growth. With exponential growth, the tumor size increases at a rate proportional to its current size. It has long been known that if observed over a long enough period of time, a tumor's growth will slow [71, 72, 208–211], because of nutrient or spatial constraints, for example. This leveling off of growth can be modeled by a logistic or a Gompertz model, both of which are commonly used to model cancer growth curves [2, 70, 72, 73, 209, 212–218]. Gompertz growth results from replacing the growth rate in an exponential model with a time dependent growth rate that decays exponentially. Similarly, logistic growth has a growth rate that decreases linearly with the size (Fig. 5.2A).

Modeling the growth curve of a patient's tumor can provide many useful clinical insights, such as optimal cancer screening timing [219], how the tumor will respond to treatment [220, 221], and estimated tumor size at a certain time [222, 223]. Some work has focused on the backward prediction problem of how long a tumor has been growing [223], while other work has focused on the forward prediction problem of how large a tumor will be at a future time [222]. When using only a few available measurements of a tumor's size, it can be difficult to fit the tumor's growth curve, which will be very sensitive to noise. However, the limited amount of data available for a single patient can be augmented by population information obtained from other patients' tumor growth measurements. A mixed effects

model combines individual-level data with population-wide information [224]. The mixed effects approach has been most commonly used in pharmaceutical studies of a drug's effect on a tumor [225–237]. Reasonably so, there is a more limited body of work using this population approach in unperturbed cancers (i.e. naturally growing, without any chemotherapy or radiation). There are some examples of mixed effects models used to study laboratory animal models of unperturbed tumor growth [73, 222, 238].

With most human cancers, treatment is applied soon after diagnosis, making long term study of natural cancer progression rare. Chronic lymphocytic leukemia (CLL), however, can often be monitored without treatment from several months to many years until a more "active" disease state is reached [56]. Measurements of tumor growth can be obtained from white blood cell (WBC) counts obtained from blood samples. As a result, CLL provides a key opportunity to study time series of cancer progression.

In this chapter I present an analysis of the overall cancer growth dynamics for the leukemic cell population detected in the peripheral blood of individual CLL patients. In this work, I show the suitability of a population approach to modeling the CLL time series data from ref. [2] with a Gompertz fit. I propose using a nonlinear mixed effects model with a Gompertz growth curve to capture the within-patient and within-population variation in growth curves. The goal of this work is to perform a forward-in-time prediction of a CLL patient's WBC times series from a few measurements taken early in the patient's disease course. This forward prediction could better inform clinical decisions about treatment and disease progression. Additionally, since many quantitative approaches for studying cancer genomics and evolution include implicit and explicit assumptions about the underlying cancer growth dynamics, it will be insiteful to assess the goodness-of-fit of the most commonly used models for cancer growth.

5.2 Results

I consider three commonly used growth curves: logistic, Gompertz, and a reduced Gompertz. The logistic is given in its differential equation form as

$$\frac{dy}{dt} = ry(1 - y/K) \tag{5.1}$$

$$y(0) = y_0 \tag{5.2}$$

with solution

$$y(t) = \frac{e^{rt} K y_0}{K - y_0 + e^{rt} y_0} \tag{5.3}$$

The Gompertz model is described by the differential equation with initial value,

$$\frac{dy}{dt} = \Big(\alpha - \beta \log(\frac{y}{y_0})\Big) y \tag{5.4}$$

$$y(t_0) = y_0 \tag{5.5}$$

with solution

$$y(t) = y_0 \exp((\alpha - \alpha \exp(-\beta t))/\beta) \tag{5.6}$$

It has been reported that the parameters α and β are highly correlated in many tumors [73, 209, 213–216]. If the parameters are linearly correlated, then a linear regression of fitted α and β parameters can allow a simpler model that only depends on β, by writing

$$\alpha = k\beta + c \tag{5.7}$$

where k is the slope of the line of best fit, and c is the fitted intercept [73].

When the logistic curve (Eq. 5.3) is fit to each patient using nonlinear least squares for all the patients, two distinct clusters are observed (Fig. 5.1). This reflects what was observed in ref. [2], where patients' WBC counts showed both exponential and finite carrying capacity growth curves. The two clusters are separated by the carrying capacity parameter K. The distribution of the fitted K values is bimodal (Fig. 5.1d). The rate and initial value

parameters do not appear to contribute to the formation of the two main clusters, but some smaller clusters seem to exist in these two dimensions (Fig. 5.1b,c). The distribution of fitted growth rates r has a heavy tailed, unimodal distribution (Fig. 5.1e) and the initial value $y0$ is more concentrated (Fig. 5.1f).

When a Gompertz function (Eq. 5.6) is fit for each patient using nonlinear least-squares, there is not as strong of a linear correlation as observed in [73] (Fig. 5.2B,C). The correlation is 0.49, with an R^2 value of 0.242. When an intercept parameter is also fit, in addition to the α and β parameters, the correlation is 0.41, with an R^2 value of 0.1684.

I assess the goodness-of-fit of these three commonly used functions for modeling tumor growth curves that can accommodate a slowing of growth after an initial period of fast growth (logistic, Gompertz, and reduced Gompertz) (Fig. 5.3). In Figure 5.2D-F I show examples of these fits for 3 patients. Using the Gompertz model with a fitted intercept has the lowest mean absolute percentage error (MAPE) and Akaike information criterion (AIC). AIC, a measure of goodness-of-fit that accounts for model complexity, is given by [239]

$$AIC = 2k - 2\log(\widehat{L}) \tag{5.8}$$

where k is the number of estimated parameters and $\widehat{L}$ is the maximum value of the model's likelihood function. The Gompertz function can account for both the exponential behavior and eventual slowing down of growth (carrying capacity) that is commonly observed in tumors.

We would now like to assess whether a population approach to modeling these growth curves would allow a better fit. The motivation for a population approach is two-fold. First, it could provide additional insights into the behavior of each patient's growth curve, and how the growth curves vary within the population. Second, using population information can help predict forward in time, when only a few measurements are available, early on in the disease course. With the nonlinear nature of these growth curves, it is difficult to predict future values of the time series from only a few early time points of a single patient. The information from the population data can serve as a prior for an individual patient's growth

curve. Fitting a growth curve to patient WBC counts is sensitive to noise in the beginning of the time series when only a few measurements are available. However, using the population information helps reduce the sensitivity of the curve fit to that noise.

Now, I will briefly introduce the mixed effects model, treated more thoroughly in refs. [224, 240, 241]. Suppose there are n patients in the population. For each patient i, there is a time series of m_i WBC count measurements $\{y_1^i, \ldots, y_{m_i}^i\}$ taken at times $\{t_1^i, \ldots, t_{m_i}^i\}$. For patient i $(i = 1, \ldots, n)$ at the jth observation $(j = 1, \ldots, m_i)$, we assume the following model

$$y_j^i = f(t_j^i; \boldsymbol{\theta}^i) + \epsilon_j^i \tag{5.9}$$

where $f(t_j^i; \boldsymbol{\theta}^i)$ is the growth curve function (e.g. exponential, logistic, or Gompertz) evaluated at time t_j^i, $\boldsymbol{\theta}^i$ is the parameter vector for patient i, ϵ_j^i is the residual error. The patient i parameters $\boldsymbol{\theta}^i$ depend on fixed effects $\boldsymbol{\mu}$ and a random effect $\boldsymbol{\nu}^i$. The fixed effects are fixed across the entire population, and the random effects are unique to each individual. There are many ways to combine the fixed and random effects, but the simplest is

$$\boldsymbol{\theta}^i = \boldsymbol{\mu} + \boldsymbol{\nu}^i \tag{5.10}$$

To implement this nonlinear mixed effects method, the `R` package `nlme` is used. For the preliminary results, the simplest parameter choice was made (Eq. 5.10), but more sophisticated models may be considered in the future if necessary. The utility of considering population data when predicting WBC counts is evident in the simulation experiment shown in Figure 5.4. To keep the experiment simple, I used the simplest growth curve model - exponential growth. I sampled random effects from a normal distribution and added a Gaussian error term to the WBC count time series. Figure 5.4 shows (for a single patient i) the comparison between an exponential curve fit using nonlinear least squares and a nonlinear mixed effects model. In each panel, I increase the number of WBC counts measured for patient i, ranging from 3 to the full time series. However, the mixed effects model also uses the other 19 patients' full time series to train its model for patient i. Thus, the mixed effect model has a much more accurate curve fit when fewer measurements are available soon after their disease began

being monitored. When using population information, the curve fit isn't as sensitive to noise in the first several points. As more of the full time series for patient i becomes available, the two methods become more similar.

I then fit a mixed effects model with the 3 parameter Gompertz growth curve (α, β, intercept) as $f(\cdot)$ in Equation 5.9. The model is fit with data from all the other patients' WBC time series, as well as the first 12 time points of the patient under consideration. Figure 5.5E,F shows examples from two patients where using population data is advantageous. G and H show examples where a simple curve fit using just that patient's first 12 WBC count measurements is sufficient.

5.3 Discussion

It can be difficult to achieve computational convergence when fitting a mixed effects model, especially when only a few points are used for the patient under consideration. Good initializations of these methods are critical for convergence. The data is fairly noisy, and some patients' WBC counts clearly don't behave like any standard growth curve (e.g. an initial increase, and then a decay). Other functions could be considered for the growth curve, such as a generalized ordinary differential equation model that allows for an increase and a decrease, or splines. However, this will require the fitting of additional parameters. Increasing the degrees of freedom in the model risks overfitting, especially since most longitudinal cancer datasets contain few timepoints. This CLL dataset has more tumor burden timepoints available than most other available datasets. Liquid cancers—like CLL—will have more samples available due to the relative ease of measuring tumor burden via peripheral blood draws. Nevertheless, even with this higher sample coverage, I show here that it still remains difficult to consistently perform forward prediction of cancer cell count. Thus, in many common clinical scenarios, forward prediction of cancer progression using tumor size alone will be difficult to apply, especially in solid tumors where serial cancer samples are rare.

Despite these caveats and challenges, the broader problem of forward-in-time predictions about cancer progression remain promising. As we showed, prediction of tumor size can

be signifantly improved by taking population information into consideration. Outside of the scope of this work focused on tumor growth functions, an even greater improvement in prediction can likely be obtained by using cancer datasets with additional features relevant to cancer progression, such as genetic risk factors, more detailed blood count tests, and other clinical indicators.

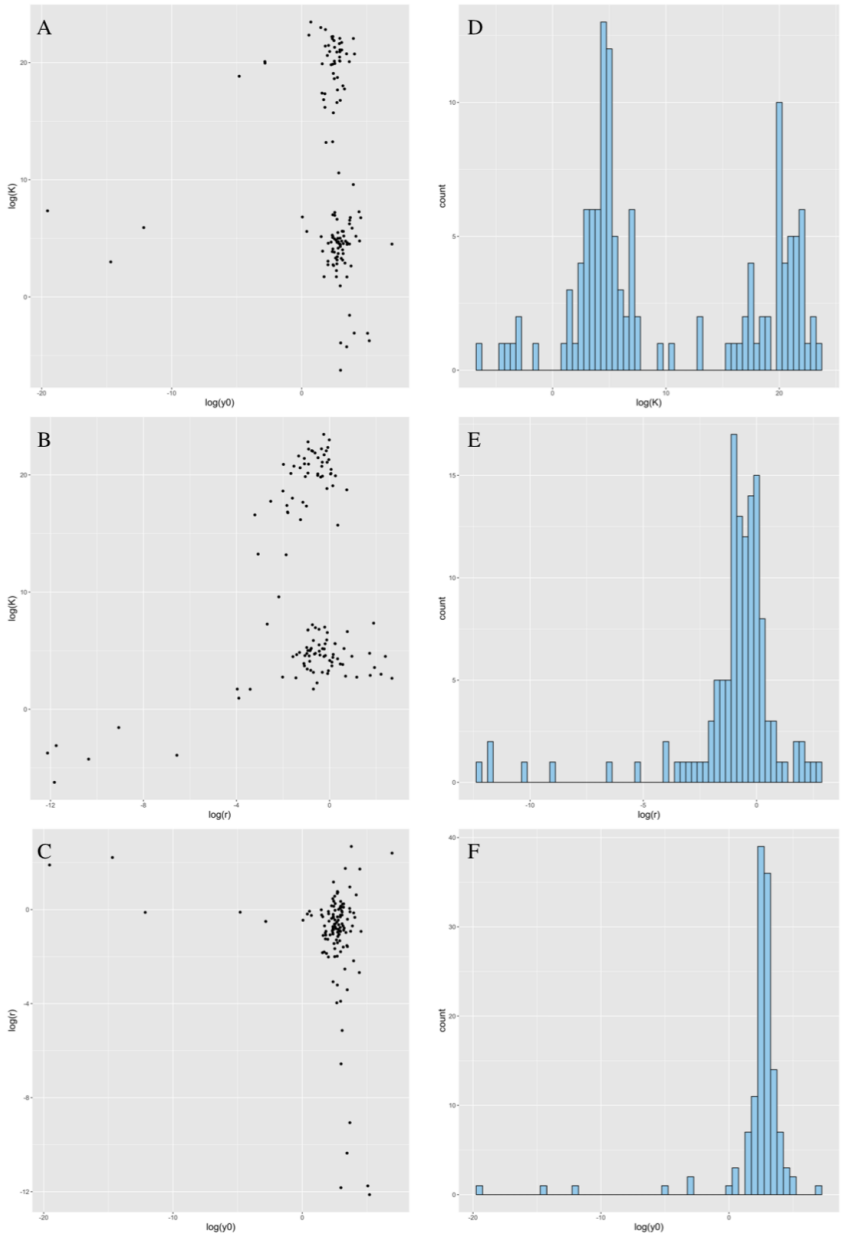

Figure 5.1: **All patients fit to a logistic growth curve.** As observed in ref [2], the patients fall into two classes: finite carrying capacity and exponential growth. (A-C) Log-log plots of the fitted logistic parameter values. (D-F) Distributions of the parameters. K is bimodal, r is unimodal with some long tails, and y0 is more concentrated.

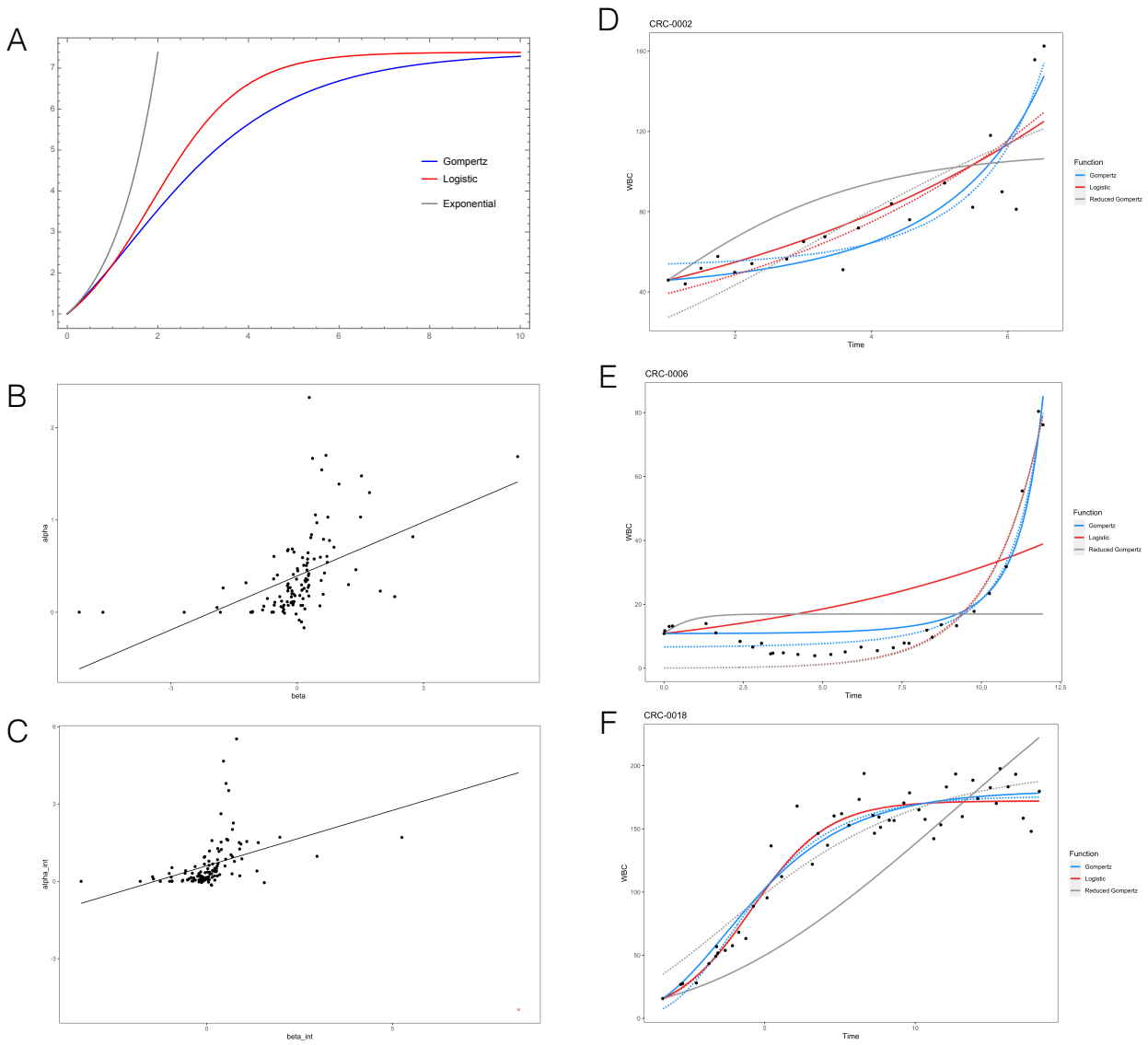

Figure 5.2: **Comparison of growth curve functions.** (A) Comparison of exponential, logistic, and Gompertz growth curves. (B) Correlation and linear regression of the α and β parameters in the Gompertz fits (C) Correlation and linear regression of the α and β parameters in the Gompertz model with a fitted intercept parameter. (D-F) For 3 patients, comparison of the Gompertz, logistic, and reduced Gompertz models. Dashed and solid lines indicate a model with and without an additional fitted parameter for the intercept, respectively.

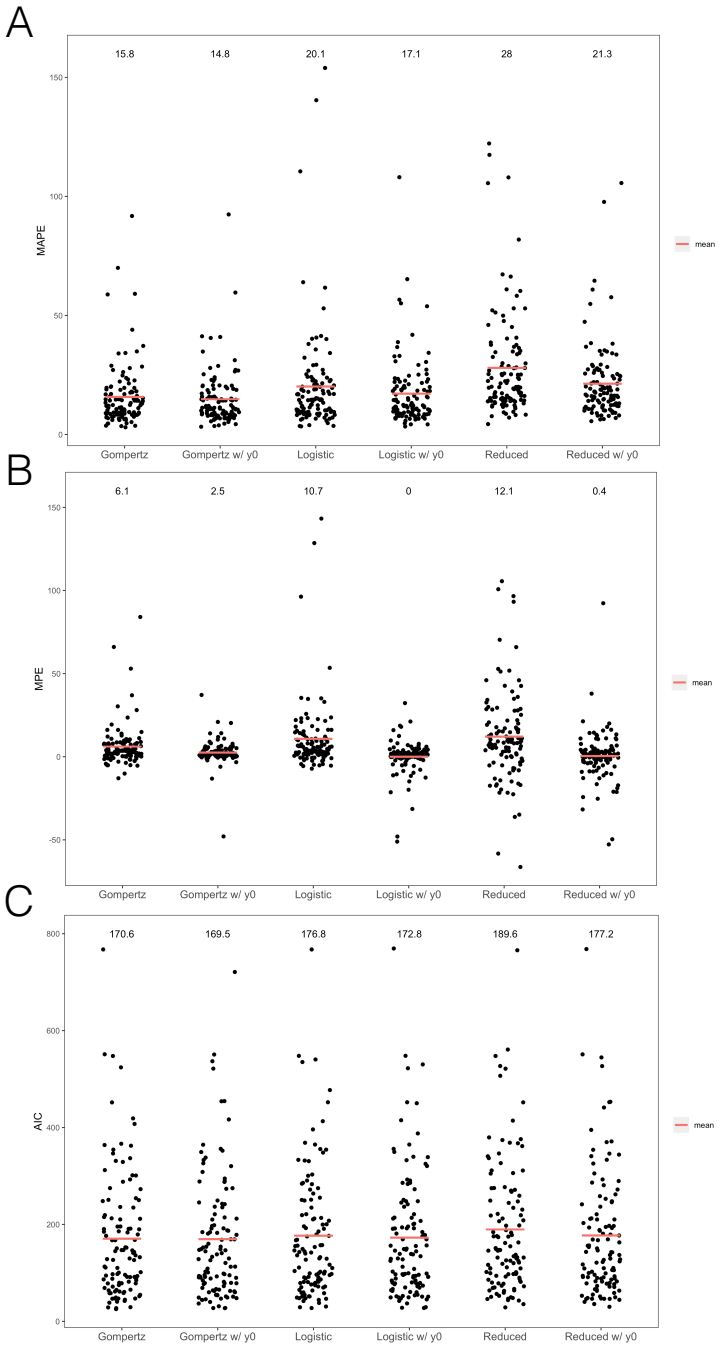

Figure 5.3: **Metrics for goodness of fit of different models for the growth curve.** Mean absolute percentage error (A), mean percentage error (B), and Akaike information criterion (C) of the fit for each patients are evaluated, where the fit for each patient is represented by a single point. The following growth curves functions were fit by nonlinear least squares, with parameters fit in parentheses: Gompertz (α, β), Gompertz with a y-intercept parameter (α, β, y_0), logistic (r, K), logistic with an intercept parameter (r, K, y_0), reduced Gompertz model (β), and reduced Gompertz model with intercept parameter (β, y_0). The mean is displayed at the top and indicated by the red line.

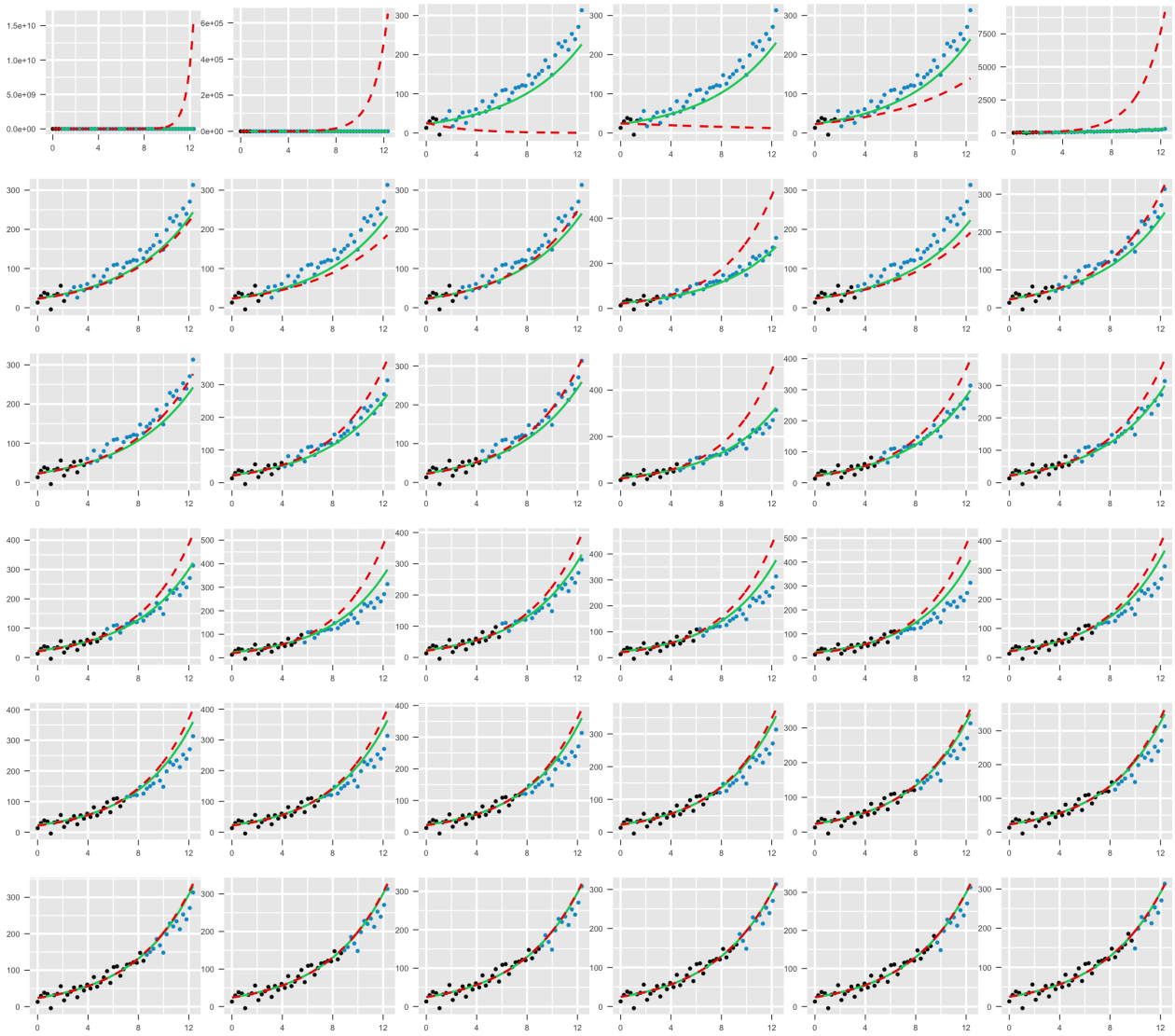

Figure 5.4: **Benefits of using population data for prediction with few measurements in simple simulated example.** Exponential phase of WBC growth curve was simulated for 20 patients. All panels shows the fitted curves for a single patient. Left to right, top to bottom, more points are added to the training set used to fit the model (black points). Blue points are not used to train the model. Training set is black points for patient under consideration, and the rest of the populations' time series. A curve fit is performed using nonlinear mixed effects (solid green line) and nonlinear least squares (dashed red line).

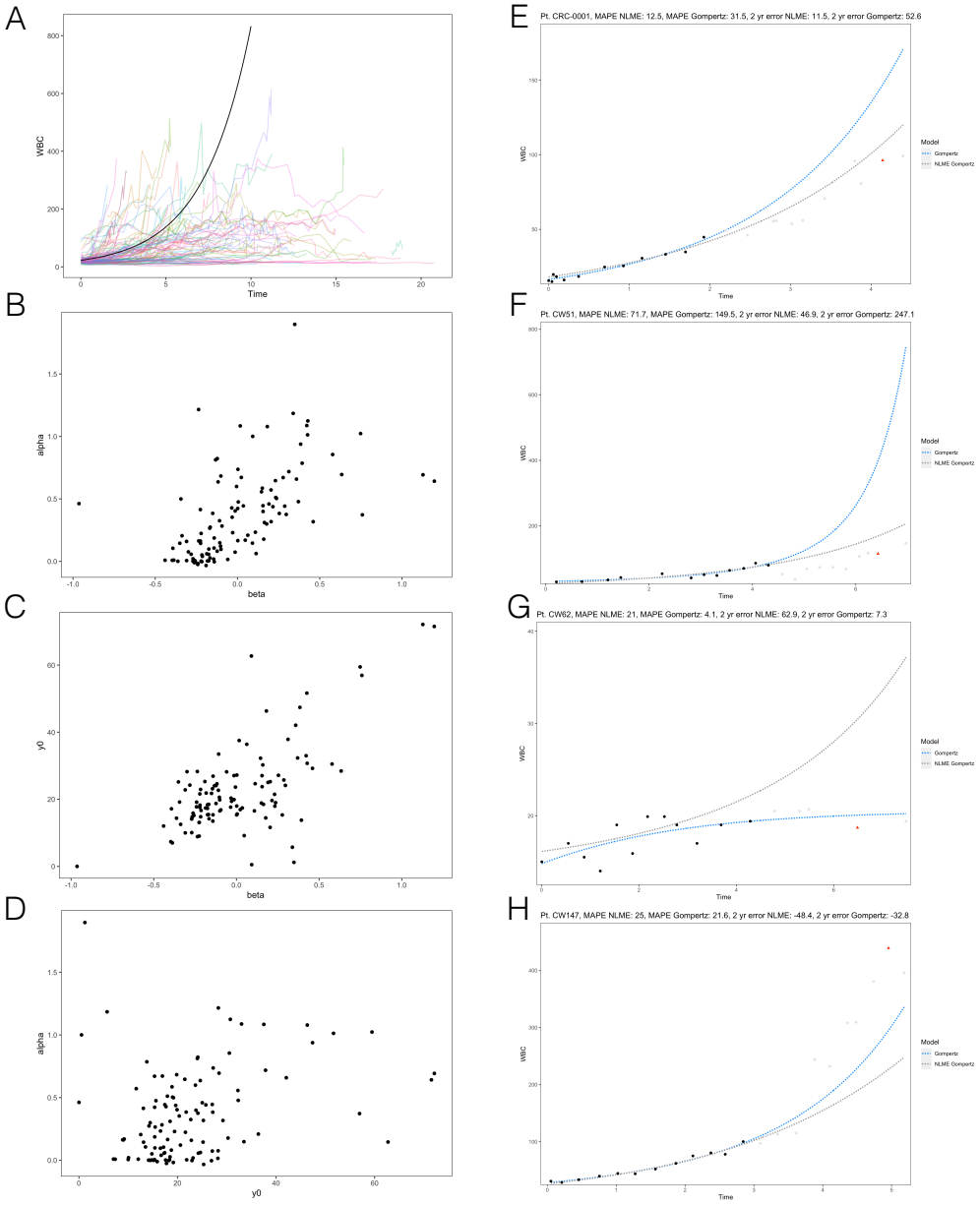

Figure 5.5: **Nonlinear mixed-effects model of Gompertz growth curve.** (A) Fixed effects curve (black solid line) for all the patients fitted with a 3 parameter Gompertz model (α, β, and intercept). All other lines are the growth curves of all the patient WBC counts. (B-D) Random effects for all the patients fitted with a 3 parameter Gompertz model. (E-H) Examples from 4 patients where the first 12 points of WBC count time series (black points) for patient under consideration are fit to 3 parameter Gompertz model using nonlinear least-squares (blue line) and nonlinear mixed effects (gray line). The mixed effects model makes use of WBC time series from the other patients. Red triangular point indicates first measurement $>$ 2 years after last measurement used to fit the model.

BIBLIOGRAPHY

[1] Lee ND, Bozic I. Inferring parameters of cancer evolution in chronic lymphocytic leukemia. PLOS Computational Biology. 2022 Nov;18(11):e1010677. Available from: https://dx.plos.org/10.1371/journal.pcbi.1010677.

[2] Gruber M, Bozic I, Leshchiner I, Livitz D, Stevenson K, Rassenti L, et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. Nature. 2019 Jun;570(7762):474–479. Available from: https://www.nature.com/articles/s41586-019-1252-x.

[3] Leshchiner I, Livitz D, Gainor JF, Rosebrock D, Spiro O, Martinez A, et al. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. Bioinformatics; 2018. Available from: http://biorxiv.org/lookup/doi/10.1101/508127.

[4] Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. British Journal of Cancer. 2004 Dec;91(12):1983–1989. Available from: http://www.nature.com/articles/6602297.

[5] Nowell PC. The Clonal Evolution of Tumor Cell Populations. Science. 1976;194(4260):23–28. Available from: http://www.jstor.org/stable/1742535.

[6] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011 Mar;144(5):646–674. Available from: https://www.cell.com/fulltext/S0092-8674(11)00127-9.

[7] Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021 May;593(7859):405–410. Available from: http://www.nature.com/articles/s41586-021-03477-4.

[8] Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Medicine. 2017 Apr;9(1):34. Available from: https://doi.org/10.1186/s13073-017-0424-2.

[9] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013 Mar;339(6127):1546–1558. Available from: https://science.sciencemag.org/content/339/6127/1546.

[10] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82–93.

[11] Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018 Apr;173(2):371–385.e18. Available from: https://www.cell.com/cell/abstract/S0092-8674(18)30237-X.

[12] Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. Science. 2019 Nov;366(6465):eaan4673. Available from: https://www.science.org/doi/10.1126/science.aan4673.

[13] Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature Genetics. 2016 Mar;48(3):238–244. Available from: https://www.nature.com/articles/ng.3489.

[14] Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetics Research. 1974 Feb;23(1):23–35. Available from: http://www.cambridge.org/core/journals/genetics-research/article/hitchhiking-effect-of-a-favourable-gene/918291A3B62BD50E1AE5C1F22165EF1B.

[15] Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. Nature Reviews Genetics. 2019 Jul;20(7):404–416. Available from: https://www.nature.com/articles/s41576-019-0114-6.

[16] Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012 Jan;481(7381):306–313. Available from: https://www.nature.com/articles/nature10762.

[17] Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968 Feb;217(5129):624–626. Available from: https://www.nature.com/articles/217624a0.

[18] Kimura M. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles*. Genetics Research. 1968 Jun;11(3):247–270. Available from: http://www.cambridge.org/core/journals/genetics-research/article/genetic-variability-maintained-in-a-finite-population-due-to-mutational-production-A74BD3A5D72ED2C52444FD99DFE483EF.

[19] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 Jul;499(7457):214–218. Available from: https://www.nature.com/articles/nature12213.

[20] Dentro SC, Wedge DC, Van Loo P. Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harbor Perspectives in Medicine. 2017 Aug;7(8):a026625.

[21] DREAM SMC-Het Participants, Salcedo A, Tarabichi M, Espiritu SMG, Deshwar AG, David M, et al. A community effort to create standards for evaluating tumor subclonal reconstruction. Nature Biotechnology. 2020 Jan;38(1):97–107. Available from: http://www.nature.com/articles/s41587-019-0364-z.

[22] Miura S, Vu T, Deng J, Buturla T, Oladeinde O, Choi J, et al. Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. Scientific Reports. 2020 Dec;10(1):3498. Available from: http://www.nature.com/articles/s41598-020-59006-2.

[23] Tarabichi M, Salcedo A, Deshwar AG, Ni Leathlobhair M, Wintersinger J, Wedge DC, et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. Nature Methods. 2021 Feb;18(2):144–155. Available from: http://www.nature.com/articles/s41592-020-01013-2.

[24] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009 Apr;458(7239):719–724. Available from: https://www.nature.com/articles/nature07943.

[25] Merlo LMF, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nature Reviews Cancer. 2006 Dec;6(12):924–935. Available from: https://www.nature.com/articles/nrc2013.

[26] Pepper JW, Findlay CS, Kassen R, Spencer SL, Maley CC. SYNTHESIS: Cancer research meets evolutionary biology. Evolutionary Applications. 2009;2(1):62–70. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-4571.2008.00063.x.

[27] Tsao JL, Yatabe Y, Salovaara R, Järvinen HJ, Mecklin JP, Aaltonen LA, et al. Genetic reconstruction of individual colorectal tumor histories. Proceedings of the National Academy of Sciences. 2000 Feb;97(3):1236–1241. Available from: https://www.pnas.org/content/97/3/1236.

[28] Jones S, Chen Wd, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, et al. Comparative lesion sequencing provides insights into tumor evolution. Proceedings of the National Academy of Sciences of the United States of America. 2008 Mar;105(11):4283–4288. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2393770/.

[29] Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010 Oct;467(7319):1114–1117. Available from: https://www.nature.com/articles/nature09515.

[30] Naxerova K, Brachtel E, Salk JJ, Seese AM, Power K, Abbasi B, et al. Hypermutable DNA chronicles the evolution of human colon cancer. Proceedings of the National Academy of Sciences. 2014 May;111(18):E1889–E1898. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1400179111.

[31] McGranahan N, Favero F, Bruin ECd, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Science Translational Medicine. 2015 Apr;7(283):283ra54–283ra54. Available from: https://stm.sciencemag.org/content/7/283/283ra54.

[32] Mitchell TJ, Turajlic S, Rowan A, Nicol D, Farmery JHR, O'Brien T, et al. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. Cell. 2018 Apr;173(3):611–623.e17. Available from: http://www.sciencedirect.com/science/article/pii/S0092867418301648.

[33] PCAWG Evolution & Heterogeneity Working Group, PCAWG Consortium, Gerstung M, Jolly C, Leshchiner I, Dentro SC, et al. The evolutionary history of 2,658 cancers. Nature. 2020 Feb;578(7793):122–128. Available from: http://www.nature.com/articles/s41586-019-1907-7.

[34] Sundermann LK, Wintersinger J, Rätsch G, Stoye J, Morris Q. Reconstructing tumor evolutionary histories and clone trees in polynomial-time with SubMARine. PLOS Computational Biology. 2021 Jan;17(1):e1008400. Available from: https://dx.plos.org/10.1371/journal.pcbi.1008400.

[35] PCAWG Evolution and Heterogeneity Working Group, PCAWG Consortium, Rubanova Y, Shi R, Harrigan CF, Li R, et al. Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. Nature Communications. 2020 Dec;11(1):731. Available from: http://www.nature.com/articles/s41467-020-14352-7.

[36] Tomasetti C, Bozic I. The (not so) immortal strand hypothesis. Stem Cell Research. 2015 Mar;14(2):238–241.

[37] Werner B, Case J, Williams MJ, Chkhaidze K, Temko D, Fernández-Mateos J, et al. Measuring single cell divisions in human tissues from multi-region sequencing data. Nature Communications. 2020 Feb;11(1):1–9. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-020-14844-6.

[38] Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences of the United States of America. 2010 Oct;107(43):18545–18550. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2972991/.

[39] Sun R, Hu Z, Sottoriva A, Graham TA, Harpak A, Ma Z, et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nature Genetics. 2017 Jul;49(7):1015–1024. Available from: https://www.nature.com/articles/ng.3891.

[40] Salichos L, Meyerson W, Warrell J, Gerstein M. Estimating growth patterns and driver effects in tumor evolution from individual samples. Nature Communications. 2020 Feb;11(1):1–14. Available from: https://www.nature.com/articles/s41467-020-14407-9.

[41] Noble R, Burri D, Le Sueur C, Lemant J, Viossat Y, Kather JN, et al. Spatial structure governs the mode of tumour evolution. Nature Ecology & Evolution. 2021 Dec;Available from: https://www.nature.com/articles/s41559-021-01615-9.

[42] Chkhaidze K, Heide T, Werner B, Williams MJ, Huang W, Caravagna G, et al. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. PLOS Computational Biology. 2019 Jul;15(7):e1007243. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007243.

[43] Fu X, Zhao Y, Lopez JI, Rowan A, Au L, Fendler A, et al. Spatial patterns of tumour growth impact clonal diversification in a computational model and the TRACERx Renal study. Nature Ecology & Evolution. 2021 Dec;Available from: https://www.nature.com/articles/s41559-021-01586-x.

[44] Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics. 2018 Jun;50(6):895. Available from: https://www.nature.com/articles/s41588-018-0128-6.

[45] Avanzini S, Kurtz DM, Chabon JJ, Hori SS, Gambhir SS, Alizadeh AA, et al. A mathematical model of ctDNA shedding predicts tumor detection size. Cancer Biology; 2020. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.02.12.946228.

[46] Bozic I, Reiter JG, Allen B, Antal T, Chatterjee K, Shah P, et al. Evolutionary dynamics of cancer in response to targeted combination therapy. eLife. 2013 Jun;2:e00747. Available from: https://elifesciences.org/articles/00747.

[47] Dinh KN, Jaksik R, Kimmel M, Lambert A, Tavaré S. Statistical Inference for the Evolutionary History of Cancer Genomes. Statistical Science. 2020 Feb;35(1):129–144. Available from: https://projecteuclid.org/euclid.ss/1583226033.

[48] Lahouel K, Younes L, Danilova L, Giardiello FM, Hruban RH, Groopman J, et al. Revisiting the tumorigenesis timeline with a data-driven generative model. Proceedings of the National Academy of Sciences. 2020 Jan;117(2):857–864. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1914589117.

[49] Bozic I, Wu CJ. Delineating the evolutionary dynamics of cancer from theory to reality. Nature Cancer. 2020 Jun;1(6):580–588. Available from: http://www.nature.com/articles/s43018-020-0079-6.

[50] Durrett R. Branching Process Models of Cancer. In: Durrett R, editor. Branching Process Models of Cancer. Mathematical Biosciences Institute Lecture Series. Cham: Springer International Publishing; 2015. p. 1–63. Available from: https://doi.org/10.1007/978-3-319-16065-8_1.

[51] Tavaré S. The linear birth-death process: an inferential retrospective. Advances in Applied Probability. 2018 Dec;50(A):253–269. Available from: https://www.cambridge.org/core/product/identifier/S0001867818000848/type/journal_article.

[52] Heyde A, Reiter JG, Naxerova K, Nowak MA. Consecutive seeding and transfer of genetic diversity in metastasis. Proceedings of the National Academy of Sciences. 2019 Jul;116(28):14129–14137. Available from: https://www.pnas.org/content/116/28/14129.

[53] Griffith M, Miller C, Griffith O, Krysiak K, Skidmore Z, Ramu A, et al. Optimizing Cancer Genome Sequencing and Analysis. Cell Systems. 2015 Sep;1(3):210–223. Available from: http://www.sciencedirect.com/science/article/pii/S2405471215001131.

[54] Haber DA, Velculescu VE. Blood-Based Analyses of Cancer: Circulating Tumor Cells and Circulating Tumor DNA. Cancer Discovery. 2014 Jun;4(6):650–661. Available from: https://cancerdiscovery.aacrjournals.org/content/4/6/650.

[55] Myers MA, Satas G, Raphael BJ. CALDER: Inferring Phylogenetic Trees from Longitudinal Tumor Samples. Cell Systems. 2019 Jun;8(6):514–522.e5. Available from: http://www.sciencedirect.com/science/article/pii/S2405471219301917.

[56] Hallek M, Cheson BD, Catovsky D, Caligaris-Cappio F, Dighiero G, Döhner H, et al. iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. Blood. 2018 Jun;131(25):2745–2760.

[57] Marionneaux SM, Keohane EM, Lamanna N, King TC, Mehta SR. Smudge Cells in Chronic Lymphocytic Leukemia: Pathophysiology, Laboratory Considerations, and Clinical Significance. Laboratory Medicine. 2021 Sep;52(5):426–438.

[58] Bozic I, Gerold JM, Nowak MA. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLOS Computational Biology. 2016 Feb;12(2):e1004731. Available from: http://dx.plos.org/10.1371/journal.pcbi.1004731.

[59] Kim J, Kim D, Lim JS, Maeng JH, Son H, Kang HC, et al. The use of technical replication for detection of low-level somatic mutations in next-generation sequencing. Nature Communications. 2019 Dec;10(1):1047. Available from: http://www.nature.com/articles/s41467-019-09026-y.

[60] Song P, Chen SX, Yan YH, Pinto A, Cheng LY, Dai P, et al. Selective multiplexed enrichment for the detection and quantitation of low-fraction DNA variants via low-depth sequencing. Nature Biomedical Engineering. 2021 Jul;5(7):690–701. Available from: http://www.nature.com/articles/s41551-021-00713-0.

[61] Fabre MA, de Almeida JG, Fiorillo E, Mitchell E, Damaskou A, Rak J, et al. The longitudinal dynamics and natural history of clonal haematopoiesis. Nature. 2022 Jun;606(7913):335–342.

[62] Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016 Oct;538(7624):260–264. Available from: http://www.nature.com/articles/nature19768.

[63] Mitchell E, Spencer Chapman M, Williams N, Dawson KJ, Mende N, Calderbank EF, et al. Clonal dynamics of haematopoiesis across the human lifespan. Nature. 2022 Jun;606(7913):343–350. Available from: https://www.nature.com/articles/s41586-022-04786-y.

[64] Auslander N, Wolf YI, Koonin EV. In silico learning of tumor evolution through mutational time series. Proceedings of the National Academy of Sciences. 2019 May;116(19):9501–9510. Available from: https://www.pnas.org/content/116/19/9501.

[65] PCAWG Mutational Signatures Working Group, PCAWG Consortium, Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94–101. Available from: http://www.nature.com/articles/s41586-020-1943-3.

[66] Wang Z, Xia Y, Mills L, Nikolakopoulos AN, Maeser N, Sheltzer JM, et al. Evolving copy number gains promote tumor expansion and bolster mutational diversification. Genomics; 2022. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.06.14.495959.

[67] Friberg S, Mattson S. On the growth rates of human malignant tumors: implications for medical decision making. Journal of Surgical Oncology. 1997 Aug;65(4):284–297.

[68] Rodriguez-Brenes IA, Komarova NL, Wodarz D. Tumor growth dynamics: insights into evolutionary processes. Trends in Ecology & Evolution. 2013 Oct;28(10):597–604.

[69] Talkington A, Durrett R. Estimating Tumor Growth Rates In Vivo. Bulletin of Mathematical Biology. 2015 Oct;77(10):1934–1954.

[70] Norton L. A Gompertzian model of human breast cancer growth. Cancer Research. 1988 Dec;48(24 Pt 1):7067–7071.

[71] Spratt JA, von Fournier D, Spratt JS, Weber EE. Decelerating growth and human breast cancer. Cancer. 1993 Mar;71(6):2013–2019.

[72] Gerlee P. The Model Muddle: In Search of Tumor Growth Laws. Cancer Research. 2013 Apr;73(8):2407. Available from: http://cancerres.aacrjournals.org/content/73/8/2407.abstract.

[73] Vaghi C, Rodallec A, Fanciullino R, Ciccolini J, Mochel JP, Mastri M, et al. Population modeling of tumor growth curves and the reduced Gompertz model improve prediction of the age of experimental tumors. PLOS Computational Biology. 2020 Feb;16(2):e1007178. Available from: https://dx.plos.org/10.1371/journal.pcbi.1007178.

[74] Carlsson G, Gullberg B, Hafström L. Estimation of liver tumor volume using different formulas?An experimental study in rats. Journal of Cancer Research and Clinical Oncology. 1983 Jan;105(1):20–23. Available from: http://link.springer.com/10.1007/BF00391826.

[75] West J, Schenck RO, Gatenbee C, Robertson-Tessi M, Anderson ARA. Normal tissue architecture determines the evolutionary course of cancer. Nature Communications. 2021 Dec;12(1):2060. Available from: http://www.nature.com/articles/s41467-021-22123-1.

[76] Marusyk A, Tabassum DP, Altrock PM, Almendro V, Michor F, Polyak K. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature. 2014 Oct;514(7520):54–58.

[77] Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell. 2019 Mar;176(6):1282–1294.e20. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419301618.

[78] Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry. 1977 Dec;81(25):2340–2361. Available from: https://doi.org/10.1021/j100540a008.

[79] Bozic I, Paterson C, Waclaw B. On measuring selection in cancer from subclonal mutation frequencies. PLOS Computational Biology. 2019 Sep;15(9):e1007368. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007368.

[80] Keohane EM, Smith LJ, Walenga JM, editors. Rodak's hematology: clinical principles and applications. Fifth edition ed. St. Louis, Missouri.: Elsevier/Saunders; 2016. OCLC: 906612363.

[81] Howlader N, Noone A, Krapcho M, Miller D, Brest A, Yu M, et al. SEER Cancer Statistics Review, 1975-2018. Bethesda, MD: National Cancer Institute; 2021. Available from: https://seer.cancer.gov/csr/1975_2018/.

[82] Elnair R, Ellithi M, Kallam A, Shostrom V, Bociek RG. Outcomes of Richter's transformation of chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL): an analysis of the SEER database. Annals of Hematology. 2021 Oct;100(10):2513–2519. Available from: https://link.springer.com/10.1007/s00277-021-04603-y.

[83] Wang Y, Tschautscher MA, Rabe KG, Call TG, Leis JF, Kenderian SS, et al. Clinical characteristics and outcomes of Richter transformation: experience of 204 patients from a single center. Haematologica. 2020 Mar;105(3):765–773. Available from: http://www.haematologica.org/lookup/doi/10.3324/haematol.2019.224121.

[84] Jain N. Selecting Frontline Therapy for CLL in 2018. Hematology. 2018 Nov;2018(1):242–247. Available from: https://ashpublications.org/hematology/article/2018/1/242/277585/Selecting-Frontline-Therapy-for-CLL-in-2018.

[85] Yeung CCS, Shadman M. How to Choose the Best Treatment and Testing for Chronic Lymphocytic Leukemia in the Tsunami of New Treatment Options. Current Oncology Reports. 2019 Aug;21(8):74. Available from: http://link.springer.com/10.1007/s11912-019-0819-x.

[86] Honigberg LA, Smith AM, Sirisawad M, Verner E, Loury D, Chang B, et al. The Bruton tyrosine kinase inhibitor PCI-32765 blocks B-cell activation and is efficacious in models of autoimmune disease and B-cell malignancy. Proceedings of the National Academy of Sciences. 2010 Jul;107(29):13075–13080. Available from: https://pnas.org/doi/full/10.1073/pnas.1004594107.

[87] Byrd JC, Furman RR, Coutre SE, Flinn IW, Burger JA, Blum KA, et al. Targeting BTK with Ibrutinib in Relapsed Chronic Lymphocytic Leukemia. New England Journal of Medicine. 2013 Jul;369(1):32–42. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1215637.

[88] Liang C, Tian D, Ren X, Ding S, Jia M, Xin M, et al. The development of Bruton's tyrosine kinase (BTK) inhibitors from 2012 to 2017: A mini-review. European Journal of Medicinal Chemistry. 2018 May;151:315–326. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0223523418303052.

[89] Dingjan GM, Middendorp S, Dahlenborg K, Maas A, Grosveld F, Hendriks RW. Bruton's Tyrosine Kinase Regulates the Activation of Gene Rearrangements at the Lambda Light Chain Locus in Precursor B Cells in the Mouse. Journal of Experimental Medicine. 2001 May;193(10):1169–1178. Available from: https://rupress.org/jem/article/193/10/1169/25882/Brutons-Tyrosine-Kinase-Regulates-the-Activation.

[90] O'Brien S, Furman RR, Coutre S, Flinn IW, Burger JA, Blum K, et al. Single-agent ibrutinib in treatment-naïve and relapsed/refractory chronic lymphocytic leukemia: a 5-year experience. Blood. 2018 Apr;131(17):1910–1919. Available from: https://ashpublications.org/blood/article/131/17/1910/36756/Singleagent-ibrutinib-in-treatmentna%C3%AFve-and.

[91] Itchaki G, Brown JR. Experience with ibrutinib for first-line use in patients with chronic lymphocytic leukemia. Therapeutic Advances in Hematology. 2018 Jan;9(1):3–19. Available from: http://journals.sagepub.com/doi/10.1177/2040620717741861.

[92] Woyach JA, Ruppert AS, Guinn D, Lehman A, Blachly JS, Lozanski A, et al. BTK^{C481S} -Mediated Resistance to Ibrutinib in Chronic Lymphocytic Leukemia. Journal of Clinical Oncology. 2017 May;35(13):1437–1443. Available from: https://ascopubs.org/doi/10.1200/JCO.2016.70.2282.

[93] Liu TM, Woyach JA, Zhong Y, Lozanski A, Lozanski G, Dong S, et al. Hypermorphic mutation of phospholipase C, Gamma2 acquired in ibrutinib-resistant CLL confers BTK independency upon B-cell receptor activation. Blood. 2015 Jul;126(1):61–68. Available from: https://ashpublications.org/blood/article/126/1/61/34349/Hypermorphic-mutation-of-phospholipase-C-%CE%B32.

[94] Furman RR, Cheng S, Lu P, Setty M, Perez AR, Guo A, et al. Ibrutinib Resistance in Chronic Lymphocytic Leukemia. New England Journal of Medicine. 2014 Jun;370(24):2352–2354. Available from: http://www.nejm.org/doi/10.1056/NEJMc1402716.

[95] Sedlarikova L, Petrackova A, Papajik T, Turcsanyi P, Kriegova E. Resistance-Associated Mutations in Chronic Lymphocytic Leukemia Patients Treated With Novel Agents. Frontiers in Oncology. 2020 Jun;10:894. Available from: https://www.frontiersin.org/article/10.3389/fonc.2020.00894/full.

[96] Ahn IE, Brown JR. Targeting Bruton's Tyrosine Kinase in CLL. Frontiers in Immunology. 2021 Jun;12:687458. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.687458/full.

[97] Byrd JC, Furman RR, Coutre SE, Flinn IW, Burger JA, Blum K, et al. Ibrutinib Treatment for First-Line and Relapsed/Refractory Chronic Lymphocytic Leukemia: Final Analysis of the Pivotal Phase Ib/II PCYC-1102 Study. Clinical Cancer Research. 2020 Aug;26(15):3918–3927. Available from: https://aacrjournals.org/clincancerres/article/26/15/3918/82635/Ibrutinib-Treatment-for-First-Line-and-Relapsed.

[98] Jain P, Keating M, Wierda W, Estrov Z, Ferrajoli A, Jain N, et al. Outcomes of patients with chronic lymphocytic leukemia after discontinuing ibrutinib. Blood. 2015 Mar;125(13):2062–2067. Available from: https://ashpublications.org/blood/article/125/13/2062/33945/Outcomes-of-patients-with-chronic-lymphocytic.

[99] Maddocks KJ, Ruppert AS, Lozanski G, Heerema NA, Zhao W, Abruzzo L, et al. Etiology of Ibrutinib Therapy Discontinuation and Outcomes in Patients With Chronic Lymphocytic Leukemia. JAMA Oncology. 2015 Apr;1(1):80. Available from: http://oncology.jamanetwork.com/article.aspx?doi=10.1001/jamaoncol.2014.218.

[100] Woyach J, Huang Y, Rogers K, Bhat SA, Grever MR, Lozanski A, et al. Resistance to Acalabrutinib in CLL Is Mediated Primarily By BTK Mutations. Blood. 2019 Nov;134(Supplement_1):504–504. Available from: https://ashpublications.org/blood/article/134/Supplement_1/504/426369/Resistance-to-Acalabrutinib-in-CLL-Is-Mediated.

[101] Brandhuber B, Gomez E, Smith S, Eary T, Spencer S, Rothenberg SM, et al. LOXO-305, A Next Generation Reversible BTK Inhibitor, for Overcoming Acquired Resistance to Irreversible BTK Inhibitors. Clinical Lymphoma Myeloma and Leukemia. 2018

Sep;18:S216. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2152265018308309.

[102] Gomez EB, Isabel L, Rosendahal MS, Rothenberg SM, Andrews SW, Brandhuber BJ. Loxo-305, a Highly Selective and Non-Covalent Next Generation BTK Inhibitor, Inhibits Diverse BTK C481 Substitution Mutations. Blood. 2019 Nov;134(Supplement_1):4644–4644. Available from: https://ashpublications.org/blood/article/134/Supplement_1/4644/428544/Loxo305-a-Highly-Selective-and-NonCovalent-Next.

[103] Naeem AS, Nguy WI, Tyekucheva S, Fernandes SM, Rai V, Ebata K, et al. LOXO-305: Targeting C481S Bruton Tyrosine Kinase in Patients with Ibrutinib-Resistant CLL. Blood. 2019 Nov;134(Supplement_1):478–478. Available from: https://ashpublications.org/blood/article/134/Supplement_1/478/426456/LOXO305-Targeting-C481S-Bruton-Tyrosine-Kinase-in.

[104] Mato AR, Shah NN, Jurczak W, Cheah CY, Pagel JM, Woyach JA, et al. Pirtobrutinib in relapsed or refractory B-cell malignancies (BRUIN): a phase 1/2 study. The Lancet. 2021 Mar;397(10277):892–901. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673621002245.

[105] Iskierka-Jażdżewska E, Obracaj A, Urbaniak M, Robak T. New Treatment Options for Newly-Diagnosed and Relapsed Chronic Lymphocytic Leukemia. Current Treatment Options in Oncology. 2022 Jun;23(6):775–795. Available from: https://link.springer.com/10.1007/s11864-022-00974-0.

[106] García-Marco JA, Jiménez JL, Recasens V, Zarzoso MF, González-Barca E, De Marcos NS, et al. High prognostic value of measurable residual disease detection by flow cytometry in chronic lymphocytic leukemia patients treated with front-line fludarabine, cyclophosphamide, and rituximab, followed by three years of rituximab maintenance. Haematologica. 2019 Nov;104(11):2249–2257. Available from: http://www.haematologica.org/lookup/doi/10.3324/haematol.2018.204891.

[107] Del Giudice I, Raponi S, Della Starza I, De Propris MS, Cavalli M, De Novi LA, et al. Minimal Residual Disease in Chronic Lymphocytic Leukemia: A New Goal? Frontiers in Oncology. 2019 Aug;9:689. Available from: https://www.frontiersin.org/article/10.3389/fonc.2019.00689/full.

[108] Ruppert AS, Yin J, Davidian M, Tsiatis AA, Byrd JC, Woyach JA, et al. Application of a sequential multiple assignment randomized trial (SMART) design in older patients with chronic lymphocytic leukemia. Annals of Oncology. 2019 Apr;30(4):542–550. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0923753419311329.

[109] Wierda WG, Rawstron A, Cymbalista F, Badoux X, Rossi D, Brown JR, et al. Measurable residual disease in chronic lymphocytic leukemia: expert review and consensus recommendations. Leukemia. 2021 Nov;35(11):3059–3072. Available from: https://www.nature.com/articles/s41375-021-01241-1.

[110] Short NJ, Kantarjian H, Kanagal-Shamanna R, Sasaki K, Ravandi F, Cortes J, et al. Ultra-accurate Duplex Sequencing for the assessment of pretreatment ABL1 kinase domain mutations in Ph+ ALL. Blood Cancer Journal. 2020 May;10(5):61. Available from: https://www.nature.com/articles/s41408-020-0329-y.

[111] Kamath-Loeb AS, Shen JC, Schmitt MW, Kohrn BF, Loeb KR, Estey EH, et al. Accurate detection of subclonal variants in paired diagnosis-relapse acute myeloid leukemia samples by next generation Duplex Sequencing. Leukemia Research. 2022 Apr;115:106822. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0145212622000480.

[112] Schmitt MW, Pritchard JR, Leighow SM, Aminov BI, Beppu L, Kim DS, et al. Single-Molecule Sequencing Reveals Patterns of Preexisting Drug Resistance That Suggest Treatment Strategies in Philadelphia-Positive Leukemias. Clinical Cancer Research. 2018 Nov;24(21):5321–5334. Available from: https://aacrjournals.org/clincancerres/article/24/21/5321/281648/Single-Molecule-Sequencing-Reveals-Patterns-of.

[113] Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nature Protocols. 2014 Nov;9(11):2586–2606. Available from: https://www.nature.com/articles/nprot.2014.170.

[114] Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, et al. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. Proceedings of the National Academy of Sciences. 2016 May;113(21):6005–6010. Available from: https://pnas.org/doi/full/10.1073/pnas.1601311113.

[115] Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences. 2012 Sep;109(36):14508–14513. Available from: https://pnas.org/doi/full/10.1073/pnas.1208715109.

[116] Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nature Reviews Genetics. 2018

May;19(5):269–285. Available from: http://www.nature.com/articles/nrg.2017.117.

[117] Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. Nature Methods. 2015 May;12(5):423–425. Available from: http://www.nature.com/articles/nmeth.3351.

[118] Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Research. 2016 Jun;44(11):e108–e108. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw227.

[119] Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019 Nov;4(43):1686. Available from: https://joss.theoj.org/papers/10.21105/joss.01686.

[120] Aslan B, Kismali G, Chen LS, Iles LR, Mahendra M, Peoples M, et al. Development and characterization of prototypes for in vitro and in vivo mouse models of ibrutinib-resistant CLL. Blood Advances. 2021 Aug;5(16):3134–3146. Available from: https://ashpublications.org/bloodadvances/article/5/16/3134/476567/Development-and-characterization-of-prototypes-for.

[121] Black GS, Huang X, Qiao Y, Tarapcsak S, Rogers KA, Misra S, et al. Subclonal evolution of CLL driver mutations is associated with relapse in ibrutinib- and acalabrutinib-treated patients. Blood. 2022 Jul;140(4):401–405. Available from: https://ashpublications.org/blood/article/140/4/401/485093/Subclonal-evolution-of-CLL-driver-mutations-is.

[122] Wang E, Mi X, Thompson MC, Montoya S, Notti RQ, Afaghani J, et al. Mechanisms of Resistance to Noncovalent Bruton's Tyrosine Kinase Inhibitors. New England Journal of Medicine. 2022 Feb;386(8):735–743. Available from: http://www.nejm.org/doi/10.1056/NEJMoa2114110.

[123] Mato AR, Pagel JM, Coombs CC, Shah NN, Lamanna N, Munir T, et al. Pirtobrutinib, A Next Generation, Highly Selective, Non-Covalent BTK Inhibitor in Previously Treated CLL/SLL: Updated Results from the Phase 1/2 BRUIN Study. Blood. 2021 Nov;138(Supplement 1):391–391. Available from: https://ashpublications.org/blood/article/138/Supplement%201/391/478198/Pirtobrutinib-A-Next-Generation-Highly-Selective.

[124] Naeem A, Utro F, Wang Q, Cha J, Vihinen M, Martindale SP, et al. Pirtobrutinib Targets BTK C481S in Ibrutinib-Resistant CLL but Second-Site BTK Mutations Lead to Resistance. Blood Advances. 2022 Oct;p. bloodadvances.2022008447. Available from: https://ashpublications.org/bloodadvances/article/doi/10.1182/bloodadvances.2022008447/486896/Pirtobrutinib-Targets-BTK-C481S-in-Ibrutinib.

[125] Rampino N, Yamamoto H, Ionov Y, Li Y, Sawai H, Reed JC, et al. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science (New York, NY). 1997 Feb;275(5302):967–969.

[126] Blombery P, Thompson ER, Chen X, Nguyen T, Anderson MA, Westerman DA, et al. BAX-Mutated Clonal Hematopoiesis in Patients on Long-Term Venetoclax for Relapsed/Refractory Chronic Lymphocytic Leukemia. Blood. 2020 Nov;136(Supplement 1):9–10. Available from: https://ashpublications.org/blood/article/136/Supplement%201/9/470031/BAX-Mutated-Clonal-Hematopoiesis-in-Patients-on.

[127] Blombery P, Lew TE, Dengler MA, Thompson ER, Lin VS, Chen X, et al. Clonal hematopoiesis, myeloid disorders and *BAX* -mutated myelopoiesis in patients receiving venetoclax for CLL. Blood. 2022 Feb;139(8):1198–1207. Available from: https://ashpublications.org/blood/article/139/8/1198/476749/Clonal-hematopoiesis-myeloid-disorders-and-BAX.

[128] Moujalled DM, Brown FC, Chua CC, Dengler MA, Pomilio G, Anstee NS, et al. Acquired mutations in BAX confer resistance to BH3-mimetic therapy in Acute Myeloid Leukemia. Blood. 2022 Oct;p. blood.2022016090. Available from: https://ashpublications.org/blood/article/doi/10.1182/blood.2022016090/486781/Acquired-mutations-in-BAX-confer-resistance-to-BH3.

[129] Malcikova J, Pavlova S, Kunt Vonkova B, Radova L, Plevova K, Kotaskova J, et al. Low-burden *TP53* mutations in CLL: clinical impact and clonal evolution within the context of different treatment options. Blood. 2021 Dec;138(25):2670–2685. Available from: https://ashpublications.org/blood/article/138/25/2670/475902/Low-burden-TP53-mutations-in-CLL-clinical-impact.

[130] Quinquenel A, Fornecker LM, Letestu R, Ysebaert L, Fleury C, Lazarian G, et al. Prevalence of BTK and PLCG2 mutations in a real-life CLL cohort still on ibrutinib after 3 years: a FILO group study. Blood. 2019 Aug;134(7):641–644. Available from: https://ashpublications.org/blood/article/134/7/641/260720/Prevalence-of-BTK-and-PLCG2-mutations-in-a.

[131] Tanaka TN, Bejar R. MDS overlap disorders and diagnostic boundaries. Blood. 2019 Mar;133(10):1086–1095. Available from: https://ashpublications.org/blood/article/133/10/1086/272719/MDS-overlap-disorders-and-diagnostic-boundaries.

[132] Platzbecker U. Treatment of MDS. Blood. 2019 Mar;133(10):1096–1107. Available from: https://ashpublications.org/blood/article/133/10/1096/272732/Treatment-of-MDS.

[133] Sperling AS, Gibson CJ, Ebert BL. The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. Nature Reviews Cancer. 2017 Jan;17(1):5–19.

[134] Vardiman JW, Harris NL, Brunning RD. The World Health Organization (WHO) classification of the myeloid neoplasms. Blood. 2002 Oct;100(7):2292–2302.

[135] Lichtman MA. Does a diagnosis of myelogenous leukemia require 20% marrow myeloblasts, and does <5% marrow myeloblasts represent a remission? The history and ambiguity of arbitrary diagnostic boundaries in the understanding of myelodysplasia. The Oncologist. 2013;18(9):973–980.

[136] Estey E, Hasserjian RP, Döhner H. Distinguishing AML from MDS: a fixed blast percentage may no longer be optimal. Blood. 2022 Jan;139(3):323–332. Available from: https://ashpublications.org/blood/article/139/3/323/476130/Distinguishing-AML-from-MDS-a-fixed-blast.

[137] Makishima H, Yoshizato T, Yoshida K, Sekeres MA, Radivoyevitch T, Suzuki H, et al. Dynamics of clonal evolution in myelodysplastic syndromes. Nature Genetics. 2017 Feb;49(2):204–212.

[138] Shiozawa Y, Malcovati L, Gallì A, Pellagatti A, Karimi M, Sato-Otsubo A, et al. Gene expression and risk of leukemic transformation in myelodysplasia. Blood. 2017 Dec;130(24):2642–2653.

[139] Ogawa S. Genetics of MDS. Blood. 2019 Mar;133(10):1049–1059. Available from: https://ashpublications.org/blood/article/133/10/1049/272730/Genetics-of-MDS.

[140] Horowitz M, Schreiber H, Elder A, Heidenreich O, Vormoor J, Toffalori C, et al. Epidemiology and biology of relapse after stem cell transplantation. Bone Marrow Transplantation. 2018 Nov;53(11):1379–1389. Available from: http://www.nature.com/articles/s41409-018-0171-z.

[141] Schmid C, Labopin M, Nagler A, Niederwieser D, Castagna L, Tabrizi R, et al. Treatment, risk factors, and outcome of adults with relapsed AML after reduced intensity conditioning for allogeneic stem cell transplantation. Blood. 2012 Feb;119(6):1599–1606. Available from: https://ashpublications.org/blood/article/119/6/1599/30200/Treatment-risk-factors-and-outcome-of-adults-with.

[142] Cornelissen JJ, Versluis J, Passweg JR, van Putten WLJ, Manz MG, Maertens J, et al. Comparative therapeutic value of post-remission approaches in patients with acute myeloid leukemia aged 40–60 years. Leukemia. 2015 May;29(5):1041–1050. Available from: https://www.nature.com/articles/leu2014332.

[143] Gyurkocza B, Storb R, Storer BE, Chauncey TR, Lange T, Shizuru JA, et al. Nonmyeloablative allogeneic hematopoietic cell transplantation in patients with acute myeloid leukemia. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2010 Jun;28(17):2859–2867.

[144] Hughes AEO, Magrini V, Demeter R, Miller CA, Fulton R, Fulton LL, et al. Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. PLoS genetics. 2014 Jul;10(7):e1004462.

[145] Klco J, Spencer D, Miller C, Griffith M, Lamprecht T, O'Laughlin M, et al. Functional Heterogeneity of Genetically Defined Subclones in Acute Myeloid Leukemia. Cancer Cell. 2014 Mar;25(3):379–392. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1535610814000543.

[146] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017 Apr;8(1):14049. Available from: http://www.nature.com/articles/ncomms14049.

[147] Quek L, Otto GW, Garnett C, Lhermitte L, Karamitros D, Stoilova B, et al. Genetically distinct leukemic stem cells in human CD34- acute myeloid leukemia are arrested at a hemopoietic precursor-like stage. Journal of Experimental Medicine. 2016 Jul;213(8):1513–1535. Available from: https://rupress.org/jem/article/213/8/1513/42086/Genetically-distinct-leukemic-stem-cells-in-human.

[148] Potter N, Miraki-Moud F, Ermini L, Titley I, Vijayaraghavan G, Papaemmanuil E, et al. Single cell analysis of clonal architecture in acute myeloid leukaemia. Leukemia. 2019 May;33(5):1113–1123.

[149] Paguirigan AL, Smith J, Meshinchi S, Carroll M, Maley C, Radich JP. Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia. Science

Translational Medicine. 2015 Apr;7(281). Available from: https://www.science.org/doi/10.1126/scitranslmed.aaa0763.

[150] Lei Y, Tang R, Xu J, Wang W, Zhang B, Liu J, et al. Applications of single-cell sequencing in cancer research: progress and perspectives. Journal of Hematology & Oncology. 2021 Dec;14(1):91. Available from: https://jhoonline.biomedcentral.com/articles/10.1186/s13045-021-01105-2.

[151] Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nature Genetics. 2021 Jun;53(6):770–777. Available from: https://www.nature.com/articles/s41588-021-00873-4.

[152] Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018 Oct;19(4):562–578. Available from: https://academic.oup.com/biostatistics/article/19/4/562/4599254.

[153] Salehi S, Steif A, Roth A, Aparicio S, Bouchard-Côté A, Shah SP. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. Genome Biology. 2017 Mar;18(1):44.

[154] Malikic S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. Nature Communications. 2019 Dec;10(1):2750. Available from: http://www.nature.com/articles/s41467-019-10737-5.

[155] Kuipers J, Moore AL, Jahn K, Schraml P, Wang F, Morita K, et al. Statistical tests for intra-tumour clonal co-occurrence and exclusivity. PLOS Computational Biology. 2021 Dec;17(12):e1009036. Available from: https://dx.plos.org/10.1371/journal.pcbi.1009036.

[156] Jerby-Arnon L, Pfetzer N, Waldman YY, McGarry L, James D, Shanks E, et al. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell. 2014 Aug;158(5):1199–1209.

[157] Kim YA, Cho DY, Dao P, Przytycka TM. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. Bioinformatics (Oxford, England). 2015 Jun;31(12):i284–292.

[158] Leiserson MDM, Wu HT, Vandin F, Raphael BJ. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biology. 2015 Aug;16(1):160.

[159] Babur O, Gonen M, Aksoy BA, Schultz N, Ciriello G, Sander C, et al. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome Biology. 2015 Feb;16(1):45.

[160] Constantinescu S, Szczurek E, Mohammadi P, Rahnenführer J, Beerenwinkel N. TiMEx: a waiting time model for mutually exclusive cancer alterations. Bioinformatics (Oxford, England). 2016 Apr;32(7):968–975.

[161] Cristea S, Kuipers J, Beerenwinkel N. pathTiMEx: Joint Inference of Mutually Exclusive Cancer Pathways and Their Progression Dynamics. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 2017 Jun;24(6):603–615.

[162] Kuipers J, Thurnherr T, Moffa G, Suter P, Behr J, Goosen R, et al. Mutational interactions define novel cancer subgroups. Nature Communications. 2018 Oct;9(1):4353.

[163] Srivastava A, Pastor-Pareja JC, Igaki T, Pagliarini R, Xu T. Basement membrane remodeling is essential for *Drosophila* disc eversion and tumor invasion. Proceedings of the National Academy of Sciences. 2007 Feb;104(8):2721–2726. Available from: https://pnas.org/doi/full/10.1073/pnas.0611666104.

[164] Igaki T, Pagliarini RA, Xu T. Loss of Cell Polarity Drives Tumor Growth and Invasion through JNK Activation in Drosophila. Current Biology. 2006 Jun;16(11):1139–1146. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0960982206015399.

[165] Uhlirova M, Bohmann D. JNK- and Fos-regulated Mmp1 expression cooperates with Ras to induce invasive tumors in Drosophila. The EMBO Journal. 2006 Nov;25(22):5294–5304. Available from: http://emboj.embopress.org/cgi/doi/10.1038/sj.emboj.7601401.

[166] Pagliarini RA, Xu T. A Genetic Screen in *Drosophila* for Metastatic Behavior. Science. 2003 Nov;302(5648):1227–1231. Available from: https://www.science.org/doi/10.1126/science.1088474.

[167] Parker TM, Gupta K, Palma AM, Yekelchyk M, Fisher PB, Grossman SR, et al. Cell competition in intratumoral and tumor microenvironment interactions. The EMBO journal. 2021 Sep;40(17):e107271.

[168] van Neerven SM, Vermeulen L. Cell competition in development, homeostasis and cancer. Nature Reviews Molecular Cell Biology. 2022 Sep;.

[169] Zhou H, Neelakantan D, Ford HL. Clonal cooperativity in heterogenous cancers. Seminars in Cell & Developmental Biology. 2017 Apr;64:79–89.

[170] Tabassum DP, Polyak K. Tumorigenesis: it takes a village. Nature Reviews Cancer. 2015 Aug;15(8):473–483. Available from: http://www.nature.com/articles/nrc3971.

[171] Caswell DR, Swanton C. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. BMC medicine. 2017 Jul;15(1):133.

[172] Krakow EF, Gyurkocza B, Storer BE, Chauncey TR, McCune JS, Radich JP, et al. Phase I/II multisite trial of optimally dosed clofarabine and low-dose TBI for hematopoietic cell transplantation in acute myeloid leukemia. American Journal of Hematology. 2020 Jan;95(1):48–56. Available from: https://onlinelibrary.wiley.com/doi/10.1002/ajh.25665.

[173] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research. 2019 Jan;47(D1):D941–D947. Available from: https://academic.oup.com/nar/article/47/D1/D941/5146192.

[174] Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biology. 2015 Dec;16(1):35. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0602-8.

[175] Deveau P, Colmet Daage L, Oldridge D, Bernard V, Bellini A, Chicard M, et al. QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. Bioinformatics. 2018 Jun;34(11):1808–1816. Available from: https://academic.oup.com/bioinformatics/article/34/11/1808/4802225.

[176] Andor N, Harness JV, Müller S, Mewes HW, Petritsch C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics (Oxford, England). 2014 Jan;30(1):50–60.

[177] Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS computational biology. 2014 Aug;10(8):e1003665.

[178] Xiao Y, Wang X, Zhang H, Ulintz PJ, Li H, Guan Y. FastClone is a probabilistic tool for deconvoluting tumor heterogeneity in bulk-sequencing samples. Nature Communications. 2020 Dec;11(1):4469. Available from: https://www.nature.com/articles/s41467-020-18169-2.

[179] Yuan K, Macintyre G, Liu W, PCAWG-11 working group, Markowetz F. Ccube: A fast and robust method for estimating cancer cell fractions. Bioinformatics; 2018. Available from: http://biorxiv.org/lookup/doi/10.1101/484402.

[180] Fischer A, Vázquez-García I, Illingworth CJR, Mustonen V. High-definition reconstruction of clonal composition in cancer. Cell Reports. 2014 Jun;7(5):1740–1752.

[181] Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West RB, Batzoglou S. Fast and scalable inference of multi-sample cancer lineages. Genome Biology. 2015 May;16:91.

[182] Zare H, Wang J, Hu A, Weber K, Smith J, Nickerson D, et al. Inferring clonal composition from multiple sections of a breast cancer. PLoS computational biology. 2014 Jul;10(7):e1003703.

[183] Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proceedings of the National Academy of Sciences. 2016 Sep;113(37). Available from: https://pnas.org/doi/full/10.1073/pnas.1522203113.

[184] Satas G, Raphael BJ. Tumor phylogeny inference using tree-constrained importance sampling. Bioinformatics. 2017 Jul;33(14):i152–i160. Available from: https://academic.oup.com/bioinformatics/article/33/14/i152/3953987.

[185] Cun Y, Yang TP, Achter V, Lang U, Peifer M. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. Nature Protocols. 2018 Jun;13(6):1488–1501. Available from: http://www.nature.com/articles/nprot.2018.033.

[186] Caravagna G, Sanguinetti G, Graham TA, Sottoriva A. The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data. BMC Bioinformatics. 2020 Dec;21(1):531. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03863-1.

[187] Gillis S, Roth A. PyClone-VI: scalable inference of clonal population structures using whole genome data. BMC Bioinformatics. 2020 Dec;21(1):571. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03919-2.

[188] Tanner G, Westhead DR, Droop A, Stead LF. Benchmarking pipelines for subclonal deconvolution of bulk tumour sequencing data. Nature Communications. 2021 Dec;12(1):6396. Available from: https://www.nature.com/articles/s41467-021-26698-7.

[189] Zheng L, Niknafs N, Wood LD, Karchin R, Scharpf RB. Estimation of cancer cell fractions and clone trees from multi-region sequencing of tumors. Bioinformatics. 2022 Jun;p. btac367. Available from: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac367/6596597.

[190] Wintersinger JA, Dobson SM, Kulman E, Stein LD, Dick JE, Morris Q. Reconstructing Complex Cancer Evolutionary Histories from Multiple Bulk DNA Samples Using Pairtree. Blood Cancer Discovery. 2022 Mar;p. OF1–OF12. Available from: https://aacrjournals.org/bloodcancerdiscov/article/doi/10.1158/2643-3230.BCD-21-0092/694089/Reconstructing-Complex-Cancer-Evolutionary.

[191] Malikic S, McPherson AW, Donmez N, Sahinalp CS. Clonality inference in multiple tumor samples using phylogeny. Bioinformatics. 2015 May;31(9):1349–1356. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv003.

[192] Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proceedings of the National Academy of Sciences. 2008 Sep;105(35):13081–13086. Available from: https://pnas.org/doi/full/10.1073/pnas.0801523105.

[193] El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics. 2015 Jun;31(12):i62–i70. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv261.

[194] Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC bioinformatics. 2014 Feb;15:35.

[195] Gatenbee CD, Schenck RO, Bravo RR, Anderson ARA. EvoFreq: visualization of the Evolutionary Frequencies of sequence and model data. BMC bioinformatics. 2019 Dec;20(1):710.

[196] Miller CA, McMichael J, Dang HX, Maher CA, Ding L, Ley TJ, et al. Visualizing tumor evolution with the fishplot package for R. BMC Genomics. 2016 Dec;17(1):880. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3195-z.

[197] Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nature Methods. 2014 Apr;11(4):396–398. Available from: http://www.nature.com/articles/nmeth.2883.

[198] Kuipers J, Jahn K, Raphael BJ, Beerenwinkel N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. Genome Research. 2017 Nov;27(11):1885–1894.

[199] Demeulemeester J, Dentro SC, Gerstung M, Van Loo P. Biallelic mutations in cancer genomes reveal local mutational determinants. Nature Genetics. 2022 Feb;54(2):128–133. Available from: https://www.nature.com/articles/s41588-021-01005-8.

[200] McPherson A, Roth A, Laks E, Masud T, Bashashati A, Zhang AW, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nature Genetics. 2016 Jul;48(7):758–767. Available from: http://www.nature.com/articles/ng.3573.

[201] Malikic S, Mehrabadi FR, Ciccolella S, Rahman MK, Ricketts C, Haghshenas E, et al. PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. Genome Research. 2019 Nov;29(11):1860–1877. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.234435.118.

[202] Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome Biology. 2017 Dec;18(1):178. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1311-2.

[203] Zafar H, Navin N, Chen K, Nakhleh L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. Genome Research. 2019 Nov;29(11):1847–1859. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.243121.118.

[204] Marass F, Mouliere F, Yuan K, Rosenfeld N, Markowetz F. A phylogenetic latent feature model for clonal deconvolution. The Annals of Applied Statistics. 2016 Dec;10(4). Available from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-10/issue-4/A-phylogenetic-latent-feature-model-for-clonal-deconvolution/10.1214/16-AOAS986.full.

[205] Satas G, Zaccaria S, Mon G, Raphael BJ. SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses. Cell Systems. 2020 Apr;10(4):323–332.e8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405471220301150.

[206] Ciccolella S, Ricketts C, Soto Gomez M, Patterson M, Silverbush D, Bonizzoni P, et al. Inferring cancer progression from Single-Cell Sequencing while allowing mutation losses. Bioinformatics. 2021 Apr;37(3):326–333. Available from: https://academic.oup.com/bioinformatics/article/37/3/326/5893545.

[207] Bonizzoni P, Ciccolella S, Vedova GD, Soto M. Does Relaxing the Infinite Sites Assumption Give Better Tumor Phylogenies? An ILP-Based Comparative Approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019 Sep;16(5):1410–1423. Available from: https://ieeexplore.ieee.org/document/8438928/.

[208] Schrek R. A Comparison of the Growth Curves of Malignant and Normal (Embryonic and Postembryonic) Tissues of the Rat. The American Journal of Pathology. 1936 Jul;12(4):525–530.

[209] Laird AK. DYNAMICS OF TUMOR GROWTH. British Journal of Cancer. 1964 Sep;13:490–502.

[210] Steel GG, Lamerton LF. The growth rate of human tumours. British Journal of Cancer. 1966 Mar;20(1):74–86.

[211] Akanuma A. Parameter analysis of Gompertzian function growth model in clinical tumors. European Journal of Cancer. 1978 Jun;14(6):681–688.

[212] Gompertz B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c. Philosophical Transactions of the Royal Society of London. 1825 Dec;115:513–583. Available from: https://royalsocietypublishing.org/doi/10.1098/rstl.1825.0026.

[213] Brunton GF, Wheldon TE. Characteristic species dependent growth patterns of mammalian neoplasms. Cell and Tissue Kinetics. 1978 Mar;11(2):161–175.

[214] Demicheli R. Growth of testicular neoplasm lung metastases: tumor-specific relation between two Gompertzian parameters. European Journal of Cancer. 1980 Dec;16(12):1603–1608.

[215] Norton L, Simon R, Brereton HD, Bogden AE. Predicting the course of Gompertzian growth. Nature. 1976 Dec;264(5586):542–545.

[216] Parfitt AM, Fyhrie DP. Gompertzian growth curves in parathyroid tumours: further evidence for the set-point hypothesis. Cell Proliferation. 1997 Sep;30(8-9):341–349.

[217] Sullivan PW, Salmon SE. Kinetics of tumor growth and regression in IgG multiple myeloma. The Journal of Clinical Investigation. 1972 Jul;51(7):1697–1708.

[218] West J, Hasnain Z, Macklin P, Newton PK. An Evolutionary Model of Tumor Cell Kinetics and the Emergence of Molecular Heterogeneity Driving Gompertzian Growth. SIAM Review. 2016 Jan;58(4):716–736. Available from: http://epubs.siam.org/doi/10.1137/15M1044825.

[219] Hart D, Shochat E, Agur Z. The growth law of primary breast cancer as inferred from mammography screening trials data. British Journal of Cancer. 1998 Aug;78(3):382–387.

[220] Hahnfeldt P, Panigrahy D, Folkman J, Hlatky L. Tumor Development under Angiogenic Signaling. Cancer Research. 1999 Oct;59(19):4770. Available from: http://cancerres.aacrjournals.org/content/59/19/4770.abstract.

[221] O'Donoghue JA. The response of tumours with Gompertzian growth characteristics to fractionated radiotherapy. International Journal of Radiation Biology. 1997 Sep;72(3):325–339.

[222] Benzekry S, Lamont C, Beheshti A, Tracz A, Ebos JML, Hlatky L, et al. Classical mathematical models for description and prediction of experimental tumor growth. PLoS computational biology. 2014 Aug;10(8):e1003800.

[223] Patrone MV, Hubbs JL, Bailey JE, Marks LB. How long have I had my cancer, doctor? Estimating tumor age via Collins' law. Oncology (Williston Park, NY). 2011 Jan;25(1):38–43, 46.

[224] Lavielle M. Mixed effects models for the population approach: models, tasks, methods and tools. Chapman & Hall/CRC biostatistics series. Boca Raton: Taylor & Francis; 2014.

[225] Bonate PL, Suttle AB. Modeling tumor growth kinetics after treatment with pazopanib or placebo in patients with renal cell carcinoma. Cancer Chemotherapy and Pharmacology. 2013 Jul;72(1):231–240.

[226] Claret L, Lu JF, Sun YN, Bruno R. Development of a modeling framework to simulate efficacy endpoints for motesanib in patients with thyroid cancer. Cancer Chemotherapy and Pharmacology. 2010 Nov;66(6):1141–1149.

[227] Claret L, Gupta M, Han K, Joshi A, Sarapa N, He J, et al. Evaluation of tumor-size response metrics to predict overall survival in Western and Chinese patients with first-line metastatic colorectal cancer. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2013 Jun;31(17):2110–2114.

[228] Claret L, Girard P, Hoff PM, Van Cutsem E, Zuideveld KP, Jorga K, et al. Model-based prediction of phase III overall survival in colorectal cancer on the basis of phase II tumor dynamics. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2009 Sep;27(25):4103–4108.

[229] Frances N, Claret L, Bruno R, Iliadis A. Tumor growth modeling from clinical trials reveals synergistic anticancer effect of the capecitabine and docetaxel combination in metastatic breast cancer. Cancer Chemotherapy and Pharmacology. 2011 Dec;68(6):1413–1419.

[230] Hansson EK, Amantea MA, Westwood P, Milligan PA, Houk BE, French J, et al. PKPD Modeling of VEGF, sVEGFR-2, sVEGFR-3, and sKIT as Predictors of Tumor Dynamics and Overall Survival Following Sunitinib Treatment in GIST. CPT: pharmacometrics & systems pharmacology. 2013 Nov;2:e84.

[231] Hansson EK, Ma G, Amantea MA, French J, Milligan PA, Friberg LE, et al. PKPD Modeling of Predictors for Adverse Effects and Overall Survival in Sunitinib-Treated Patients With GIST. CPT: pharmacometrics & systems pharmacology. 2013 Dec;2:e85.

[232] Houk BE, Bello CL, Poland B, Rosen LS, Demetri GD, Motzer RJ. Relationship between exposure to sunitinib and efficacy and tolerability endpoints in patients with cancer: results of a pharmacokinetic/pharmacodynamic meta-analysis. Cancer Chemotherapy and Pharmacology. 2010 Jul;66(2):357–371.

[233] Maitland ML, Wu K, Sharma MR, Jin Y, Kang SP, Stadler WM, et al. Estimation of renal cell carcinoma treatment effects from disease progression modeling. Clinical Pharmacology and Therapeutics. 2013 Apr;93(4):345–351.

[234] Ribba B, Kaloshi G, Peyre M, Ricard D, Calvez V, Tod M, et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research. 2012 Sep;18(18):5071–5080.

[235] Stein A, Wang W, Carter AA, Chiparus O, Hollaender N, Kim H, et al. Dynamic tumor modeling of the dose-response relationship for everolimus in metastatic renal cell carcinoma using data from the phase 3 RECORD-1 trial. BMC cancer. 2012 Jul;12:311.

[236] Tham LS, Wang L, Soo RA, Lee SC, Lee HS, Yong WP, et al. A pharmacodynamic model for the time course of tumor shrinkage by gemcitabine + carboplatin in non-small cell lung cancer patients. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research. 2008 Jul;14(13):4213–4218.

[237] Wang Y, Sung C, Dartois C, Ramchandani R, Booth BP, Rock E, et al. Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. Clinical Pharmacology and Therapeutics. 2009 Aug;86(2):167–174.

[238] Parra-Guillen ZP, Mangas-Sanjuan V, Garcia-Cremades M, Troconiz IF, Mo G, Pitou C, et al. Systematic Modeling and Design Evaluation of Unperturbed Tumor Dynamics in Xenografts. The Journal of Pharmacology and Experimental Therapeutics. 2018 Jul;366(1):96–104.

[239] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974 Dec;19(6):716–723. Available from: http://ieeexplore.ieee.org/document/1100705/.

[240] Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge; New York: Cambridge University Press; 2007. OCLC: 646068240. Available from: https://doi.org/10.1017/CBO9780511790942.

[241] LaMotte LR. Fixed-, Random-, and Mixed-Effects Models. In: Balakrishnan N, Colton T, Everitt B, Piegorsch W, Ruggeri F, Teugels JL, editors. Wiley StatsRef: Statistics Reference Online. Chichester, UK: John Wiley & Sons, Ltd; 2014. p. stat03169. Available from: http://doi.wiley.com/10.1002/9781118445112.stat03169.