Quantitative fitness modelling in cancer using single cell timeseries population dynamics

by

Sohrab Salehi

B.Sc., Amirkabir University of Technology, 2011
M.Sc., The University of British Columbia, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

February 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Quantitative fitness modelling in cancer using single cell timeseries population dynamics

submitted by **Sohrab Salehi** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Bioinformatics**.

Examining Committee:

Sohrab Shah, Department of Pathology and Laboratory Medicine, UBC

Co-supervisor

Alexandre Bouchard-Côté, Department of Statistics, UBC

Co-supervisor

Samuel Aparicio, Department of Molecular Oncology, UBC

Supervisory Committee Member

Ryan Brinkman, Department of Medical Genetics, UBC

University Examiner

Geoffrey Schiebinger, Department of Mathematics, UBC

University Examiner

Additional Supervisory Committee Members:

Sarah Otto, Department of Zoology, UBC

Supervisory Committee Member

Abstract

Tumour fitness landscapes underpin selection in cancer, impacting evolution and response to treatment. Quantitative fitness modelling of cancer cells has numerous and diverse implications: attributing clonal dynamics to drift or selection, identifying the determinants of clonal expansion, and forecasting tumour growth trajectories. Why and how drug resistance evolves is among the key unresolved areas of investigation that require advanced understanding of fitness in cancer. Longitudinal xenoengraftment interrogated via next generation single cell sequencing (SCS) has enabled more accurate, quantitative measurements of tumours as they evolve. This process generates timestamped samples comprising thousands of cells each measured at thousands of copy number aberrations (CNA). Major analytical challenges introduced by this new datatype include (i) how to identify biologically meaningful groups of cells (i.e., clones) across multiple timepoints, and (ii) how to quantitatively reason about the underlying evolutionary forces acting on the clones via their observed dynamics.

To address the first problem, we describe and provide supplementary tools for `sitka`, a scalable Bayesian phylogenetic inference method in Chapter 2. It resolves the clonal structure of a heterogeneous tumour cell population sampled over multiple timepoints by reconstructing the evolutionary relationship between single cells from their inferred CNA profiles. We then develop `Lumberjack`, a tree-cutting algorithm, and use it to assign cells to clones. We address the second problem in Chapter 3 by developing `fitClone`, a Bayesian probabilistic framework that ascribes quantitative selection coefficients to individual cancer clones and forecasts competitive clonal dynamics over time. In Chapter 4, we exemplify the computational models introduced above on real-world data collected from cancer cells over a multi-year period to verify two key hypotheses, that (i) clonal dynamics in a pre-treatment triple negative breast cancer (TNBC) tumour is quantifiably reproducible, and that (ii) the fitness landscape is reversed under early response to cisplatin treatment. Our results show that population genetic modelling of timeseries tumour measurements to predict clonal evolution is tractable. Further study with timeseries modelling will provide insight into therapeutic strategies promoting early intervention, drug combinations and evolution-aware approaches to clinical management.

Lay Summary

Cells accrue mutations that in time result in uncontrolled growth and in some cases the ability to invade the surrounding tissue and seed metastasis. This thesis develops a statistical framework that characterises the subpopulations that exit in an evolving tumour monitored over a period of time at a single-cell resolution. This information is then used to generate a model that predicts how cancerous tumours evolve over time, with or without treatment. We expect that this research provides insight into therapeutic strategies promoting early intervention, drug combinations and evolution-aware approaches to clinical management of human cancers.

Preface

A version of Chapter 2 is a manuscript under preparation and has a draft on the *bioRxiv* preprint repository [1]. Professor Alexandre Bouchard-Côté developed and implemented the `sitka` model. I developed the visualisation methods and performed the benchmarking experiments and generated the figures.

Chapter 3 and Chapter 4 are based on a manuscript under review (*bioRxiv* preprint [2]). I have developed and implemented the `fitClone` model and generated the figures in Chapters 2, 3, and 4. I have performed the phylogenetic and fitness analysis of all the datasets introduced in Chapters 2 and 4. Farhia Kabeer has generated the xenograft biological systems. Dr. Marc Williams has done the missegregation analysis. Professor Sohrab Shah and Samuel Aparicio have contributed to the original writing of the manuscript on which these two chapters are based.

Table of Contents

Appendix

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisors, Dr. Sohrab Shah and Dr. Alexandre Bouchard-Côté for their unyielding support both in my academic and personal life. Thank you to my supervisory committee, Dr. Samuel Aparicio and Dr. Sarah Otto. I learned a lot in terms of biological and evolutionary fundamentals of the problems investigated in this work. I consider myself utterly fortunate to have had a supervisory committee that were dedicated and engaged in my project.

I would like to thank members of the Shah and Aparicio labs, specially Farhia Kabeer, Kieran Campbell, Andrew McPherson, and Mirela Andronescu for their help and stimulating conversations.

Dedication

To my family, my gratitude for your unconditional support. To Monique, thank you for believing in me. Without you I could not have done this.

Chapter 1

Introduction

Cellular fitness underpins the tissue population dynamics of cancer progression and treatment response. Yet, quantifying fitness in heterogeneous cell populations and identifying causal mechanisms shaping fitness landscapes remain open problems, impeding progress in developing effective and durable therapeutic strategies. In particular, quantitative fitness modelling of cancer cells has numerous and diverse implications: attributing clonal dynamics to drift or selection, identifying the determinants of clonal expansion, enabling causal inference, and forecasting growth trajectories. Why and how drug resistance evolves is among the key unresolved areas of investigation that require advanced understanding of fitness in cancer. For example, drug resistance mechanisms are commonly attributed to phenotypic plasticity encoded via epigenetic changes [3, 4] or evolutionary selection of pre-existing genomic clones [5]. However, the relative contribution of these processes when studied in tandem is poorly understood and requires integrated genome-transcriptome investigation. Moreover, how changes in genomic architecture brought about by copy number alterations (CNA) drive tumour progression remains understudied [6]. Genetic editing of cancer related genes and treatment with pharmacological drugs are among possible ways to induce fitness changes. The effects of these perturbations can be monitored to profile tumour growth progression and other therapeutic responses. We contend that quantitatively ascribing fitness values to clonal dynamics over long-range timeseries in the context of such perturbations would provide higher order insights in drug resistance than current models allow.

Previous work has established models of fitness through interpreting allelic measurements of single snapshots [7–11] from bulk sequencing over large patient cohorts [12], timeseries monitoring of cell free DNA [13], multiregion sequencing [8, 14–17] and estimating fitness landscapes of clonal haematopoiesis [18]. However, the cancer field has generally lacked serial measurements from patient derived tissues to directly observe cancer evolution over realistic timescales. This has impeded a thorough understanding of factors driving selection, achieved in other biological systems through studying granular timeseries with population genetic modelling [19]. The majority of work in cancer has focused on bulk tumour sequencing, where it is computationally difficult or sometimes impossible to accurately resolve the subclonal composition of tumours. Single cell genome measurements to scalably define clonal populations in cancer over thousands of cells have only recently emerged [20, 21], enabling identification of rare populations, precise tracking of clones and robust clone-specific measurements suitable for population genetic modelling.

In this dissertation we present a body of work that aims to (i) identify clones, (ii) measure their abundances over time and (iii) infer evolutionary fitness parameters for populations measured at the single cell resolution to help investigate how human cancer cells evolve at the copy number level, and how establishing baseline fitness measures helps to interpret selection under drug administration.

1.1 Tumour evolution and models of fitness

Cells accrue mutations that in time result in uncontrolled growth and in some cases the ability to invade the surrounding tissue and seed metastasis. The stochastic nature of mutation accumulation and subsequent selection dictates that the tumour growth process is, to some degree, random. This in turn may result in treatment-resistant subpopulations and ultimately cause relapse. Advances in next generation sequencing (NGS) combined with statistical deconvolution methods have enabled investigation of tumour subpopulation structures at a high resolution [22, 23], where clusters of mutations existing at the same abundance can be identified. Many resistance mechanisms to drugs that occur reproducibly in patients, and thus are predictors of response, are known [24, 25]. For instance, the emergence of T790M mutations in *EGFR* in patients with non-small-cell lung cancers after treatment with tyrosine kinase inhibitors (TKI) [26], secondary mutation in *KRAS* in colorectal cancer patients post mono-clonal antibody (mAb) anti-EGFR therapy [27, 28], *ESR1* mutations induced acquired-resistance in ER+ breast cancer patients undergoing endocrine therapy (e.g., tamoxifen) [29, 30] and *BRCA1* revertant mutations in platinum and PARPi treated ovarian cancers [31]. These illustrate that targeted selective pressures induce selection of specific alleles. Despite this, the literature is incomplete and we do not yet know what contributes to relapse in most cases. Little progress has been made in the capacity to predict the genetic makeup of a tumour sample over time and under drug administration.

1.1.1 Detection and measurement of evolutionary forces

One way to quantify the strength of selective pressure at the gene level is by measuring the rate of non-synonymous to synonymous mutations (dN/dS also known as Ka/Ks) [32]. Assuming that the infinite sites model holds, the probability of a mutation is uniform across the genome. We would expect that in the absence of any selection, dN/dS ≈ 1, since no functional change should be advantageous or deleterious to the cell. dS (Ks) could be thought of as representing the rate of background mutation. Accurately estimating dN/dS is challenging. Multiple biases have to be corrected for, including the differential rate of mutation based on the upstream and downstream nucleotides of a site (i.e., context), structural changes, and in the absence of normal matching tissue, polymorphic sites. Additionally, very deep targeted sequencing would be required to ensure a reasonable number of mutations are recovered [33, 34]. A recent method demonstrating these issues uses a tri-nucleotide context-aware substitution model with 192 parameters to overcome some of the mentioned biases, but it is only able to achieve statistical power at the level of cohorts of patients and not individuals, limiting its clinical feasibility [12].

Another approach is to assume neutral evolution in an exponentially growing population when the inverse allele frequency, $1/f$, has a linear relationship with the expected number of mutations M [7, 35]. According to this framework, deviation from linearity between $1/f$ and M is evidence of the presence of selection. An extension of this work allows for quantifying selection magnitudes for up to two subpopulations [8]. Frequency of mutations are detected using bulk WGS of a single timepoint, the state of the tumour at the time of a biopsy. The method in [7] is predicated upon the observation that branching processes [36], a family of stochastic models intensely used in population genetics, in the presence of some modes of selection, fails to generate data that show a linear $1/f$ and M relation.

A promising study concerns timeseries monitoring of cell free DNA in 45 patients with colorectal cancer and wildtype *RAS*, treated with Cetuximab [13]. Plasma samples were collected monthly until disease progression state was determined. One application was to predict time to disease progression. The tumour in each patient was partitioned into sensitive and resistant subpopulations. The tumour burden (the total number of cancerous cells) was modelled as the sum of the exponential decay and the exponential growth of the sensitive and resistant subpopulations respectively, each parameterized by the initial population size and a decay (growth) rate. A key promise of this study is the relative ease of obtaining plasma samples from patients as compared to solid biopsies. This allows more frequent sampling and forming of longitudinal timeseries. A drawback is that the sequencing of the plasma samples was limited (a panel of predefined 77 oncogenes and genomic regions), which in turn hindered the ability to detect all existing subpopulations. The results were also focused on one type of cancer and the generalisability to other cancer types is unknown.

An experimental system, the CRISPR-knockout screens [37, 38], allows the abundance of subpopulations (genotypes) of isogenic backgrounds to be observed, where each harbours specific mutations induced by CRISPR-Cas9. Including a set of non-targeting sgRNAs establishes a baseline for neutral growth, i.e., the expected change in the abundance of putatively neutral genotypes. In such assays, depletion or expansion of a genotype hints at negative or positive selection respectively. An improved version of these screens incorporates unique molecular identifiers (UMIs) [39] into the sgRNAs allowing for quantitative interpretation of the effective population size of each expanded genotype, unaffected by PCR duplicates [40].

Inferring population genetics parameters with high confidence based on only one timepoint is difficult and the associated uncertainty may render making meaningful conclusions implausible [41]. For one, as shown in [41], models of evolution with selection can also produce a linear $1/f$ and M relationship. Birth, death and mutation rates, population size and number of generations are all unknown and comprise too many degrees of freedom to fit to one noisy observation of the endpoint of a stochastic process. The signal for clones existing at similar prevalences is obscured in bulk WGS data and the frequency of alleles is further skewed by CNAs, which are difficult to resolve using bulk WGS data [23, 42]. Analysis based on a single sample is particularly prone to sampling bias due to the spatial heterogeneity of the tumour and may result in loss of signal. Low sequencing depth, characteristic of public libraries denies access to more recent events which further dilutes the signal. These limitations combined suggest that bulk WGS at a single-timepoint may be insufficient to quantify fitness values with high confidence.

1.2 Measurement of clonal abundance over time

Longitudinal studies, including immortalised cell lines and xenoengraftment experiments, allow tracking changes in the relative abundances of clusters of mutations [43], as well as Rx dosage experiments to induce resistance with subclinical dosing. These dynamics appear to be reproducible, raising hope that they may be predictable. Patient derived xenograft (PDX) systems are an effective model to study timeseries of a human tumour cell population [44, 45]. By serially transplanting a patient derived tumour sample into highly immunodeficient mice, it is possible to continuously monitor how tumour composition may evolve in a patient over time. CRISPR-Cas9 has made directed evolution experiments possible in both cell lines and PDX systems [40, 46]. Specific genomic contexts could be set up to study gene interactions in the presence of genetic perturbation induced by CRISPR-Cas9.

1.2.1 Leveraging single cell sequencing

Advances in SCS permits the investigation of tumour genetic composition with unprecedented accuracy and resolution. SCS obviates the need for complex computational deconvolution methods and allows the direct observation of the genotypes of the extant subclones. This in turn allows the use of models from the well established discipline of population genetics to be adapted to power quantitative reasoning about the future trajectory, dominance, and depletion of subpopulations. Reproducible clonal dynamics from serially propagated PDX models with validated clonal genomes at single cell resolution has been established [43]. Motivated by this observation, we propose to investigate the hypothesis that clones that grow reproducibly have higher fitness.

Real-time monitoring of evolution of cancerous cells at the single cells CNA resolution is now feasible [20, 21]. These advances in single cell whole genome sequencing enables population genetics analysis of cancer progression. It has been shown that single cell sequencing is essential for resolving genomically defined clones in high-grade serous ovarian cancer (HGSC) [17] and in patient derived breast cancer xenografts [43]. These studies employed multiplexed PCR approaches on amplified material targeting small sets of mutations assayed with Fluidigm Access arrays. To enable whole genome single cell sequencing, bias limitations in genomic depth [47] and breadth [48] have been mitigated. A direct tagmentation approach [21] (direct library prep (DLP)) implemented in a microlithography microfluidic device obviates pre-amplification and results in a superior representation of the copy number architecture of single cancer cells to below 1Mb resolution [20].

Quantitative attributes of fitness are central to developing predictive models of cancer evolution. Such models currently elude the field due to a lack of appropriate timeseries measurements and generative analytic methods. The goal is to predict unobserved clonal trajectories over time by inferring the evolutionary dynamics underlying disease progression.

1.2.2 Clone identification via phylogenetic analysis

A key step in determining the quantitative fitness attributes of cancer cells is to resolve the subpopulation structure. Cancer cells are evolutionarily related and phylogenetic reconstruction aims at recovering this relationship. A fundamental assumption is that cells that descend from the same ancestor harbour similar mutations. While single cell whole genome sequencing (scWGS) has an advantage over bulk WGS in that it obviates the need to computationally determine which mutations co-occur, the evolutionary history of the cells is not directly observed. Sorting cells into groups enables us to track the change in their abundance over time. A principled way to group cells is according to their phylogenetic relationships. Many phylogenetic inference methods are tailored for using point mutations as input and assume a small number of leaf nodes [49]. However, the emerging single cell data produce up to thousands of single cell genomes and are suitable for determining CNAs [20, 21].

1.3 Research contributions

This dissertation is organised as follows. In Chapter 2, we review and benchmark `sitka`, a scalable Bayesian phylogenetic tree inference method and introduce `Lumberjack`, a tree-cutting algorithm to identify clones from `sitka` inferred trees. In Chapter 3, we introduce and develop the `fitClone` model, a Bayesian fitness inference framework for timeseries data and establish its performance via simulation studies. Finally, in Chapter 4, we apply methods developed in the previous chapters on timeseries data from in vitro and in vivo model systems with and without treatment. Figure 1.1 shows a diagrammatic representation of this dissertation. Below we give a summary of the chapters to come.

1.3.1 Chapter 2: `sitka`: benchmarking and visualisation of a scalable Bayesian phylogenetic inference method

A resolved clonal structure of a heterogeneous tumour cell population sampled over multiple timepoints is a key input to the `fitClone` model. In many modern single cell DNA sequencing scDNAseq platforms including DLP+ used in this work, the evolutionary relationship between single cells is not directly observed. `sitka` is a scaleable method that approximates the phylogeny in a Bayesian framework. In this chapter we first review the `sitka` probabilistic model and its inference procedure. We then describe its workflow, starting from preprocessing its inputs to visualising its outputs. Next we exemplify `sitka` on three real-world cancer datasets, namely two from TNBC-PDX timeseries and one multi-region pre and post treatment HGSC sample. We review some existing baseline methods and compare them to `sitka` using an accuracy test that we derive. Finally we introduce `Lumberjack`, a method for cutting the phylogenetic tree to assign cells to clones determined by their CNA profiles.

1.3.2 Chapter 3: Modelling fitness in longitudinal data via an approximation to the diffusion model

In this chapter we develop a probabilistic approach that is a mechanistic model of tumour growth based on the Wright-Fisher diffusion approximation. We introduce the `fitClone` probabilistic framework, its observation model and inference procedure. We then describe an extension, the conditional sampler, that would allow inference for a larger number of clones. We review parameter estimation and key posterior summarisation procedures, specifically the posterior ordering matrix that can be used for pairwise comparison of clones and in a principled way establish whether a clone has higher fitness than another. We conclude by exemplifying the method over simulation studies and a real-world dataset.

1.3.3 Chapter 4: Application of the Wright-Fisher diffusion approximation to cancer model systems interrogated at single cell level

In this chapter we exemplify the computational framework that we have developed in previous chapters on real-world dataset to make inferences about how human cancer cells evolve at the copy number level. We address two key hypotheses, that (i) clonal dynamics in a pre-treatment triple negative breast cancer (TNBC) tumour is quantifiably reproducible, and that (ii) the fitness landscape is reversed under early response to cisplatin treatment. For all datasets, we use `sitka` and `Lumberjack` to infer phylogenetic trees and assign cells to clones, and then use `fitClone` to ascribe quantitative selection coefficients to individual cancer clones.

As a proof of principle, we first look at the dynamics of three related cell lines. We then examine two independent TNBC PDX samples and in one, TNBC-SA609, we validate our selection coefficient predictions in-vivo via a competition experimental design that addresses hypothesis (i). To tackle hypothesis (ii) we look at three independent TNBC PDX systems, with and without treatment. We then explore in more detail the dynamics of the TNBC-SA609 system. We also look at a drug holiday experimental design, in which treatment is stopped after a few timepoints, that suggests that withholding treatment may reset the clonal composition of the tumour to a pre-treatment state.

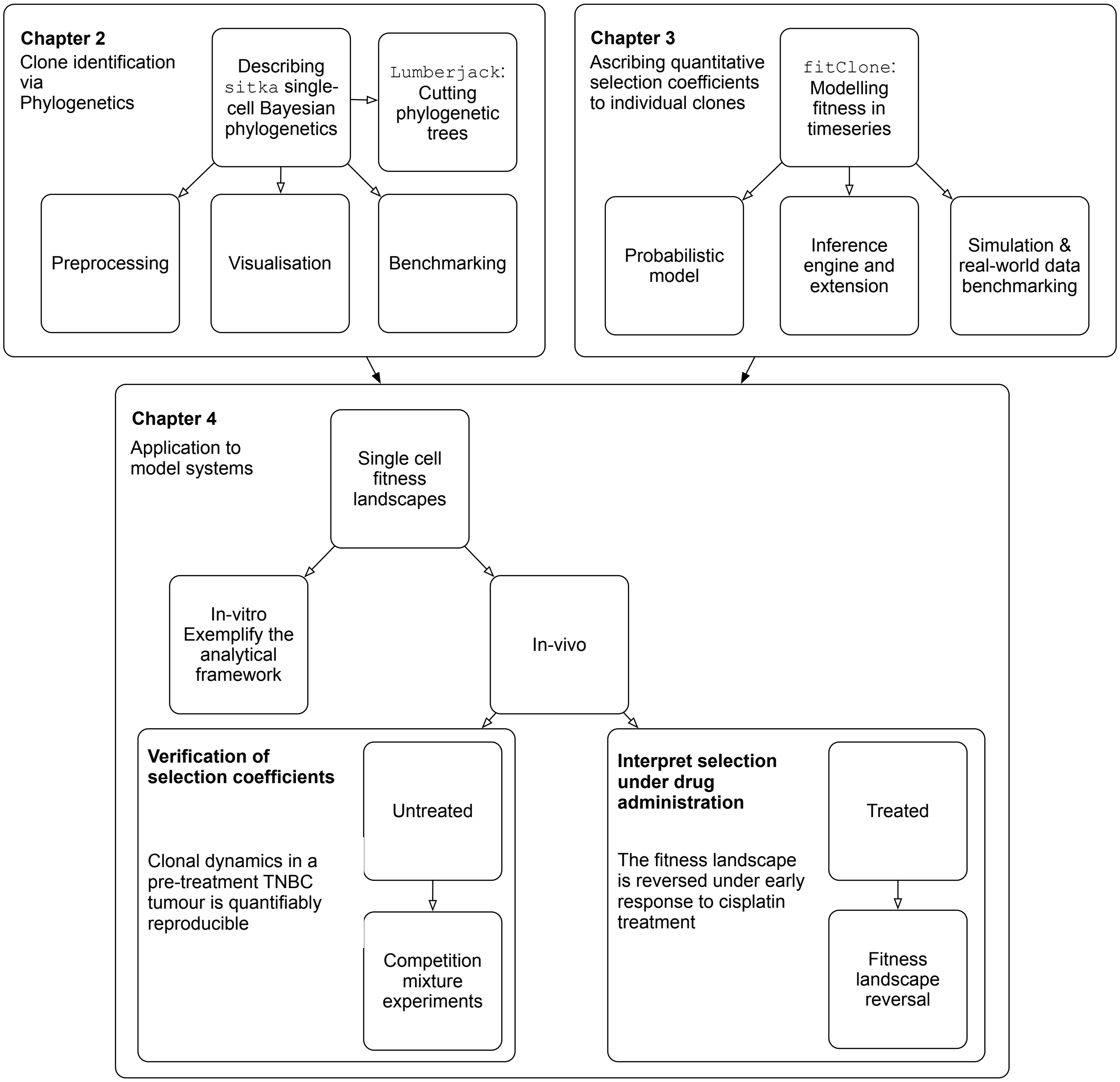

Chapter 2
Clone identification via Phylogenetics
Describing sitka single-cell Bayesian phylogenetics
Lumberjack: Cutting phylogenetic trees
Preprocessing
Visualisation
Benchmarking
Chapter 3
Ascribing quantitative selection coefficients to individual clones
fitClone: Modelling fitness in timeseries
Probabilistic model
Inference engine and extension
Simulation & real-world data benchmarking
Chapter 4
Application to model systems
Single cell fitness landscapes
In-vitro Exemplify the analytical framework
In-vivo
Verification of selection coefficients
Clonal dynamics in a pre-treatment TNBC tumour is quantifiably reproducible
Untreated
Competition mixture experiments
Interpret selection under drug administration
The fitness landscape is reversed under early response to cisplatin treatment
Treated
Fitness landscape reversal

Figure 1.1: A diagram summarising the organisation of the thesis.

Chapter 2

Benchmarking and Visualisation of a Scalable Bayesian Phylogenetic Inference Method

2.1 Introduction

A main challenge in investigating cancer evolution is the need to resolve the subpopulation structure of a heterogeneous tumour sample. Advances in next generation SCS as well as longitudinal xenoengraftment has enabled more accurate, quantitative measurements of tumours as they evolve. Phylogenetic reconstruction is central to identifying clones over multiple sampled time-points. Moreover, the ability to sequence tens of thousands of single genomes at high resolution per experiment [20] calls for tailored phylogenetic models with scalable inference algorithms.

Single cell cancer phylogenetics is an evolving field. Multiple approaches, spanning different study designs and data sources are reviewed in [50]. Many phylogenetic inference methods are tailored for using point mutations as input and assume a small number (on the order of 10s) of leaf nodes [49, 51–53]. However, the emerging single cell data produce up to thousands of single cell genomes and are suitable for determining CNAs [20, 21]. Distance based and agglomerative clustering methods such as neighbour joining are scalable and are used to elucidate hierarchical structures over cells [54, 55]. While useful heuristics, these methods are statistically sub-optimal relative to likelihood based methods [56].

We describe `sitka`, a phylogenetic model and associated Bayesian inference procedure which exploits the specifics of scWGS data. `sitka` enjoys a number of attractive characteristics, including (i) a novel phylogenetic encoding of CN data providing a statistical-computational trade-off by simplifying the site dependencies induced by rearrangements while still forming a sound foundation to phylogenetic inference, and (ii) an innovative phylogenetic tree exploration move which makes the cost of MCMC iterations bounded by $O(|C|+|L|)$, where $|C|$ is the number of cells and $|L|$ is the number of loci. In contrast, existing off-the-shelf likelihood-based methods incur an iteration cost of $O(|C|\ |L|)$, and (iii) the novel move considers an exponential number of neighbouring trees whereas off-the-shelf moves consider a polynomial size set of neighbours. A key piece is to add visualisation tools to the `sitka` toolbox and this is a main contribution of this chapter. Visualisation techniques help explore, summarise and communicate the tree inference results.

In this chapter, we will first give a high level description the `sitka` model (Section 2.2) and review its analysis workflow (Section 2.3). We will present the quality control (QC) of single cells, different encodings of the input space (preprocessing), and then formally describe the model (Section 2.3.4) and an MCMC based inference procedure (Section 2.3.5). We will introduce two novel visualisation views to summarise the tree inference results (Section 2.3.7). We will compare `sitka` with other tree-inference methods that scale to scWGS datasets over three real-world datasets (Sections 2.4 and 2.5). Finally we introduce `Lumberjack`, an algorithm for cutting a phylogenetic tree to determine clonal subpopulations (Section 2.6).

2.2 The sitka model

`sitka` is based on lossy transformation of single cell copy number matrices retaining only presence or absence of changes in copy number profiles. This transformation turns a complex evolutionary process (integer-valued copy numbers, prone to a high degree of homoplasy and dense dependence structure across sites) into a simpler one which can be approximated by a probabilistic version of a perfect phylogeny (see Figure 2.1). We leverage the special structure created by the change point transformation to build a special purpose MCMC kernel which has better computational scalability per move compared to classical phylogenetic kernels.

The input data for `sitka` can be visualised in a colour-coded matrix exemplified in Figure 2.1-**a**. Each row in the matrix corresponds to an individual cell that has been sequenced in a single-cell platform. Each column in the matrix is a locus that is represented by a bin, a contiguous set of genomic positions. We assume that the integer copy number of each bin has been estimated

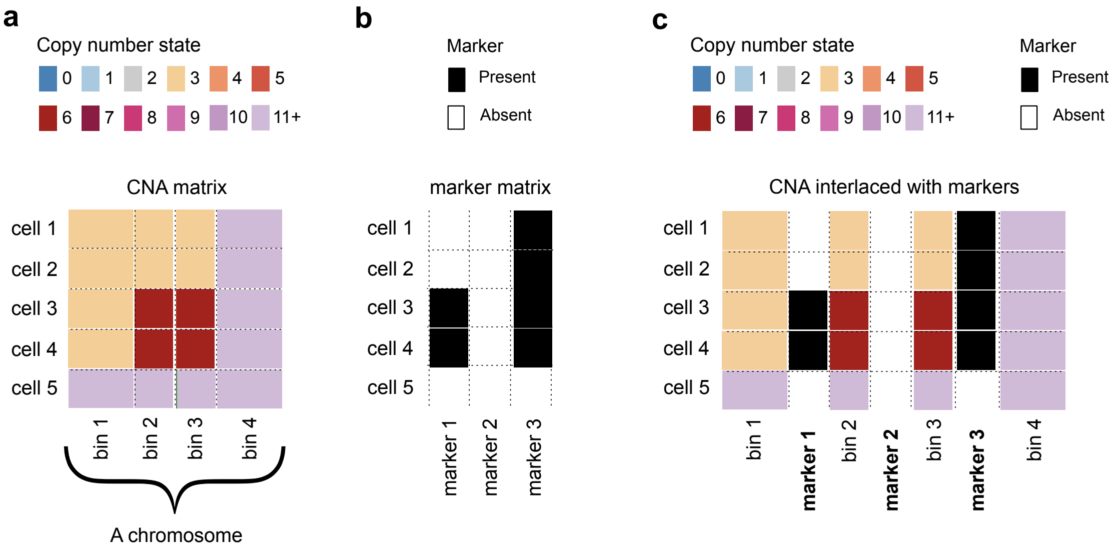


Figure 2.1: Description of the process involved in construction of *markers*, the input to the `sitka` model. A *bin* is a contiguous set of genomic positions. Each pair of consecutive bins (e.g., bins 1 and 2 in (**a**) is associated with a *marker* (e.g., marker 1) that measures for each individual cell, whether there is a difference between the CNA states of the two bins. (**a**) The observed CNA matrix for a subset of bins on a chromosome. The rows are sequenced single cells, and the columns are bins. The CN states are colour-coded. (**b**) The three markers shown are associated with the four bins. Each marker records the presence (black) or absence (white) of a CN state change between a pair of consecutive bins. Note that in the CNA matrix, there is a CN change at row 3 from bin 1 to bin 2 (CN state 3 to 6). This is reflected in the marker matrix, at row 3 of marker 1 with a black square. There are no changes between bins 2 and 3 across any rows in the CNA matrix. This is reflected in marker 2 comprising all white squares. (**c**) For visualisation purposes, the CNA matrix can be interlaced with the marker matrix to more clearly show where the CNA changes occur. Each column of the marker matrix is inserted between the associated pair of columns in the CNA matrix. The resulting matrix is an example of an *augmented* view that combines data from two or more sources (here the CNA matrix and the marker matrix). In an augmented view, we call columns from each source a *channel*.

(i.e., called) as a preprocessing step, e.g., using a hidden Markov model tool such as in [21]. In Figure 2.1-**a** the copy number state is encoded by the colour of each entry in the matrix.

The output of `sitka` includes two types of trees. The first is the tree used for MCMC sampling in the inference procedure (type I), and the second is the tree used in visualisation (type II) and can be derived from the first type. Briefly, the first type is a directed tree that spans all cells and CN change points (markers) under study in addition to a virtual root node. If a cell vertex is attached to a marker vertex, we hypothesize that the cell harbours only markers in the shortest path between its parent and the root node in the tree. A cell vertex, attached to the root node, is hypothesized to harbour none of the markers. If a marker vertex is a descendent of another marker vertex, we hypothesize that the trait that the former indexes, emerged in a cell that was a descendant of the cell that the latter indexes. Note that in this type of tree, some marker vertices could be leaf nodes. Such markers are hypothesised to violate the assumptions of the model and were not consistent with any cells. All cell vertices will be leaf nodes. We describe the type I tree more formally in Section 2.3.4 when we introduce priors on trees.

Type II tree is a transformation of type I tree. We remove from the type I tree all marker nodes that are leaf nodes, i.e., markers that are not present in any cells. We also collapse into a single node, the list of connected marker nodes that have exactly one descendent (i.e., chains). Figure 2.2 shows a small *type I tree*, its transformation to a *type II tree* and the respective marker matrix. We visualise the input matrix and the estimated tree simultaneously by sorting the individual cells (rows of the matrix) in such a way that they line up with the position of the corresponding leaves of the tree.

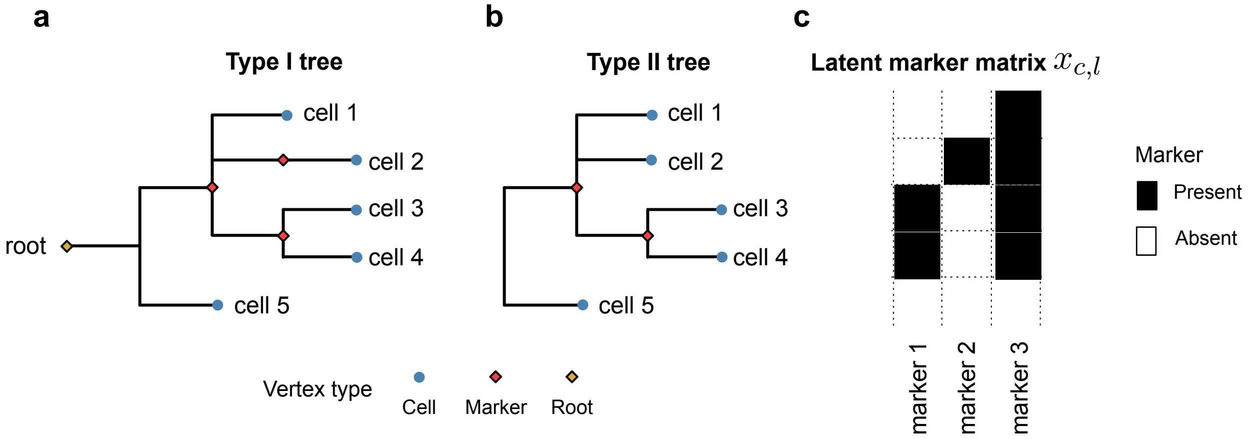


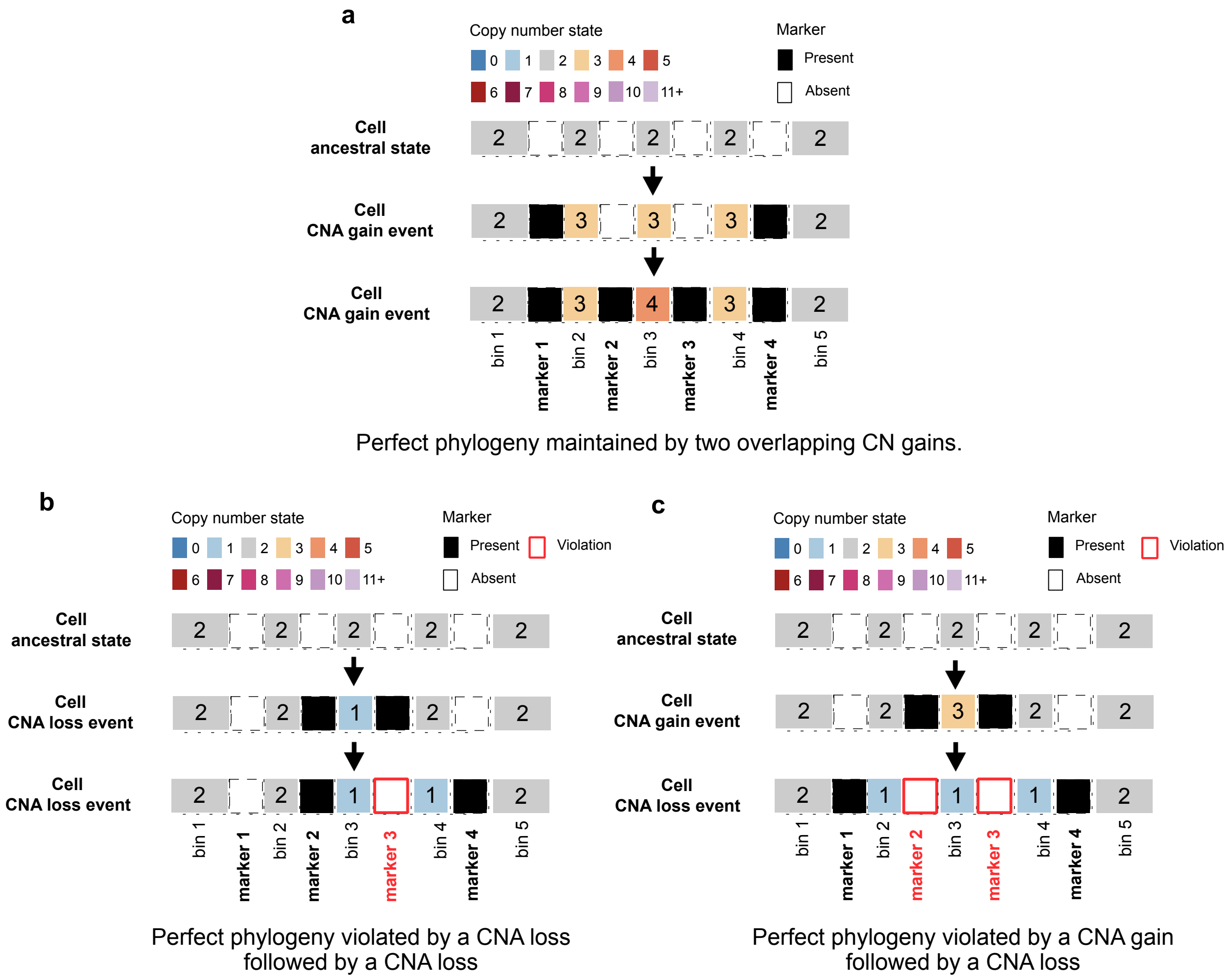
Figure 2.2: Visualisation of a small type I tree t (**a**), its transformation into a type II tree (**b**), and the corresponding marker matrix $x = (x_{c,l})$ (**c**). Given a tree t, the latent marker matrix x is a deterministic function $x = x(t)$. We compute $x : t \to \{0,1\}^{C \times L}$ by setting $x_{c,l} = 1$ if the single-cell c is a descendent of the marker node l in tree t, and otherwise $x_{c,l} = 0$. Note that the clade comprising single-cells 3 and 4 has support in both markers 1 and 3. For clarity, we do not visualise type I trees, but plot their transformation, i.e., type II trees as follows. We remove from the type I tree all marker nodes that have $x_{c,l} = 0$ for all single-cells c. Lists of connected edges that have exactly one descendent (i.e., chains) are also collapsed into a single edge, e.g., the edge corresponding to markers 2 and 3 are collapsed into one edge (since marker 2 has only one descendent, namely single-cell 2).

`sitka` is predicated on the perfect phylogeny assumption where each phylogenetic trait arises exactly once on a rooted tree topology and that all individual cells descending from that position on the tree will inherit that trait. The perfect phylogeny model allows us to use CNA change points, i.e., markers, instead of integer CNA states as phylogenetic traits. When using change points as phylogenetic traits, non-overlapping CNA events do not break the perfect phylogeny assumption. Figure 2.3 shows examples of overlapping CNA events and their effect on markers. The two scenarios that can lead to the violation of the perfect phylogeny assumption are (i) when a CNA gain event is followed by an overlapping loss event, or (ii) when a loss event is followed by an overlapping loss event, and the second event removes either end-point of the first event. Note how for a violation to occur, the second overlapping event has to happen on the same copy as the first event.

Imposing a perfect phylogeny on the *observed* change points is restrictive, as violations of the assumptions (e.g., due to homoplasy) or an error in observing a single trait can lead to considerable change to the inferred tree. To improve the robustness of the perfect phylogeny construction, we use an observation model described in Section 2.3.4. A central modelling assumption is that the perfect phylogeny holds for the latent states (denoted by $x_{c,l} = x_{c,l}(t)$ in Section 2.3.4), but not for the observed markers ($y_{c,l}$ in Section 2.3.4). Using this formulation, incorrect change point estimates (as a result of noisy CNAs calls) as well as markers inconsistent with the perfect phylogeny assumption can be accounted for as noise in the observations.

2.3 Workflow

We begin by describing the preprocessing and lossy transformation of the input CNA states and then briefly describe the inference procedure. Figure 2.4 shows the workflow of the phylogenetic inference in `sitka`.



Perfect phylogeny maintained by two overlapping CN gains.

Perfect phylogeny violated by a CNA loss followed by a CNA loss

Perfect phylogeny violated by a CNA gain followed by a CNA loss

Figure 2.3: The effects of overlapping CNA events on the perfect phylogeny assumption. A segment of a chromosome with five consecutive bins and their four corresponding markers are shown. Each panel follows the CN states interlaced with markers for a cell at the ancestral state (top), after a CNA event (middle), and after a second overlapping CNA event (bottom). The numbers in the CNA squares show the integer CN state (e.g., the ancestral state has two copies of the 5-bins long segment). (**a**) Two overlapping CNA gains maintain the perfect phylogeny assumption. By the infinite site argument, it is unlikely for the end-points of the two gain events to exactly match. The same argument holds for a CNA loss followed by a CNA gain event. Note that in these cases, once a change point is acquired, it is not lost. (**b**) If a loss event is followed by another loss event in which either end-points of the first event is removed, the perfect phylogeny assumption will be violated (e.g., marker 3 is lost after the second loss event). Note that a violation does not occur if the loss events hit different copies of a segment. (**c**) Similarly, if a gain event is followed by a loss event, only if the latter erases the end-points of the former is the perfect phylogeny violated. Note how marker 2 and marker 3 are lost after the second CNA event.

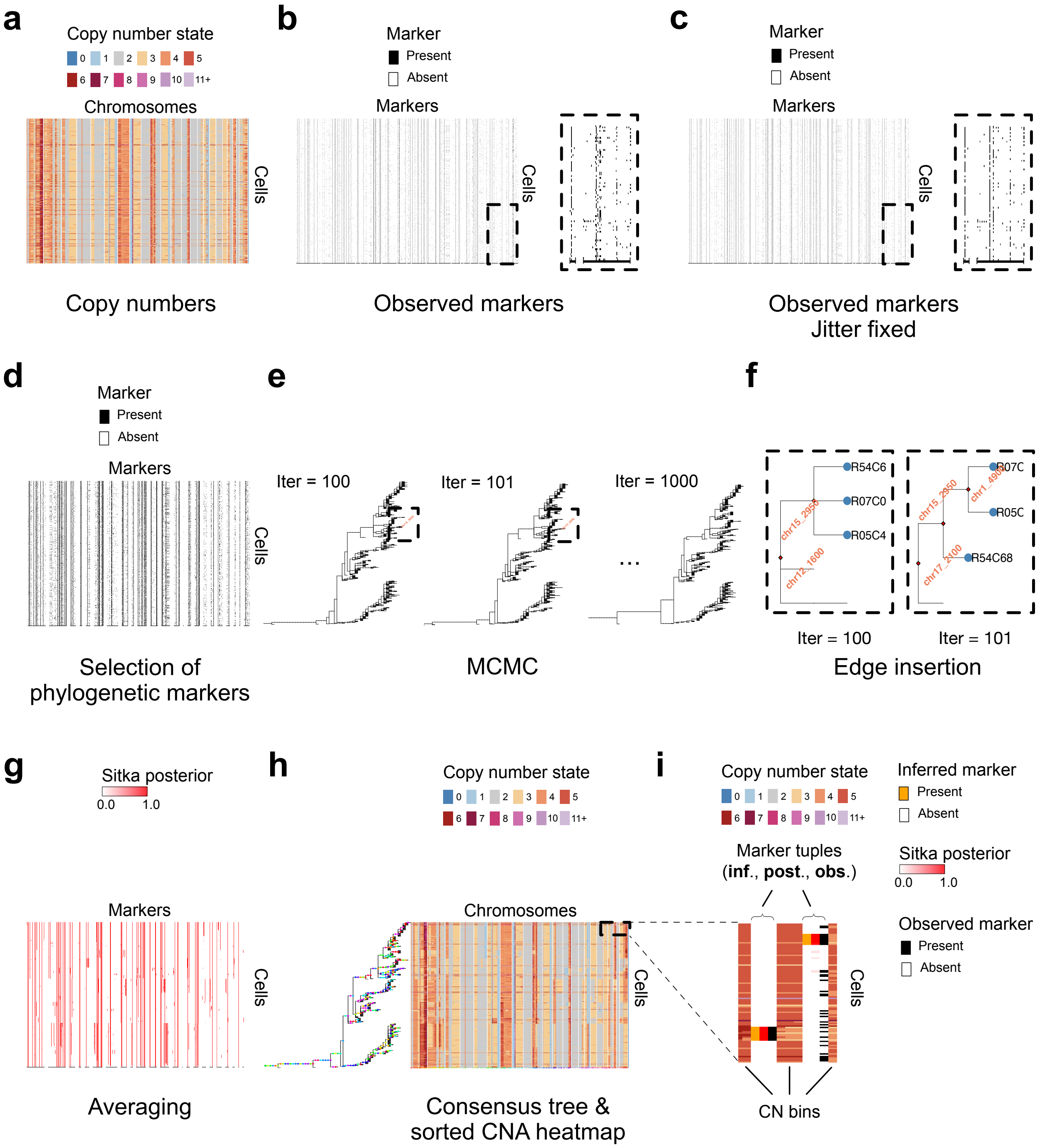

a
Copy number state
0 1 2 3 4 5
6 7 8 9 10 11+
Chromosomes
Cells
Copy numbers
b
Marker
Present
Absent
Markers
Cells
Observed markers
c
Marker
Present
Absent
Markers
Cells
Observed markers
Jitter fixed
d
Marker
Present
Absent
Markers
Cells
Selection of
phylogenetic markers
e
Iter = 100
Iter = 101
Iter = 1000
...
MCMC
f
R54C6
R07C0
R05C4
chr15_2950
chr12_1600
R07C
R05C
R54C68
chr15_2950
chr1_4900
chr17_2100
Iter = 100
Iter = 101
Edge insertion
g
Sitka posterior
0.0 1.0
Markers
Cells
Averaging
h
Copy number state
0 1 2 3 4 5
6 7 8 9 10 11+
Chromosomes
Cells
Consensus tree &
sorted CNA heatmap
i
Copy number state
0 1 2 3 4 5
6 7 8 9 10 11+
Inferred marker
Present
Absent
Sitka posterior
0.0 1.0
Observed marker
Present
Absent
Marker tuples
(inf., post., obs.)
Cells
CN bins

Figure 2.4 *(previous page)*: Workflow of `sitka`. (**a**) The input is the CNA data from a heterogeneous single-cell population. The rows of the CNA matrix are permuted at random. (**b**) CNA change points, i.e., markers are obtained by applying a lossy binary transformation to the CNA matrix (Section 2.3.2 and Figure 2.1). Note that each single-cell is now represented by the presence or absence of CN changes between consecutive bins. (**c**) A shared change point is sometimes shifted by a few bins across different cells. We use a heuristic (Section 2.3.3) to correct for this marker misalignment. Note how the columns in the inset in panel-**c** are less noisy than their counterpart in panel-**b**. (**d**) As a computational trade-off, only a subset of markers present in at least 5% of cells are chosen for phylogenetic inference. (**e**) An MCMC algorithm explores the tree space. (**f**) An example of an edge insertion MCMC move. (**g**) The posterior probability of latent state $x_{c,l}$ is computed by averaging over all MCMC trees, summing over the number of samples in which a single-cell c was a descendant of a marker node l (Section 2.3.6). (**h**) The rows in the CNA matrix in panel-**a** are sorted according to the inferred consensus (type II) tree, shown on the left of the matrix. Note that the block structure of the CNs are clearly visible as single-cells with similar CN profiles are grouped together. (**i**) The inset shows tuples of marker columns, namely **inf**. (inferred markers, i.e., latent state $x_{c,l}$), **post**. (posterior probability of the latent state $x_{c,l}$), and **obs**. (observed markers), interlaced with the CN columns (similar to Figure 2.1). The results are from the $SA535$ dataset, a triple negative breast cancer patient derived xenograft sample (Section 2.4).

2.3.1 Filtering

The raw data contains cells that are either contaminated (e.g., contains biological material from mice) or have undesired sequencing artefacts. These include cells that were captured for DNA sequencing when undergoing mitosis. Since the `stka` model does not account for such phenomena, the filtering is an important step. Figure 2.5 shows the steps taken from pulling the raw data to the CNA integer matrix ready for `sitka` transformation. See Section A.4 for detailed descriptions of each step in the filtering process. Briefly, we remove control cells, cells with highly-noisy CN calls, and cells that have very few mapped reads. We also remove copy number bins that lie in difficult to sequence regions of the genome (bins with low-mapability). Finally, we drop cells that, based on their CNA profile, are suspected to be cycling cells that were not detected in previous steps.

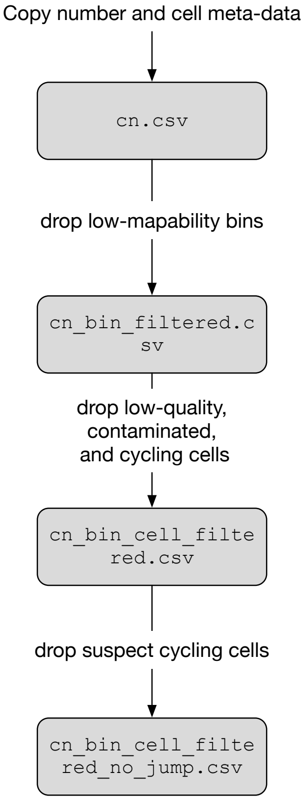

Copy number and cell meta-data
cn.csv
drop low-mapability bins
cn_bin_filtered.c sv
drop low-quality, contaminated, and cycling cells
cn_bin_cell_filte red.csv
drop suspect cycling cells
cn_bin_cell_filte red_no_jump.csv

Figure 2.5: Filtering the CNA data for tree inference.

2.3.2 Binarising

To obtain the $C \times L_{\text{Markers}}$ phylogenetic markers matrix y that comprises the input to the `sitka` model, we apply a lossy transformation to the $C \times L_{\text{Bins}}$ CNA matrix a that involves computing the change in copy number state between two consecutive bins. Below we introduce two different ways to compute the binarisation which differ on whether any artificial bins are included or not. Figure 2.1 shows a small CNA matrix and its corresponding transformation into the marker matrix. For brevity, in what follows we assume that only one chromosome is used, so that $L_{\text{Bins}} = L$ and $L_{\text{Markers}} = L_{\text{Bins}} - 1$. In practice, we use all available chromosomes, and $L_{\text{Markers}} = L_{\text{Bins}} - N_{\text{Chr}}$ where N_{Chr} denotes the total number of chromosomes used.

Basic

In the basic binarisation process, the markers are computed as follows. Given a filtered cell-by-locus matrix a (`cn_bin_cell_filtered_no_jump.csv`), we sort bins by `(chr-name, start-position, end-position)`. Then in each chromosome, we compute markers as the binarised difference between consecutive bins, and drop duplicated markers, i.e., keep one marker in a set of markers with identical binary patterns across cells. In other words, $y = (y_{c,l'})$ and $l' \in \{1, \ldots, L-1\}$, and

$$y_{c,l'} \coloneqq \mathbf{1}\left(\left|a_{c,l'} - a_{c,l'+1}\right| > 0\right) \tag{2.1}$$

where $\mathbf{1}(x)$ is the indicator function.

Sync

Some datasets exhibit synchronous copy number changes in which at least two distinct cell subpopulations have a copy number change at a bin across a subset of cells, where either the direction or the magnitude of the change is different in at least one cell. That is, there exists i, c, c' such that:

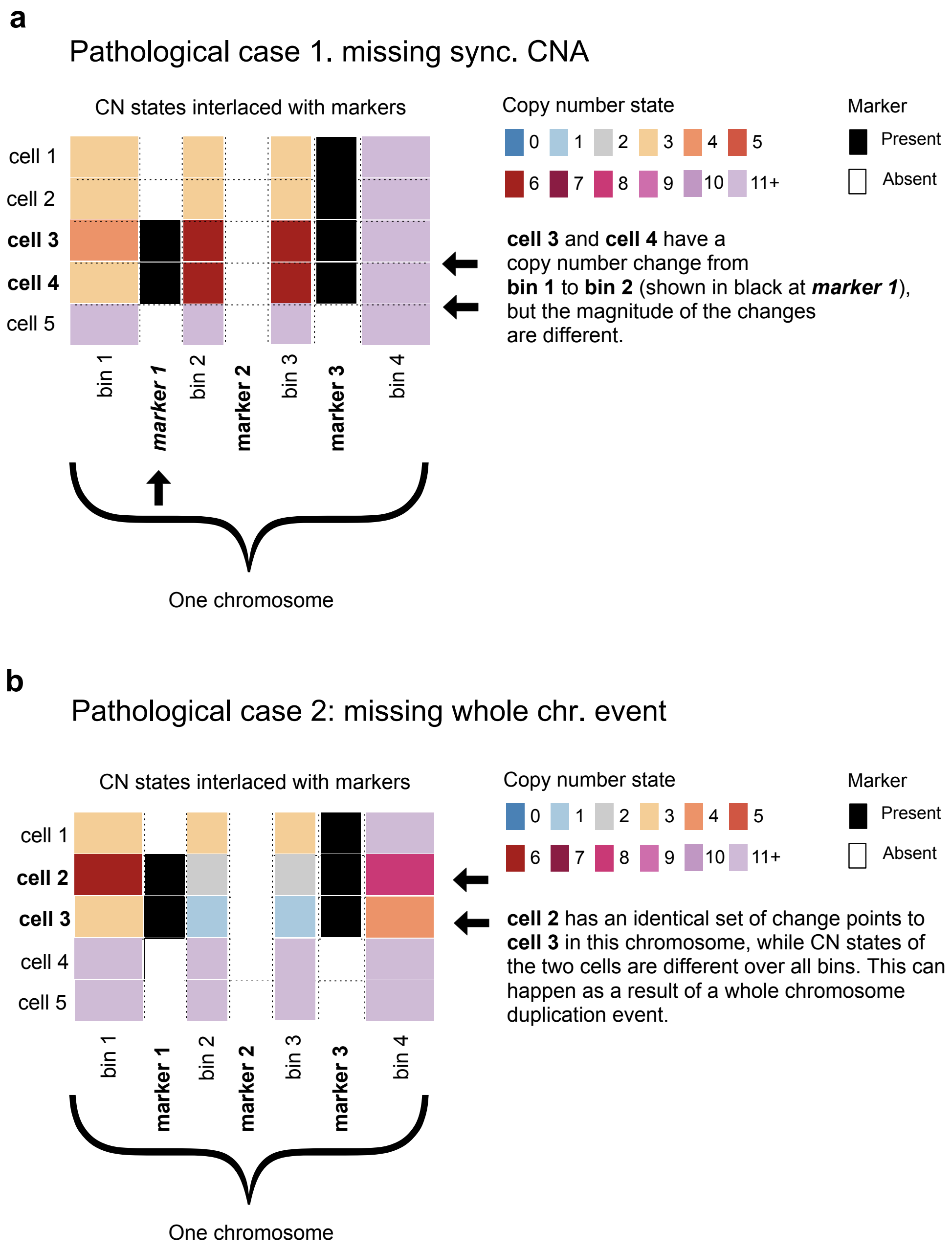
$$\big(y_{c,i} = y_{c',i} = 1\big) \quad \wedge \quad \big(a_{c,i} - a_{c,i+1} \neq a_{c',i} - a_{c',i+1}\big) \tag{2.2}$$

for $i \in \{1, \ldots, L-1\}$ and $c, c' \in \{1, \ldots, C\}$ and a and y are the CNA and marker matrices respectively. Figure 2.6 shows examples of synchronous copy number changes. The basic binarisation method will be unable to distinguish such copy number induced subpopulations. The *Sync* procedure aims to alleviate this by inserting the following additional marker columns in matrix y (the result of the basic encoding above):

1. Find columns $b \coloneqq a_{1:C,i}$ that satisfy Equation (2.2).

2. For each b found above, add a virtual column b' with its CN state set to the most frequent CN state in b.

3. Add to the matrix y a column e whose elements are computed as the binarised difference between b and b', that is, for $c \in \{1, \ldots, C\}$ set $e_{c,1} \coloneqq \mathbf{1}\left(\left|b_{c,1} - b'_{c,1}\right| > 0\right)$.



a

Pathological case 1. missing sync. CNA

cell 3 and **cell 4** have a copy number change from **bin 1** to **bin 2** (shown in black at ***marker 1***), but the magnitude of the changes are different.

b

Pathological case 2: missing whole chr. event

cell 2 has an identical set of change points to **cell 3** in this chromosome, while CN states of the two cells are different over all bins. This can happen as a result of a whole chromosome duplication event.

Figure 2.6: Synchronous CNA changes. If two distinct cell subpopulations have a copy number change at a bin across a subset of cells, where either the direction or the magnitude of the change is different in at least one cell (a synchronous CNA change), the basic binarisation method will be unable to distinguish the induced subpopulations. Black and white represent the presence or absence of a CN change (marker) respectfully. Panels (**a**) and (**b**) show cases where synchronous or whole chromosome events are missed.

2.3.3 Fixing jitter and selection of phylogenetic markers

The copy numbers available to us in this work are estimated independently for each cell. This is one reason why the start position (bin) of the same CN change event may be slightly different across cells, generating some *jitter*. We address this by enumerating each change point column and moving the jitter back to the most frequent bin. Concretely,

Algorithm 1 JitterFix

```
procedure JITTER-FIX($y$, $k$)
    column-queue $\leftarrow$ OrderByDensityDecreasing($y$)
    columns-visited $\leftarrow \{\}$
    for column-index $c$ in column-queue do
        neighbours $\leftarrow$ neighbours $(c, y, k)$
        for column-index $n$ in neighbours do
            if $n \notin$ columns-visited then
                $y_{1:C,c} \leftarrow y_{1:C,c} \vee y_{1:C,n}$
                $y_{1:C,n} \leftarrow 0$
                columns-visited $\leftarrow$ columns-visited $\cup\ n$
    return $y$
```

The function neighbours (i, y, k) returns $2k$ columns (if they exist), half, immediately to the left, and half, immediately to the right of column i in matrix y. We use $k = 2$ based on visual inspection. An example of the result of the jitter correction heuristic is shown in Figure 2.4 panel **c**. As a computational trade-off, only a subset of markers present in at least a minimum number of cells are chosen for phylogenetic inference. That is, we removed columns l in y with relative density $\sum_{c \in C} y_{c,l}/|C|$ less than a threshold, set to 5%. Larger values of this threshold may lead to less resolved clades in the inferred tree.

2.3.4 Model

The `sitka` model starts with the perfect phylogeny assumption for the latent variables $x_{c,l}$ but allows deviation from it by allowing noisy observations $y_{c,l}$. In a perfect phylogeny model, each phylogenetic trait arises only once on the rooted tree topology, and all cells descending from that position will inherit that trait. Recall that we assume the latent states matrix is a deterministic function of the tree t, i.e., $x = x(t)$ (see Figure 2.2).

Let C and L denote the disjoint sets of cells and loci respectively. Assuming that a^* is the true CNA state matrix, then informally, we can think of the trait $x_{c,l}$ for cell $c \in C$ at marker $l \in L$ as in Equation (2.1), with a replaced with a^*. But a^* is not directly observed and therefore neither is $x_{c,l}$. We assume noisy estimates of CNA calls, $a_{c,l}$ are available. Then we define $y_{c,l}$ as in Equation (2.1). Informally, $y_{c,l}$ is a noisy version of $x_{c,l}$. Therefore, we posit an observation probability model $p(y \mid x, \theta)$, where x is the matrix $x = (x_{c,l})$, y is the data or a data summary, and θ are model parameters. To model errors in copy number states, we introduce false positive and negative rate parameters $r^{\mathrm{FP}} \in (0, 1)$ and $r^{\mathrm{FN}} \in (0, 1)$ respectively, and an error matrix

$$e^{r^{\mathrm{FP}},r^{\mathrm{FN}}} = \begin{bmatrix} 1 - r^{\mathrm{FP}} & r^{\mathrm{FP}} \\ r^{\mathrm{FN}} & 1 - r^{\mathrm{FN}} \end{bmatrix}, \tag{2.3}$$

$$p\left(y_{c,l} \mid x_{c,l}, r^{\mathrm{FP}}, r^{\mathrm{FN}}\right) = e^{r^{\mathrm{FP}},r^{\mathrm{FN}}}_{x_{c,l},y_{c,l}}, \tag{2.4}$$

from which we set:

$$p(y \mid x, \theta) = \prod_{l \in L} \prod_{c \in C} p\left(y_{c,l} \mid x_{c,l}, r^{\mathrm{FP}}_{c,l}(\theta), r^{\mathrm{FN}}_{c,l}(\theta)\right),$$

where the scalar $e^{r^{\mathrm{FP}},r^{\mathrm{FN}}}_{x_{c,l},y_{c,l}}$ takes values as in Table 2.1. Here we use a global parameterisation where the false positive and false negative functions $r^{\mathrm{FP}}_{c,l}(\theta)$ and $r^{\mathrm{FN}}_{c,l}(\theta)$ are shared across markers and cells. We considered a marker-specific parameterisation, but it incurs a higher computational cost and results in a similar performance and that is why we focus here on a global parameterisation.

$x_{c,l}$	$y_{c,l}$	$p_{c,l}$
0	0	$1 - r^{FP}$
0	1	r^{FP}
1	0	r^{FN}
1	1	1 - r^{FN}

Table 2.1: The false positive and false negative errors in `sitka`'s observation model.

In the global parameterisation, we have $\theta = (r^{\mathrm{FP}}_{\mathrm{global}}, r^{\mathrm{FN}}_{\mathrm{global}})$. Using a uniform prior distribution for both error rates can lead to pathological cases as shown in Figure 2.7. To avoid that, we use the following:

$$r^{\mathrm{FP}}_{\mathrm{global}} \sim \mathrm{Uniform}\left(0, \overline{r^{\mathrm{FP}}}\right),$$

$$r^{\mathrm{FN}}_{\mathrm{global}} \sim \mathrm{Uniform}\left(0, \overline{r^{\mathrm{FN}}}\right),$$

We set $\overline{r^{\mathrm{FP}}} = 1/10$ and $\overline{r^{\mathrm{FN}}} = 1/2$ as defaults in our experiments. We use a larger $\overline{r^{\mathrm{FN}}}$; although errors in approximating copy number states can lead to both types of errors, there are more mechanisms that can cause a false negative, including a copy number gain, followed by an overlapping copy number loss.

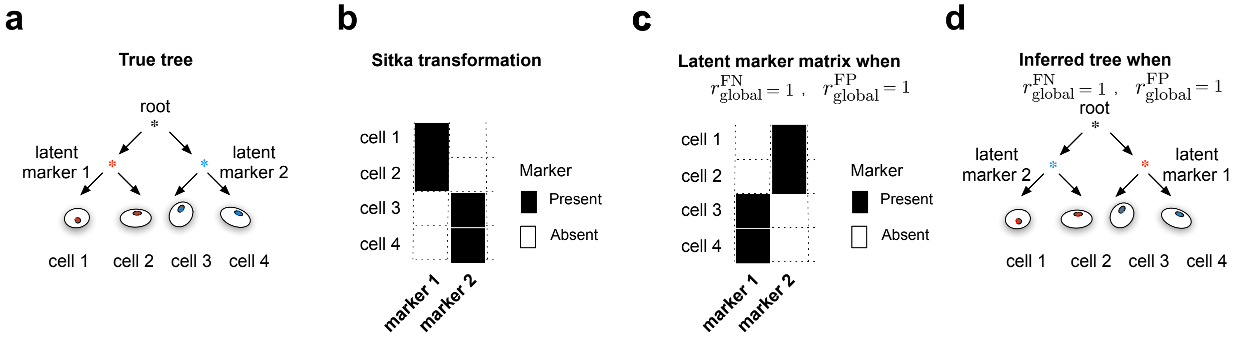


Figure 2.7: Pathological tree reconstruction under default observation prior. (**a**) The true tree reconstruction in a simple example with a balanced phylogeny with two clades of size two, and two unique markers, coloured red and blue, that distinguish the left and right clades respectively. (**b**) The binarised input matrix corresponding to the four cells at the two markers. The desired observation error rates should be zero and the latent and observed marker matrices should match exactly, as the perfect phylogeny assumption holds. If the observation error parameters are set to one, that is $r^{\mathrm{FP}}_{\mathrm{global}} = 1$ and $r^{\mathrm{FN}}_{\mathrm{global}} = 1$, then the latent marker matrix with all entries flipped as shown in (**c**) will have an equal likelihood under this setting as the desired latent matrix has when error rates are set to zero (see Equation (2.4)). (**d**) The incorrect tree reconstruction where the left and right clades are erroneously assigned to the blue and red markers.

We represent the prior on perfect phylogenetic trees via the following two step generating process: (i) sample a mutation tree topology t (i.e., a type I tree with all cell vertices removed), (ii) assign all cells c to t. Recall that the tree t comprises all markers $l \in L$ as well as an additional imaginary marker v^* that is the root. If a marker l' is an ancestor of marker l in t, we assume that all cells that are descendants of l have the trait l' as well. If a cell c is attached to marker l, we posit that it has a copy number change at all the markers in the shortest path from l to the root marker v^* and has no copy number changes across the other markers in L.

For simplicity, we use a uniform prior on tree topologies as follows:

$$p(t) = \frac{\mathbf{1}\left[t \in \mathcal{T}\right]}{(|L| + 1)^{|L| + |C| - 1}},$$

where $\mathcal{T}$ is the set of all perfect phylogenetic trees that result from the two step generative process described above. Simulation from the prior can be performed using Wilson's algorithm [57], followed by independent categorical sampling to simulate the cell assignments.

2.3.5 Inference

We aim to approximate the posterior distribution of the tree t and the parameters θ via MCMC

$$\pi(t, \theta) \propto p(t)p(\theta)p(y|x(t), \theta).$$

First we describe a tree exploration MCMC move that considers an exponential number of neighbouring trees, then we review how parameters $r^{\text{FP}}_{\text{global}}$ and $r^{\text{FN}}_{\text{global}}$ are sampled.

`sitka` uses a tree sampling move to explore a large neighbourhood of a given tree. Given a tree t and marker l, we remove l from t, and add a new marker such that the new tree t' remains a perfect phylogeny, i.e., $t' \in N^l(t) \subset \mathcal{T}$. After removing l (*edge contraction*), adding a new marker that is an *edge insertion move* can be described as follows:

1. Pick a non-cell vertex v from $R = \{v^*\} \cup L \backslash \{l\}$ where v^* is the root node.

2. Pick any subset of v's descendent subtrees and remove it from v.

3. Add a new node v' under v and move the selected nodes from step 2 above and attach them to v'.

Note that the choice of node v in the first step above, partitions $N^l(t)$ into blocks of $N^l_v(t_{\backslash l})$, that is $N^l(t_{\backslash l}) = \cup_v N^l_v(t_{\backslash l})$. Figure 2.4-**e** (right) shows an example of an edge insertion move. A marker named *chr15_5950*, coloured red, has three children, all cell vertices in the sampled tree at MCMC iteration 100. This would be node v. In the next iteration, two of its children, namely cells *RC07C* and *RC05C4* are chosen and removed from v. They are then inserted under marker *chr1_4900*, v', which is now a child of marker *chr15_5950*.

The probabilities required in step 1 above are of the form:

$$\bar{\rho}_v = \frac{\rho_v}{\sum_{\tilde{v} \in R} \rho_{\tilde{v}}},$$

where:

$$\rho_v = \sum_{t \in N_v^l(t_{\setminus l})} p(t) p(y \mid x(t), \theta), \tag{2.5}$$

and $t_{\setminus l}$ denotes the tree t with node l removed. To compute ρ_v, we start with the following recursion for all vertices v in $t_{\setminus l}$: first, for all vertices c corresponding to a cell and $b \in \{0, 1\}$, define:

$$p_c^b = p\left(y_{c,l} \mid b, \theta\right),$$

where $p\left(y_{c,l} \mid b, \theta\right)$ is defined in Equation (2.4). Next, we compute these distributions for all subtrees of $t_{\setminus l}$. This can be done efficiently via a bottom-up recursion on the tree $t_{\setminus l}$: for all $v \in R$, $b \in \{0, 1\}$,

$$p_v^b = \prod_{v'' \in \,\text{children}(v)} p_{v''}^b,$$

where children(v) denotes the list of children of vertex v.

It can be shown that we can compute the probabilities required in step 1 above as follows ([1]):

$$\begin{align}\bar{\rho}_v &= \frac{\rho_v}{\sum_{\tilde{v} \in R} \rho_{\tilde{v}}} \tag{2.6}\\ &= \frac{\left(\frac{\prod_{v_i \in \text{children}(v)} \left(p_{v_i}^0 + p_{v_i}^1\right)}{p_v^0}\right)}{\sum_{\tilde{v} \in R} \left(\frac{\prod_{v_i' \in \text{children}(\tilde{v})} \left(p_{v_i'}^0 + p_{v_i'}^1\right)}{p_{\tilde{v}}^0}\right)}. \tag{2.7}\end{align}$$

Once v is sampled, we choose a subset of its children to move to v' by sampling k independent Bernoulli random variables with the probability for each child v_i of v as below and selecting children with corresponding Bernoulli realisations of 1:

$$\frac{p_{v_i}^1}{p_{v_i}^0 + p_{v_i}^1}.$$

To resample the parameters θ we use a slice sampling algorithm [58] and condition on the hidden state matrix x in a Metropolis-within-Gibbs framework. Briefly, we compute two sufficient statistics from the matrix x, (i) the number of false positive instances, n^{FP}, and (ii) the number of false negative instances, n^{FN} as follows:

$$\begin{aligned} n^{\mathrm{FP}} &= n^{\mathrm{FP}}(x) = \sum_{c \in C} \sum_{l \in L} \mathbf{1}[x_{c,l} = 0, y_{c,l} = 1] \\ n^{\mathrm{FN}} &= n^{\mathrm{FN}}(x) = \sum_{c \in C} \sum_{l \in L} \mathbf{1}[x_{c,l} = 1, y_{c,l} = 0]. \end{aligned}$$

Based on these sufficient statistics, we obtain:

$$p(y \mid x, \theta_{\mathrm{global}}) = \left(r^{\mathrm{FP}}_{\mathrm{global}}\right)^{n^{\mathrm{FP}}} \left(r^{\mathrm{FN}}_{\mathrm{global}}\right)^{n^{\mathrm{FN}}} \left(1 - r^{\mathrm{FP}}_{\mathrm{global}}\right)^{n^{\mathrm{A}} - n^{\mathrm{FN}}} \left(1 - r^{\mathrm{FN}}_{\mathrm{global}}\right)^{n^{\mathrm{P}} - n^{\mathrm{FP}}}, \tag{2.8}$$

where n^{P} and n^{A} denote the number of present and absent markers in the observed data and can be pre-computed as follows:

$$\begin{aligned} n^{\mathrm{P}} &= \sum_{c \in C} \sum_{l \in L} \mathbf{1}[y_{c,l} = 1] \\ n^{\mathrm{A}} &= |C||L| - n^{\mathrm{P}}. \end{aligned}$$

The model is implemented in the Blang probabilistic programming language [59]. We use a parallel tempering (PT) algorithm for inference (see [1] for a detailed explanation of the annealed distributions) and initialise it via a tree sampled uniformly from the prior.

2.3.6 Summarising the posterior distribution

Here we approximate the Bayes estimator by minimising the Bayes risk [60]:

$$\underset{t \in \mathcal{T}}{\operatorname{argmin}} \sum_{t' \in \mathcal{T}} \int L(t, t') \pi(t, \mathrm{d}\theta). \tag{2.9}$$

where $\pi(t, \theta)$ is the posterior distribution of the tree t and the parameters θ and $L(t, t')$ is a loss function. We use the L1 metric on the matrices of induced indicators $x(t)$ as the loss-function:

$$L(t, t') = \sum_{l \in L} \sum_{c \in C} |x_{c,l}(t) - x_{c,l}(t')|.$$

It is useful to define marginals $m_{c,l}$ that could be conceptualised as the posterior probability of cell c to have trait l:

$$m_{c,l} = \sum_{t \in \mathcal{T}} \int \mathbf{1}[x_{c,l}(t) = 1] \pi(t, \, \mathrm{d}\theta),$$

Using the MCMC samples $t^1, t^2, \ldots, t^N$ to average marginals $m_{c,l}$ we get a Monte Carlo approximation:

$$\frac{1}{N} \sum_{i=1}^{N} x_{c,l}(t^i) \to m_{c,l} \text{ a.s.}$$

Figure 2.4-**g** shows an example of the matrix m, each element of which is one of the approximated $\bar{m}_{c,l}$. We can now write the objective function of Equation (2.9) via the above marginals:

$$\begin{aligned} \sum_{t' \in \mathcal{T}} \int L(t, t') \pi(t, \, \mathrm{d}\theta) &= \sum_{t' \in \mathcal{T}} \int \sum_{l \in L} \sum_{c \in C} |x_{c,l}(t) - x_{c,l}(t')| \pi(t, \, \mathrm{d}\theta) \\ &= \sum_{l \in L} \sum_{c \in C} \sum_{t' \in \mathcal{T}} \int |x_{c,l}(t) - x_{c,l}(t')| \pi(t, \, \mathrm{d}\theta) \\ &= \sum_{l \in L} \sum_{c \in C} \left\{ m_{c,l}(1 - x_{c,l}(t)) + (1 - m_{c,l}) x_{c,l}(t) \right\} \\ &= \sum_{l \in L} \sum_{c \in C} \left\{ x_{c,l}(t) - 2 m_{c,l} x_{c,l}(t) \right\} + \text{constant} \qquad (2.10) \end{aligned}$$

We use a greedy algorithm to approximately minimise the Equation (2.10). Briefly, we start with a star-shaped tree with leaves C rooted at v^* and add markers from L one by one from a marker queue sorted by priority score. The priority score of each marker l is computed as

$$\text{priority}(l) = \max_{t' \in N^l(t)} \frac{q(t')}{\sum_{t'' \in N^l(t)} q(t'')}$$

where

$$q(x) = \prod_{c \in C} \prod_{l \in L(x)} q_{c,l}(x_{c,l})$$

$$q_{c,l}(x_{c,l}) = 2m_{c,l}x_{c,l} - x_{c,l}.$$

The quantities in the priority queue can be computed as in Section 2.3.5. We take the result of minimising the Bayes risk as the consensus tree.

2.3.7 Consensus tree and CNA heatmap visualisation

In visualising the output of `sitka`, we always use the type II trees. In what follows we introduce two different ways of visualising `sitka` trees that associate the tree to their input and parameter space.

Marker-centric view

This view maps the location of a marker to its position on the genome. As each marker is associated with a pair of consecutive bins, we take the convention of setting the genomic position of a marker to that of its associated leftmost bin. The plots comprise a type II tree t (Section 2.2) on the left next to a $(L-1) \times (L-1)$ matrix w to the right. In this plot, each node on the tree corresponds to a row. The matrix w has exactly $L-1$ non-empty elements. Each element $w_{i,j}$ is non-empty if marker i is on the i-th row of t and the j-th bin. The direction of the change in CN corresponding to a marker is shown by a left-pointing or right-pointing arrow for increase or decrease in CN respectively. Figure 2.8 shows the marker-centric view for the $SA535$ dataset. The rows of the matrix can be annotated to show information about specific internal nodes. For instance, in Figure 2.8, the *All* and *Immed* columns show what fraction of the genome is altered between a node and its parent on the tree. The former constructs a summary genotype for the parent node based on cells that are directly attached to it (if they exist), while the latter uses all cells that are descendent of the parent node to do so. We use the CNA profiles to compute the summary genotype. Let a^{sub} be an integer valued $C_{\text{sub}} \times D_{\text{sub}}$ matrix that encodes the CN state of a set of C_{sub} cells over D_{sub} copy number bins. We define s as a column vector of size D_{sub} where $s_i = f_{\text{summary}}\left(a_{1:C_{\text{sub}},i}\right)$. That is, s is an artificial cell whose CNA profile reflects that of those in a^{sub}. We set $f_{\text{summary}}(.)$ to the median function. To compute the percentage of the genome altered, we compute the Manhattan distance between the median genotypes of each node and its parent, normalised by the number of copy number bins. We acknowledge that the summary genotype at an internal node computed as above is a heuristic for data exploration purposes only as the phylogenetic tree is based on the CN changes, and not the CNs. In contract, the marker presence or absence profile can be reconstructed at any internal node v via the function $x(v)$.

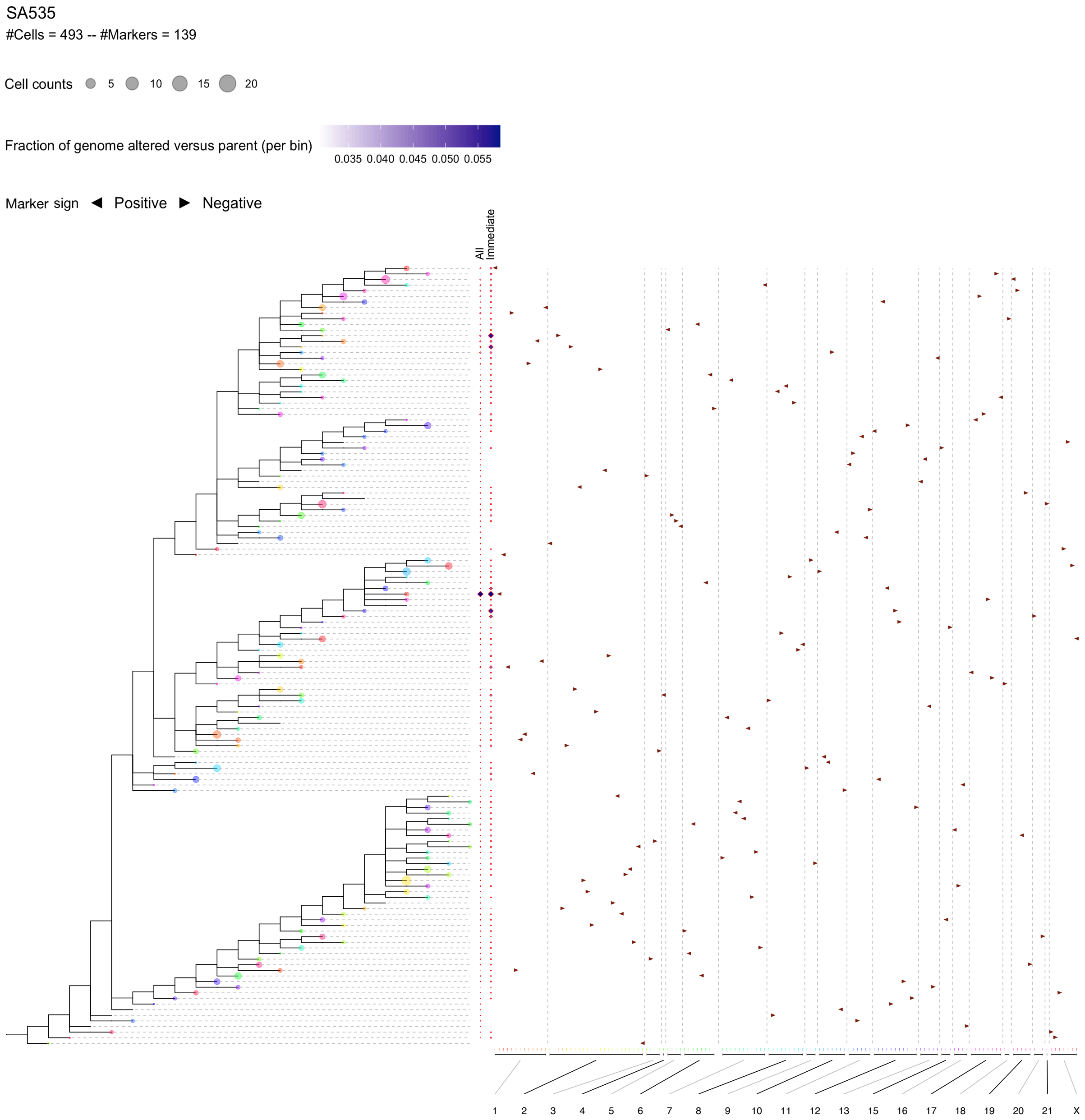


Figure 2.8: Marker-centric view for $SA535$. A type II tree where all cell nodes are dropped and chains are collapsed. The terminal nodes on the tree represent markers and their colours map to genomic positions as indicated on top of the chromosome (horizontal) axis (e.g., yellow maps to chromosome 2). Note that there are more negative (right-facing triangle) than positive (left-facing triangle) CN changes on chromosome 12 (7 and 4 respectively). One possible explanation is that one end of at least 3 CN changes are unobserved; potentially the imbalance may be due to sequencing errors. Another explanation is that the change points are located or stretch over to the end of the chromosome and therefore are not reflected in `sitka`'s lossy binary transformation.

Multi-channel view

One way to visualise the tree inference results is to arrange the tree and the cell-by-locus CN matrix side by side where the rows of the matrix correspond to the position of individual cells on the tree and the loci are arranged by their genomic position. This is shown in Figure 2.4-**h**. Figure 2.12-**a-c** are examples of the multi-channel visualisation where each marker is represented by a tuple of three different data-types, namely, (i) the markers, (ii) the `sitka` posterior, and (iii) the `sitka` transformation. We describe each channel for a fixed marker l in tree t. The first channel corresponds to the induced matrix $x(t)$ and indicates which cells in the consensus tree are descendants of l. In other words, whether or not in the latent matrix, the inferred marker is present in each cell. The second channel corresponds to the marginal matrix $\bar{m}$ as defined in Section 2.3.6, and shown in Figure 2.4-**g**. It denotes the posterior probability of the trait l to be present in each cell. Finally, the third channel corresponds to the transformed marker matrix y (see Section 2.3.3) and shows whether the trait was observed in each cell.

This view can be used to quickly assess the discrepancy between the input data and the inferred tree. It may illuminate cases in which the perfect phylogeny assumption is violated. An example is ChrX in the $OV2295$ dataset (Figure 2.12-**a**). ChrX has a long orange band (inferred marker) not matched by a black band (jitter-fixed observed marker) suggesting that a perfect phylogeny violation may have occurred. The pattern in this marker is consistent with the presence of an ancestral event followed by a deletion.

In Figure 2.12-**b**, a set of diploid cells are attached to the root of the tree. These are control cells included in the experiment and correspond to a marker region in the bottom of the matrix with no inferred markers (orange bands) and almost no observed markers (black bands). In this dataset, there are multiple change points where the observed marker has a high density (black band), but the tree is reconstructed with the marker absent (no matching orange band). Examples include Chr1, Chr7 and Chr16. One possible explanation could be that the end-points of each event were detected as slightly shifted in each cell. For instance, in Figure 2.11 there are two loci with an amplification (CN state equal to three) in the p-arm of Chr1 where the first locus is amplified in fewer cells than the second locus. In particular, cells that harbour a mutation in the first locus, appear to not have a mutation in the second locus, suggesting that the same event was called in the first locus in some cells, and in the second locus in others. A solution may be to fix jitter with a larger neighbourhood size (Section 2.3.3). An alternative hypothesis is that the cells in this dataset have a mutator phenotype that promotes de novo CN mutations in these loci.

Another locus suspected of violating the perfect phylogeny assumption is in the SA535 dataset on Chr3 (Figure 2.12-**c**). A black band is not matched by an orange one. This pattern could be a result of two insertion events, or two deletion events. For instance, in the former, the marker is hypothesised to be absent in the ancestral state, and it is gained in two subsequent clades (the subset of the tree aligned with the black band). While in the latter, the converse holds.

Define the mismatch rate as the fraction of cells that have a discrepancy between the jitter fixed input and the latent tree. Figure A.4 shows the distribution of mismatch rate for each dataset. The distribution for OV2295 and SA535 have an approximately bimodal shape while the SA501 distribution exhibits a heavy tail. In OV2295, 41 markers (11%) have a mismatch rate of over 50%, where marker *chr15_67000001_67500000* has the highest mismatch rate at 70%. In SA501, 30 markers (11%) have a mismatch rate of over 50%, 13 of which (5%) have a mismatch rate of over 75%. SA535 has the lowest maximum mismatch rate at 49% (marker *15_72000001_72500000*). While in the OV2295 dataset, and to a lesser extent, the SA535 dataset, there is some evidence of a cluster of markers with a high mismatch rate, the distribution is not concentrated at the minimum and maximum mismatch rates. This suggests that a few markers do not explain a large portion of the observed noise.

2.4 Overview of datasets

Here we introduce three real-world datasets and look at their type II trees. The first dataset, SA535 from [20], contains 679 cells from three passages of a triple negative breast cancer (TNBC) patient derived xenograft sample. Passages X1, X5, and X8 had 62, 369, and 231 cells post quality filtering respectively. We also included 17 mostly diploid control cells. These cells were combined to generate the input to the analysis pipeline. Figure 2.9 shows the phylogenetic tree and the heatmap encoding of the CNA profiles where the cells are sorted by their order appearance on the tree. The second dataset labelled OVA consists of cells from three samples taken from a patient with high grade serous (HGS) ovarian cancer; the first sample, SA1090, was from an ascites pre-treatment, while $SA922$ was from an ascites post-treatment. The third sample, $SA921$, was taken from the ovary. See Figure 2.10 for the tree and the CNA profile heatmap for this dataset. The final dataset, $SA501$ [43], is another TNBC xenograft tumour from 6 untreated passages, namely X2, X5, X6, X8, X11, and X15. After filtering we had 515, 236, 328, 189, 836, and 308 cells in each passage respectively (for a total of 2,412 cells, see Figure 2.11). Table A.5 shows the attrition after each step of filtering cells per passage in each dataset. We discuss how CNA rate changes over time with the passage information in Section A.5.

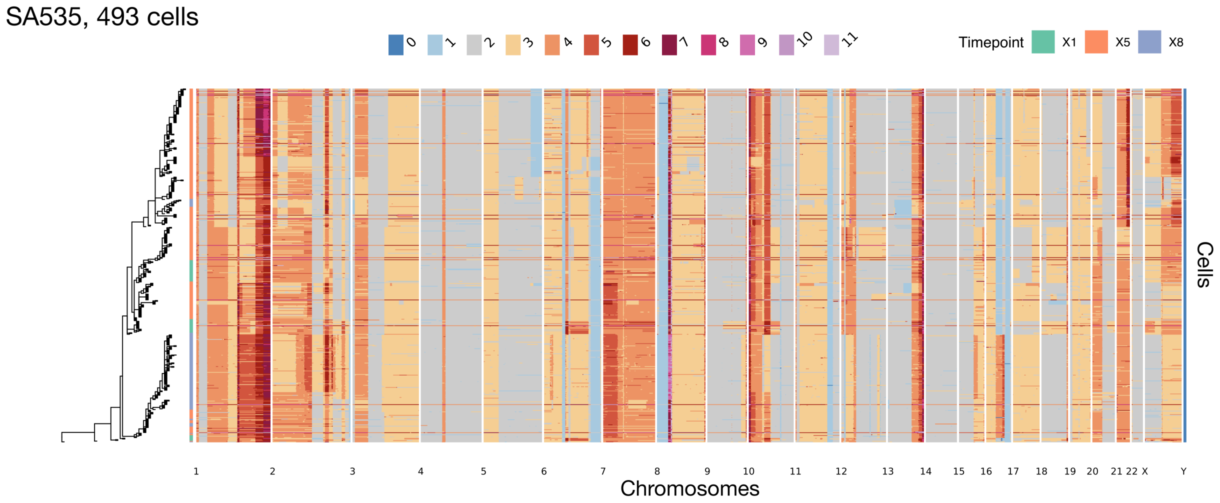


Figure 2.9: Phylogenetic tree and CNA profile heatmap for the SA535 dataset. The rows of the heatmap are sorted according to the placement of cells on the phylogenetic tree. The columns of the heatmap are sorted by their genomic position.

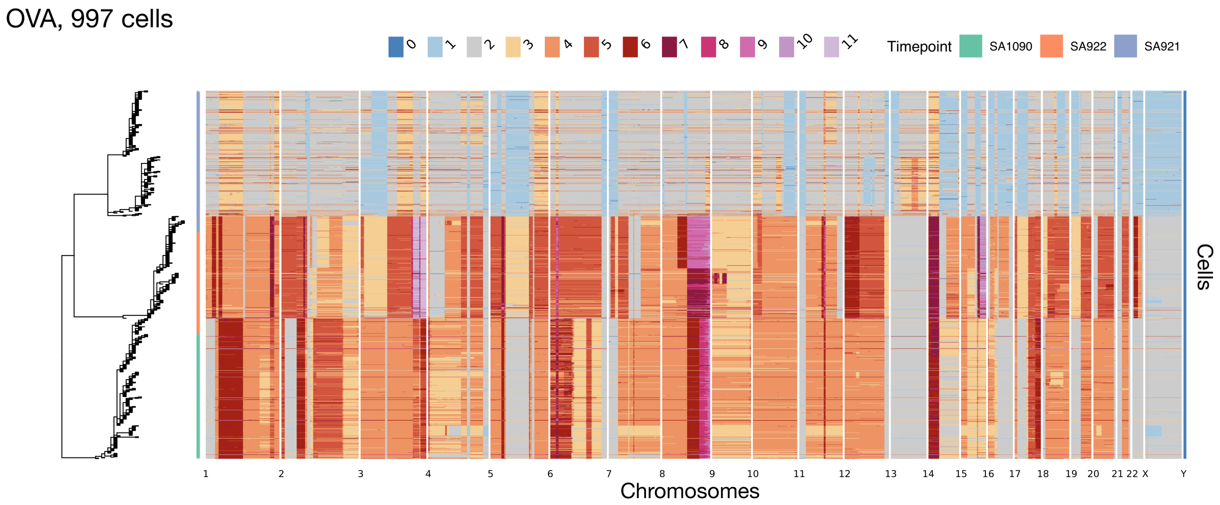


Figure 2.10: Phylogenetic tree and CNA profile heatmap for the *OVA* dataset. The nearly diploid cells with the loss of heterozygosity on chromosome X are from SA1090. The cells with an amplification on chromosome 22 are from SA922. The rest belong to SA921.

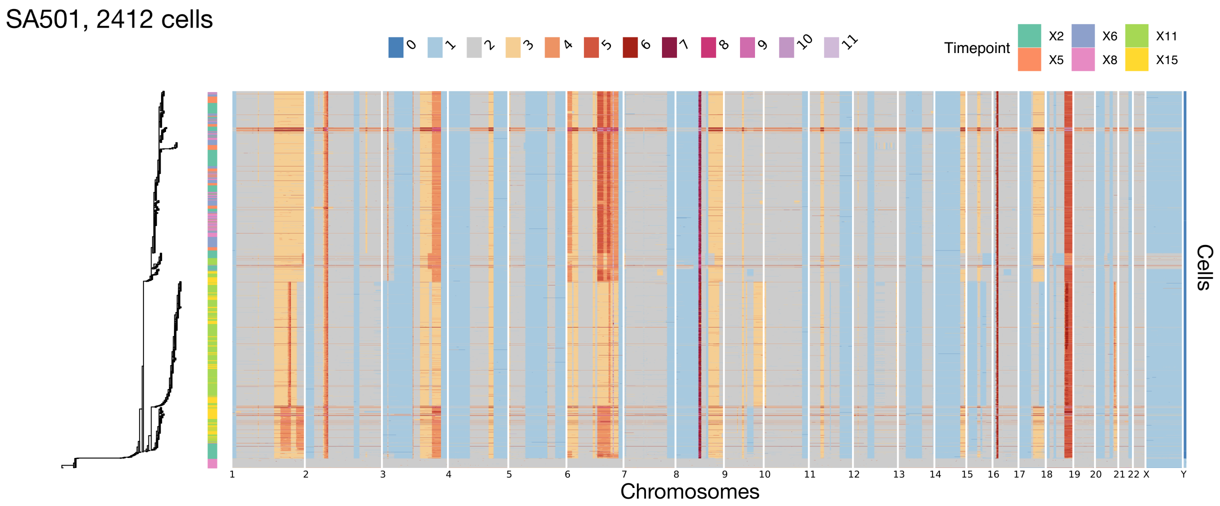


Figure 2.11: Phylogenetic tree and CNA profile heatmap for the SA501 dataset. Note that the diploid cells at the bottom of the heatmap are control cells that were included in the experiment.

2.5 Tree evaluation

2.5.1 Predictive test

To evaluate the inferred trees, we suggest a test that involves predicting the entries in the input binary matrix given to the tree inference method. We take the binarised input matrix y, the input matrix to the `sitka` algorithm as described in Section 2.3.3 as ground truth. Consider an inferred tree, t, and the corresponding matrix $g = x(t)$. In general the inferred trees from the baseline methods do not have named internal nodes, nor do they have the same number of internal nodes as the number of marker L. Therefore we do not know which marker in the inferred tree t corresponds to which column in the matrix y. We note that this is not the case with trees inferred from `sitka` where the internal nodes of the tree t correspond to the columns of the induced matrix $x(t)$. As a result, for methods other than `sitka`, for each column in the input data matrix, we pick a clade in t that has the highest prediction accuracy for the entries in that column.

For each method, we report Youden's J index [61]. We define below the function h to be a binary classification counts matrix, i.e., for two column vectors w and z of size C, it forms the confusion matrix. $h : \{0,1\}^C \times \{0,1\}^C \to \{0,1\}^{2\times 2}$ where

$$h_{i,j}(w, z) = \sum_{c \in C} \mathbf{1}\left[w_c = i, z_c = j\right].$$

For example $h_{0,0}(w, z)$ would count the number of times both elements of w and z were equal to zero (or the number of *true negatives*). We define accuracy for a given confusion matrix ζ computed from the h map above as:

$$\text{acc}(\zeta) := \frac{\zeta_{0,0} + \zeta_{1,1}}{\sum_{i \in \{0,1\}} \sum_{j \in \{0,1\}} \zeta_{i,j}}$$

We further define sensitivity and specificity as

$$\begin{aligned}
\text{sensitivity}(\zeta) &:= \frac{\zeta_{1,1}}{\zeta_{1,1} + \zeta_{1,0}} \\
\text{specificity}(\zeta) &:= \frac{\zeta_{0,0}}{\zeta_{0,0} + \zeta_{0,1}} \\
\text{youden}(\zeta) &:= \text{sensitivity}(\zeta) + \text{specificity}(\zeta) - 1
\end{aligned}$$

For a given tree t and its latent matrix $g = x(t)$ we compute the Youden's index as follows:

1. For all marker l in y, compute $\zeta_l = \text{argmax}_{\zeta_{l'}, l' \in \text{columns(g)}} \text{acc}(\zeta_{l'})$.

2. For tree t, set compute the confusion matrix as $\zeta_t = \sum_{l' \in \text{columns(g)}} \zeta_{l'}$.

3. Define the Youden's index for t as $\text{youden}_t := \text{youden}(\zeta_t)$.

That is for each marker in y, we take the clade that among all possible clades in t maximises the accuracy in predicting which cells are present in the l-th column of y. We then sum over all these scores to compute a confusion matrix for t and use this agglomerative matrix to compute the Youden's index for the tree. We use the delta method to calculate confidence intervals. Figure 2.12-**d** shows the Youden's index, its 95% confidence interval for `sitka` and 6 baseline methods over 3 different real-world datasets. `sitka` has a higher score than all competing methods. In Section 2.5.2 we review these results in detail.

A similar idea can be used to compare the different binarisation methods discussed in Section 2.3.2. Briefly, we report a leave-one-out score that represents how well the model predicts $y_{(c,l)} \,|\, y_{\setminus(c,l)}$ where $y_{\setminus(c,l)}$ denotes all the elements of the matrix except one entry. In other words, how well the model predicts one entry of the y matrix given all the other entries. This is averaged over all cell-marker pairs (c, l). It is computed using importance sampling (in the context of efficient leave-one-out estimation [62]). To compare different encodings of the input matrix, a common set of markers should be used. For example to compare preprocessing methods 1 and 2 that result in sets of markers L_1 and L_2, respectively, we ensure $L_1 \cap L_2 \neq \emptyset$, then the leave-one-out score is averaged over this intersection only.

2.5.2 Comparison to baseline methods

Here we compare the performance of `sitka` against a number of baseline methods over real-world datasets. We first briefly overview the baseline methods. The comparisons are summarised in Figure 2.12-**d**.

Overview of baseline methods

Here we briefly review the baseline phylogenetic methods. For all agglomerative clustering methods, the distance function $d(c_1, c_2)$ between two single-cells c_1 and c_2 was set to the Euclidean distance between their CNA profiles.

`UPGMA` (Unweighted Pair-Group Method with Arithmetic mean) is an agglomerative clustering method that, given a similarity matrix, constructs a rooted phylogenetic tree [63]. We start by putting each datapoint in a separate cluster. At each step of the algorithm, we merge the two clusters with minimum inter-cluster distance. Inter-cluster distance between clusters X and Y is defined by:

$$d\left(X,Y\right) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d\left(x,y\right) \tag{2.11}$$

where $|.|$ denotes the cardinality of a set. In other words, the average pairwise distance between elements of X and Y.

`WPGMA` (Weighted Pair-Group Method with Averaging) is similar to `UPGMA` but uses the following to compute the distance between clusters X and Y as defined by [63]:

$$d\left(X,Y\right) = \frac{1}{2} \sum_{x \in X, y \in Y} d\left(x,y\right) \tag{2.12}$$

In other words, it ignores the cardinality of clusters when computing the distance between them.

`NJ` (Neighbour joining) is a similar agglomerative clustering method that accepts an initial distance matrix d [64]. It keeps a running distance matrix q between each node of the tree and updates it over each iteration. In each iteration, `NJ` (i) re-calculates q, (ii) picks (f,g) to join under a new node u such that $f \neq g$ and $q(f,g)$ is the minimum entry in q, and (iii) calculates branch-length for the two newly created edges (u,f) and (u,g). The running distance matrix q is updated as follows:

$$q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d\left(j,k\right)$$

The entries for nodes (f,g) in matrix d is then replaced by that of the newly created node u where for each node $k \notin \{f,g\}$ we have:

$$d(u,k) = \frac{1}{2}\left(d(f,k) + d(f,g) - d(f,g)\right)$$

`HDBSCAN` first computes a 2-dimensional representation of the copy number matrix, i.e., casts the $C \times L$ CNA matrix a into a $C \times 2$ embedding matrix u. This step uses UMAP [65]. Secondly, a hierarchical clustering is computed via the method described in [66].

`MrBayes` is a Bayesian phylogenetics framework that implements multiple evolutionary models and uses MCMC to approximate the posterior distribution of trees and model parameters [67].

We ran MrBayes version 3.2.6 on two types of copy number input matrices: (i) integer copy numbers from 0 to 9 across the entire genome where CNs greater than 9 were assigned a value of 9 (MrBayes-np2 or MrBayes-np8), and (ii) the same as sitka's input of binary copy number changes at a reduced set of genomic locations (MrBayesWithBinaryInput-np2 or MrBayesWithBinaryInput-np8). We ran MrBayes using either two runs with one chain each (MrBayes-np2, MrBayesWithBinaryInput-np2), or two runs with four chains each (MrBayes-np8 or MrBayesWithBinaryInput-n8).

MrBayesWithBinaryInput-np8 did not finish running on any of our datasets. This is expected as the model based assessment of likelihood usually leads to undesirably long execution time. In some cases it is possible to speed up the run time of MrBayes via the use of graphical processing units (GPUs) and parallelization ([68]). However, it is unlikely to scale up to the sizes of datasets that we have in this work, and anticipate would be soon available. One reason is that the parallelization occurs across sites (markers) and not taxa (cells). For brevity, we will denote MrBayesWithBinaryInput-np2 by MrBayesWithBinaryInput. Following [43], we ran 10 million iterations with 50 percent burn-in fraction. All parameters of the likelihood model were left at the default values except lset rates=adgamma in order to account for correlated rates for adjacent sites. When using integer CN states as input to MrBayes, we set datatype=standard.

Benchmarking results using the predictive test

Now we use the predictive test derived in Section 2.5.1 to compare sitka to baseline methods reviewed above, over three real-world datasets $SA535$, OVA, and $SA535$. Figure 2.12-**d** shows the Youden's J index and its 95% confidence interval for each method.

sitka has the highest Youden's index across all three datasets. UPGMA and WPGMA perform similarly on $SA501$ and $SA535$. UPGMA does slightly better than WPGMA on the OVA dataset. HDBSCAN has a close but slightly smaller Youden's index than UPGMA over the $SA535$ and OVA datasets, but performs marginally better on $SA501$. NJ trails WPGMA on $SA501$ and the OVA datasets, and has the lowest Youden's index on $SA535$. MrBayes does well on the smallest dataset, $SA535$, with MrBayes-np2 and MrBayes-np8 performing similar to WPGMA, and MrBayesWithBinaryInput having achieved the second highest Youden's index. On the OVA data, MrBayesWithBinaryInput and MrBayes-np2 trail behind NJ, while MrBayes-np2 has the lowest Youden's index among all methods on all datasets. Similar to the OVA case, MrBayesWithBinaryInput and MrBayes-np2 trail behind NJ over the $SA501$ dataset. MrBayes-np8 did not finish running on $SA501$ after several days. This suggests that the algorithm used in MrBayes needs prohibitively more computational budget to achieve a Youden's index on par with the other methods. The results in this comparison suggest that sitka performs better than the baseline methods and therefore we chose it to infer clonal phylogenies in Chapter 4.

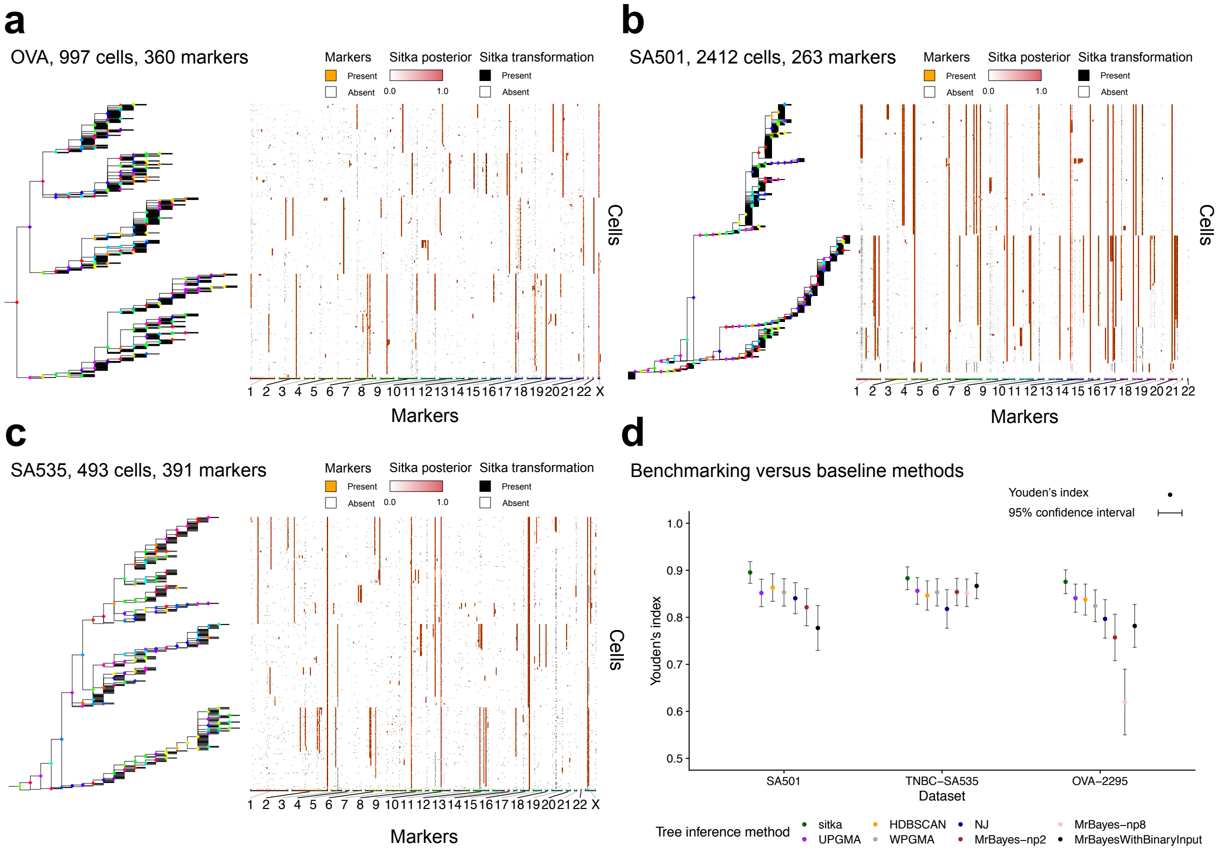


Figure 2.12: Results over real-world datasets and benchmarking against baseline methods. (**a**), (**b**), and (**c**) show the consensus tree and the multi-channel marker matrix for the OVA, SA501, and SA535 datasets respectively. In each panel, the tree is the consensus tree constructed as described in Section 2.3.6 and the matrix is the multi-channel visualisation of Section 2.3.7. Each marker is represented by a marker tuple, comprising three consecutive columns we call channels (as in Figure 2.4-**i**). The three channels encode for each marker in each cell (i) whether it is inferred to be present, (ii) the posterior probability of its presence, and (iii) whether it was present in the transformed input matrix. The number of rows (cells) and markers (columns) are noted on top of each panel. The position of the internal nodes on the genome is colour-coded to match the colour on the chromosome axis. (**d**) Comparison of methods. The dot shows the Youden's J index, while the bar represents the 95% confidence interval.

2.6 Identifying clones through phylogenetic analysis

Using `sitka` we can establish the evolutionary relationships of cells in a heterogeneous sample. To investigate cancer evolution we need to determine the abundance of subpopulations over time. To this end, we introduce `Lumberjack`, a tree-cutting algorithm that we use to define clonal subpopulations. In the output tree of `sitka`, cells are part of the terminal leaf nodes of the phylogenetic topology. We post-process the inferred trees to identify clonal populations from major clades. When clonal populations are defined, their abundances can be counted as a function of timeseries and these can be used for fitness inference (see Chapter 3). Clones are constructed by identifying connected components (each a clade or a paraphyly) in the phylogenetic tree reconstruction. The tree is 'cut' into discrete populations according to the following procedure (see Algorithm (2)). We review and build on the notation used in this chapter to facilitate the introduction of the `Lumberjack` algorithm. Let L be a set of markers and C be a set of cells. Define $t = (L, C, E)$ to be a rooted phylogenetic tree with E its set of directed edges. We assume that t is a type II tree (Section 2.2). Let $|t| = |C|$, that is the number of cells that belong to tree t. Let $t_l = (L_l, C_l, E_l)$ denote the subtree rooted at node l. Define parent(l) to be the parent of node l, and let descendants(l) comprise all its descendants.

The inputs to the algorithm are the rooted phylogenetic tree t, the CN states of its cells and the minimum M_{min} and maximum M_{max} allowed clone sizes. A clone is defined as a connected component (each a clade or a paraphyly) in the graph tree t composed of cells of sufficient genomic homogeneity. The degree of homogeneity can be tuned by limiting the number of markers and the difference in the CN of sub-clades in a clone. The algorithm first finds the coarse structure, that is, it divides the tree into major clades, and then looks for fine structures within each clade by traversing the tree in a bottom up manner and merging markers that are sufficiently similar. The remaining markers constitute the roots of detected clades (clones).

To obtain the coarse structure from the reconstructed phylogenetic tree we use a two step procedure: (i) identify monophyletic clades via Algorithm (2), (ii) then remove the cells comprising the clades found in step one from the tree and repeat Algorithm (2). We note that these new clades (if any) could be paraphyletic.

To find the fine structures within the initial clades we use the following procedure. For each clade

Algorithm 2 Heuristic (Top-down)

```
procedure SITKA-LUMBERJACK($t$, $M_{\min}$, $M_{\max}$)
    marker-queue $\leftarrow$ depthFirstSearch($t$)
    for marker $l$ in marker-queue do
        if $m \leq |t_l| \leq M$ then
            CUTS $\leftarrow$ CUTS $\cup\, l$
            Remove all markers below $l$ from $t$
        else
            for marker $l'$ in descendants($l$) do
                if $M_{\min} \leq |t_{l'}| \leq M_{\max}$ then
                    CUTS $\leftarrow$ CUTS $\cup\, l'$
                    Remove all markers below $l'$ from $t$
            if no eligible markers were found then
                CUTS $\leftarrow$ CUTS $\cup\, l$
                Remove $t_l$ from $t$
    return CUTS
```

ρ, and its corresponding sub-tree t_ρ, denote by L_ρ a set of markers l for which $M_{\text{min}} \leq |t_l| \leq M_{\text{max}}$. In a bottom-up traverse of the tree, for each node $l \in L_\rho$, remove l from L_ρ if $|t_{\text{parent}(l)}| - |t_l| \leq M_{\text{diff}}$, otherwise remove t_l from t_ρ. M_{diff} is a user-defined constant that intuitively results in merging clades that are too similar. At the end of the tree traversal, the set L_ρ contains new candidate roots for each initial clade. For each $l \in L_\rho$ define the summary CN profile as a vector whose i-th element is the median of the copy number states of the i-th bin for all cells in t_l. Compute the distance between two subclades as the mean absolute difference of their median genotypes. Merge subclones induced by L_ρ if their summary CN profiles are too similar. We can do this by computing a t-test over the pairwise distances to exclude outlier subclades and merge the rest.

Once clones are identified, we set the abundance of each clone at a specific timepoint as the fraction of cells in that clone from that timepoint. We note that for the data from WGS bulk sequencing [43] we used the following procedure to estimate clonal fractions: (i) let ν denote the mutational cellular prevalence (rows) estimated over multiple timepoints (columns) using the multi-sample `PyClone` [22] model, (ii) define β as the genotype matrix (which mutation-cluster (rows) is present in which clones (columns)), (iii) then we set $\beta\gamma = \nu$ where $\gamma = \beta^{-1}\nu$ are the clonal fractions over time, and (iv) we solve for γ using QR-decomposition.

2.7 Conclusions

In this chapter we introduced the `sitka` analysis pipeline, and described a number of visualisation methods that supplement the `sitka` phylogenetic reconstruction model. We briefly reviewed parts of the sitka model's mathematical formulation that was related to the visualisation toolkit. We also compared sitka over three real-world datasets to some benchmark methods using a posterior-predictive test that we derived. We then introduced a tree-cutting procedure that we call Lumberjack that is used to identify major clades in the trees inferred by sitka.

In this work we use data in which the genome of the single cells CNA profiles are partitioned into fixed-500Kb bins. Each bin is assigned a constant integer CN state. The size of these bins indicates that we cannot preclude that the same bin harbours multiple CNA events. Biological processes that result in complex DNA rearrangements could further increase the probability of this possibly [69, 70]. These cases will violate the perfect phylogeny assumptions.

It may be possible to take a data-driven approach to determining bin sizes. One approach would be to use smaller bin sizes for regions of the genome where more sequenced reads are available. In such regions, the copy number states can be more accurately determined for the smaller bins. While in this work we have used the CNA states of pre-defined fixed size bins as input, segmenting the genome, and calling the copy number jointly with inferring the phylogeny may be possible, although at a higher computational cost.

Characterising existing subpopulations in a population of cells is a major goal of this work. The preprocessing step removes multiple cells (up to 90% of the sequenced cells, see Table A.5). We filter out a fraction of cells to remove contaminated cells, either doublets or mouse cells, cells with too many erroneous sequencing artefacts, and cycling cells. Removing a portion of the sequenced cells will decrease the statistical power to determine the subclonal structure of the population. Our current cell filtering step is focused on removing cycling cells (Section A.4). This may bias our sampling against clones that have a higher division rate. It would be beneficial to retain more sequenced cells without negatively impacting the quality of the downstream analysis. One possible way is to retain false positive doublet cells by reviewing the images taken by the DLP+ platform's microscope prior to sequencing. As each library may contain up to thousands of cells, this approach is labour intensive and will benefit from automatic image processing frameworks. Another area to explore is to decrease the error rate in the cell quality classifier [20]. This may be done by taking a data-driven approach to choosing the features used for scoring cells in the classifier and to retrain the classifier on a larger dataset. A related idea consists in quantifying the error rate in different regions of the genome; if in a cell with a low quality score, the genomic bins can be classified into high-quality and low-quality regions, it may be possible to use the high-quality regions as phylogenetic markers and assign the cells to an already inferred phylogenetic tree.

Structural variations such as Chromothripsis that affect multiple segments of the genome at the same time make it difficult to determine the rate of CNA events and suggest that CNA events may not be suitable molecular clocks to estimate branch lengths. One possible remedy is to first infer the tree topology via markers based on CNA events and then conditioned on this topology, add SNVs to the tree. The number of SNVs on each edge of the tree may give an indication of the mutation rate in each clade.

We note that the trees inferred in Figure 2.12-**a-c** are unbalanced. Unbalanced tree topologies have been observed and are expected in adapting populations [71]. However this may be partially due to the fact that the copy number states of each cell is determined individually and independently from the tree inference step and that our jitter fix heuristic method cannot completely correct the CN calls. It may be improved by the joint inference of the CN states, the genomic segments, and the tree topology.

The visualisation toolkit developed here is static. Adding new annotations requires re-rendering the entire plot. It would be useful to create an interactive toolkit where individual nodes on the tree could be selected to view more detailed annotations. This may include relevant genomic information such as genes affected by copy number changes and if available, data on point mutations.

Evaluating the performance of a phylogenetic reconstruction method over real-world datasets is difficult, mainly due to a lack of ground through. One promising area of research is the use of CRISPR-Cas9 based lineage tracing [72]. In absence of ground truth data, we developed a predictive test (Section 2.5.1) that to our knowledge enables a first of a kind benchmarking of phylogenetic inference methods over real-world SCG CNA datasets.

Phylogenetic tree reconstruction is a principled way to identify subpopulations in a heterogeneous single-cell population. In Chapter 4, we use `sitka` to infer phylogenetic trees in multiple model systems from which scDNAseq is available through the DLP+ platform. We then use a tree-cutting algorithm to cut the `sitka` trees and assign single-cells to clones. This enables us to track the abundance of subpopulations over multiple timepoints. In the next chapter we introduce `fitClone`, a statistical framework that uses the timeseries clonal abundance data to infer evolutionary parameters of clones in model systems.

Chapter 3

Modelling Fitness in Longitudinal Data Via an Approximation to the Wright-Fisher Diffusion Model

3.1 Introduction

An important but open question in the treatment of patients with cancer is why and how resistance to therapy is developed. In many patients, despite initial positive response to therapy, the cancer eventually relapses. Advances in next generation SCS as well as longitudinal xenoengraftment has enabled more accurate, quantitative measurements of tumours. By serially transplanting a tumour sample into highly immunodeficient mice, a system called patient derived xenograft, one can continuously monitor how tumour composition may evolve in a patient over time. This process generates timestamped samples comprising thousands of cells that through SCS can be measured at thousands of genomic features including CNAs. This is in contrast to the standard of care where often only up to two timepoints are acquired and measured at an aggregate course level. The new higher resolution datatype introduces major challenges in extracting meaningful knowledge from the data, including (i) how to identify biologically meaningful groups of cells (i.e., clones) across multiple timepoints, and (ii) how to quantitatively reason about the underlying evolutionary forces acting on the clones via their observed dynamics. In Chapter 2, we developed `sitka` and `Lumberjack` to address the first question. Here we set out to tackle the second question.

Tracking the relative abundances of the clones over time produces a timeseries reminiscent of allele frequencies. Over time evolutionary forces act on a population, changing the allele frequencies through adaptation, mutation and neutral evolution [19, 73]. In particular, clonal abundances are similar to the fraction of the individuals in a population that harbour a specific allele. We adopted ideas from the population genetics literature to model the evolution of tumour growth via a state space model, the dynamics of which is dictated by the diffusion approximation to the Wright-Fisher process. The algorithm infers evolutionary fitness coefficients for each clone by simulating trajectories consistent with their observed clonal abundances. To the best of our knowledge, the only exact algorithm for the multivariate case (where we track more than one clone at a time) has infinite expected run time [74]. Therefore we decided to develop an approximation and its respective inference engine. These models will quantify whether observed dynamics are significantly determined by selection versus purely driven by random stochastic processes (i.e., genetic drift) - a key determinant on the path to predictive models. Those clones under selection then provide hypotheses over mechanisms of disease progression through analysis of mutational or molecular features.

We describe in this chapter, two approaches to fitness modelling. The first is a Bayesian state-space model (`fitClone`) based on the Wright-Fisher diffusion with selection (Section 3.1.1). The second is a deterministic logistic growth model (Section 3.9). We compare the two methods in Section 3.10. The comparison favours the Bayesian model, hence this is the model used in our results unless we specify otherwise.

3.1.1 fitClone: a Bayesian fitness model for timeseries data

We developed a Bayesian model and associated inference algorithm based on a diffusion approximation to K-allele Wright-Fisher model with selection. We start with timeseries clonal abundance measurements over a fixed number of clones and estimate two key unknown parameters of interest: *fitness coefficients* s_i for clone i which represents a quantitative measure of the growth potential of a given clone; and *distributions over continuous-time trajectories*, a latent (unobserved) population structure trajectory in 'generational' time.

After briefly reviewing and setting notation for Wright-Fisher diffusions with selection (Section 3.2), we introduce the Bayesian model we used to infer quantitative fitness of clones from timeseries data (Section 3.3). We then describe a novel algorithmic extension that scales simultaneous inference to dozens of clones (Section 3.5), ancillary methods for effective population size estimation (Section 3.6), and reference clone selection (Section 3.8).

Statistical methods for inference in the Wright-Fisher model is reviewed in [75]. In the bi-allelic case, one method is to estimate N_e first and then use approximate Bayesian computation [76] to target the posterior of S, using trajectories sampled from a Binomial distribution [77]. It is possible to use the continuous transformation of the discrete Wright-Fisher model in which both time and states (number of alleles/individuals) are scaled by a multiple of N_e, and then approximate it using discretisation in both time and allele-frequency states [78–80]. An interesting direction is the exact simulation of diffusions [81, 82]. For the one-dimensional Wright-Fisher diffusion with a general drift function, an exact simulation algorithm exists, although each sampled trajectory can take up to minutes to generate [83].

3.2 Wright-Fisher diffusions with selection

We take the convention of using uppercase symbols to denote random variables and lowercase symbols for realisations of random variables. Let K denote the number of clones and denote by $Z_t = \left(Z_t^1, \dots, Z_t^K\right)$ the relative abundance of each of the K clones at time t in the population. The process Z_t satisfies, for all t, the constraints $\sum_{i=1}^K z_t^i = 1$ and $z_t^i \geq 0$ for $i \in \{1, \dots, K\}$. In other words it is a $(K-1)$ simplex. We model the process Z_t using a Wright-Fisher diffusion with selection.

A Wright-Fisher diffusion can be written in stochastic calculus notation as

$$dZ_t = \mu^{s,N_\mathrm{e}}(Z_t)\mathrm{d}t + \sigma(Z_t)\mathrm{d}W_t \tag{3.1}$$

where $\{W_t\}$ is a K-dimensional Brownian motion, and the functions μ and σ, defined below, respectively control the deterministic and stochastic aspects of the dynamics where we assume this maps to the selection and genetic-drift components of the observed data. For $z = \left(z^1, z^2, \dots, z^K\right)$, the vector-valued function $\mu^{s,N_\mathrm{e}} : \mathbb{R}^K \to \mathbb{R}^K$ is defined as

$$\begin{aligned} \mu^{s,N_\mathrm{e}}(z) &= \left(\mu_1^{s,N_\mathrm{e}}(z), \dots, \mu_K^{s,N_\mathrm{e}}(z)\right) \\ \mu_i^{s,N_\mathrm{e}}(z) &= N_\mathrm{e} z^i (s_i - \langle s, z \rangle), \end{aligned}$$

where $\langle x, y \rangle$ is the inner product of vectors x and y; N_e, the *effective population size*, discussed in more details in Section 3.6; and the parameters $s = (s_1, s_2, \ldots, s_K)$ are called *fitness coefficients*. The interpretation of the fitness parameters is that if $s_i > s_j$, then subpopulation i has higher growth potential compared to subpopulation j. In practice, small values of $s_i - s_j$ can result in a large difference in the observed clonal abundances (e.g., see clones A and B from the HER2+ PDX timeseries analysed in Chapter 4). The matrix-valued function $\sigma : \mathbb{R}^K \to \mathbb{R}^{K\times K}$ is defined as

$$\begin{aligned}\sigma^2(z) &= [\sigma^2_{i,j}(z)]_{i,j\in\{1,\ldots,K\}}\\ \sigma^2_{i,j}(z) &= z^i\left(\delta_{i,j} - z^j\right)\end{aligned}$$

where $\delta_{i,j}$ is the Kronecker delta. Given an initial value z, we denote the marginal distribution of the process at time t by $Z_t \sim \mathrm{WF}(s, N_\mathrm{e}, t, z)$. Equation (3.2) shows an example of a two dimensional WF process with $s = (s_1, 0.0)$, $Z_t = \left(Z^1_t, Z^2_t\right)$, and $W_t = \left(W^1_t, W^2_t\right)$.

$$dZ^1_t = N_\mathrm{e} Z^1_{t-1}\left(s_1\left(1 - Z^1_{t-1}\right)\right)dt + \sqrt{Z^1_{t-1}\left(1 - Z^1_{t-1}\right)}dW^1_t \tag{3.2}$$

3.3 fitClone model specification

Given timeseries data measuring the relative abundances of K subpopulations at a finite number of timepoints, the output of the `fitClone` model is a posterior distribution over the unknown parameters of interest: the fitness parameters s described in the previous section, and the continuous-time trajectories interpolating and extrapolating the discrete set of observations.

To do this, `fitClone` places a prior on the fitness parameters s, uses a state space model (Figure 3.1) in which the latent Markov chain is distributed according to a Wright-Fisher diffusion, and encodes the noisy sampling from the population at a discrete set of timepoints via the observation model.

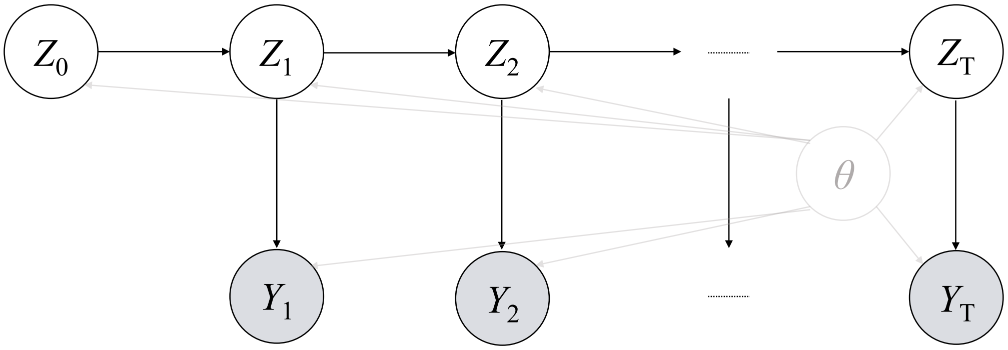


Figure 3.1: A state space graphical model. The observed quantity Y_t denotes the noisy clonal abundances while Z_t denotes the latent clonal fractions. $\theta = (N_{\mathrm{e}}, S)$ comprises the parameters of the model. The Wright-Fisher process controls the transition between states.

We model the fitness parameter with a uniform prior over range I:

$$S_k \sim \mathrm{Uniform}(I), k > 1,$$

where we set $S_1 = 0$ to make the model identifiable (see Section 3.8 for details). We used $I = (-10, 10)$ in our experiments. Note that the posterior is contained far from the boundaries of this prior range in all experiments.

The initial distribution, i.e., the distribution of the value of the process at time zero, is a Dirichlet distribution with hyper-parameter $(1, 1, \ldots, 1)$,

$$Z_0 \sim \mathrm{Dirichlet}(1, 1, \ldots, 1).$$

This can equivalently be seen as a uniform distribution over the K-simplex.

Let $t_1 < t_2 < \cdots < t_{T-1} < t_T$ denote a set of process times at which measurements of relative abundance of clones are available. Ideally, we would like the latent transition kernels to be given by the marginal transitions of the Wright-Fisher diffusion from the last section,

$$Z_{t_m} | Z_{t_{m-1}}, S \sim \mathrm{WF}(S, N_\mathrm{e}, t_m - t_{m-1}, Z_{t_{m-1}}), \tag{3.3}$$

where N_e is estimated as a preprocessing step (Section 3.6). In practice we resort to approximating the distribution in Equation (3.3) via an Euler-Maruyama scheme (Section 3.4.1).

Finally, for each $t \in \{t_1, t_2, \ldots, t_T\}$, let $Y_t = \left(Y_t^1, \ldots Y_t^K\right)$ denote a noisy observation of the clonal prevalences at process time t. In the single-cell context, this is obtained by counting, for each clone defined in Section 2.6, the number of cells coming from each passage, and normalizing by the number of cells sequenced in that passage. In the bulk sequencing context, see Section 2.6. When observations are counts, a Multinomial observation model or the Normal distribution approximation to it can be used. For simplicity, in both cases we use a Normal observation model, i.e., $y_t^i \mid z_t^i \sim \mathcal{N}\left(z_t^i, \sigma^2_{\mathrm{obs}}\right)$, where $\sigma^2_{\mathrm{obs}} = np_i\left(1 - p_i\right)$ and $n = \sum_j y_t^j$ and $p_i = y_t^i / n$.

3.4 Posterior inference under the fitClone model

Since the marginal distributions of the Wright-Fisher diffusion do not admit closed form expressions, and previous work on exact simulation does not scale to high values of K, we resort to discretisation using an Euler-Maruyama scheme [84].

3.4.1 Euler-Maruyama discretisation

In datasets available to us, the latent state of the Wright-Fisher process, i.e., the relative clonal fractions are observed at a few timepoints. In other words the system is sparsely observed. Note that the Wright-Fisher diffusion is a continuous time process where time is measured in units of N_e. Time represents the number of generations scaled by $1/N_\mathrm{e}$ and each generation comprises a cell division event. However, in order to perform inference using this model we need to discretise time. Intuitively, this discretisation in time can be thought of as an intermediate approximation between sparse observations.

We discretise the interval (t_m, t_{m+1}) using N equidistant intermediate timepoints $\left(t_{m,1}, t_{m,2}, \ldots, t_{m,N}\right)$

such that $h = t_{m,i} - t_{m,i-1} = (t_{m+1} - t_m)/N$ and h is the discretisation step size. For simplicity, we used the same h between all observed time steps and moved the observed timepoints to their closest grid point. We augment the state space of the process Z_t and add between Z_{t_m} and $Z_{t_{m+1}}$, N imputed intermediate states $\left(Z_{t_{m,1}}, Z_{t_{m,2}} \ldots, Z_{t_{m,N}}\right)$. Using these intermediate states, the marginal transition kernels in Equation (3.3) can be written as:

$$p\left(Z_{t_{m+1}}|Z_{t_m}, S\right) =$$
$$\int \ldots \int p\left(Z_{t_{m+1}}|Z_{t_{m,n}}, S\right) \{\prod_{i=2}^{N} p(Z_{t_{m,i}}|Z_{t_{m,i-1}}, S)\} p(Z_{t_{m,1}}|Z_{t_m}, S) dZ_{t_{m,1}} \ldots dZ_{t_{m,N}}$$

where the *intermediate* transition kernels, i.e., the Euler-Maruyama steps are approximated via a Normal distribution as follows:

$$Z_{t_{m,i}}|Z_{t_{m,i-1}}, S \approx \mathcal{N}\left(Z_{t_{m,i-1}} + \mu^{S,N_\mathrm{e}}(Z_{t_{m,i-1}})h, \sigma^2(Z_{t_{m,i-1}})h\right) \tag{3.4}$$

In Chapter 4 we apply this discretisation to timeseries data from in vitro and in vivo systems. For instance, in Section 4.3, a TNBC PDX timeseries dataset passaged over 1,002 days is analysed. We estimate the diffusion time $t_{10} - t_1 = 0.90$ and use a discretisation step of $h = 0.05$ to yield 8 intermediate steps.

3.4.2 Inference

We used a particle Markov chain Monte Carlo (pMCMC) method called Particle Gibbs with Ancestor Sampling and particle rejuvenation [85, 86] to sample from Z_{t_m} and the intermediate Euler-Maruyama steps, and a Metropolis within Gibbs sampler to sample the selection parameters S.

In sampling S, and for a fixed trajectory $\boldsymbol{z} = (z_{t_1}, z_{t_2}, \ldots, z_{T-1}, z_T)$, we use a truncated Normal random walk proposal with variance σ_p and boundaries (I_l, I_u) that are set to the boundaries of the prior on S. This proposal is non-symmetric and yields the following MH acceptance ratio:

$$a = \min\left[1.0, \frac{p\left(s'|\boldsymbol{z}\right)\Phi_{\mathcal{TN}}\left(s|s', \sigma_p, I_l, I_u\right)}{p\left(s|\boldsymbol{z}\right)\Phi_{\mathcal{TN}}\left(s'|s, \sigma_p, I_l, I_u\right)}\right]$$

where s' denotes the proposed value and $\Phi_{\mathcal{TN}}\left(. \mid s, \sigma_p, I_l, I_u\right)$ is the probability density function of a truncated Normal distribution parameterised by its mean, variance, and lower and upper bounds.

We note that from $p\left(s|\boldsymbol{z}\right) = p\left(\boldsymbol{z}|s\right)p(s)/p(\boldsymbol{z})$ only the likelihood term $p\left(\boldsymbol{z}|s\right)$ remains after cancelling identical factors in the numerator and denominator of the acceptance ratio above. This likelihood term factors into marginal distributions of the Wright-Fisher diffusion:

$$p\left(\boldsymbol{z}|s\right) = \prod_{t=1}^{T-1} p\left(z_{t+1}|z_t, s\right)$$

In practice we keep the Euler-Maruyama intermediate steps and Equation (3.4) can be used to evaluate $p(\boldsymbol{z}; s)$ for a given s.

3.5 Conditional sampler

In sampling a trajectory between a pair of consecutive latent states $z_{t_{m-1}}$ and z_{t_m}, the sampler initialises a set of particles and evolves them according to the distribution of the Euler-Maruyama intermediate steps to form potential trajectories. Each particle represents a potential latent state. A trajectory is valid only if it ends inside a K-dimensional sphere centred at the observation at time t_m. We call this sphere an $\epsilon_{t_m}^{\text{ball}}$. This is essentially a soft-bridging problem and becomes more difficult with increasing K as more components of a potential trajectory have to pass through the $\epsilon_{t_m}^{\text{ball}}$. Given a valid trajectory between $z_{t_{m-1}}$ and z_{t_m}, it is possible to evolve only a few components of the potential trajectory conditioned on the other components remaining constant. In other words, instead of sampling the clonal fraction of all clones at the same time, we keep the abundance of a subset of clones fixed and conditioning on those, update the clonal fraction of the rest. The discretised Wright-Fisher diffusion transition kernel at time t is normally distributed so we can use the conditional distribution form of a multivariate Normal distribution.

Without loss of generality we focus on a single intermediate Euler-Maruyama transition step, namely $Z_{t_{m,i}}|Z_{t_{m,i-1}}, S$. In the following it is useful to distinguish the subset of parameters that are sampled from those that are fixed. Let $B = (B_1, B_2)$ partition the index set $1{:}K = \{1, 2, \ldots, K\}$ into two subsets, that is $B_1 \subset 1{:}K$, $B_1 \cup B_2 = 1{:}K$ and $|B_1| \neq 0$. We denote by B_1 the indexes of the subset of parameters that are being sampled.

Define $Z_{t_m}^{B_j} := \left\{ Z_{t_m}^i \mid i \in B_j \right\}$, then we can write $Z_{t_m} = \begin{pmatrix} Z_{t_m}^{B_1} \\ Z_{t_m}^{B_2} \end{pmatrix}$. Now the problem can be stated as sampling from the following conditional distribution:

$$Z_{t_{m+1}}^{B_1} | Z_{t_m}^{B_1}, Z_{t_{m+1}}^{B_2}, Z_{t_m}^{B_2}, S$$

We use a two step procedure to achieve this: First, we sample $\bar{Z}_{t_{m+1}}^{B_1}$ from

$$\bar{Z}_{t_{m+1}}^{B_1} | Z_{t_m}^{B_1}, Z_{t_{m+1}}^{B_2}, Z_{t_m}^{B_2}, S$$

Second we ensure $Z_{t_{m+1}}^{B_1}$ respects the simplex criterion and set a realisation of $Z_{t_{m+1}}^{B_1}$ to

$$z_{t_{m+1}}^{B_1} = f^{z_{t_{m+1}}^{B_2}} \left(\bar{Z}_{t_{m+1}}^{B_1} \right)$$

where the vector-valued function $f^y : [0,1]^J \to [0,1]^J$ for the vectors x and y of sizes J and $K - J$ is defined as $f^y(x) = \big(f_1(x;y), f_2(x;y), \ldots, f_J(x;y) \big)$ and each component is equal to

$$f_j(x;y) = \begin{cases} \left| x_j \right|, & j < i, \\ \max \left[0, 1 - \sum_{r=1}^{K-J} y_r - \sum_{r=1}^{i-1} |x_r| \right], & j = i, \\ 0.0 & \text{otherwise.} \end{cases}$$

where

$$i = \underset{i \in \{1,\ldots,J\}}{\operatorname{argmin}} \left(\sum_{r=1}^{i} |x_r| \geq 1.0 - \sum_{r=1}^{K-J} y_r \right).$$

The effect of the mapping $f(.)$ becomes negligible as the discretisation step size goes to zero, but it is needed in a finite discretisation to ensure that the simplex constraint is satisfied. Now we describe sampling from $\bar{Z}_{t_{m+1}}^{B_1}$. To reduce clutter we drop the dependence on parameters s and N_e and assume they are fixed to some given value. Then, from the conditional distribution of a multivariate Normal we get:

$$\bar{Z}_{t_{m+1}}^{B_1} \mid Z_{t_m}^{B_1}, Z_{t_{m+1}}^{B_2}, Z_{t_m}^{B_2}, S \sim \mathcal{N} \left(\bar{\mu}, \bar{\Sigma} \right)$$

where

$$\bar{\mu} = \mu^{B_1} + \Sigma^{B_{1,2}} \left(\Sigma^{B_{2,2}}\right)^{-1} \left(z_{t_{m+1}}^{B_2} - \mu^{B_2}\right)$$

and

$$\bar{\Sigma} = \Sigma^{B_{1,1}} - \Sigma^{B_{1,2}} \left(\Sigma^{B_{2,2}}\right)^{-1} \Sigma^{B_{2,1}}$$

$\bar{\Sigma}$ is the Schur's complement of $\Sigma^{B_{2,2}}$ in Σ.

$$\mu\left(z_{t_m}\right) = \mu = \begin{pmatrix} \mu^{B_1} \\ \mu^{B_2} \end{pmatrix} \tag{3.5}$$

$$\Sigma\left(z_{t_m}\right) = \Sigma = \begin{pmatrix} \Sigma^{B_{1,1}} & \Sigma^{B_{1,2}} \\ \Sigma^{B_{2,1}} & \Sigma^{B_{2,2}} \end{pmatrix} \tag{3.6}$$

where in a slight abuse of notation, $\Sigma^{B_{i,j}} := \left[\Sigma_{p,q}\right]_{p\in B_i, q\in B_j}$. For a more efficient computation, we precompute and cache the value of $\left(\Sigma^{B_{2,2}}\right)^{-1}$ the inverse of $\Sigma^{B_{2,2}}$. For data analysed in this work, we only update one component at a time, that is we set $|B_1| = 1$.

3.6 Estimating the effective population size

Following [87] we use $F_s^{'}$ an unbiased moment-based estimator of the N_e where $N_\mathrm{e} = \frac{1}{F_S^{'}}$ and t is the number of generations between each passage.

$$F_s^{'} = (1/t)\frac{F_s(1 - 1/(2\tilde{n})) - 1/\tilde{n}}{(1 + F_S/4)(1 - 1/n_y)} \tag{3.7}$$

where $F_s = \frac{(x-y)^2}{z(1-z)}$ and $z = (x+y)/2$ and $\tilde{n} = \frac{2n_y n_x}{n_y + n_x}$, the harmonic mean of the sample size (initial population size at the passage) n_x and n_y at the two timepoints. x and y are the minor allele frequencies at the two timepoints.

In the multi-allelic case, we have:

$$F_s = \frac{1}{K}\sum_{i=1}^{K} \frac{\left(x_i - y_i\right)^2}{z_i\left(1 - z_i\right)}$$

This is equivalent to plan 2 in [87], sampling before reproduction and without replacement. In our formulation, the multi-allelic maps to multi-clones and this is what we have used to estimate N_{e} in this work.

We used the sum of clone sizes as the approximate initial population size at each timepoint/passage. Table A.3 lists the resulting N_{e} estimates. Since `fitClone` is robust to the choice of N_{e} in this range (Figure 3.2-**c**), we set $N_{\mathrm{e}} = 500$ for all datasets analysed in this work. A smaller N_{e} results in more stochasticity in the estimation of selection coefficients. We have chosen this value for N_{e} in part to account for phenomena that we do not explicitly model in our formulation of the Wright-Fisher diffusion, including clonal interference. We do not claim that $N_{\mathrm{e}} = 500$ represents the population structure.

We note that in our model we assume that the effective population size remains constant over all timepoints. This does not take into account the potential changing population growth rate or the bottleneck effect due to passaging. These phenomena may scale the diffusion time and bias our estimates of evolutionary events, including fixation or extinction times (see also Section 4.3.2). This stretching and compressing of time could be accounted for by adding random effects to the number of generations in the model, for example by taking $N_{\mathrm{e},t}$ as a piece-wise constant random variable that can vary between passages. Informally, the estimator introduced in this section can be thought of as a corrected harmonic mean estimator of the census population sizes over time that can to some degree account for the chaning population size [88]. Table A.4 shows the other parameters used in the inference over the real-world datasets.

3.7 Summarising the posterior distribution

In all real-world data that we analysed in Chapter 4, 10,000 particles and a burn-in equal to 10% of the MCMC samples were used. For the trajectories, we reported $\hat{z}_{1:T}$ where $\hat{z}_t = \left(\hat{z}_t^1, \ldots, \hat{z}_t^K\right)$ and $\hat{z}_t^k$ encodes the posterior mean of the clonal fraction of clone k at time t.

The posterior of the selection coefficient vector was summarised by $\hat{s} = (\hat{s}_2, \ldots, \hat{s}_K)$ where $\hat{s}_k$ denotes the posterior mean of the selection coefficient of clone k (see Figure 4.9). To compare the selection coefficients of two clones we used a $(K-1) \times (K-1)$ posterior ordering matrix P (Figure 3.2-**b,c**). $P_{i,j} = P(S_i \leq S_j \mid \boldsymbol{y})$ shows the posterior probability that clone i has a higher selection coefficient than clone j. We estimate this quantity by $P_{i,j} = \sum_{m=1}^{M} \mathbf{1}\left(s_{m,i} > s_{m,j}\right)$ where $s_{m,k}$ is the sampled selection coefficient of clone k at MCMC iteration m. Figure A.6 shows the visualisations of the posterior ordering matrices (P) for the real-world datasets analysed in Chapter 4. In these visualisations we colour coded the value of $P_{i,j}$, with the stronger purple hues (close to 1.0) representing a higher confidence that clone i dominates clone j, and conversely the stronger grey hues (close to 0.0) denote that clone j dominates clone i. Colours closer to white (0.5) represent no dominance. Note that for the lower diagonal elements $P\left(S_j \leq S_i \mid \boldsymbol{y}\right) = 1 - P\left(S_i \leq S_j \mid \boldsymbol{y}\right)$ and are omitted for clarity. The diagonal entries are to guide the eyes only.

3.8 Selecting the reference clone

In our formulation of the Wright-Fisher diffusion, one reference clone with a selection coefficient of zero has to be chosen. The selection coefficient of the other clones are reported relative to this value. For instance, if the fittest clone is chosen as the reference, the other clones will have negative selection coefficients. We chose to set the reference to a clone with an approximately monotonically decreasing trajectory (clonal abundance over time). This choice was motivated by a desire to infer a non-negative value for the fittest clones. Figure 3.2-**b** shows that the model is robust to the choice of the reference clone. We ran the inference procedure over the same dataset multiple times, each time changing the reference. The posterior ordering of clones over different choices of clones remained mostly identical.

3.9 The deterministic logistic growth model

We developed a closely related population genetics model which incorporates selection via deterministic differential equations, but has closed form solutions. In this model [36], the solution of a deterministic DE, the frequency of each population c out of possible K populations, at time t, with selection coefficient s_c is proportional to its starting prevalence $p_0(c)$ multiplied by a power of the relative fitness coefficient $w_c = (s_c + 1)$,

$$f(c, t, s) = \frac{(s_c + 1)^t p_0(c)}{\sum_{k=1}^{K} (s_k + 1)^t p_0(k)} \tag{3.8}$$

To estimate fitness coefficients $w_{1:T}$ from observed clonal fractions $y_{1:T}$, we solve the optimization problem in Equation (3.9) using a limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization procedure with box constraints [89]. Note that $f_{t,c}(w)$ is the clonal fraction estimate from Equation (3.8) and $y_{t,c}$ is the observed clonal fraction for clone c at time t.

$$\min_{w \in \mathcal{W}} \left(\sum_{t=1}^{T} d_t(w) \right) \tag{3.9}$$

where $d(w) = \big(d_1(w), d_2(w), \ldots, d_T(w)\big)$ is a vector valued function whose elements are $d_t(w) = \Big(\sum_{c=1}^{K} (f_{t,c}(w) - y_{t,c})\Big)^{\frac{1}{2}}$ and $\mathcal{W} = \big(\mathcal{R} \cup \{0\}\big)^K$ and $w_1 = 1.0$.

3.10 Simulation benchmarking

We forward simulated $L = 40$ datasets from the joint distribution of the Wright-Fisher model with $K = 5$, and $L = 40$ datasets with $K = 11$. For each simulated dataset l, we sampled the initial clonal abundance vector $Z_{1,l} \sim \text{Dirichlet}\,(\alpha_{1:K})$ where $\alpha_i = 1$ and selection coefficients from a Normal distribution with $s_{i,l} \sim \mathcal{N}(0.0, 0.3)$ truncated at $(-0.5, 1)$ for $i \in \{2, \ldots, K\}$ assuming the index of the reference clone is $i = 1$ and $s_1 = 0$. Discretisation constant (step size) $\Delta\tau = 0.001$ and the standard deviation of the emission model was set to $\sigma_{\text{obs,simul}} = 0.001$. At simulation and inference, we set $N_{\text{e}} = 500$. The simulation was continued to a diffusion time of 0.1 after which 10 equidistant samples were recorded as observed values for the process. In all models except the Logistic growth, we put a uniform prior on each component of the s vector, that is, $s_i \sim \text{Uniform}(-5.0, 5.0)$, and a Dirichlet prior on the initial clonal distributions. We set step size $\Delta\tau = 0.001$, $\sigma_{\text{obs,infer}} = 0.01$ and used 10,000 particles for 10,000 MCMC iterations.

We ran 5 different models on the simulated dataset as follows (Figure 3.2): (i) WFda is the Wright-Fisher model with diffusion approximation. (ii) Logistic growth is the deterministic differential equation Wright-Fisher model. (iii) One-step is identical to the WFda model but applies no discretisation for the time between observations. (iv) Single Clone is the WFda model with $K = 2$. See Section 3.9 for how the logistic model was fit. For the other models, we used the mean absolute error (MAE) of the post burn-in mean posterior of marginal selection coefficients for each method and the selection coefficient used to generate the simulated data. For each dataset l, the Single

Clone model is run K times, once for each clone k, where its input consists in the observed clonal fraction of only one clone. For this model we reported the averaged MAE across clones per dataset.

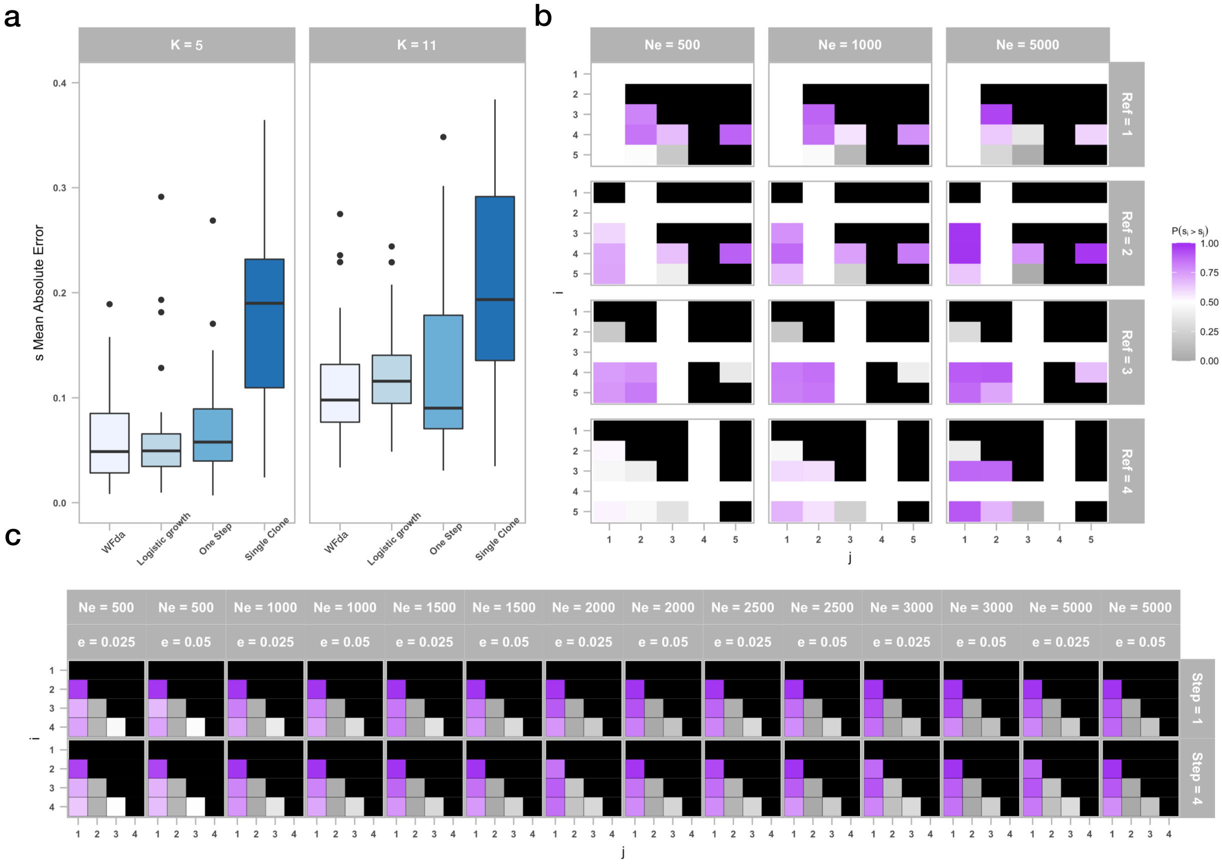


Figure 3.2: Simulation studies for the fitClone model. (**a**) Comparison to baseline methods for $K = 5$ clones (left) and $K = 11$ clones (right). (**b**) Posterior ordering of clones based on their inferred posterior selection coefficients across three values of effective population size (columns) and different choice of the reference clone (rows). (**c**) Posterior ordering of clones across different hyper parameters in the `fitClone` model. Effective population size in the range of 500 to 5,000 (top column), and observation error (bottom column). Minimum number of interpolations between two observations (rows).

The WFda and the logistic growth methods result in comparable error profiles, with a slight advantage for the WFda method in the $K = 11$ regime. This may be due to accounting for stochasticity in the dynamics via the diffusion process. Since the logistic growth is our baseline, a performance in line with this method suggests that our implementation of WFda is correct.

Recall that in the Euler-Maruyama scheme, N denotes the number of discretisation steps between two consecutive timepoints (t_m, t_{m+1}) where clonal abundance measurements are available. The One-step method uses $N = 0$. When K is small, this setting is in line with that of WFda and logistic growth. This suggests that with fewer clones, a coarser discretisation (smaller N) can be used. However, when K is large, the One-step setting results in an error distribution with a large variance where the inferred selection coefficients in a large subset of datasets have higher mean absolute errors. This suggests that even though the computational cost of inference rises in general with increasing the number of clones, decreasing the discretisation budget may result in inaccurate s estimates and should be avoided.

The Single Clone model has the highest estimation error. This suggests that the $K = 2$ and $K > 2$ regimes can generate similar clonal abundance trajectories with different underlying s values. Therefore treating each clone independently when estimating s is sub-optimal.

Comparing $K = 5$ to $K = 11$ across the methods compared here, we note that increasing the number of clones shifts the error distribution for s upwards. This may be due to unidentifiability where as K increases, more configurations of s are compatible with the observed clonal abundance trajectories. Observing the same timeseries for longer, or having more parallel repeats of the timeseries, for instance, establishing a parallel branch, may help improve the estimation accuracy (see Chapter 4).

Figure 3.2-**c** shows the model fits over a range of parameters for a dataset with $K = 5$. We repeated the inference procedure with $N_{\mathrm{e}} \in \{500, 1000, 1500, 2000, 2500, 3000, 5000\}$, and the observation error $\sigma^2_{\mathrm{obs}} \in \{0.025, 0.05\}$, and $N^* \in \{1, 4\}$ where N^* is the minimum number of discretisation steps between two consecutive timepoints. The posterior ordering matrices show that in general the model is more confident in the difference between the s coefficients as the N_{e} rises. Overall the model fits were robust to the choice of the effective population size and the discretisation step size. Together these simulations established a rationale for systematic modelling of all clones in a unified approach with a generative process.

3.11 Proof of principle application on TNBC-SA501 PDX

We applied `fitClone` to previously published data, wherein experimentally derived reproducible clonal dynamics had been reported in breast cancer PDXs [43]. From one of these lines, the abundance of five major clones (A-E), as determined by single cell genotyping was measured over six serial passages. Note that in this dataset, unlike those analysed in Chapter 4, clones were identified by their point mutation profiles. One clone (clone E) was found in the original study to undergo a selective sweep in repeat and independent passaging [43], suggesting selection due to higher fitness.

The input to `fitClone` are the relative abundances of the five clones at every timepoint, along with an estimate of the number of generations up to each timepoint (Figure 3.3-**a**). From this, we infer N_{e} as described in Section 3.6. For each timepoint, we set the number of generations to the number of days from tumour transplantation to tumour collection for in vivo systems, and the number days in culture for in vitro systems. In Section 2.6 we explain how the clonal fractions were derived from the bulk WGS data in more detail.

`fitClone` projects the observations from the generation time into the diffusion time which is in units of N_{e} (Section 3.4.1). It then applies the Euler-Maruyama scheme to discretise time and imputes the clonal fractions at unobserved timepoints between the observed timepoints. The outputs of the model include the clonal fractions at imputed timepoints consistent with the observations, and their estimated selection coefficients. Figure 3.3-**b** shows the observed and imputed clonal fractions in diffusion time. The dashed white vertical lines correspond to the discretised times at which clonal fractions were observed. Figure 3.3-**c,d,e** shows the distribution of the estimated selection coefficient of each clone, the credible intervals, and the posterior ordering matrix.

`fitClone` estimates converged with Clone E bearing the highest fitness ($1 + s$=1.03 ± 0.01), consistent with positive selection over the timeseries (Figure 3.3), and thus representing a proof of principle application of `fitClone` on real-world data.

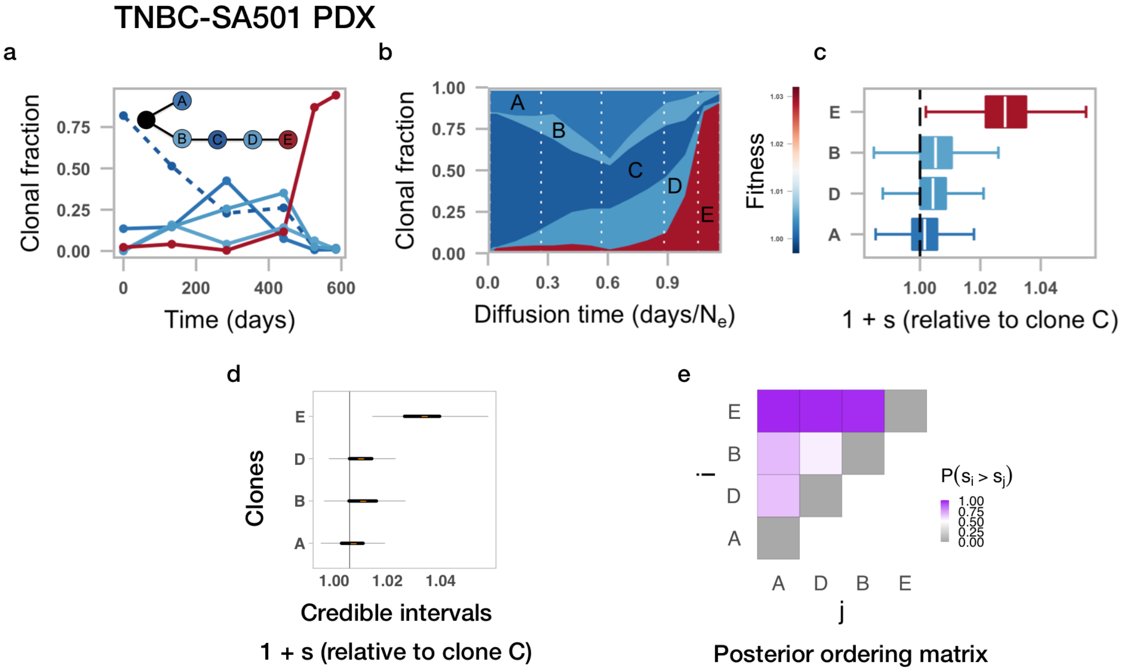


Figure 3.3: The fitness analysis for clones derived from bulk WG. The dataset is SA501. Application of `fitClone` to previously published clonal dynamics in PDX showing observed data (**a**), inferred trajectories (**b**), posterior distributions of fitness coefficients (**c**), the credible intervals for fitness coefficients (**d**) and pairwise comparison of the selection coefficients (**e**).

3.11.1 Software and implementation

The software implementation of `fitClone` is available at: [`https://github.com/UBC-Stat-ML/fitclone`]

3.12 Conclusions

In this chapter, we introduced a statistical framework, `fitClone`, that models timeseries clonal abundance observations using an implementation of the Wright-Fisher diffusion process. It simultaneously estimates growth trajectories, Z_i and fitness coefficients, S_i for each clone i in the population (Figure 3.3). We note that increasing $1 + S_i$ indicates positive selection and higher growth potential. The model accounts for drift as well as selection, with fitness estimated relative to a reference population, where $S = 0$ by construction. As a generative process, the model can be used for forecasting evolutionary trajectories of specific clones. In the next chapter, we will apply our framework to multiple model systems of cancer, measured at the single cell level. We will use `sitka` to infer the clonal structure of heterogeneous single cell populations, and then will apply `fitClone` to model the clonal dynamics.

f

Chapter 4

Application of the Wright-Fisher Diffusion Approximation to Cancer Model Systems Interrogated at the Single Cell Level

4.1 Introduction

Tumour fitness landscapes underpin selection in cancer evolution and response to treatment. However, quantifying fitness in heterogeneous cell populations remains an open problem that hinders progress in developing effective therapeutic strategies. In Chapter 1, we posited that two key technological advances, namely the emerging single cell sequencing platforms and model systems, in particular longitudinal patient derived xenoengraftment, may in principle facilitate the monitoring of change in tumour cellular composition over time and in response to therapeutic interventions. Measurements taken from these systems constitute novel datatypes that introduce major analytical challenges including (i) how to identify biologically meaningful groups of cells (i.e., clones) across multiple timepoints, and (ii) how to quantitatively reason about the underlying evolutionary forces acting on the clones via their observed dynamics. In Chapters 2 and 3, we described two main computational ingredients to address challenges above, namely `sitka`, a phylogenetic inference method and accompanying toolkits to assign cells to clones, and `fitClone`, a probabilistic framework that ascribes quantitative selection coefficients to individual cancer clones and forecasts competitive clonal dynamics over time

In this chapter we apply the methods developed in the previous chapters to real-world datasets to make inferences about how human cancer cells evolve at the copy number level, and that how establishing baseline fitness measures helps to interpret selection under drug administration. We have developed an experimental and computational platform consisting of three major components: scaleable phylogenetics for single cell genomes to identify clones (`sitka` and `Lumberjack`), timeseries sampling of immortal cell lines and patient derived xenografts to observe clonal dynamics, and a mathematical model for inferring clone-specific fitness measures (`fitClone`). Figure 4.1 shows the experimental setup and the data analysis, from the experimental setup in the cell lines and the PDX timeseries up to the clonal fraction measurements over time, the input to the `fitClone` model.

For normal human breast epithelial cells [90] *in vitro* and in breast cancer PDX [43, 91] (Figure 4.1-**a**), we sequenced >60,000 cells over interval passaging (Figure 4.1-**b**) with single cell whole genome sequencing [20] measuring single cell copy number profiles, and computing phylogenetic trees (via `sitka`) to identify genotypic clones (using `Lumberjack`) and their relative abundances as a function of time. Genetic (p53 biallelic inactivation for cell lines) and pharmacologic (cisplatin dosing in PDX models) perturbations were applied to determine their impact on fitness landscapes. `fitClone` was used to measure the selection coefficient for each clone (s) which we hypothesise to indicate growth potential. The larger the value of s, the more fit the clone is relative to the chosen reference clone.

The chapter is organised as follows. We first investigate the evolution of clonal trajectories in vitro in three related cell lines (Section 4.2) and in vivo with two independent breast cancer PDX timeseries (Section 4.3.1). Next we turn our attention to in vivo experiments where three different independent PDX lines are analysed to (i) explore the predictive power of our model through establishing physical mixture lines (Section 4.3.2), and (ii) to determine the dynamics of selection under early response to cisplatin treatment (Section 4.4.2).

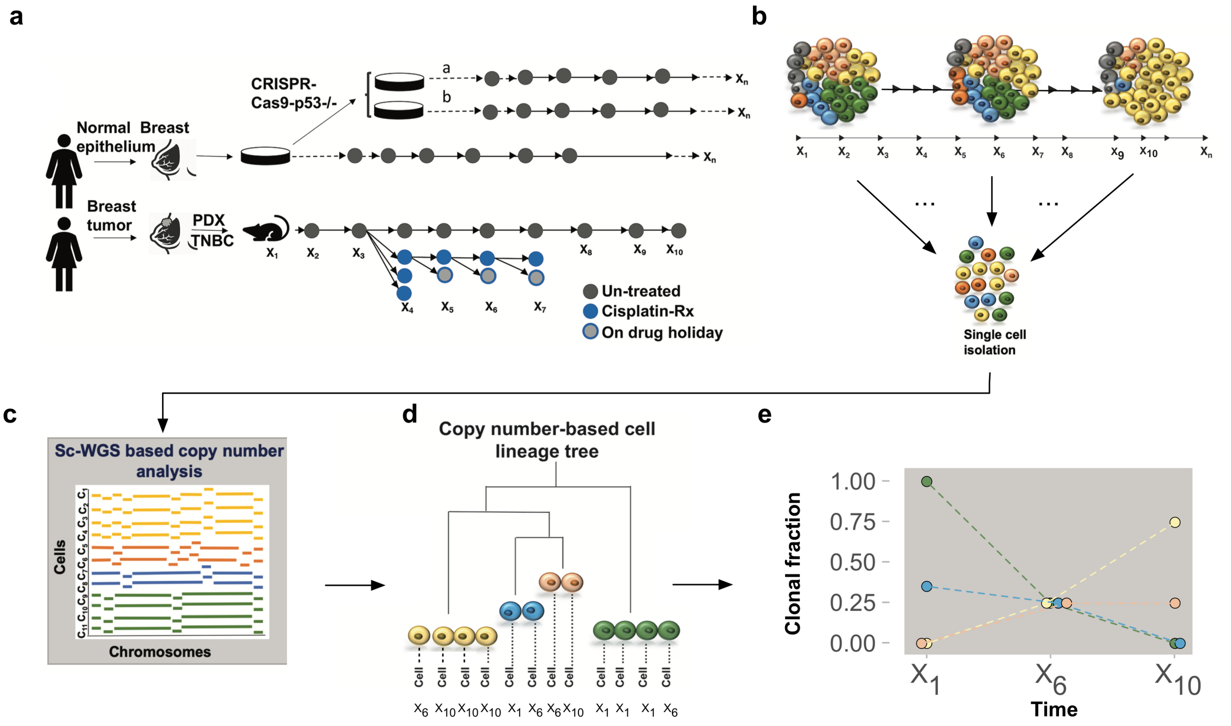

a
b
CRISPR-
Cas9-p53-/-
a
b
X_n
Normal Breast epithelium
Breast tumor
PDX
TNBC
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}
Un-treated
Cisplatin-Rx
On drug holiday
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_n
Single cell isolation
c
Sc-WGS based copy number analysis
Cells
Chromosomes
d
Copy number-based cell lineage tree
Cell
X_6 X_{10} X_{10} X_{10} X_1 X_6 X_6 X_{10} X_1 X_1 X_1 X_6
e
Clonal fraction
1.00
0.75
0.50
0.25
0.00
X_1 X_6 X_{10}
Time

Figure 4.1: Schematic overview of experimental design for quantitatively modelling clone-specific fitness. (**a**) Timeseries sampling from in vitro or PDX systems. (**b**) Clonal dynamics of cell populations observed over time. (**c**) Whole genome single cell sequencing of timeseries samples. (**d**) Phylogenetic tree inference using `sitka`. (**e**) Clonal fractions over time constituting the input to the `fitClone` framework, the mathematical modelling of fitness with diffusion approximations to the K-type Wright-Fisher model.

4.2 In vitro systems

4.2.1 Serial passaging of immortalized 184hTert diploid breast epithelial cell lines

We applied our framework to immortalized 184hTert diploid breast epithelial cell lines [90] to determine mechanisms by which *TP53* mutation induces clonal expansions and fitness trajectories. Cell lines, relative to PDX systems, are easy to establish and perturb. In particular, they are well suited for high-throughput drug screening [92, 93]. We note that our analysis here is a proof of principle and more replications are required to generalise our observations and conclusions. We investigate three related timeseries (*branches*). Their relationship is shown in Figure 4.1-**a**.

TP53 is the most abundantly mutated gene in all human cancers [94] and specifically in breast cancers [10, 95]. Known to be permissive of genomic instability, *TP53* loss is often acquired early in evolution and results in profound alteration of the copy number landscape [17, 96–98]. We asked whether specific clonal expansions could be observed, and moreover if selective fitness advantages could be quantified as a function of *TP53* ablation, thereby modelling copy number driven etiologic processes in a controlled system of immortalized mammary epithelial cells. *TP53* wildtype (*p53 WT*) timeseries sampling (60 passages over 300 days, 4 samples) was contrasted with two isogenic *TP53* deficient `NM_000546(TP53):c.[156delA];[156delA]` [20] parallel branches (*p53-/-a* and *p53-/-b*), each passaged over 60 generations (285 and 220 days, respectively) and sampled 7 times. A median of 1,231 cells per passage were sequenced yielding a total of 6,620, 7,935, 9,615 single cell genomes for each timeseries, respectively (Figure A.7-**a**,**b**).

4.2.2 Phylogenetic and fitness analysis

For each of *p53 WT*, *p53-/-a*, *p53-/-b* we inferred single cell copy number profiles, constructed a phylogenetic tree to establish clonal lineages (Section 2.6, A.8) and measured clonal abundances as a function of time (Figure 4.2-**a,b,c**). For each timeseries (branch), we pulled all single cells and used `sitka` to infer a phylogenetic tree. The tree was cut using `Lumberjack` to yield clones and their abundances over time. In the following, we summarise the `fitClone` inferred selection coefficient of a clone by reporting the posterior mean $1 + s$ followed by its standard deviation. See Figure A.5 for highest posterior density credible intervals for the selection coefficient of all datasets that we analyse in this chapter.

Modelling the abundances with `fitClone` (Tables A.3, A.4) revealed *p53 WT* clonal trajectories consistent with small differences over the posterior distributions of fitness coefficients amongst four major clones (Figure 4.2-**d**). In contrast, p53 mutant branches each showed expansions of clones with aneuploid genotypes, where the founder diploid population was out-competed. Relative to *p53 WT*, rates of expansion of p53 mutant, aneuploid clones were higher, leading to rapid depletion of diploid cells (Figure 4.2-**d**). Pairwise difference of selection coefficients $s_i - s_j$ (Δs) between clones in the `fitClone` inference process was larger in the *p53-/-* lines relative to *p53 WT* (Figure 4.2-**e**). This suggests that p53 mutation permits the expansion of clones at higher rates, and that these clones have measurably higher positive selection coefficients.

The p53 mutant lines harboured 11 (size range 47 to 1,474 cells, median 204), and 10 (size range 158 to 997 cells, median 404) distinct clones for *p53-/-a* and *p53-/-b*, respectively. In each series the diploid founder clones, devoid of detectable copy number alterations, were systematically out-competed by populations that had acquired at least one copy number alteration (Figure 4.2-panels **f** to **m**). Some similarities in copy number events were observed during the two replicate timeseries. These included gains in chromosomes 13, 19p and 20, and losses on chromosomes 8p and 19q (Figure A.8). However, by the end of the timeseries, the genotypes in the two lines had diverged considerably. Selection coefficients were highest in clones with localised amplifications of known prototypic oncogenes in breast cancer [10, 95, 98, 99] including in *MDM4*, *MYC* (Figure 4.2-**f**) and *TSHZ2* (Figure 4.2-**j**), in some cases on a whole genome doubled background. Clone A, the highest fitness clone in *p53-/-a* (57% of cells at last timepoint, $1+s = 1.05 \pm 0.09$) exhibited a whole genome doubling event (18 chromosomes with four copies) and harboured a focal, high level amplification at the *MDM4* locus on Chr1q (Figure 4.2-**f**). Clone G (27% of cells at last timepoint, $1+s = 1.03 \pm 0.03$), the next highest fitness clone in *p53-/-a* remained diploid, with the exception of a focal high level amplification precisely at the *MYC* locus on Chr8q (Figure 4.2-**g**). By contrast clone K, chosen here as the reference clone for modelling (see Section 3.8), remained entirely diploid and exhibited a monotonically decreasing trajectory (from 90% to 0% of cells over the timeseries, Figure 4.2-**h,i**). In *p53-/-b*, two clones exhibited non-neutral, positive selection coefficients (Figure 4.2-**j**). Clone D (52% of cells at last timepoint, $1 + s = 1.05 \pm 0.02$) harboured a Chr20q single copy gain with an additional high level amplification at the *TSHZ2* locus, while clone E (35% of cells at last timepoint, $1 + s = 1.05 \pm 0.04$) harboured a Chr4 loss, Chr19p gain/19q loss and Chr20q single copy gain (Figure 4.2-**f**). As seen in *p53-/-a*, the 'root' clone I that remained diploid was systematically outcompeted, diminishing from 68% to 0% abundance over the timeseries (Figure 4.2-**k**,**l**). We next tested whether the genomes of individual cells became progressively more aberrant over time and whether this correlated with measurements of clonal fitness. We estimated both sample and clone specific mutation rates at the copy number breakpoint and point mutation level (as previously described [20]). Both *p53-/-a* and *p53-/-b* exhibited increased mutations and breakpoints over time relative to *p53 WT* (Figure 4.3**a-e**). Cells accumulated 0.08 additional breakpoints ($p<0.0001$) and an average of 0.4 additional mutations (0.17 in *p53-/-a* and 0.67 in *p53-/-b*) per generation (Figure 4.3-**k**,**l**), while the *p53 WT* line accumulated 0.03 additional breakpoints per generation (Figure 4.3-**k**). This cell line was used as the reference for clone-specific point mutation detection in *p53-/-* cell lines and as such we did not call or analyse any point mutations in *p53 WT*. Clone level distributions of breakpoints and mutations were positively correlated with inferred fitness coefficients in both *p53-/-* lines (Figure 4.3-**g**-**j**), but not in *p53 WT* (Figure 4.3-**f**).

4.2.3 Segmental aneuploidies correlate with positive selection in diploid p53 deficient cells

These results indicate that the impact of genetic perturbation on selection can be measured and modelled in clonal populations. In particular, p53 deletion, known to be an early event in the evolution of many cancers [100], yields clonal expansions driven by whole genome, chromosomal and segmental aneuploidies, conferring quantitative fitness advantages over cells that maintain diploid genomes. Modes of positive selection involving high level amplification of proto-oncogenes often seen in human breast cancer [101–104], and aneuploidies in general, suggest that in vitro genetic manipulations can induce fitness-enhancing genomic copy number changes consistent with etiologic roles in cancer.

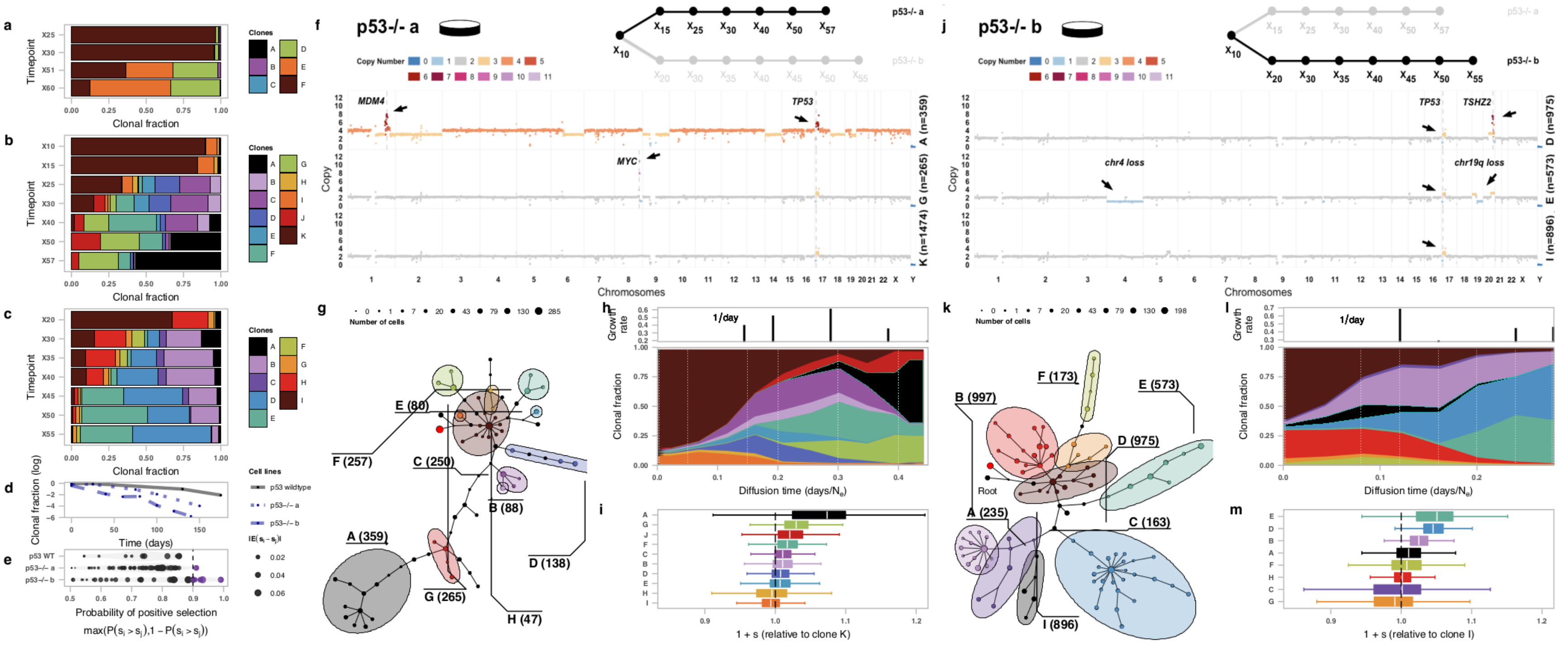


Figure 4.2: Impact of p53 mutation on modes of selection on 184hTert cells. (a-c) Clonal dynamics of p53 wildtype, and two independent timeseries of p53 mutant 184hTert mammary epithelial cell lines. (**d**) Clonal fraction of diploid reference over time. (**e**) Distribution over magnitude of difference between selection coefficients of pairs of clones. Each point is analogous to an element in the posterior ordering matrix of Section 3.7. (**f**) Clonal genotypes of three representative clones for *p53-/-a* showing high level amplification of *MDM4* in clone A and *MYC* in clone G. Reference diploid clone K shown for comparison. (**g**) Phylogeny (simplified type II `sitka` tree) of cells over the timeseries *p53-/-a* where nodes are groups of cells (scaled in size by number) with shared copy number genotype and edges represent distinct genomic copy number change points (`sitka` markers). (**h**) Inferred trajectories and (**i**) quantiles of the posterior distributions over selection coefficients of `fitClone` model fits to *p53-/-a*. See Figure A.5 for the credible intervals and Figure A.6 for pairwise comparison of the selection coefficients. (**j**) Clonal genotypes of three representative clones for *p53-/-b* showing high level amplification of *TSHZ2* in clone D, Chr4 loss in clone E. Reference diploid clone I shown for comparison. (**k-m**) Analogous to (**f-i**) but for *p53-/-b*.

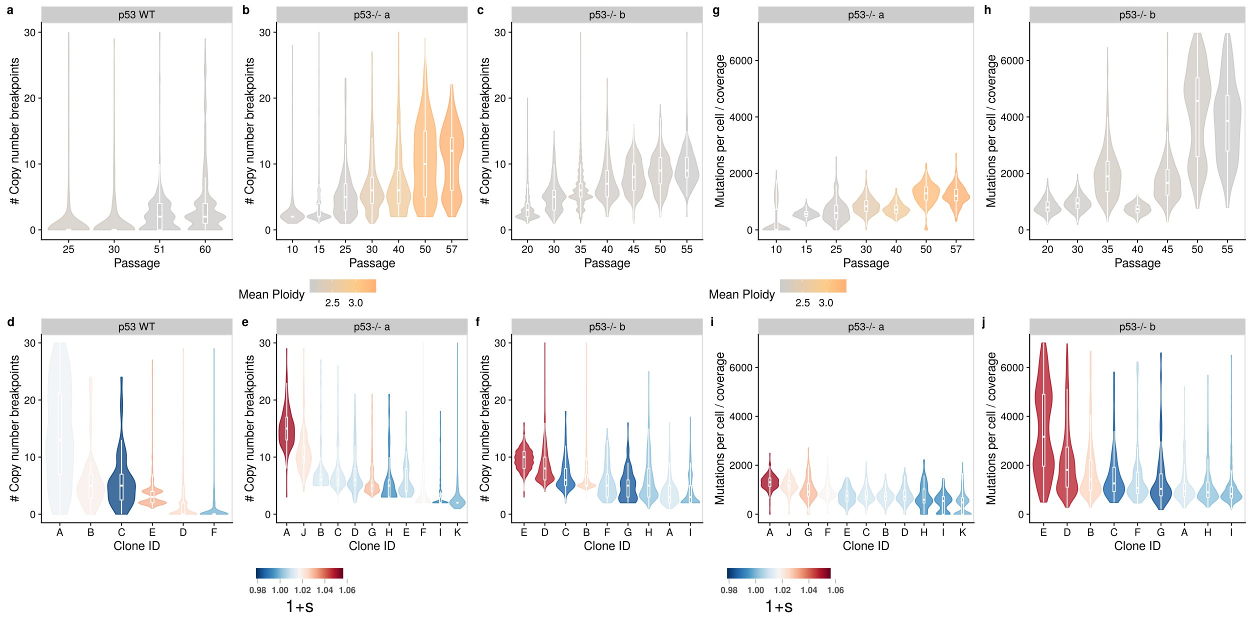


Figure 4.3: Structural variant and mutation rates of 184hTert cells. Distribution over copy number breakpoints/cell as a function of generation for (**a**) *p53 WT* (**b**) *p53-/-a* (**c**) *p53-/-b*. Distribution over point mutations/cell as a function of generation for (**d**) *p53-/-a* and (**e**) *p53-/-b*. Clone specific distributions over copy number breakpoints/cell, coloured by fitness coefficients for (**f**) *p53 WT* (**g**) *p53-/-a* and (**h**) *p53-/-b*. Clone specific distributions over point mutations/cell, coloured by fitness coefficients for (**i**) *p53-/-a* and (**j**) *p53-/-b*.

4.3 In vivo system in the untreated regime

4.3.1 Serial passaging of *TP53* mutant PDX tumours

We next modelled clonal expansions observed during serial passaging of *TP53* mutant PDX tumours to determine the predictive capacity of fitness coefficients (Figure 4.4). PDX systems are more difficult and time consuming to establish than cell lines, but may represent a more realistic model of tumour evolution as cells are located in their micro-environment. Here we exemplify our framework on two PDX timeseries, namely HER2-positive-SA532 and TNBC-SA609, and contrast their evolutionary characteristics.

We generated single cell genomes from 8 serial PDX transplants over 721 days from a HER2 positive (HER2+) breast cancer with a *TP53* p.A159P missense mutation (SA532), and contrasted this with 10 serial samples over 1,002 days from a TNBC PDX (SA609) with a *TP53* p.R213* nonsense mutation (Figure A.7-**c**,**e**). A median of 907 single cell genomes were sequenced per passage for a total of 11,705 and 10,553 single cell genomes from the HER2+ and TNBC series, respectively. Both series exhibited progressively higher tumour growth rates over time (Figures 4.4-**f**,**j**, A.7-**d**). Data were analysed as per the *in vitro* lines described above and modelled with `fitClone` (Tables A.3, A.4). The HER2+ series exhibited 4 distinct clones ranging in size from 134 to 1,421 cells (median 319, Figure 4.4-**e**), and the TNBC series exhibited 8 distinct clones with 18 to 680 cells (median 556, Figure 4.4-**i**). In the HER2+ model, clonal trajectories were consistent with selection coefficients with small relative differences in fitness (Figure 4.4-**f,g**). Clones B, C, and D show similar but distinct copy number profiles. The median genotypes of Clones B and C are different across 26 genomic bins, that of clones B and D are different in 24 genomic bins, and those of clones C and D have differences along 21 genomic bins. By contrast, the TNBC model trajectories resulted in a high positive selection coefficient for a minority of clones (Figure 4.4-**j,k**). Consistent with increased dynamics in the TNBC series, we found an initial increase of 0.1 breakpoints per cell per generation in the first 4 passages (Figure A.10-**a**). After this initial increase the average number of breakpoints per cell remained mostly constant. In the HER2+ line we observed a small decrease of 0.04 copy number breakpoints per generation. We note that clone E in TNBC swept through the population over the last 3 timepoints (Figure 4.4-**j**,**k**, n=541 over the timeseries). Clone E had the highest selection coefficient ($1 + s = 1.08 \pm 0.043$), having grown from undetectable proportions in earlier timepoints to 58% of cells by the end of the timeseries. Clone E also had the highest number of breakpoints with 12.8 additional copy number breakpoints per cell, relative to the reference clone C with the lowest (Figure A.10-**c**).

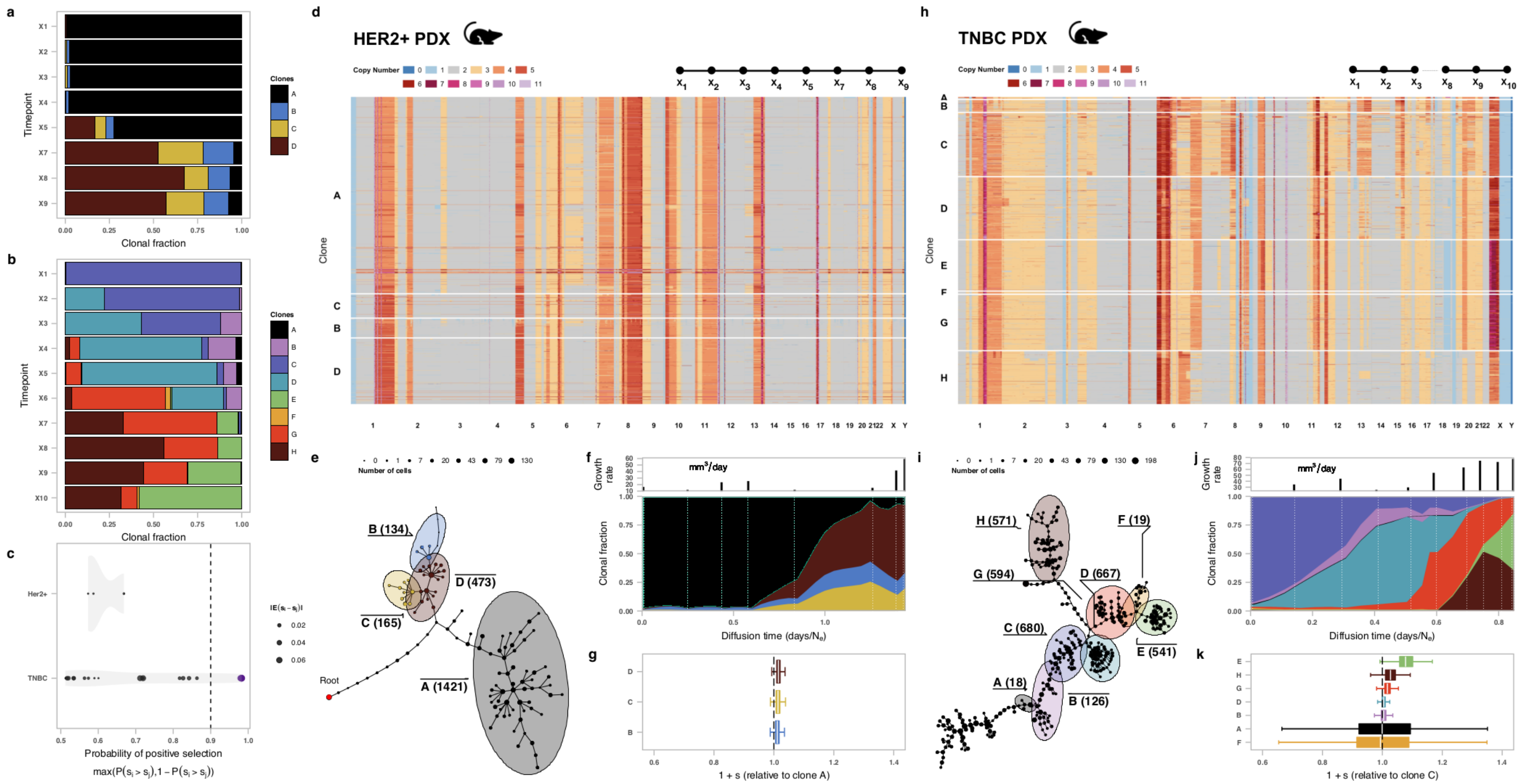


Figure 4.4: Comparison of fitness landscapes of breast cancer PDX models. (**a-c**) Clonal dynamics of HER2+ and TNBC models. (**d**) Heatmap representation of copy number profiles of 2,193 cells, grouped in 4 phylogenetic clades. (**e**) Phylogeny as per Figure 4.2 for HER2+. (**f**) Inferred `fitClone` trajectories and (**g**) selection coefficients for the HER2+ model. See Figure A.5 for the credible intervals and Figure A.6 for pairwise comparison of the selection coefficients. (**h-k**) Analogous plots for the TNBC model (n=3,216 cells).

4.3.2 Mixture experiments

In the two previous sections, we used our framework to explore the evolutionary dynamics in three related cell lines and two independent PDX timeseries. Here we aim to examine the predictive capabilities of our model. Specifically we are interested in verifying the selection coefficients estimated by `fitClone`. As such we return to the following hypothesis: given a resolved starting configuration, and inferred selection coefficients in a sufficiently similar context, it is possible to estimate subclonal trajectories. Testing this hypothesis motivates an experimental design that we call the *mixture experiments*. In this setup, we extract cells from an early (X3) and a late (X8) passage of the TNBC-SA609 series and then physically mix them. We established two such lines. In the first line (branch a) we aimed to have a mixture comprising equal proportions from the two timepoints (a ration of 1.0 to 1.0). In the second line (branch b) we aimed to have 71.0% of the cells come from the X3 passage and 29.0% of cells from the X8 passage (a ratio of 1.0 to 0.4). We did so to measure the effects of different starting population configurations on the selection coefficients. We call the initial physical mixture in each branch M0. For each branch, we describe the forecasting study and then the results from serially passaging the PDX line measured by DLP+.

Mixture branch a We forward-simulated trajectories from `fitClone` using the median of the posterior distribution of the estimated selection coefficients (F = -0.01 ± 0.13, A = -0.00 ± 0.16, B = 0.00 ± 0.01, D = 0.00 ± 0.01, G = 0.02 ± 0.02, H = 0.03 ± 0.03, E = 0.08 ± 0.10) and starting-clonal proportions of (A=0.000 (no observed cells), B=0.07, C=0.25, D=0.51, E=0.02, F=0.000 (no observed cells), G=0.08 H=0.07) Figures A.11-**a**, and 4.5-**a,d**. The starting clonal proportions were inferred by adding the cells from the initial passage (M0) to the tree inferred from the cells in the original series (TNBC-SA609) and assigning them to their respective clones. We generated 10,000 trajectories from the model. Figure 4.5-**c** (top) shows the simulated trajectories in black. The mean clonal fraction at each step is shown in red. All clones except for clones E, D, and F are predicted to vanish to clonal fractions of below 1%. We have combined the trajectories for clones E and F since clone F (i) had fewer than 19 total cells in the original series and consequently its selection coefficient had a high variance, (ii) is phylogenetically proximal to clone E (Figure 4.5-**b**) and thus likely represented a biologically similar population and (iii) finally, is not observed above a threshold of 20 cells in any other line in the TNBC-SA609 family.

We experimentally tested these predictions by initiating a new PDX line with the remixed population (from M0), serially passaged over 4 timepoints (Figure 4.5-**a** (top), and sequenced with DLP+ (7,839 single cell genomes, median 1,354.5 per library)). After placing the cells from all timepoints of this mixture experiment on the tree, they got assigned to seven clones from the original timeseries (all but clone A) with between 26 to 499 (median 162) cells (Figure 4.5-**c** (middle)). In Figure 4.5-**c** (top) blue dots show the observed clonal fractions at each timepoint in PDX branch a. Clones with higher selection coefficients swept through the mixture timeseries by passage 4. In the last timepoint, the clade composed of clones E and F comprised 94% of cells, outcompeting low-fitness clones. The estimated selection coefficients were relatively strongly correlated (Pearson correlation of 0.795, considering only clones that reached overall prevalence of over 1% in the original series, i.e., all clones except A and F).

Mixture branch b We forward simulated 10,000 trajectories using identical selection coefficient values as branch a, but different clonal proportions, estimated from adding cells from timepoint X1 in mixture branch b to the TNBC-SA609 phylogenetic tree and assigning them to the corresponding clones (C = 0.02, D = 0.00, E = 0.05, F = 0.00, G = 0.06, H = 0.87). Unlike mixture branch a, DLP+ data was not available for timepoint M0 in mixture branch b. We initiated a PDX line passaged over 5 timepoints that were sequenced using the DLP+ platform to generate 6,730 single cell genomes (median 1,270 per library). We observed 6 clones from the original series, namely all clones except A and B. However, we only captured over the 5 timepoints, 1 cell from clone D, 8 cells from clone F, and 13 cells from clone C. As predicted, clone E, which had the highest predicted selection coefficient in the original series despite having started from a very low clonal fraction (0.05), rises to an abundance of 0.28, while clone H steadily falls from 0.87 to about 0.67.

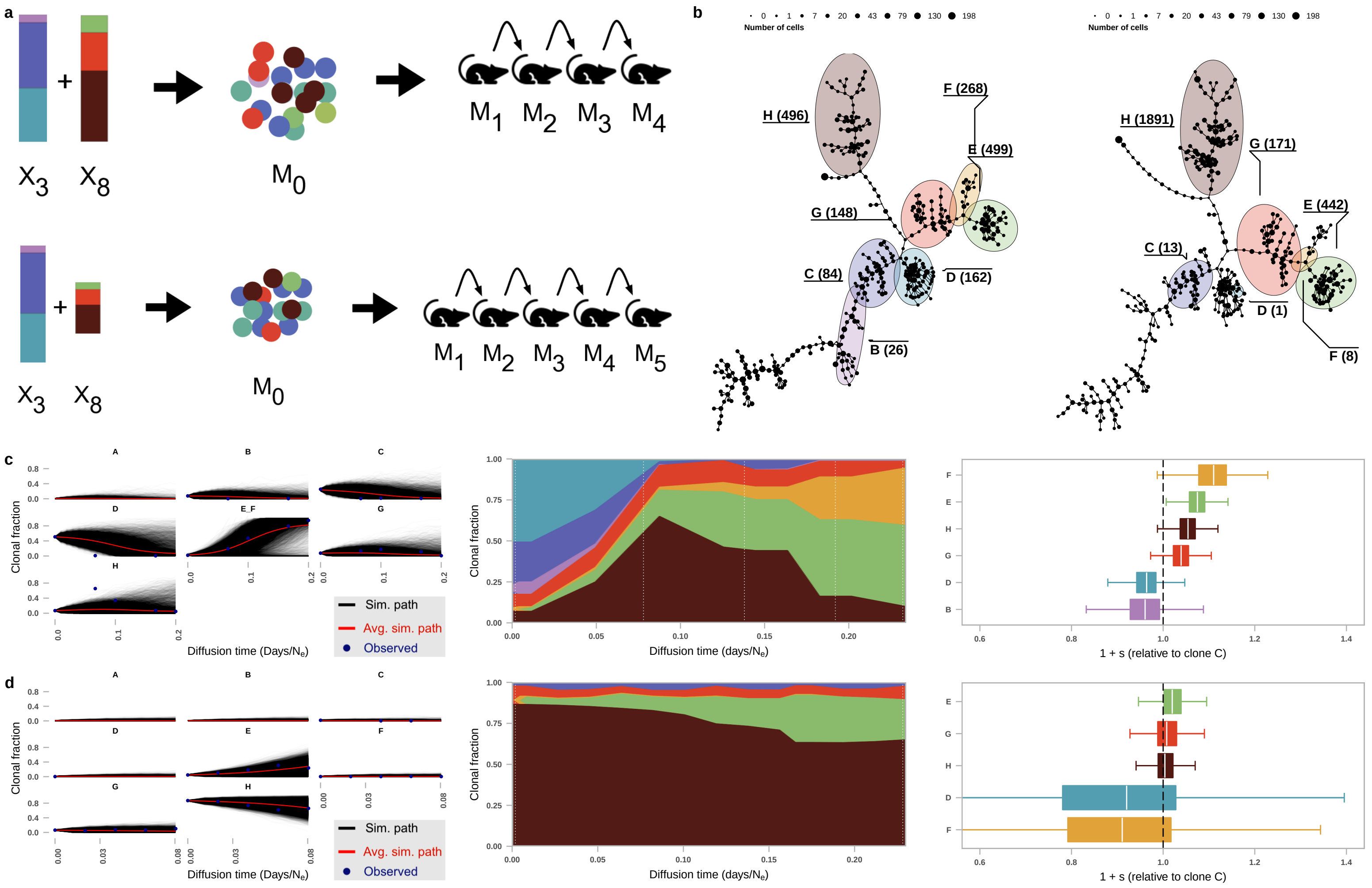


Figure 4.5: Mixture experiments on TNBC-SA609 PDX. In each panel, mixture a is followed by mixture b. (**a**) Clonal proportions of X3 and X8 to generate the initial mixture M0 and subsequent serial passaging, yielding 4 samples M1-M4 for mixture a and 5 samples M1-M5 for mixture b. (**b**) Phylogenies showing cells observed in the mixture a (left) and mixture b (right) timeseries. (**c**) For mixture a: (left) Forward simulations using inferred selection coefficients and starting population proportions in the initial experimental mixture. Simulated trajectories are shown. (middle) Inferred trajectories of mixture timeseries. (right) Selection coefficients of `fitClone` fit to M1-M4 clonal abundance observations. See Figure A.5 for the credible intervals and Figure A.6 for pairwise comparison of the selection coefficients. (**d**) as in (**c**) but for mixture b. Note that in the prediction trajectories plot (Figure 4.5-**c**,**d**-(left)), the time-axis for the observations is shrunk to best match the mean predicted trajectories line (red line).

Clone-specific fitness estimates forecast clonal competition trajectories

The analysis of the two mixture series presented here suggests that (i) we can validate the selection coefficients estimated from a timeseries using `fitClone` and (ii) it is possible make quantitative predictions about the likely trajectories of tumour subpopulations at least at the clade level. We note that in the prediction trajectories plot (Figure 4.5-**c**,**d**-left), the time-axis for the observations is shrunk to best match the mean predicted trajectories line (red line). The diffusion time horizon that is obtained by dividing the generation time measured in days by the N_e estimate results in trajectories that are ahead of the biological system. While in branch a the original time horizon is 0.24 and the best matching time is 0.20, in branch b the values are 0.23 and 0.08 respectively.

4.4 In vivo systems under cisplatin treatment

So far we have applied our framework to real-world datasets to track human cancer clones as identified by their CN genotypes over time and to reason about their likely abundances in certain experimental conditions. Now we investigate whether establishing baseline fitness measures could help to interpret selection under drug administration. We first present the TNBC-SA609 series in detail and make an observation about early response to cisplatin treatment. We then analyse two additional TNBC PDX lines to check the reproduabiblty of our observation.

4.4.1 TNBC-SA609 PDX

We tested how pharmacologic perturbation with cisplatin impacted the stability of the fitness landscape of the TNBC series. Using frozen material from the third timepoint (X3) in the original TNBC-SA609 untreated line introduced in Section 4.3, we established a new line that was passaged 4 times (X4, X5, X6, and X7) comprising 6,323 cells, 841 of which passed the QC step (Figure A.9). Figure A.1 shows how the different timepoints are related to one another. We also formed a separate branch from the same starting material (frozen X3) and administered cisplatin (2mg/kg, *Q3Dx8* i.p. max) serially over four successive passages to induce drug resistance (Figure A.11-**b**,**c**). For each serially treated tumour in mice, a parallel set of transplanted mice were left untreated, establishing corresponding drug 'holiday' samples (Figure 4.1-**a**). We coded the treated passages with 'T' and untreated with 'U', initialised by the X3 untreated (U) passage. The first treatment passage (*X4*

UT) exhibited rapid tumour shrinkage (>50% of initial size). However *X5 UTT*, *X6 UTTT* and *X7 UTTTT* had progressively less response, indicating drug resistance and positive growth kinetics (Figure A.11-**e**). Decomposing the growth dynamics over (*X3 U*; *X4 UT*; *X5 UTT*; *X6 UTTT*; *X7 UTTTT*) into clonal trajectories with DLP+ analysis suggested that sustained cisplatin treatment inverted the fitness landscape. A new clone R, derived from clone A in the phylogeny, but with a distinct clonal genotype (fewer copies of *MYC* and deletions at *RB1*, *PRDM9* and *NUDT15* loci (Figure 4.6-**a**,**b**)), swept to fixation comprising 48% (*X4 UT*), 98% (*X5 UTT*), 100% (*X6 UTTT*) and 100% (*X7 UTTTT*) of cells across the treated series (Figure 4.6-**c**). Notably, the high fitness clones E, H, G, D from the untreated series exhibited low fitness coefficients in the treatment series and were no longer detected (Figures 4.6-**d**, 4.9). Conversely, clones A, B, and C, comprising a low fitness phylogenetic superclade, distinct from high fitness clones E and F in the untreated series, were the precursors to the resistant clone R (Figure 4.6-**e**). Thus, cisplatin perturbation resulted in a near complete reversal in the fitness landscape.

Drug holiday

Next, we asked whether the clonal dynamics in the presence of cisplatin were reversible by examining the drug holiday samples (Figure 4.1-**a**, *X5 UTU*; *X6 UTTU*; *X7 UTTTU*). In the first drug holiday *X5 UTU*, clonal composition reverted to consist predominantly of precursor clone B with 90% abundance, and only 10% abundance from clone R (Figure 4.9-(left)). However, in *X6 UTTU* and *X7 UTTTU* no reversion was detected, and these populations consisted of >99% clone R, similar to their on-treatment analogues. Thus, clonal competition in the absence of drug led to clones derived from the A-B clade outcompeting clone R, and clone-specific cisplatin resistance has a fitness cost. Moreover, the genotype specificity of reversion between *X4 UT* to *X5 UTU* indicates that the clonal dynamics can be attributed to selection of genomically defined clones with differential fitness.

Together, these data suggest the impact of cisplatin selective pressure on the starting tumour cell population is reversible while genomic clonal competition with precursor clones is still possible, but dominates the population once the evolutionary bottleneck narrows and purifies the population.

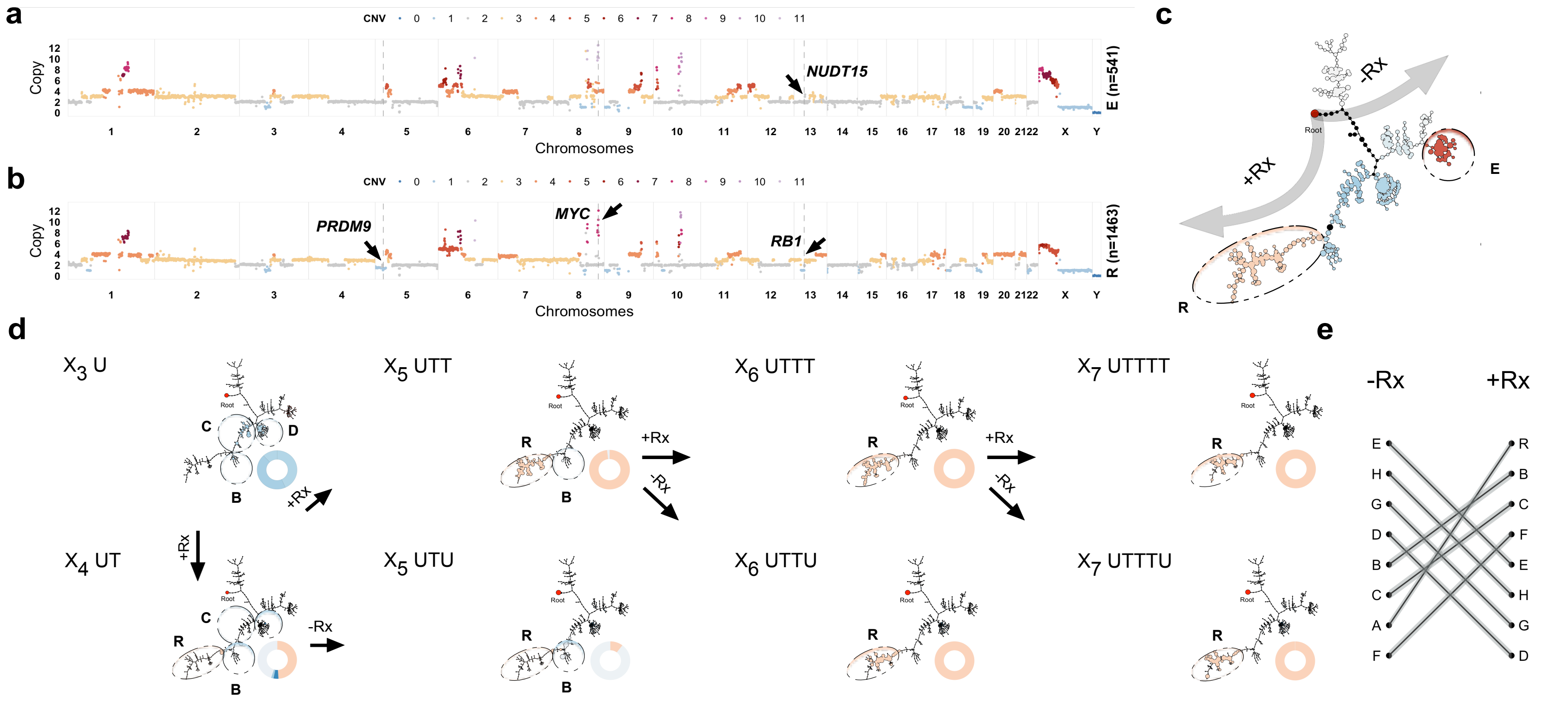


Figure 4.6: Impact of pharmacologic perturbation with cisplatin on fitness landscapes TNBC-SA609 PDX. (**a**) Copy number genotype of clone E from untreated timeseries. (**b**) Copy number genotype of clone R from treated timeseries (arrows indicate differences to clone E). (**c**) Clades with higher fitness in the untreated (-Rx) and treated (+Rx) series. (**d**) Evolution as a function of drug treatment and drug holiday. Arrows indicate the path of serial passaging. For each sample, the phylogeny with clonal abundance from DLP+ is shown, reflecting selection. (**e**) Reversal in the fitness landscape - note clone A maps to clone R in post treatment since R is derived from A, but only emerges in the treatment series.

4.4.2 Fitness landscape reversal in early response to cisplatin

We then asked if the reversal in the fitness landscape observed in the TNBC-SA609 system under cisplatin treatment is reproducible. To address this, we established two independent TNBC PDX systems as follows.

TNBC-SA535 We generated a total of 15,302 single cells out of which 4,023 passed our quality filters. TNBC-SA535 is a BRCA1 deficient patient derived tumour. We established a timeseries with treated and untreated branches similar to TNBC-SA609 (see Figure A.2). We generated 5 consecutively transplanted timepoints (X5, X6, X7, X8, X9) left untreated for a total of 1,341 single cells (mean $= 335$, $\sigma = 84.4$ per timepoint). Simultaneously, we established a cisplatin treated timeseries starting from timepoint X6, and continued cisplatin treatment for 5 cycles up to X10, generating a total of 1,425 cells from scWGS (mean $= 356$, $\sigma = 159$ per timepoint) from 4 cycles. A cut of the phylogenetic tree inferred over all cells in this series, resulted in 7 clones. In the untreated line, clonal fractions were A(0.003), B(0.702), C(0.034), D(0.006), E(0.006), F(0.013), and G(0.237). Clone B was chosen as the reference clone as it had a monotonically decreasing clonal fraction trajectory in the untreated branch. Clonal trajectories were consistent with selection coefficients with small relative differences in fitness (Figure 4.7-**d**). Clone G had the highest fitness ($1 + s = 1.01 \pm 0.00751$) closely followed by clone C ($1 + s = 1.00 \pm 0.0282$) and the reference clone. In the treated branch, clonal fractions were A(0.156), B(0.066), C(0.140), D(0.194), E(0.182), F(0.151), and G(0.112). In this regime, clone A emerged with the highest selection coefficient ($1 + s = 1.03 \pm 0.0152$) followed by clone D ($1 + s = 1.02 \pm 0.0116$). Notably clones A and D had low fitness values under no treatment, whereas clones G ($1 + s = 1.01 \pm 0.0115$) and C ($1 + s = 1.02 \pm 0.0119$) had low fitness coefficients under treatment.

TNBC-SA1035 Another independent PDX system with 14,170 single cells were generated where 4,444 passed the quality filters. The experimental design diagram is shown in Figure A.3. An untreated branch with five serial passages (X4, X5, X6, X7, and X8) was established with a total of 2,015 single cells. A parallel branch was treated with cisplatin starting at X5, X6, X7, and X8 comprising 1,596 filtered cells. 833 cells belonged to the drug-holiday timepoints and are not analyzed here. Phylogenetic inference followed by cutting the tree yielded 11 clones (Figure 4.8). Clonal fractions over all timepoints in the untreated branch were A(0.097), B(0.140), C(0.087), D(0.160), E(0.266), F(0.010), G(0.058), H(0.053), I(0.065), J(0.018), and K(0.047). Clone A's abundance fell over time and it was chosen as the reference clone. Clone E rose from a clonal fraction of 0.028 at X4 to 0.69 at X8 and had the highest selection coefficient ($1+s = 1.06 \pm 0.0367$). In the treated branch, clonal fractions were A(0.065), B(0.129), C(0.132), D(0.066), E(0.055), F(0.018), G(0.205), H(0.144), I(0.094), J(0.014), and K(0.078). In this regime, G ($1+s = 1.01 \pm 0.0123$) and H ($1+s = 1.02 \pm 0.0135$), which were among the clones with lower fitness in absence of treatment, rose to occupy 73% at X8 while clone E ($1+s = 0.993 \pm 0.0344$) fell from about 10% at X5 to undetectable at X8.

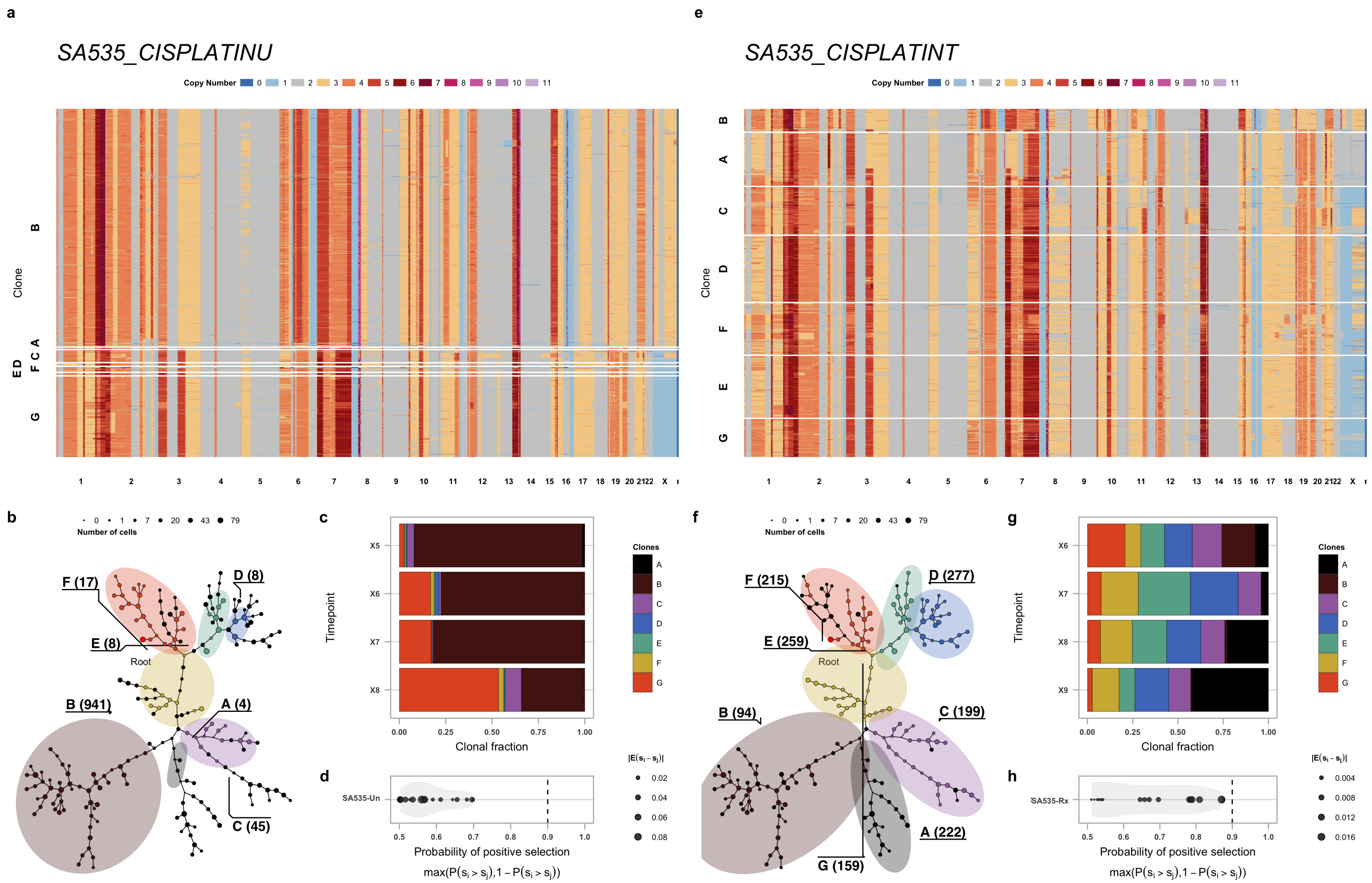


Figure 4.7: TNBC-SA535 PDX clonal dynamics with and without treatment. (**a**) Heatmap representation of copy number profiles of 1,341 cells, grouped in 7 phylogenetic clades. (**b**) Phylogeny as per Figure 4.2 for TNBC-SA535 PDX untreated branch. (**c**) Observed clonal abundances. (**d**) Distribution over magnitude of difference between selection coefficients of pairs of clones. (**e-h**) Analogous plots for the treated branch (n=1,425 cells).

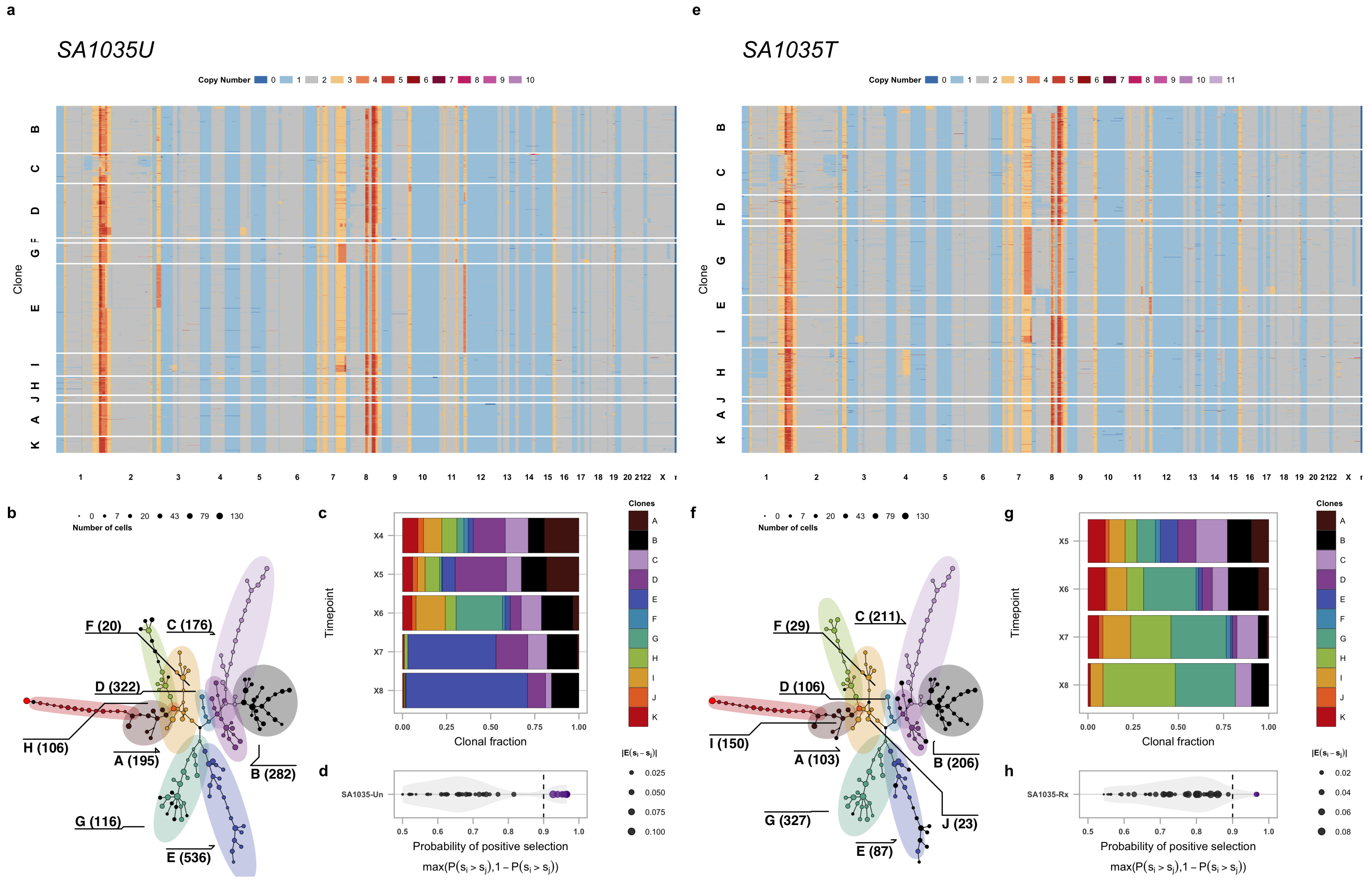


Figure 4.8: TNBC-SA1035 PDX clonal dynamics with and without treatment. (**a**) Heatmap representation of copy number profiles of 2,015 cells, grouped in 11 phylogenetic clades. (**b**) Phylogeny as per Figure 4.2 for TNBC-SA1035 PDX untreated branch. (**c**) Observed clonal abundances. (**d**) Distribution over magnitude of difference between selection coefficients of pairs of clones. (**e-h**) Analogous plots for the treated branch (n=1,596 cells).

Figure 4.9 summarises the reversal in the fitness landscape in response to cisplatin treatment in TNBC-PDX model systems. In the TNBC-SA609 system the fitness landscape is inverted wherein clones more fit in the untreated regime (H, D) are less fit in the treated regime, whereas less fit clones in the untreated regime (A, B) are the most fit clones under treatment. This pattern is mirrored in two independent TNBC PDX lines treated with cisplatin, namely TNBC-SA535 and TNBC-SA1035. In TNBC-SA535, clones G, C, and B are drug-sensitive, while A and D are drug-resistant. The former have higher relative fitness in untreated versus treated regimes, while the latter exhibit an inverted fitness pattern. Similarly, in TNBC-SA1035, drug-sensitive clones consist of clones E, B, and A, while the drug-resistant group comprises clones H, I, and G. From untreated to treated, the first group goes from high to low fitness, while the second group goes from low to high fitness.

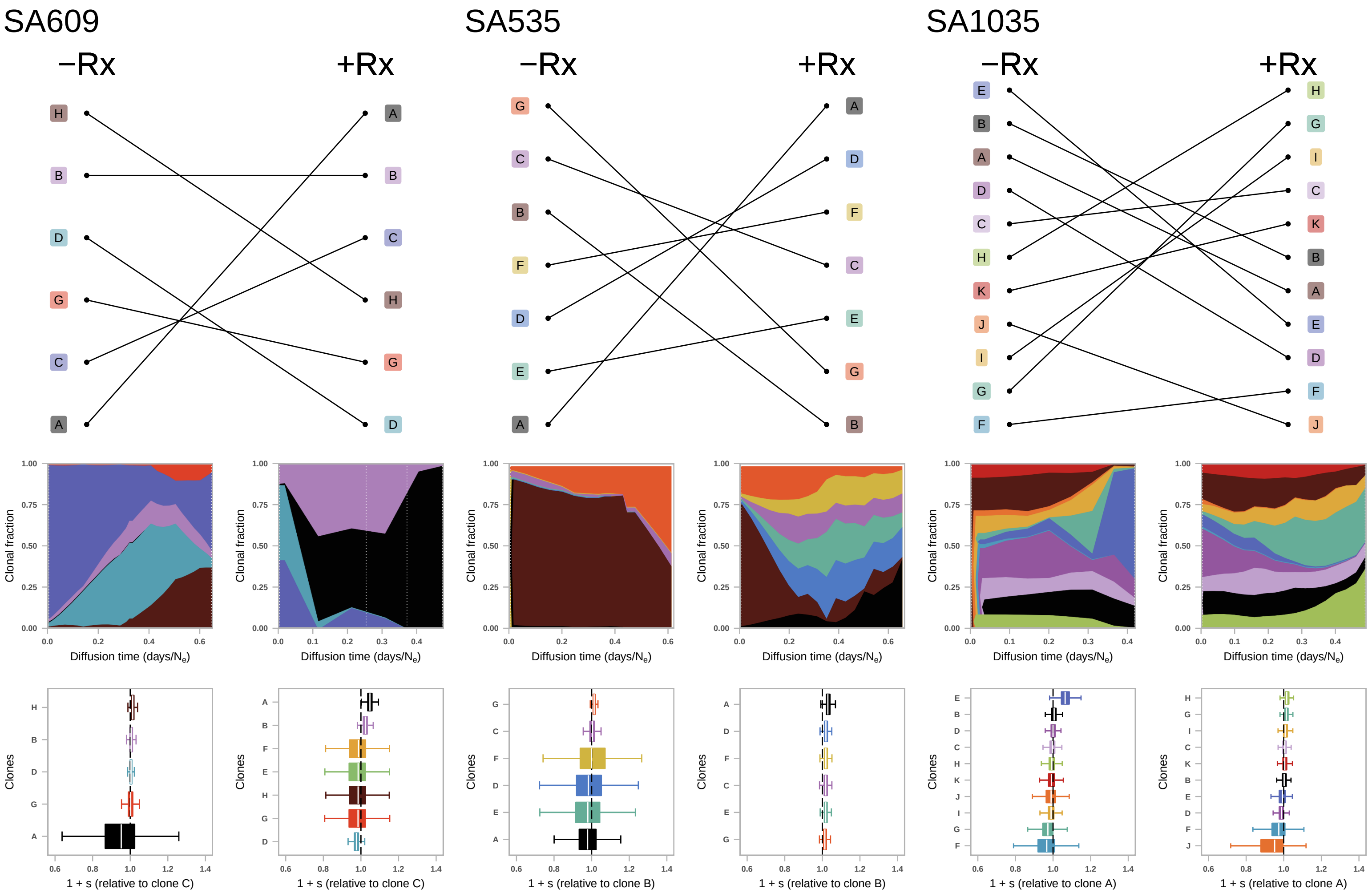


Figure 4.9: Fitness landscape reversal in early cisplatin treatment in TNBC PDX models. We observe that clone specific resistance to cisplatin treatment arises in 3 independent TNBC PDX lines. In all three cases, clones with low fitness under no treatment exhibit high fitness under the treatment regime. In each panel, the left and right sub-panels are from the untreated and treated branches respectively. (**top**) Clones sorted by their median selection coefficient in -Rx and +Rx regimes. (**middle**) Inferred trajectories, and (**bottom**) selection coefficients of `fitClone` model fits to each branch. See Figure A.5 for the credible intervals and Figure A.6 for pairwise comparison of the selection coefficients.

4.5 Discussion

Here we show that decoding the contributions of clonal competition landscapes in the course of tumour growth with and without drug selection can be achieved by measuring and modelling cellular dynamics with granular timeseries over several months-years. Single cell whole genome sequencing allowed for the robust application of population genetic statistical models. *TP53* ablation in diploid mammary epithelial cells resulted in clonal expansions driven by copy number alterations, with whole genome, whole chromosome and more localised alterations, all leading to fitness advantages over diploid baseline. Positive selection attributed to copy number changes of all scales may be under-investigated [6]. This has implications for interpreting etiologic processes of *TP53* driven cancers where the rates of structural variation acquisition and deviation away from diploid configurations confer quantitative fitness advantages. In tumour evolution within patients, copy number alterations, as a key biological process, is revealing impact on treatment outcomes and co-morbidities [105], innate and adaptive immune response [106, 107], the root cause of major genomic reconfigurations [108] and evolutionary plasticity [16]. Our work here suggests that *TP53* mutation can lead to structural alterations *in vitro* that are also observed in breast cancer patients, thereby representing a realistic model for studying how the impact of driver mutations inducing genomic instability leads to clonal expansions and evolutionary selection. Variously through single cell approaches [109] or computational reconstruction of evolutionary histories [9], the evolutionary impact of structural variations is still a work in progress. Over successive generations in vitro and in PDX, in the context of pharmacologic and mutational perturbation, emergent copy number changes contribute to the kinetics of the fitness landscape, consistent with a continual diversifying mechanism that induces competitive clonal advantages.

The impact of drug intervention on cancer evolution is a key determinant of patient outcomes across all human cancers. As clonal drug resistance was consistent with a fitness cost, we suggest this could be exploited in future therapeutic strategies. Forecasting the trajectories of cancer clones is of immediate importance to understanding therapeutic response in cancer and for deploying adaptive approaches [5]. The presence within a tumour of lineage precursors to resistant genotypes may define time windows within which clonal competition could mediate plasticity to treatment. We suggest that population genetic modelling of timeseries tumour or tumour cell-free

DNA measurements to predict clonal evolution is tractable, but will require complementary measurements of genotypic clonal abundance to gain comprehensive understanding. Further study with timeseries modelling will provide insight into therapeutic strategies promoting early intervention, drug combinations and evolution-aware approaches to clinical management [110].

4.6 Conclusions

In this chapter, we exemplified the computational methods that we developed in Chapters 2 and 3 on cell lines and PDX systems generated over a multi-year period. We used `sitka` and `Lumberjack` to identify the clonal subpopulations in heterogenous single cell subpopulations and their abundances over time based on their copy number profile. We then used `fitClone` to ascribe quantitative selection coefficients to individual cancer clones and forecast competitive clonal dynamics over time. Specifically we used a mixture competition design to experimentally verify the selection coefficients inferred by our model. We observed that fitness landscape is reversed in early response to treatment with cisplatin in three independent TNBC PDX timeseries. We hope that our results would help the research community map the fitness landscape of tumour development and ultimately aid in selecting better treatment regimes for patients with cancer.

4.6.1 Limitations and future research

In the context of pharmacological intervention, we are ultimately interested in mechanisms underlying sensitivity and resistance to drug. We acknowledge that establishing causal relationships is very difficult and requires controlled experiments and larger sample sizes. However, exploring the potential correlates of differential fitness may aid in designing follow up experiments, and in eventually finding possible druggable molecular targets. In the TNBC-SA609 PDX series, we did not find a single nucleotide variant (SNV) that was private to a clone and that was implicated in the literature as oncogenic. It is possible that such a mutation exists but was missed either at sequencing or by our computational pipelines. Follow up replication experiments sequenced at higher depth may be required to rule out the existence of driving mutations. Other biological mechanisms including epigenetic changes may influence the fitness of clones and interrogating the system via orthogonal data types such as chromatin accessibility and DNA methylation assays is the subject of future research.

In asexually reproducing populations, multiple subpopulations each with a set of beneficial mutations can coexist [111, 112]. TNBC-SA609 PDX series shows evidence of persistence of multiple clones over time 4.4-**f**. This competition between clones, called clonal interference, reduces the rate of adaptation and may lead to an underestimation of the selection coefficients of individual clones. In other words, in absence of competition from other clones, each clone may have a higher innate selection coefficient. Clonal interference can also result in more stochasticity in the evolutionary process [112]. As a consequence, the clones detected in one timeseries, may behave differently in absence of one or more of the original clones.

Exponential growth is a good fit to the initial stage of tumour growth. Once the tumour has grown to a certain size, it is expected that the growth would be boundary-driven. In an exponentially growing scenario, the observed clonal fraction trajectories more accurately represent the innate selection coefficients of clones, whereas in a boundary-driven growth scenario, spatial limitations bias the inferred selection coefficients; for instance, a highly fit clone can emerge in the middle of the tumour and due to space constraints will not grow as much a neutral clone growing at the boundary of the tumour [113].

We note that in the formulation of `fitClone`, no pair of clones can be assigned an equal selection coefficient. As a consequence, `fitClone` cannot delineate clones that are identical from those that are close in terms of evolutionary fitness. One remedy is to use a spike and slab prior that puts a point mass on $s_i = 0$.

Chapter 5

Conclusions

5.1 Summary of contributions

In this work we have presented a novel approach to investigate clonal evolution in longitudinal studies in vitro and in vivo model systems. The methodological contributions include (i) the preprocessing, benchmarking and development of visualisations for a novel Bayesian phylogenetics method (`sitka`) and its application in sorting single cells into clones in copy number space, along with an algorithm to cut the tree (`Lumberjack`), (ii) the adaptation of the Wright-Fisher diffusion model to track the trajectories of clonal fractions over time and (iii) the design and implementation of an inference engine to estimate evolutionary parameters and forecast in this model. We have applied these methods to investigate the predictability of clonal trajectories and quantify the response to cisplatin treatment in PDX models. Table 5.1 lists the repositories that host the source code for the software implemented in this work.

	Name	Description	URL
1	sitka-viz	Visualisation suite for Bayesian phylogenetics via `sitka` [1]	`https://github.com/UBC-Stat-ML/sitka-viz`
2	`fitClone` [2]	The inference engine for the Wright-Fisher model SDE	`https://github.com/UBC-Stat-ML/fitclone`
3	Lumberjack	The tree-cutting algorithm for `sitka`	`https://github.com/UBC-Stat-ML/tree_cutting`
4	sitka-material	Tools for preprocessing, post-processing, and benchmarking of `sitka`	`https://github.com/UBC-Stat-ML/sitka-material`

Table 5.1: List of repositories hosting the source code for software developed as part of this thesis.

An important open question in the treatment of patients with cancer is why and how resistance to therapy is developed. Advances in next generation single cell molecular sequencing as well as longitudinal xenoengraftment has enabled us to more accurately interrogate and make quantitative measurements of tumours. By serially transplanting a tumour sample into highly immunodeficient mice, a system called patient derived xenograft, one can continuously monitor how tumour composition may evolve in a patient over time. This process generates timestamped samples comprising thousands of cells that through SCS can be measured at thousands of genomic features including CNA profiles. This is in contrast to the standard of care where often only up to two timepoints are acquired and measured at an aggregate course level. The new higher resolution datatype introduces major challenges in extracting meaningful knowledge from the data, including (i) how to identify biologically meaningful groups of cells (i.e., clones) across multiple timepoints, and (ii) how to quantitatively reason about the underlying evolutionary forces acting on the clones via their observed dynamics. We have developed tools to generate meaningful summaries, and analyse and visualise thousands of cells over thousands of CNAs. These include combining timeseries phylogenetic trees and copy number state plots, identifying clones, and inferring and comparing their evolutionary fitness over time.

To ascertain the evolutionary relationships of the cells comprising the tumour samples, a new Bayesian phylogenetic inference method called `sitka` was developed in our research group. We developed a pipeline that accepts single cell CNAs as input, performs quality control and filtering defined by the researcher, and transforms it into a binary space and a format ready for phylogenetic analysis. The outputs of the analysis include the inferred consensus evolutionary tree, the predicted posterior binary genotype, and the error-corrected input matrices. We devised and implemented `Lumberjack`, an algorithm that cuts a phylogenetic tree by grouping subsets of cells that are genomically distinct based on both the topology of the tree and the cell's CNAs. These subsets, called clones, are biologically important entities that are hypothesised to encode unique phenotypes including differential response to therapeutic interventions. The phylogenetic inference outputs are then visualised in two modes, detailed and publication quality. In the former, the tree is aligned and plotted next to a heatmap representation of the CNAs displaying the genome of individual cells and their placement on the tree, interlaced with the posterior binary genotype state and the error-corrected input matrices. This has helped in interpreting and inspecting the inference results. The latter mode produces summarised trees organised in time steps reflective of the longitudinal experimental designs where changes in the tumour composition can be immediately gleaned via the change in colour (clone assignment) and size (number of constituent cells) of the nodes on the tree.

We used this visualisation to help characterise a unique mode of acquired reversible drug-resistance in breast cancers in response to cisplatin resistance.

Tracking the relative abundances of the clones over time produces a timeseries reminiscent of allele frequencies. We adopted ideas from the population genetics literature to model the evolution of tumour growth via a state space model, the dynamics of which is dictated by the diffusion approximation to the Wright-Fisher process. This framework, called `fitClone` induces an intractable likelihood that makes inference difficult. To address this, we extended a state of the art pseudo-marginal sequential Monte Carlo algorithm and implemented the inference engine. Briefly, the algorithm infers evolutionary fitness coefficients for each clone by simulating trajectories that are consistent with their observed clonal abundances. We devised a novel algorithmic extension that scaled simultaneous inference to dozens of clones. We then detected computational bottlenecks and re-implemented them in an efficient vectorised manner.

5.2 Future research directions

The Wright-Fisher model investigated in this work assumes all clones exist at the initial timepoint. If their observed clonal fraction is zero at the first timepoint, the model assumes that the clone exists at fractions under the threshold of detection. This ignores the possibility of de novo mutations and clones that arise only at later timepoints. One drawback is that estimated selection coefficients would be biased as the model has to account for the absence and then sudden rise of such clones by decreasing the selection coefficient. One potential remedy is to incorporate mutations using the Wright-Fisher model with selection and mutation.

The selection coefficient of each clone is also assumed to be constant over time. A step model that allows for change in the selection coefficient may produce a better fit to the data. The `sitka` phylogenetics model does not infer branch lengths. One potential direction is to use the timestamps of the single cells under each edge to infer a relative evolutionary change rate, that is to create a clocked model.

In this work we are restricted to measurements of clonal fractions over time, and are agnostic to the tumour microenvironment (TME), the spatial placement of clones, their epigenetic profile, protein levels, and clinical covariates, to name a few. Cancerous cells are situated in their TME and their interactions can shape tumour fate through mechanisms such as immune escape [114–116]. More broadly other evolutionary paradigms, clonal competition, cooperation, etc, are not investigated here. Furthermore we do not take into account any medical correlates including patient's age, smoking status, and family history. With the availability of larger longitudinal datasets, it may be possible to elucidate the effects of these covariates, perhaps via embedding the `fitClone` model in the framework of generalised linear models.

In Chapter 4 we identified clones using their genomic CN profiles. Establishing what genotypic features incur higher fitness is necessary for reliable rational treatment strategies. Epistasis in which multiple mutations across multiple genomic loci interact to control a phenotype are more difficult to identify. Further controlled experiments in much larger cohorts could be used to characterise such relationships.

In this work we only considered biological substrates acquired from solid tumours. Difficulties of obtaining biopsies from patients may hinder continual monitoring of tumour growth and clonal composition of patients over time. One potential way to alleviate this is to test for circulating tumour DNA (ctDNA) in blood samples. It has been shown that ctDNA has good fidelity to the original tumour in some cancers [117, 118]. It would be interesting to test the robustness and applicability of our framework to this type of data.

Advances in non-invasive imaging [119] establish a localised view at micro (histopathology) and macro (radiology) levels. Most diagnosis and prognosis routines involve the use of one family of data, e.g., one or more molecular-assays including genomics/transcriptomics/proteomics. Some effort has been made to leverage more than one data family, often combining imaging and molecular assays [120–122]. This line of research has established the complementary nature of multi-modal data, where for instance multiple genomic states of tumours such as aneuploidy are highly correlated with quantitative features extracted from histopathological images, and resulted in characterising some aspects of the tumour-microenvironment interaction [114]. While these models outperform diagnosis solely based on one family of datasets, the gain over the standard of care, typically involving grading and staging by a pathologist based on histology slides and input from molecular assays is negligible [121]. An interesting research direction is exploring the integrative causal multimodal frameworks which may improve the accuracy of diagnosis and prognosis [123–126].

5.3 Concluding remarks

Cancer is a complex phenomenon where the laws of Darwinian evolution can lead to a unique disease in each patient. This calls for a personalised rational treatment regime where the specific genomic aberrations of the patient's cancer are taken into account. Advancements in scalable molecular sequencing technologies allow for more accurate and more frequent monitoring of tumour development. Also, in the case of solid tumours where obtaining biopsies could be difficult, using ctDNA instead shows promise as a less invasive method to continuously track the clonal makeup of tumours. It can then be determined whether a tumour is relapsing and if so, which constituting clone(s) is responsible for the resurgence. Computational methods that model the dynamics of tumour development can be employed to prioritize the target and timing for treatment administration. The body of work developed here has been used in a study [2] to elucidate the early mechanism of tumour resistance to cisplatin. We hope that with availability of more longitudinal data, our framework would help the research community to map the fitness landscape of tumour development with and without intervention. This could aid in selecting better treatment regimes for patients with cancer.

Bibliography

[1] Fatemeh Dorri, Sohrab Salehi, Kevin Chern, Tyler Funnell, Marc Williams, Daniel Lai, Mirela Andronescu, Kieran R Campbell, Andrew McPherson, Samuel Aparicio, Andrew Roth, Sohrab P Shah, and Alexandre Bouchard-Côté. Efficient Bayesian inference of phylogenetic trees from large scale, low-depth genome-wide single-cell data. *bioRxiv*, 2020.

[2] S. Salehi, F. Kabir, N. Ceglia, M. Andronescu, M. Williams, K. R. Campbell, T. Masud, B. Wang, J. Biele, J. Brimhal, J. Ting, A. W. Zhang, C. O'Flanagan, F. Dorri, N. Rusk, E. Laks, H. Lee, T. Algara, S. Lee, B. Y. C. Cheng, P. Eirew, T. Kono, J. Pham, D. Grewel, D. Lai, R. Moore, A. J. Mungall, M. A Marra, IMAXT Consertium, A. McPherson, A. Bouchard-Côté, S. Aparicio, and S. P. Shah. Single cell fitness landscapes induced by genetic and pharmacologic perturbations in cancer. *To be published*, 2020.

[3] Sydney M Shaffer, Margaret C Dunagin, Stefan R Torborg, Eduardo A Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A Brafford, Min Xiao, Elliott Eggan, Ioannis N Anastopoulos, Cesar A Vargas-Garcia, Abhyudai Singh, Katherine L Nathanson, Meenhard Herlyn, and Arjun Raj. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431–435, June 2017.

[4] Chong Sun, Liqin Wang, Sidong Huang, Guus Jje Heynen, Anirudh Prahallad, Caroline Robert, John Haanen, Christian Blank, Jelle Wesseling, Stefan M Willems, and Others. Reversible and adaptive resistance to BRAF (V600E) inhibition in melanoma. *Nature*, 508(7494):118–122, 2014.

[5] Neil Vasan, José Baselga, and David M Hyman. A view on drug resistance in cancer. *Nature*, 575(7782):299–309, November 2019.

[6] Devon A. Lukow, Erin L. Sausville, Pavit Suri, Narendra Kumar Chunduri, Justin Leu, Jude

Kendall, Zihua Wang, Zuzana Storchova, and Jason M. Sheltzer. Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. *bioRxiv*, 2020.

[7] Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature genetics*, 48(3):238, 2016.

[8] Marc J Williams, Benjamin Werner, Timon Heide, Christina Curtis, Chris P Barnes, Andrea Sottoriva, and Trevor A Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature genetics*, page 1, 2018.

[9] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C Boutros, David D Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D Morris, PCAWG Evolution & Heterogeneity Working Group, Paul T Spellman, David C Wedge, Peter Van Loo, and PCAWG Consortium. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, February 2020.

[10] S P Shah, A Roth, R Goya, A Oloumi, G Ha, Y Zhao, G Turashvili, J Ding, K Tse, G Haffari, A Bashashati, L M Prentice, J Khattra, A Burleigh, D Yap, V Bernard, A McPherson, K Shumansky, A Crisan, R Giuliany, A Heravi-Moussavi, J Rosner, D Lai, I Birol, R Varhol, A Tam, N Dhalla, T Zeng, K Ma, S K Chan, M Griffith, A Moradian, S W Cheng, G B Morin, P Watson, K Gelmon, S Chia, S F Chin, C Curtis, O M Rueda, P D Pharoah, S Damaraju, J Mackey, K Hoon, T Harkins, V Tadigotla, M Sigaroudinia, P Gascard, T Tlsty, J F Costello, I M Meyer, C J Eaves, W W Wasserman, S Jones, D Huntsman, M Hirst, C Caldas, M A Marra, and S Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399, June 2012.

[11] S Nik-Zainal, P Van Loo, D C Wedge, L B Alexandrov, C D Greenman, K W Lau, K Raine,

D Jones, J Marshall, M Ramakrishna, and Others. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

[12] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues, 2017.

[13] Khurum H Khan, David Cunningham, Benjamin Werner, Georgios Vlachogiannis, Inmaculada Spiteri, Timon Heide, Javier Fernandez Mateos, Alexandra Vatsiou, Andrea Lampis, Mahnaz Darvish Damavandi, Hazel Lote, Ian Said Huntingford, Somaieh Hedayat, Ian Chau, Nina Tunariu, Giulia Mentrasti, Francesco Trevisani, Sheela Rao, Gayathri Anandappa, David Watkins, Naureen Starling, Janet Thomas, Clare Peckitt, Nasir Khan, Massimo Rugge, Ruwaida Begum, Blanka Hezelova, Annette Bryant, Thomas Jones, Paula Proszek, Matteo Fassan, Jens C Hahne, Michael Hubank, Chiara Braconi, Andrea Sottoriva, and Nicola Valeri. Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-C phase II colorectal cancer clinical trial, 2018.

[14] Marco Gerlinger, Stuart Horswell, James Larkin, Andrew J Rowan, Max P Salm, Ignacio Varela, Rosalie Fisher, Nicholas McGranahan, Nicholas Matthews, Claudio R Santos, Pierre Martinez, Benjamin Phillimore, Sharmin Begum, Adam Rabinowitz, Bradley Spencer-Dene, Sakshi Gulati, Paul A Bates, Gordon Stamp, Lisa Pickering, Martin Gore, David L Nicol, Steven Hazell, P Andrew Futreal, Aengus Stewart, and Charles Swanton. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.*, 46(3):225–233, March 2014.

[15] Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas BK Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

[16] Saioa López, TRACERx Consortium, Emilia L Lim, Stuart Horswell, Kerstin Haase, Ariana Huebner, Michelle Dietzen, Thanos P Mourikis, Thomas B K Watkins, Andrew Rowan, Sally M Dewhurst, Nicolai J Birkbak, Gareth A Wilson, Peter Van Loo, Mariam Jamal-

Hanjani, Charles Swanton, and Nicholas McGranahan. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution, 2020.

[17] Andrew McPherson, Andrew Roth, Emma Laks, Tehmina Masud, Ali Bashashati, Allen W Zhang, Gavin Ha, Justina Biele, Damian Yap, Adrian Wan, Leah M Prentice, Jaswinder Khattra, Maia A Smith, Cydney B Nielsen, Sarah C Mullaly, Steve Kalloger, Anthony Karnezis, Karey Shumansky, Celia Siu, Jamie Rosner, Hector Li Chan, Julie Ho, Nataliya Melnyk, Janine Senz, Winnie Yang, Richard Moore, Andrew J Mungall, Marco A Marra, Alexandre Bouchard-Côté, C Blake Gilks, David G Huntsman, Jessica N McAlpine, Samuel Aparicio, and Sohrab P Shah. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.*, 48(7):758–767, 2016.

[18] Caroline J Watson, A L Papula, Gladys Y P Poon, Wing H Wong, Andrew L Young, Todd E Druley, Daniel S Fisher, and Jamie R Blundell. The evolutionary dynamics and fitness landscape of clonal hematopoiesis, 2020.

[19] Benjamin H Good, Michael J McDonald, Jeffrey E Barrick, Richard E Lenski, and Michael M Desai. The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50, November 2017.

[20] Emma Laks, Andrew McPherson, Hans Zahn, Daniel Lai, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Jerome Ting, Diljot Grewal, Cydney Nielsen, Samantha Leung, Viktoria Bojilova, Maia Smith, Oleg Golovko, Steven Poon, Peter Eirew, Farhia Kabeer, Teresa Ruiz de Algara, So Ra Lee, M Jafar Taghiyar, Curtis Huebner, Jessica Ngo, Tim Chan, Spencer Vatrt-Watts, Pascale Walters, Nafis Abrar, Sophia Chan, Matt Wiens, Lauren Martin, R Wilder Scott, T Michael Underhill, Elizabeth Chavez, Christian Steidl, Daniel Da Costa, Yussanne Ma, Robin J N Coope, Richard Corbett, Stephen Pleasance, Richard Moore, Andrew J Mungall, Colin Mar, Fergus Cafferty, Karen Gelmon, Stephen Chia, CRUK IMAXT Grand Challenge Team, Marco A Marra, Carl Hansen, Sohrab P Shah, and Samuel Aparicio. Clonal decomposition and DNA replication states defined by scaled Single-Cell genome sequencing. *Cell*, 179(5):1207–1221.e22, November 2019.

[21] Hans Zahn, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah,

Samuel Aparicio, and Carl L Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature methods*, 14(2):167, 2017.

[22] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014.

[23] Sohrab Salehi, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome biology*, 18(1):44, 2017.

[24] Yi Kan Wang, Ali Bashashati, Michael S Anglesio, Dawn R Cochrane, Diljot S Grewal, Gavin Ha, Andrew McPherson, Hugo M Horlings, Janine Senz, Leah M Prentice, et al. Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nature genetics*, 49(6):856, 2017.

[25] Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, 13(10):714, 2013.

[26] Xueli Nan, Chao Xie, Xueyan Yu, and Jie Liu. EGFR TKI as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer. *Oncotarget*, 8(43):75712, 2017.

[27] Sandra Misale, Rona Yaeger, Sebastijan Hobor, Elisa Scala, Manickam Janakiraman, David Liska, Emanuele Valtorta, Roberta Schiavo, Michela Buscarino, Giulia Siravegna, et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*, 486(7404):532, 2012.

[28] William A Hammond, Abhisek Swaika, and Kabir Mody. Pharmacologic resistance in colorectal cancer: a review. *Therapeutic advances in medical oncology*, 8(1):57–84, 2016.

[29] Tomas Reinert, Everardo D Saad, Carlos H Barrios, and José Bines. Clinical implications of ESR1 mutations in hormone receptor-positive advanced breast cancer. *Frontiers in oncology*, 7:26, 2017.

[30] Lindsay Angus, Nick Beije, Agnes Jager, John WM Martens, and Stefan Sleijfer. ESR1 mutations: Moving towards guiding treatment decision-making in metastatic breast cancer patients. *Cancer treatment reviews*, 52:33–40, 2017.

[31] Benjamin G Bitler, Zachary L Watson, Lindsay J Wheeler, and Kian Behbakht. PARP inhibitors: clinical utility and possibilities of overcoming resistance. *Gynecologic oncology*, 2017.

[32] Dan Graur and Wen-Hsiung Li. *Fundamentals of Molecular Evolution*. Sinauer Associates Inc., 2nd edition, 2000.

[33] Haiwei Luo and Austin L Hughes. dN/dS does not show positive selection drives separation of polar-tropical SAR11 populations. *Molecular systems biology*, 8(1):625, 2012.

[34] Malachi Griffith, Christopher A Miller, Obi L Griffith, Kilannin Krysiak, Zachary L Skidmore, Avinash Ramu, Jason R Walker, Ha X Dang, Lee Trani, David E Larson, et al. Optimizing cancer genome sequencing and analysis. *Cell systems*, 1(3):210–223, 2015.

[35] Andrea Sottoriva, Haeyoun Kang, Zhicheng Ma, Trevor A Graham, Matthew P Salomon, Junsong Zhao, Paul Marjoram, Kimberly Siegmund, Michael F Press, Darryl Shibata, et al. A Big Bang model of human colorectal tumor growth. *Nature genetics*, 47(3):209, 2015.

[36] Sarah P Otto and Troy Day. *A biologist's guide to mathematical modeling in ecology and evolution*. Princeton University Press, 2011.

[37] Jason Moffat, Dorre A Grueneberg, Xiaoping Yang, So Young Kim, Angela M Kloepfer, Gregory Hinkle, Bruno Piqani, Thomas M Eisenhaure, Biao Luo, Jennifer K Grenier, et al. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, 124(6):1283–1298, 2006.

[38] Ophir Shalem, Neville E Sanjana, Ella Hartenian, Xi Shi, David A Scott, Tarjei S Mikkelsen, Dirk Heckl, Benjamin L Ebert, David E Root, John G Doench, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–87, 2014.

[39] Bernhard Schmierer, Sandeep K Botla, Jilin Zhang, Mikko Turunen, Teemu Kivioja, and Jussi Taipale. CRISPR/Cas9 screening using unique molecular identifiers. *Molecular systems biology*, 13(10):945, 2017.

[40] Zoë N Rogers, Christopher D McFarland, Ian P Winters, Santiago Naranjo, Chen-Hua Chuang, Dmitri Petrov, and Monte M Winslow. A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression in vivo. *Nature methods*, 14(7):737, 2017.

[41] Thomas O. McDonald, Shaon Chakrabarti, and Franziska Michor. Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. *Nature Genetics*, 2018.

[42] Timon Heide, Luis Zapata, Marc J. Williams, Benjamin Werner, Giulio Caravagna, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva. Reply to 'Neutral tumor evolution?'. *Nature Genetics*, 2018.

[43] Peter Eirew, Adi Steif, Jaswinder Khattra, Gavin Ha, Damian Yap, Hossein Farahani, Karen Gelmon, Stephen Chia, Colin Mar, Adrian Wan, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, 2015.

[44] Juliet A Williams. Using PDX for preclinical cancer drug discovery: the evolving field. *Journal of clinical medicine*, 7(3):41, 2018.

[45] Christopher D Willey, Ashley N Gilbert, Joshua C Anderson, and George Yancey Gillespie. Patient-derived xenografts as a model system for radiation research. In *Seminars in radiation oncology*, volume 25, pages 273–280. Elsevier, 2015.

[46] Gaelen T Hess, Laure Frésard, Kyuho Han, Cameron H Lee, Amy Li, Karlene A Cimprich, Stephen B Montgomery, and Michael C Bassik. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nature methods*, 13(12):1036, 2016.

[47] Kaston Leung, Anders Klaus, Bill K Lin, Emma Laks, Justina Biele, Daniel Lai, Ali Bashashati, Yi-Fei Huang, Radhouane Aniba, Michelle Moksa, et al. Robust high-performance nanoliter-volume single-cell multiple displacement amplification on planar substrates. *Proceedings of the National Academy of Sciences*, 113(30):8484–8489, 2016.

[48] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155, 2014.

[49] E. M. Ross and F. Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17:69, 2016.

[50] Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017.

[51] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):1–17, 2016.

[52] Christopher A Miller, Brian S White, Nathan D Dees, Malachi Griffith, John S Welch, Obi L Griffith, Ravi Vij, Michael H Tomasson, Timothy A Graubert, Matthew J Walter, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665, 2014.

[53] J. Singer, J. Kuipers, K. Jahn, and N. Beerenwinkel. SCIΦ: Single-cell mutation identification via phylogenetic inference. *bioRxiv*, page 290908, March 2018.

[54] Fang Wang, Qihan Wang, Vakul Mohanty, Shaoheng Liang, Jinzhuang Dou, Jincheng Han, Darlan Conterno Minussi, Ruli Gao, Li Ding, Nicholas Navin, and Ken Chen. Single-cell copy number lineage tracing enabling gene discovery. *bioRxiv*, 2020.

[55] X. Xu, Y. Hou, X. Yin, L. Bao, A. Tang, L. Song, F. Li, S. Tsang, K. Wu, H. Wu, W. He, L. Zeng, M. Xing, R. Wu, H. Jiang, X. Liu, D. Cao, G. Guo, X. Hu, Y. Gui, Z. Li, W. Xie, X. Sun, M. Shi, Z. Cai, B. Wang, M. Zhong, J. Li, Z. Lu, N. Gu, X. Zhang, L. Goodman, L. Bolund, J. Wang, H. Yang, K. Kristiansen, M. Dean, Y. Li, and J. Wang. Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. *Cell*, 148(5):886–895, March 2012.

[56] T. L. Williams and B. M. E. Moret. An investigation of phylogenetic likelihood methods. In *Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings.*, pages 79–86, March 2003.

[57] D.B. Wilson. Generating Random Spanning Trees More Quickly Than the Cover Time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 296–303, New York, NY, USA, 1996. ACM.

[58] Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[59] A. Bouchard-Côté, K. Chern, D. Cubranic, S. Hosseini, J. Hume, M. Lepur, Z. Ouyang, and G. Sgarbi. Blang: Bayesian declarative modelling of arbitrary data structures. *arXiv:1912.10396 [stat]*, 2019.

[60] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[61] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[62] Alan E Gelfand. Model determination using sampling-based methods. *Markov chain Monte Carlo in practice*, pages 145–161, 1996.

[63] R.R. Sokal, C.D. Michener, and University of Kansas. *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas, 1958.

[64] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[65] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[66] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[67] John P Huelsenbeck and Fredrik Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[68] Daniel L Ayres, Michael P Cummings, Guy Baele, Aaron E Darling, Paul O Lewis, David L Swofford, John P Huelsenbeck, Philippe Lemey, Andrew Rambaut, and Marc A Suchard.

Beagle 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Systematic biology*, 68(6):1052–1061, 2019.

[69] Kijong Yi and Young Seok Ju. Patterns and mechanisms of structural variations in human cancer. *Experimental & Molecular Medicine*, 50(8):98, 2018.

[70] Sweta Mishra and Johnathan R. Whetstine. Different facets of copy number changes: Permanent, transient, and adaptive. *Molecular and Cellular Biology*, 36(7):1050–1063, 2016.

[71] Richard A Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences*, 110(2):437–442, 2013.

[72] Jeffrey J. Quinn, Matthew G. Jones, Ross A. Okimoto, Shigeki Nanjo, Michelle M. Chan, Nir Yosef, Trever G. Bivona, and Jonathan S. Weissman. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science*, 2021.

[73] Jeffrey E Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E Lenski, and Jihyun F Kim. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243–1247, 2009.

[74] Jose Blanchet. Exact simulation for multivariate Itô diffusions. 2017.

[75] Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical inference in the Wright–Fisher model using allele frequency data. *Systematic biology*, 66(1):e30–e46, 2017.

[76] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[77] Matthieu Foll, Yu-Ping Poh, Nicholas Renzette, Anna Ferrer-Admetlla, Claudia Bank, Hyunjin Shim, Anna-Sapfo Malaspinas, Gregory Ewing, Ping Liu, Daniel Wegmann, et al. Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genet*, 10(2):e1004185, 2014.

[78] Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. Estimation of 2Nes from temporal allele frequency data. *Genetics*, 179(1):497–502, 2008.

[79] Anna-Sapfo Malaspinas, Orestis Malaspinas, Steven N Evans, and Montgomery Slatkin. Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):599–607, 2012.

[80] Anna Ferrer-Admetlla, Christoph Leuenberger, Jeffrey D Jensen, and Daniel Wegmann. An approximate markov model for the Wright–Fisher diffusion and its application to time series data. *Genetics*, 203(2):831–846, 2016.

[81] Alexandros Beskos, Gareth O Roberts, et al. Exact simulation of diffusions. *The Annals of Applied Probability*, 15(4):2422–2444, 2005.

[82] Murray Pollock, Adam M Johansen, Gareth O Roberts, et al. On the exact and ε-strong simulation of (jump) diffusions. *Bernoulli*, 22(2):794–856, 2016.

[83] Paul A Jenkins, Dario Spano, et al. Exact simulation of the Wright–Fisher diffusion. *The Annals of Applied Probability*, 27(3):1478–1509, 2017.

[84] Simon JA Malham and Anke Wiese. An introduction to SDE simulation. *arXiv preprint arXiv:1004.0646*, 2010.

[85] Fredrik Lindsten, Michael I Jordan, and Thomas B Schön. Particle Gibbs with ancestor sampling. *The Journal of Machine Learning Research*, 15(1):2145–2184, 2014.

[86] Fredrik Lindsten, Pete Bunch, Sumeetpal S Singh, and Thomas B Schön. Particle ancestor sampling for near-degenerate or intractable state transition models. *arXiv preprint arXiv:1505.06356*, 2015.

[87] PE Jorde, S Palm, and N Ryman. Estimating genetic drift and effective population size from temporal shifts in dominant gene marker frequencies. *Molecular Ecology*, 8(7):1171–1178, 1999.

[88] Masatoshi Nei and Fumio Tajima. Genetic drift and estimation of effective population size. *Genetics*, 98(3):625–640, 1981.

[89] Jorge Nocedal and Stephen Wright. *Numerical optimization.* Springer Science & Business Media, 2006.

[90] Angela Burleigh, Steven McKinney, Jazmine Brimhall, Damian Yap, Peter Eirew, Steven Poon, Viola Ng, Adrian Wan, Leah Prentice, Lois Annab, J Carl Barrett, Carlos Caldas, Connie Eaves, and Samuel Aparicio. A co-culture genome-wide RNAi screen with mammary epithelial cells reveals transmembrane signals required for growth and differentiation. *Breast Cancer Res.*, 17:4, January 2015.

[91] Alejandra Bruna, Oscar M Rueda, Wendy Greenwood, Ankita Sati Batra, Maurizio Callari, Rajbir Nath Batra, Katherine Pogrebniak, Jose Sandoval, John W Cassidy, Ana Tufegdzic-Vidakovic, Stephen-John Sammut, Linda Jones, Elena Provenzano, Richard Baird, Peter Eirew, James Hadfield, Matthew Eldridge, Anne McLaren-Douglas, Andrew Barthorpe, Howard Lightfoot, Mark J O'Connor, Joe Gray, Javier Cortes, Jose Baselga, Elisabetta Marangoni, Alana L Welm, Samuel Aparicio, Violeta Serra, Mathew J Garnett, and Carlos Caldas. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell*, 167(1):260–274.e22, September 2016.

[92] Peter M Haverty, Eva Lin, Jenille Tan, Yihong Yu, Billy Lam, Steve Lianoglou, Richard M Neve, Scott Martin, Jeff Settleman, Robert L Yauch, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature*, 533(7603):333–337, 2016.

[93] Wesley Tansey, Kathy Li, Haoran Zhang, Scott W Linderman, Raul Rabadan, David M Blei, and Chris H Wiggins. Dose-response modeling in high-throughput cancer drug screenings: A case study with recommendations for practitioners. *arXiv preprint arXiv:1812.05691*, 2018.

[94] Montserrat Rojo de la Vega, Montserrat Rojo de la Vega, Eli Chapman, and Donna D Zhang. NRF2 and the hallmarks of cancer, 2018.

[95] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, October 2012.

[96] Ann-Marie Patch, Elizabeth L Christie, Dariush Etemadmoghadam, Dale W Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J Bailey, Karin S Kassahn, Felicity Newell, Michael C J Quinn, Stephen Kazakoff, Kelly Quek, Charlotte Wilhelm-Benartzi, Ed Curry, Huei San Leong, Australian Ovarian Cancer Study Group, Anne

Hamilton, Linda Mileshkin, George Au-Yeung, Catherine Kennedy, Jillian Hung, Yoke-Eng Chiew, Paul Harnett, Michael Friedlander, Michael Quinn, Jan Pyman, Stephen Cordner, Patricia O'Brien, Jodie Leditschke, Greg Young, Kate Strachan, Paul Waring, Walid Azar, Chris Mitchell, Nadia Traficante, Joy Hendley, Heather Thorne, Mark Shackleton, David K Miller, Gisela Mir Arnau, Richard W Tothill, Timothy P Holloway, Timothy Semple, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Senel Idrisoglu, Timothy J C Bruxner, Angelika N Christ, Barsha Poudel, Oliver Holmes, Matthew Anderson, Conrad Leonard, Andrew Lonie, Nathan Hall, Scott Wood, Darrin F Taylor, Qinying Xu, J Lynn Fink, Nick Waddell, Ronny Drapkin, Euan Stronach, Hani Gabra, Robert Brown, Andrea Jewell, Shivashankar H Nagaraj, Emma Markham, Peter J Wilson, Jason Ellul, Orla McNally, Maria A Doyle, Ravikiran Vedururu, Collin Stewart, Ernst Lengyel, John V Pearson, Nicola Waddell, Anna deFazio, Sean M Grimmond, and David D L Bowtell. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–494, May 2015.

[97] Yilong Li, Nicola D Roberts, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O Korbel, James E Haber, Marcin Imielinski, PCAWG Structural Variation Working Group, Joachim Weischenfeldt, Rameen Beroukhim, Peter J Campbell, and PCAWG Consortium. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, February 2020.

[98] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B Brinkman, Sandro Morganella, Miriam R Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A Foekens, Moritz Gerstung, Gerrit K J Hooijer, Se Jin Jang, David R Jones, Hyung-Yong Kim, Tari A King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A Purdie, Keiran Raine, Kamna Ramakrishnan, F Germán Rodríguez-González, Gilles Romieu, Anieta M Sieuwerts, Peter T Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux,

Gert G Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura Van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T Ueno, Christos Sotiriou, Alain Viari, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, John W M Martens, Anne-Lise Børresen-Dale, Andrea L Richardson, Gu Kong, Gilles Thomas, and Michael R Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534:47, May 2016.

[99] C Curtis, S P Shah, S F Chin, G Turashvili, O M Rueda, M J Dunning, D Speed, A G Lynch, S Samarajiwa, Y Yuan, S Gräf, G Ha, G Haffari, A Bashashati, R Russell, S McKinney, METABRIC Group, A Langerod, A Green, E Provenzano, G Wishart, S Pinder, P Watson, F Markowetz, L Murphy, I Ellis, A Purushotham, A L Borresen-Dale, J D Brenton, S Tavaré, C Caldas, and S Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, June 2012.

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

[101] Mieke R Van Bockstal, Marie Colombe Agahozo, Ronald van Marion, Peggy N Atmodimedjo, Hein FBM Sleddens, Winand NM Dinjens, Lindy L Visser, Esther H Lips, Jelle Wesseling, and Carolien HM van Deurzen. Somatic mutations and copy number variations in breast cancers with heterogeneous HER2 amplification. *Molecular oncology*, 14(4):671–685, 2020.

[102] Zaibing Li, Xiao Zhang, Chenxin Hou, Yuqing Zhou, Junli Chen, Haoyang Cai, Yifeng Ye, Jinping Liu, and Ning Huang. Comprehensive identification and characterization of somatic copy number alterations in triple-negative breast cancer. *International journal of oncology*, 56(2):522–530, 02 2020.

[103] Arne V. Pladsen, Gro Nilsen, Oscar M. Rueda, Miriam R. Aure, Ørnulf Borgan, Knut Liestøl, Valeria Vitelli, Arnoldo Frigessi, Anita Langerød, Anthony Mathelier, Tone F. Bathen, Elin Borgen, Anne-Lise Børresen-Dale, Olav Engebråten, Britt Fritzman, Øystein Garred, Jürgen Geisler, Gry Aarum Geitvik, Solveig Hofvind, Vessela Kristensen, Rolf Kåresen, Ole Chris-

tian Lingjærde, Gunhild Mari Mælandsmo, Bjørn Naume, Hege G. Russnes, Kristine Kleivi Sahlberg, Torill Sauer, Helle Kristine Skjerven, Ellen Schlichting, Therese Sørlie, David C. Wedge, Peter Van Loo, Carlos Caldas, Hege G. Russnes, and OSBREAC. DNA copy number motifs are strong and independent predictors of survival in breast cancer. *Communications Biology*, 3(1):153, 2020.

[104] Xin Shao, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical Genetics*, 20(1):175, 2019.

[105] Craig M Bielski, Ahmet Zehir, Alexander V Penson, Mark T A Donoghue, Walid Chatila, Joshua Armenia, Matthew T Chang, Alison M Schram, Philip Jonsson, Chaitanya Bandlamudi, Pedram Razavi, Gopa Iyer, Mark E Robson, Zsofia K Stadler, Nikolaus Schultz, Jose Baselga, David B Solit, David M Hyman, Michael F Berger, and Barry S Taylor. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.*, 50(8):1189–1195, August 2018.

[106] Samuel F Bakhoum, Bryan Ngo, Ashley M Laughney, Julie-Ann Cavallo, Charles J Murphy, Peter Ly, Pragya Shah, Roshan K Sriram, Thomas B K Watkins, Neil K Taunk, Mercedes Duran, Chantal Pauli, Christine Shaw, Kalyani Chadalavada, Vinagolu K Rajasekhar, Giulio Genovese, Subramanian Venkatesan, Nicolai J Birkbak, Nicholas McGranahan, Mark Lundquist, Quincey LaPlant, John H Healey, Olivier Elemento, Christine H Chung, Nancy Y Lee, Marcin Imielenski, Gouri Nanjangud, Dana Pe'er, Don W Cleveland, Simon N Powell, Jan Lammerding, Charles Swanton, and Lewis C Cantley. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature*, 553(7689):467–472, January 2018.

[107] Teresa Davoli, Hajime Uno, Eric C Wooten, and Stephen J Elledge. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, 355(6322), January 2017.

[108] Neil T Umbreit, Cheng-Zhong Zhang, Luke D Lynch, Logan J Blaine, Anna M Cheng, Richard Tourdot, Lili Sun, Hannah F Almubarak, Kim Judge, Thomas J Mitchell, Alexander Spektor,

and David Pellman. Mechanisms generating cancer genome complexity from a single cell division error. *Science*, 368(6488), April 2020.

[109] Ruli Gao, Alexander Davis, Thomas O McDonald, Emi Sei, Xiuqing Shi, Yong Wang, Pei-Ching Tsai, Anna Casasent, Jill Waters, Hong Zhang, Funda Meric-Bernstam, Franziska Michor, and Nicholas E Navin. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.*, 48(10):1119–1130, October 2016.

[110] Ahmet Acar, Daniel Nichol, Javier Fernandez-Mateos, George D Cresswell, Iros Barozzi, Sung Pil Hong, Nicholas Trahearn, Inmaculada Spiteri, Mark Stubbs, Rosemary Burke, Adam Stewart, Giulio Caravagna, Benjamin Werner, Georgios Vlachogiannis, Carlo C Maley, Luca Magnani, Nicola Valeri, Udai Banerji, and Andrea Sottoriva. Exploiting evolutionary steering to induce collateral drug sensitivity in cancer, 2020.

[111] Kathleen Sprouffske, Lauren MF Merlo, Philip J Gerrish, Carlo C Maley, and Paul D Sniegowski. Cancer in light of experimental evolution. *Current Biology*, 22(17):R762–R771, 2012.

[112] Gregory I Lang, Daniel P Rice, Mark J Hickman, Erica Sodergren, George M Weinstock, David Botstein, and Michael M Desai. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574, 2013.

[113] Ketevan Chkhaidze, Timon Heide, Benjamin Werner, Marc J Williams, Weini Huang, Giulio Caravagna, Trevor A Graham, and Andrea Sottoriva. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLoS computational biology*, 15(7):e1007243, 2019.

[114] Allen W Zhang, Andrew McPherson, Katy Milne, David R Kroeger, Phineas T Hamilton, Alex Miranda, Tyler Funnell, Nicole Little, Camila PE de Souza, Sonya Laan, et al. Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. *Cell*, 173(7):1755–1769, 2018.

[115] Shang-Hung Chen and Jang-Yang Chang. New insights into mechanisms of cisplatin resis-

tance: from tumor cell to microenvironment. *International journal of molecular sciences*, 20(17):4136, 2019.

[116] Hadi T. Nia, Lance L. Munn, and Rakesh K. Jain. Physical traits of cancer. *Science*, 370(6516), 2020.

[117] Luca Cavallone, Adriana Aguilar-Mahecha, Josiane Lafleur, Susie Brousse, Mohammed Aldamry, Talia Roseshter, Cathy Lan, Najmeh Alirezaie, Eric Bareke, Jacek Majewski, Cristiano Ferrario, Saima Hassan, Federico Discepola, Carole Seguin, Catalin Mihalcioiu, Elizabeth A. Marcus, André Robidoux, Josée-Anne Roy, Manuela Pelmus, and Mark Basik. Prognostic and predictive value of circulating tumor DNA during neoadjuvant chemotherapy for triple negative breast cancer. *Scientific Reports*, 10(1):14704, 2020.

[118] Rabih Said, Nicolas Guibert, Geoffrey R Oxnard, and Apostolia M Tsimberidou. Circulating tumor DNA analysis in the era of precision oncology. *Oncotarget*, 11(2):188–211, 01 2020.

[119] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16(11):703–715, 2019.

[120] Francisco Azuaje, Sang-Yoon Kim, Daniel Perez Hernandez, and Gunnar Dittmar. Connecting histopathology imaging and proteomics in kidney cancer through machine learning. *bioRxiv*, 2019.

[121] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *bioRxiv*, 2019.

[122] Mario Zanfardino, Monica Franzese, Katia Pane, Carlo Cavaliere, Serena Monti, Giuseppina Esposito, Marco Salvatore, and Marco Aiello. Bringing radiomics into a multi-omics framework for a comprehensive genotype–phenotype characterization of oncological diseases. *Journal of Translational Medicine*, 17(1):337, 2019.

[123] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine

learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[124] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, (just-accepted):1–71, 2019.

[125] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[126] Lucy L Gao, Jacob Bien, and Daniela Witten. Are clusterings of multiple data views independent? *Biostatistics*, 21(4):692–708, 2020.

[127] Qian Du, Saul A Bert, Nicola J Armstrong, C Elizabeth Caldon, Jenny Z Song, Shalima S Nair, Cathryn M Gould, Phuc-Loi Luu, Timothy Peters, Amanda Khoury, et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nature communications*, 10(1):1–15, 2019.

Appendix

A.1 TNBC PDX experimental design diagrams.

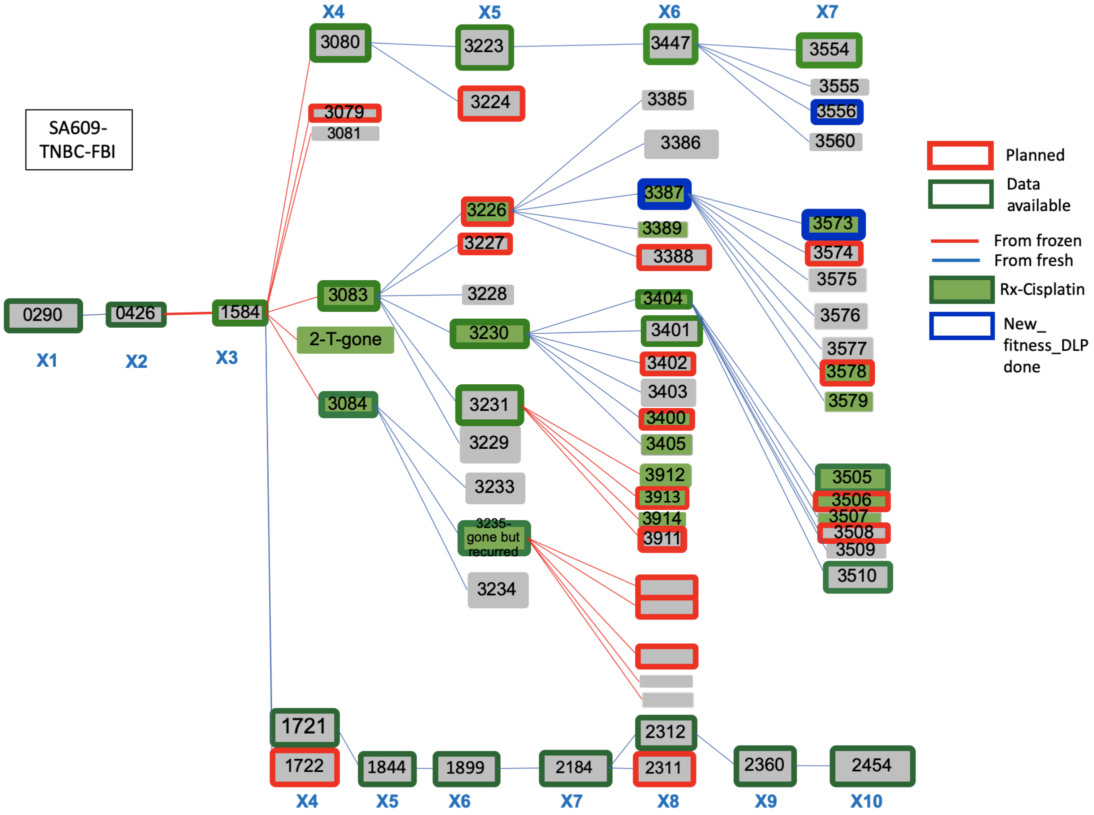


Figure A.1: The experimental design diagram for TNBC-SA609 PDX. Each node represents a mouse/timepoint and edges from left to right denote parent-child relationships where tumour material from a parent node is extracted and transplanted onto the child node. The 4-digit number on a node describes its sample ID or *barcode*. Nodes with a green background are treated with Cisplatin while all others are left untreated. The nodes border colour denotes the type of sequencing assay available. All mice designated X4 or later share ancestry with node 1584. This graph represents a tree. The nodes on the path from 0290 to 2454 comprise the 10 timepoints in the original TNBC-SA609 branch. Note that 1584 to 1721 is established from fresh material while all other branches at timepoint X4 are derived from frozen material at least 1 calendar year later. The original timeseries was

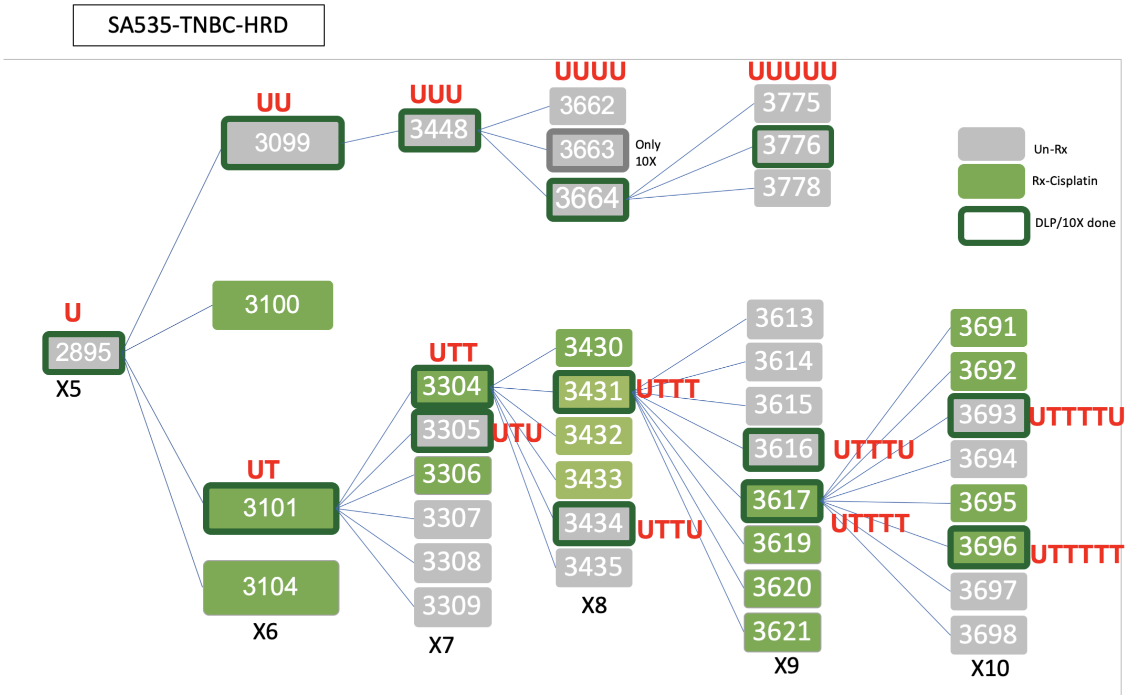


Figure A.2: The experimental design diagram for TNBC-SA535 PDX. The design is similar to Figure A.1. Note that one major untreated branch (2895 to 3776) and one major treated branch are established 2895 to 3696.

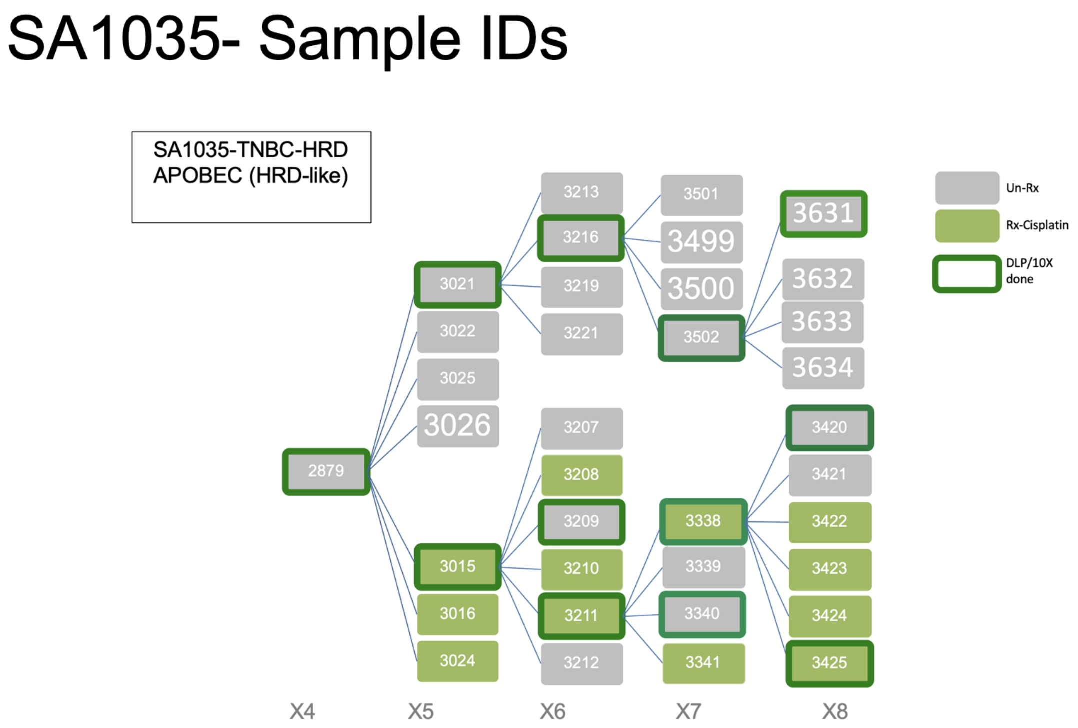


Figure A.3: The experimental design diagram for TNBC-SA1035 PDX. The design is similar to Figure A.1. In Chapter 4, the untreated branch contains nodes from 2879 to 3631 while the treated branch comprises nodes 3015 to 3425.

A.2 Figures

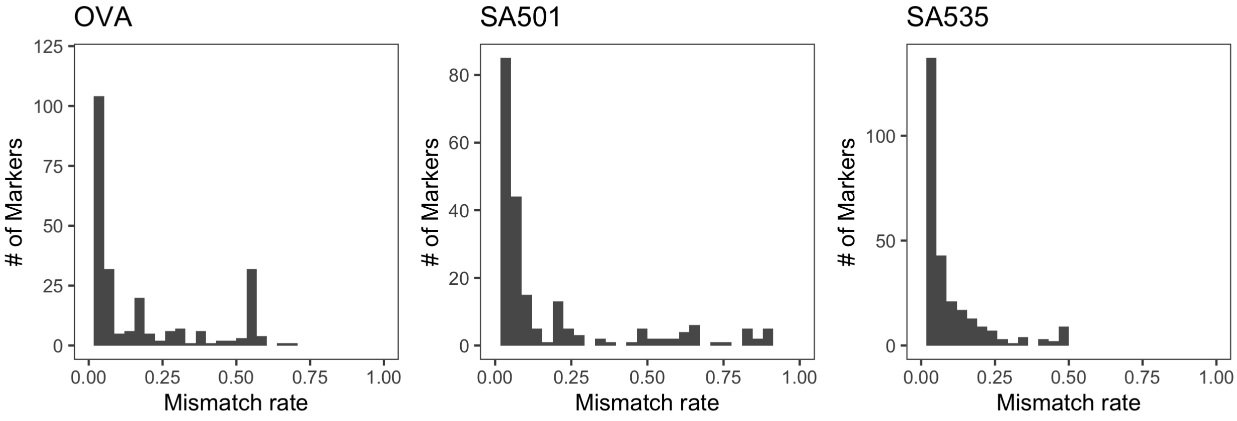


Figure A.4: the distribution of mismatch rate defined as the fraction of cells that have a mismatch between the inferred and jitter-fixed value of a marker.

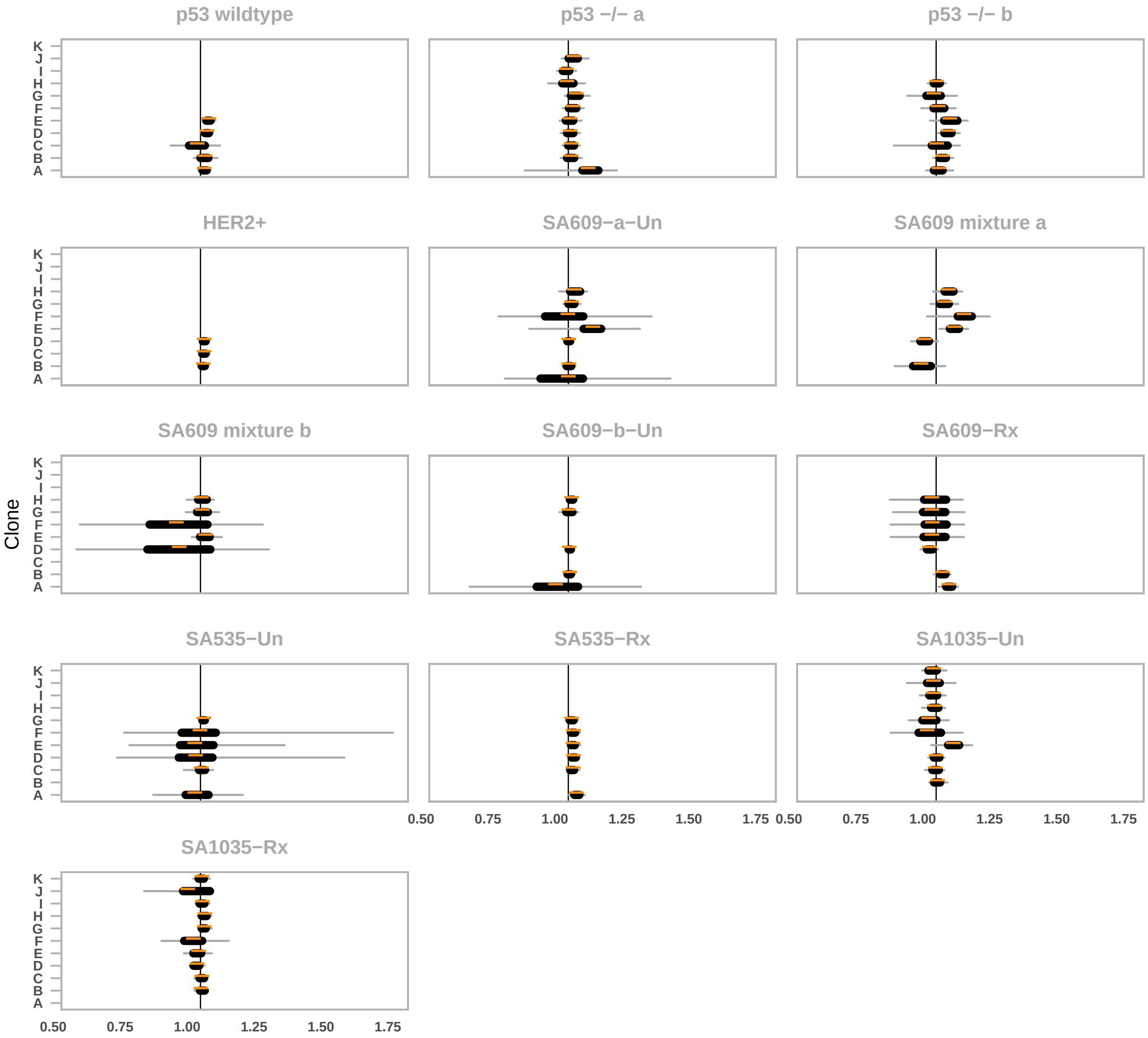


Figure A.5: Credible interval for the selection coefficients of all clones across all datasets. The dark orange colour shows the posterior median value. The narrow grey line shows the 95% highest posterior density interval (HDI) while the thicker dark grey line shows the 50% HDI.

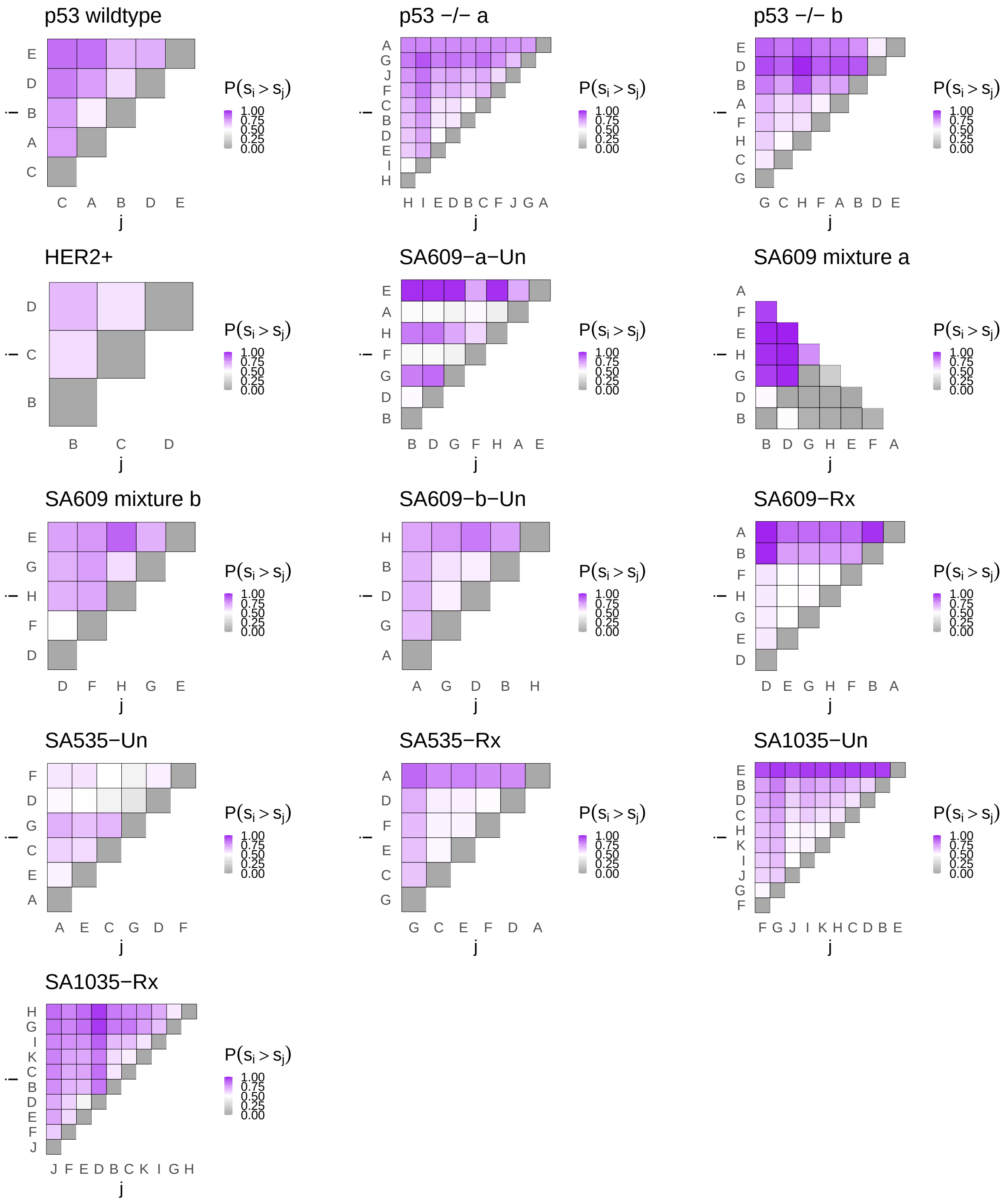


Figure A.6: Posterior order matrix for selective coefficients of all clones in all datasets. $P_{i,j} = P(s_i \leq s_j)$ shows the posterior probability that clone i has higher selective coefficient than clone j, with the stronger purple hues (close to 1.0) representing a higher confidence that clone i dominates clone j, and conversely the stronger grey hues (close to 0.0) denote that clone j dominates clone i. Colours closer to white (0.5) represent no dominance. Note that for the lower diagonal elements $P(s_j \leq s_i) = 1 - P(s_i \leq s_j)$ and are omitted for clarity. The diagonal entries are to guide the eyes only.

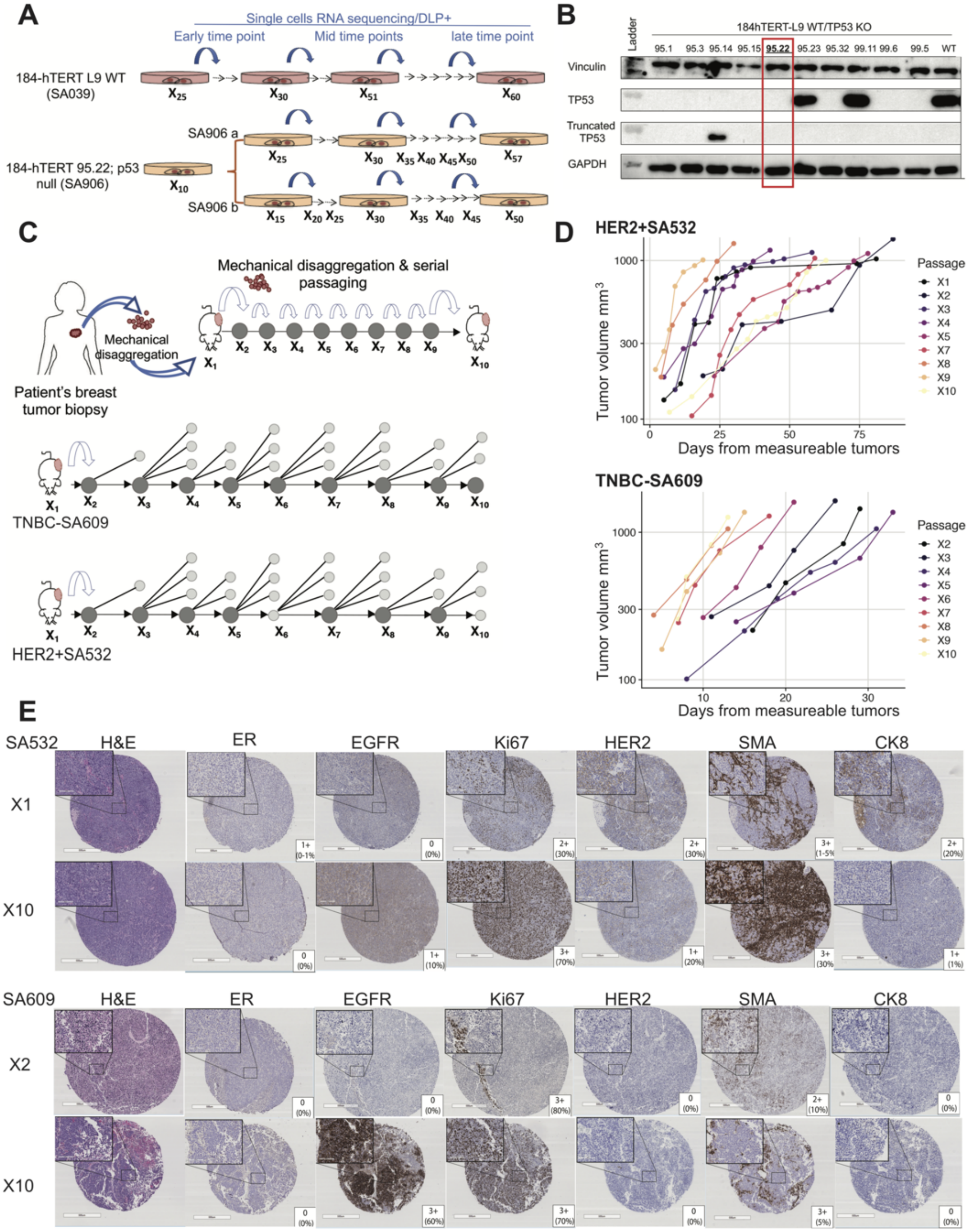


Figure A.7: Overview of experimental design and PDX growth curves. **A)** Top: Serially passaged 184-hTERT L9 WT cell line; Bottom: 184-hTERT L9 95.22; p53 null cell line by CRISPR technology. Parallel branches $P53^{-/-}$ a and $P53^{-/-}$ b were derived from the same tenth passage. **B)** Western blot confirming knock out of TP53 from 184-hTERT WT cell line. Clones shown along top. **C)** Top: Schematic for PDX timeseries; Bottom: Serial sampling of HER2+ and TNBC PDX tumours; Dark grey circles represent each sampled mouse for scWGS. The light grey circles represent the replicates of tumour-bearing mice at the same timepoint. **D)** Individual tumour growth from each passage of TNBC and HER2+ PDXs. **E)** IHC of HER2+ and TNBC tumours at early and late passages, 4x and 20x (insets). Scale bars 500 µm and 100 µm (insets). Antibodies and TMA scores.

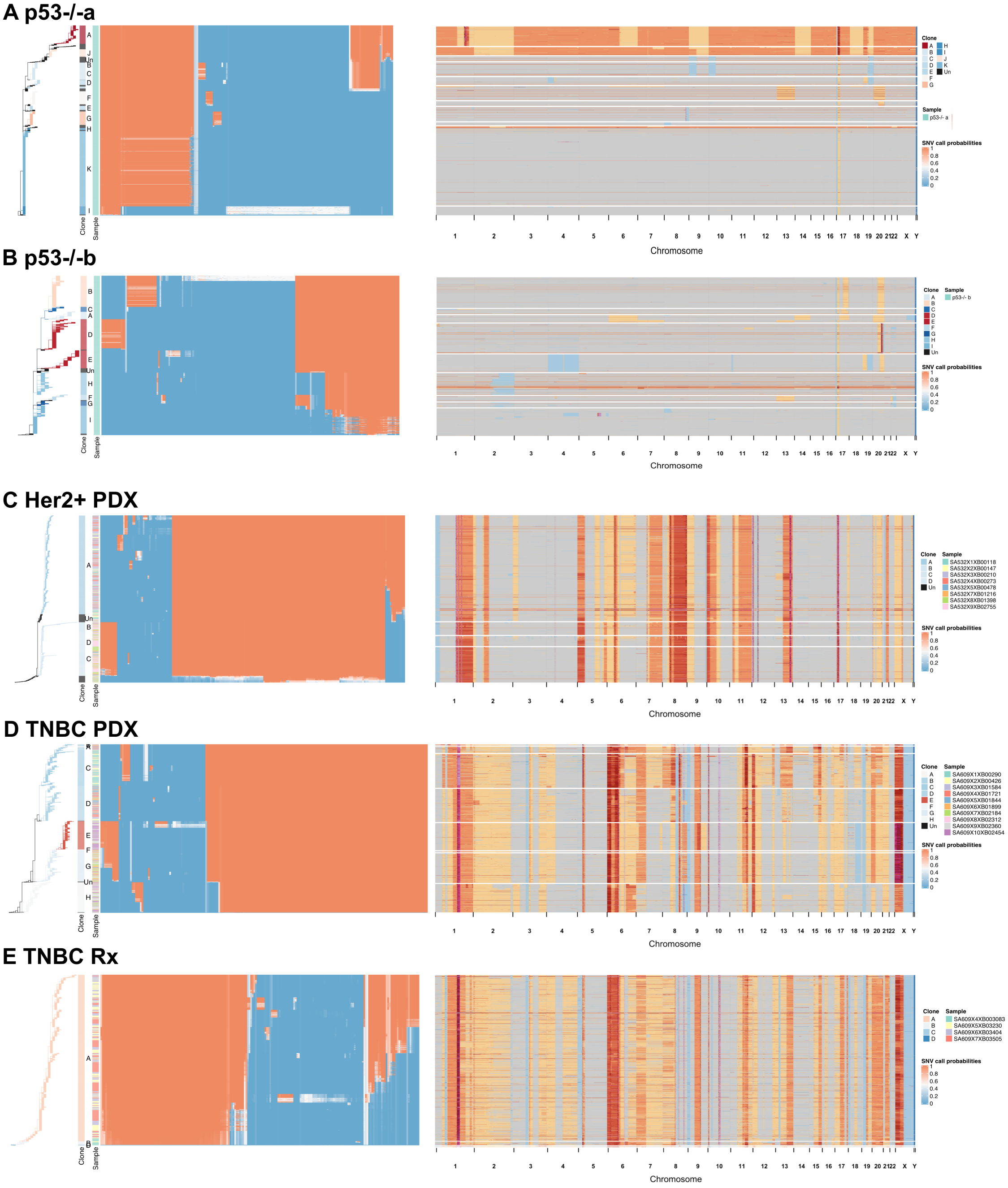


Figure A.8: Single nucleotide variant (SNV) and Copy number profiles of the cell lines and PDX libraries. **A)** Phylogenetic tree with clone labels next to SNV and copy number profile heatmaps for *p53-/-a*. Rows are individual cells and columns are SNVs (left) and copy number alterations (right). For SNVs, orange indicates high probability while blue indicates low probability. For copy number, legend encoding copy number colours is shown (right). **B-E)** Similar to **A**) but for *p53-/-b*, HER2+, TNBC, and TNBC Rx respectively

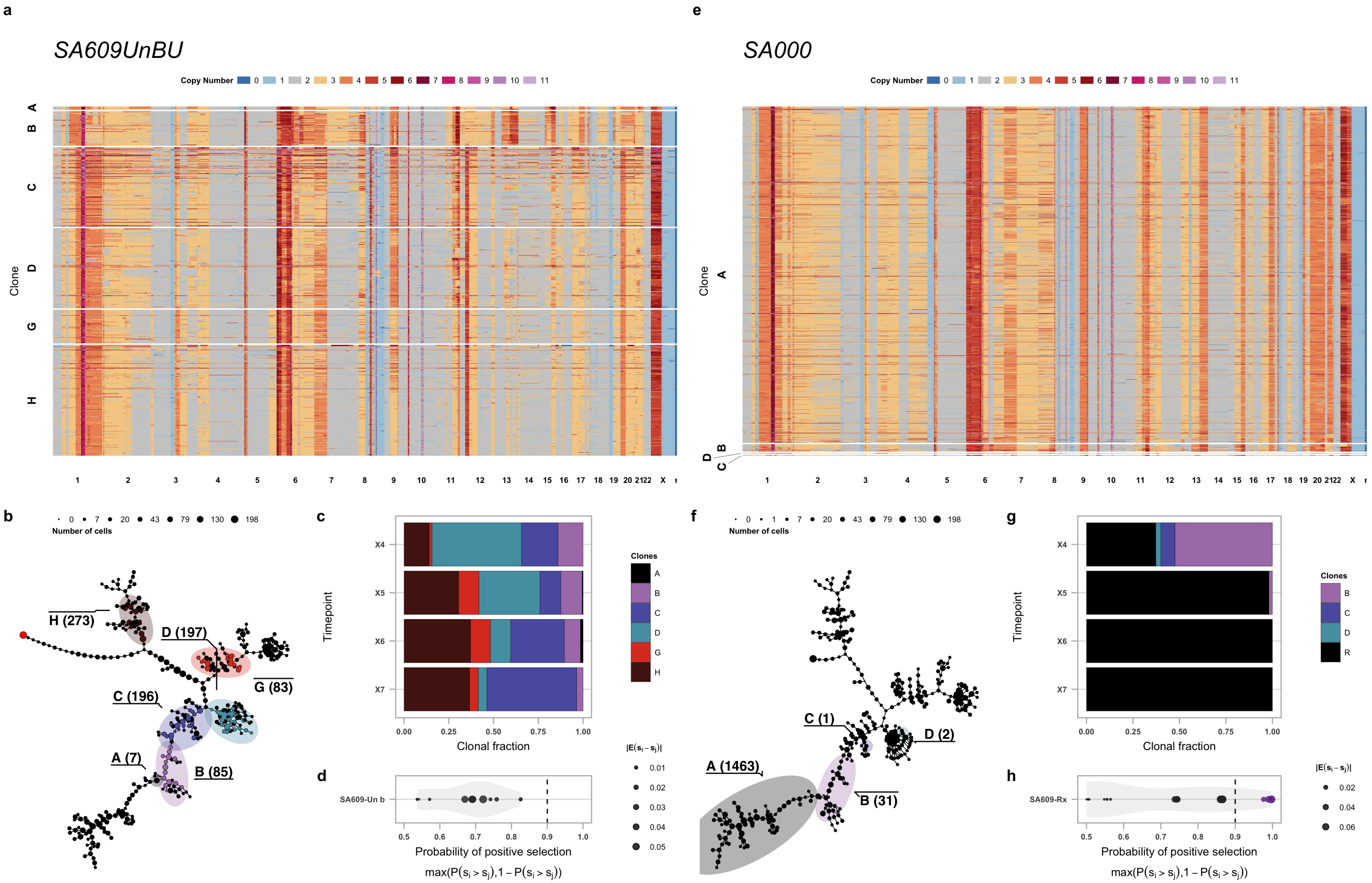


Figure A.9: TNBC-SA609 PDX clonal dynamics with and without treatment. **a)** Heatmap representation of copy number profiles of 841 cells, grouped in 6 phylogenetic clades. **b)** Phylogeny as per Figure 4.2 for TNBC-SA609 PDX untreated branch. **c)** Observed clonal abundances, and **d)** distribution over magnitude of difference between selective coefficients of pairs of clones. **e-h)** Analogous plots for the treated branch (n=1,593 cells).

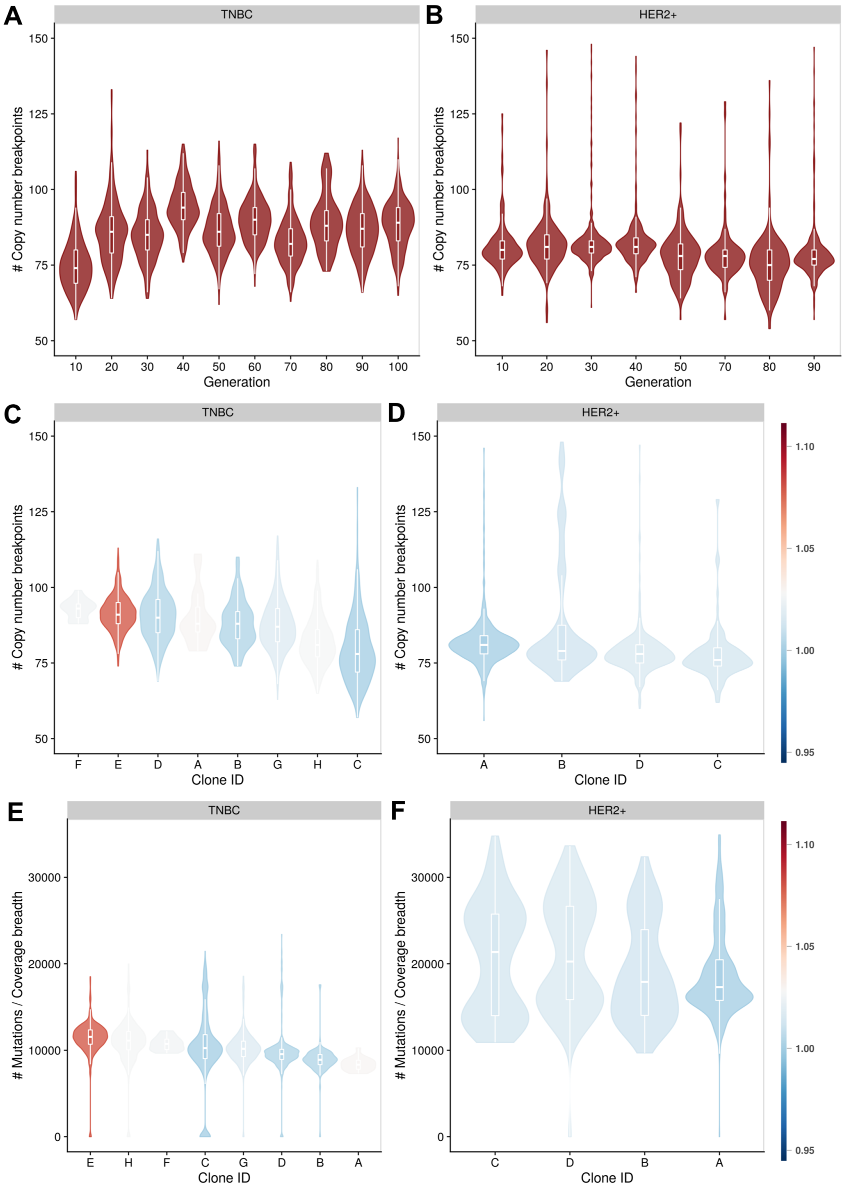


Figure A.10: Structural variant and mutation rates of PDX lines. Distribution over copy number breakpoints/cell as a function of generation for **A)** TNBC **B)** HER2+. Clone specific distributions over copy number breakpoints/cell, coloured by fitness coefficients for C) TNBC **D)** HER2+. Clone specific distributions over point mutations/cell, coloured by fitness coefficients for **E)** HER2+ and F) TNBC

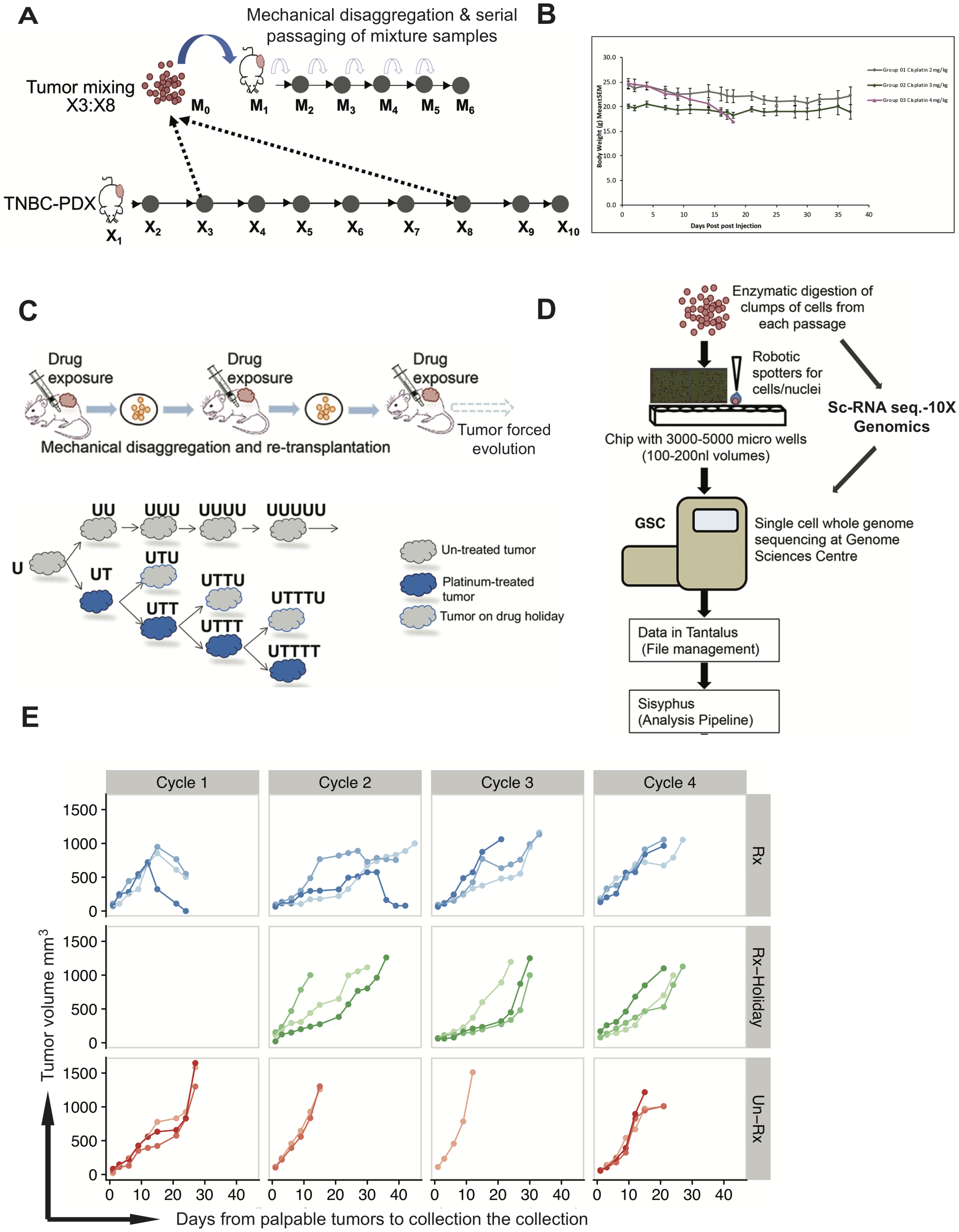

A
Mechanical disaggregation & serial passaging of mixture samples
Tumor mixing X3:X8
M_0
M_1
M_2
M_3
M_4
M_5
M_6
TNBC-PDX
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_{10}
B
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Body Weight (g) Mean±SEM
0
5
10
15
20
25
30
35
40
Days Post post Injection
Group 01 Cisplatin 2 mg/kg
Group 02 Cisplatin 3 mg/kg
Group 03 Cisplatin 4 mg/kg
C
Drug exposure
Drug exposure
Drug exposure
Mechanical disaggregation and re-transplantation
Tumor forced evolution
U
UU
UUU
UUUU
UUUUU
UT
UTU
UTT
UTTU
UTTT
UTTTU
UTTTT
Un-treated tumor
Platinum-treated tumor
Tumor on drug holiday
D
Enzymatic digestion of clumps of cells from each passage
Robotic spotters for cells/nuclei
Chip with 3000-5000 micro wells (100-200nl volumes)
Sc-RNA seq.-10X Genomics
GSC
Single cell whole genome sequencing at Genome Sciences Centre
Data in Tantalus (File management)
Sisyphus (Analysis Pipeline)
E
Cycle 1
Cycle 2
Cycle 3
Cycle 4
Rx
Rx−Holiday
Un−Rx
1500
1000
500
0
Tumor volume mm^3
0
10
20
30
40
Days from palpable tumors to collection the collection

Figure A.11: Fitness validation with tumour mixing and drug perturbation. **A)** Schematic overview of clonal mixture experiment showing source samples from the original timeseries and serial propagation into a new line. **B)** Mouse body weight graph recorded during maximum tolerated dose (MTD) evaluation of cisplatin in NRG mice (n=3 in each study cohort). **C)** Experimental design of cisplatin treatment in PDX.The residual tumour from one treated mouse was re-transplanted in the next (n=4). The solid blue colour represent cisplatin treated tumours *(UT,UTT,UTTT,UTTTT)*; blue outlined in grey represents drug holiday *(UTU,UTTU,UTTTU)*. Grey represent the untreated series *(U,UU,UUU,UUUU,UUUUU)*. **D)** Tumour response curves in each cycle of cisplatin treatment.

A.3 Tables

	Dataset	Timepoint1	Timepoint2	Normalised counts	Copy number rate
1	SA501	X2	X5	0.02	0.01
2	SA501	X5	X6	0.02	0.02
3	SA501	X6	X8	0.19	0.09
4	SA501	X8	X11	0.05	0.02
5	SA501	X11	X15	0.04	0.01
6	SA535	X1	X5	0.17	0.04
7	SA535	X5	X8	0.34	0.11

Table A.2: Estimated CNA rate over time in timeseries datasets used in Chapter 2. Note that the Copy number rate is not normalised by the length of the passage (in number of generations) or mean ploidy.

	Dataset	N_e	Number of SNVs
1	*p53 WT*	985.79	NA
2	*p53-/-a*	614.93	6,645
3	*p53-/-b*	422.14	7,905
4	HER2+	1,461.70	40,309
5	TNBC	468.30	33,178
6	TNBC-Mixture	177.01	NA
7	TNBC-Rx	333.34	25,685

Table A.3: Parameters per dataset. Effective population size estimates and number of loci used in the SNV analysis.

	Dataset	ϵ	runtime	niter	$\Delta\tau$	avg_ESS	min_ESS	rejection_rate
1	*p53 WT*	0.01	2:57:22.	10,000	0.05	911.90	448.19	0.77
2	*p53-/-a*	0.01	35:18:46.	100,000	0.05	3209.54	506.66	0.61
3	*p53-/-b*	0.03	31:29:55.	100,000	0.04	2077.27	906.95	0.89
4	HER2+	0.01	48:57:27.	100,000	0.05	26303.53	25299.69	0.38
5	TNBC	0.03	40:05:32.	100,000	0.05	1379.81	223.08	0.91
6	TNBC-mixture	0.01	29:13:01.	100,000	0.04	3745.76	1268.09	0.73
7	TNBC-Rx	0.03	30:05:33.	100,000	0.10	4362.86	2768.55	0.73
8	WGS-bulk-TNBC	0.012	35:12:29.	100,000	0.10	14560.09	11754.76	0.71

Table A.4: Real-world data parameters.

	dataset	sample.id	lib.id	tp	total	!mouse	qual.	!sphase	!lmr	final	finalr
1	SA501	SA501X2XB00096	A95621B	X2	623	570	524	580	568	498	498
2	SA501	SA501X2XB00097	A96171A	X2	492	36	21	474	37	20	20
3	SA501	SA501X2XB00097	A96109A	X2	656	59	37	399	96	34	34
4	SA501	SA501X5XB00877	A95670A	X5	615	326	253	575	252	238	228
5	SA501	SA501X6XB00294	A96213A	X6	1000	326	152	886	168	62	62
6	SA501	SA501X6XB00969	A95670B	X6	636	383	325	602	300	319	272
7	SA501	SA501X8XB01694	A90696ABC	X8	3672	3517	273	3482	1985	261	241
8	SA501	SA501X11XB00529	A96187A	X11	1063	1024	878	888	958	779	763
9	SA501	SA501X11XB00529	A96174A	X11	1234	1192	291	1007	258	116	58
10	SA501	SA501X15XB00929	A96173A	X15	1388	997	558	1131	667	399	399
11	TNBC-SA535	SA535X1XB00174	A96165B	X1	379	208	73	218	277	62	62
12	TNBC-SA535	SA535X5XB00517	A95732A	X5	928	613	372	776	464	332	324
13	TNBC-SA535	SA535X8XB00143	A95736A	X8	1072	410	194	956	354	154	152
14	OVA-2295	SA921	A90554A	TOV2295(R)	628	506	415	524	525	392	392
15	OVA-2295	SA922	A90554B	OV2295(R2)	596	466	303	518	467	296	296
16	OVA-2295	SA1090	A96213A	OV2295	1484	1438	1144	1092	1402	850	838

Table A.5: Summary of real-world datasets used. `final` is the final number of cells after all filters except for `!lmr` are applied. `final` additionally filters out `lmr` cells, those that have total mapped reads fewer than 500,000. Abbreviations used are tp: time point; qual. : quality; !sphase: not S-phase; lmr: low mapped reads; !lmr: not low mapped reads.

A.4 Preprocessing of the CNA data for phylogenetic inference

Get copy number values

CNA states are stored in `cn.csv` and this file is the input to our preprocessing pipeline. For this work the data was stored in the cloud (Microsoft Azure) and the `scgenome` application programming interface (API) suite was used to access and download the data. Please see `https://github.com/shahcompbio/scgenome` for documentation. This API was developed by authors in reference [20]. The `scgenome` API ensures that only cells with the correct `sample_ids` are selected, and removes control cells and cells that have fewer than 10,000 mapped reads. The API also provides for each cell a quality score that we call *SCG-score*. This score ranges from zero to one where values closer to one indicate higher quality cells. Lower quality cells are expected to be noisy, that is to have many non-integer CNA values and very low read counts. See the [20] for more details.

Drop low-mappability bins

Some copy number bins are located at parts of the genome where sequencing is difficult, for example due to inaccessibility of the genome at that position. These bins are sometimes assigned imputed CN states that represent a sequencing artefact. This is reflected in their mappability score. We filter the CNA matrix to keep high-mapability bins `cn_bin_filtered.csv`. In this work we use a cutoff threshold of `map >=.99` that yields 4375/6206 or 70.5% of the bins. The list of kept bins is identical across all datasets. Low-mapbaility bins may constitute phylogenetic markers that are not reflective of the underlying biology and result in grouping of cells that share these sequencing artefacts but are not evolutionarily related. Moreover, using the 0.99 cutoff, we still recover hundreds of phylogenetic markers that are present in at least 5% of cells. A potential pitfall can occur in datasets that have very few copy number changepoints to accurately distinguish subpopulations. In such cases filtering too many genomic bins may result in missing some subpopulations. A lack of ground-truth datasets makes it difficult to set this threshold in a principled way.

Drop low-quality cells

In this step, a second round of quality control is done. Cells with SCG-scores of over 0.75 are kept [20], those that are suspected to be contaminated (e.g., mouse cells) or cycling cells are removed. This results in `cn_bin_cell_filtered.csv`. The 0.75 cutoff is chosen as it balances sensitivity and specificity of the single cell scoring algorithm.

Drop cells with excess CNA changes

Some cells show a *jumpy* CNA profile in which there are excessive copy number changes. It appears that these cells are either in early or late stages of division and were missed by the `scgenome` API. The CNA profile of early replicating cells is patterned by seemingly scattered focal amplifications while the late replicating cells show scattered focal deletions. Note that not all parts of the genome duplicate at the same time during mitosis [127]. Regions that start duplicating later will show as having focal deletions in cells captured at their later replicating stage; these regions would not have started to duplicate by the time the sample was prepared for sequencing. One possible explanation is how the read counts in cells are normalised to integer CN states in the DLP+ platform. When estimating the CN states of cells, the normalizing procedure assumes that the most likely ploidy for all cells is diploid [20]. For each cell, the observed read counts are rescaled as closely to the diploid state as possible, while retaining integer CN states. These scattered patterns (Figure A.12) are hypothesised to not directly reflect the evolutionary history of the cells and are detrimental to phylogenetic tree inference.

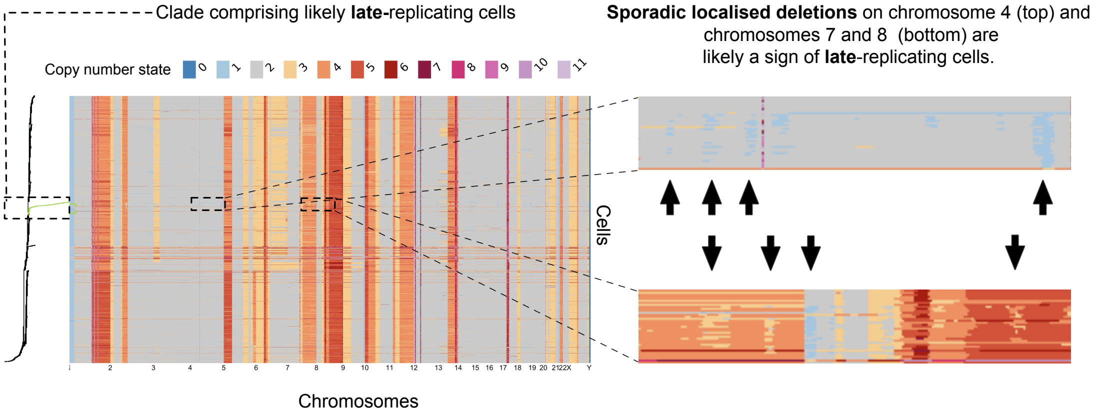


Figure A.12: An example of replicating cells. The CNA matrix and the consensus tree show cells from a HER2+ PDX dataset. Cells with excess CNA changes are *not* filtered out (see Section A.4). Note that a subset of cells, corresponding to the green clade on the tree, exhibit scattered localised deletions (the arrows on the right) across the genome. These late replicating cells form a *finger*-like clade in the tree. The similarity in the CNA profiles is not due to common evolutionary history and therefore they should be removed (see Section A.4). The top inset shows chromosome 4 while the bottom inset spans chromosomes 7 and 8.

Here we rank cells by the number of changes in their copy number states (a change is measured between consecutive bins) and pick the bottom 90-th percentile. This threshold was set via visual inspection. The file `cn_bin_cell_filtered_no_jump.csv` contains the integer copy number state with the final list of cells and genomic bins. An example input matrix is shown in Figure 2.4-**a** where the integer copy numbers are colour-coded in a heatmap. Attrition rate due to filtering of cells is shown in Table A.5.

A.5 Computing CNA rate over time

Computing the CNA rate is non-trivial. The reference timepoint is heterogenous and therefore to avoid conflating selection with mutation rate, we pick a set of markers that are not present in the reference timepoint (cellular prevalence <0.001) and call it L_0. In each timepoint i (other than the reference timepoint), find the subset of markers in L_0 that are present in that timepoint (e.g., have a cellular prevalence >0.01) and call this subset L_i. Now for each pair of consecutive timepoints, count the number of markers that are not identical (i.e., $L_i \setminus L_{i+1}$), normalised by the number of markers in each dataset. Table A.2 shows the normalised CNA rates for the timeseries datasets used in Chapter 2.