Evaluating the Use of Single-cell and Bulk Sequences in Phylogenetic Analyses of Multi-tumour Evolution

Jonathan Fu

Department of Science

The University of Auckland

Supervised by Prof. Allen Rodrigo and Dr. Teng Li

A thesis submitted in partial fulfilment of the requirements for the degree of Masters of Science in Biological Sciences, The University of Auckland, 2024.

Abstract

Studying the phylogenetic reconstruction of somatic evolution can be challenging due to constraints in resources, the influence of various biological factors, and technical limitations. Many methods concentrate on the analysis of single-cell sequencing data. This approach, while generally more accurate than using bulk-sequencing data, can be limiting due to its computational complexity and potential technical artefacts.

We integrate two simulation software tools to facilitate the modelling of evolutionary histories between multiple metastatic sites within a patient. The first tool is used to generate phylogenetic trees, providing a framework that represents the true evolutionary histories both at a multi-tumour and individual cell level. Subsequently, the second tool generates single-cell tumour sequences based on the true cell-cell trees. By combining these two software tools, cells are assigned to specific tumours, and therefore simulated under a structured population. Our study evaluates the viability of pooling single-cell data into consensus sequences by comparing their accuracy in reconstructing the multi-tumour tree against pseudo-bulk data. This approach addresses the challenge of obtaining multi-tumour level evolutionary histories from single-cell data. We aim to provide guidance for researchers when choosing their preferred sequencing analysis method or for those looking to trace multi-tumour evolution.

Under various biological conditions, we simulate single-cell tumour sequences with a predefined multi-tumour tree and its corresponding cell-cell tree replicates. We construct consensus sequences, pooling cell sequences based on shared tumour lineage. For comparison, we construct pseudobulk data. Calculations of tree distance between initial trees against reconstructed trees show that consensus sequences do not perform as well as a pseudobulk dataset.

We explore the process of reconstructing the true multi-tumour tree by integrating existing data, compiled as sets of replicates. First, we perform tree reconstruction using a species estimation method on single-cell data. Additionally, we explore supertree construction as well as the use of concatenated sequences, leveraging pooled data. Through this analysis, we find that using single-cell data directly or utilising pseudobulk data for reconstructing multi-tumour evolution yields the best results.

Acknowledgments

I would like to thank my supervisor Prof. Allen Rodrigo. His support, guidance, patience, and understanding have been constant throughout the entire duration of this project.

I extend my thanks to my co-supervisor Dr. Teng Li, for his endless advice and invaluable knowledge.

I am thankful for the friends I have made on this team. Good luck for this year and beyond.

I thank my family for their endless support.

Finally, I thank my partner, Sirithi.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

Significant strides have been made in enhancing cancer management. Research efforts have underscored the role of lifestyle factors, while expanded access to healthcare and enhanced screening capabilities have led to earlier diagnoses resulting in more favourable treatment outcomes (Boddy, 2023). Moreover, ongoing advancements in treatment have demonstrated substantial improvement in patient outcome across several cancer types (Siegel et al., 2018). Despite this, globally, cancer is still a leading cause of death (Bray et al., 2021) and persists as a significant public health issue, with mortality reaching 10 million deaths in 2020 (Sung et al., 2021). Trend-based projections indicate a constant rise in incidence, with an estimated 34 million new cases expected to emerge in 2070 (Soerjomataram & Bray, 2021). Such statistics emphasise the ongoing imperative for research directed towards developing newer understanding and solutions for the disease.

In recent decades, cancer research has been viewed through the lens of evolutionary theory. The application of phylogenetics has been instrumental in tracing the evolutionary histories of cancer, particularly since the advent of genomic technologies (Somarelli et al., 2017). This has allowed for next-generation sequencing analysis of bulk sets of cells, such as those obtained through tissue resection, or individual single cells. While both bulk and single-cell approaches have significantly contributed to understanding cancer evolution, both have their limitations (Lähnemann et al., 2020; Vendramin et al., 2021). These limitations can largely be attributed to the heterogeneity within tumours, in which diverse subpopulations co-exist as a result of dynamic selective pressures acting on accumulated mutations over time (Zhu et al., 2021). When bulk samples are extracted from tumours, they will inherently capture cells from various subpopulations. Consequently, samples may not correctly represent the complexity of the tumour's evolutionary history (Alves et al., 2017). Consider, for instance, the exploration of evolutionary histories across multiple metastatic sites within a single patient. Conversely, when examining the relationships among individual cells, it becomes apparent that these cells may have emerged from distinct subpopulations characterised by diverse selective traits and therefore possess the capacity to diverge and propagate to different metastatic sites (Rogiers et al., 2022). This phenomenon can lead to inconsistencies between the observed evolution of individual cell lineages and the actual progression of the metastases from which the cells were sampled. This results in discordance, whereby the phylogenies obtained from the individual cells differ from the phylogenies of the metastatic sites. Through the use of simulations, this particular scenario is one we intend to explore further.

1.1.1 Genetic Basis of Cancer

Cancer, observed ubiquitously across the animal kingdom (Vincze et al., 2022), originates from alterations in the DNA sequence of the cell genome. Germline mutations are inherited DNA alterations present in every cell of an individual's body, potentially predisposing them to cancer development. Therefore, they are not directly involved in the ongoing evolution of cancer within affected tissues (Wang, 2016). The DNA sequence of a cancer cell genome, has undergone alterations during the lifetime of the individual. These are known as somatic mutations, resulting from errors in DNA replication during cell division. They can occur and remain unresolved, passed down to daughter cells through subsequent divisions. Such mutations arise spontaneously in somatic cells and are not inherited by an individual's offspring (Stratton et al., 2009). Somatic mutations serve as molecular signatures, providing researchers with a means to trace the progression of the disease (Alexandrov et al., 2020).

Normal, healthy somatic cells will undergo a regulated process of growth, division, and death (Werner et al., 2020). In cancer, this process breaks down. Tumourigenesis occurs, whereby alterations in the regulatory mechanisms of somatic cell growth, such as mutations and the dysregulation of signalling pathways, contribute to the uncontrolled formation of a mass lesion of tissue called a tumour (Aktipis et al., 2015; Luzzatto, 2011). Tumours can take divergent paths. Some may maintain a benign nature, characterised by non-cancerous growth that does not spread beyond its original location. Conversely, others may progress to invade neighbouring tissues and disseminate to distant sites in the body, a phenomenon known as metastasis (Sarkar et al., 2013). The invasive growth of malignant tumours carries profound implications, involving the infiltration of nearby tissues and organs, posing fatal risks such as organ failure, or compromised bodily functions (Martin et al., 2013).

1.1.2 Sequencing in Cancer Research

Given that cancer is fundamentally a disease of genetic alterations, research has significantly focused on studying somatic mutations through the analysis of cancer genomes, involving the examination of DNA sequences (Meyerson et al., 2010). Through bioinformatic approaches, researchers can extract differing types of mutations observed in cancer sequences. These mutations are subjected to thorough analysis to identify patterns, genetic pathways implicated in cancer development, as well as to correlate them with clinical data for insights into areas of interest, such as prognosis and treatment response (Dimitrakopoulos & Beerenwinkel, 2017). These mutations are classified into distinct categories, including insertions or deletions of DNA segments, rearrangements characterised by the breaking and rejoining of DNA segments, copy number alterations, which encompass both increases and reductions of the DNA sequence, as well as single nucleotide variations (SNVs), the basic substitution of one DNA base with another (Stratton et al., 2009).

Next-generation sequencing (NGS) technology, a term that broadly encompasses technologies capable of simultaneously sequencing millions of DNA fragments in parallel, has revolutionised genomic research. In recent years, NGS technologies has advanced significantly, leading to a continuous reduction in sequencing costs (McCombie et al., 2019). Consequently, the scope of viable applications has expanded, particularly in cancer genome research. The integration of these techniques offers functionalities such as tumour classification, predictions regarding disease progression, and the identification of clinically significant mutations for treatment selection (Raphael et al., 2014).

One such application is bulk sequencing (bulkSEQ), a traditional method of compiling DNA sequence data for analysing the genetic diversity within a tumour. One common approach to bulkSEQ, involves collectively sequencing a sample of resected tissue, which comprises an admixture of cells. This approach enables the identification of genetic variations across a sample of cells, revealing common mutations and genetic features present within the tumour (Kyrochristos et al., 2019). BulkSEQ is generally considered cost-effective, as sample preparation does not require the isolation and processing of individual cells. The computational requirements for processing and analysing the data are also less intensive. This is because bulkSEQ analysis pipelines do not need to account for the complexities associated with analysing data from individual cells, resulting in a more streamlined process.(Li & Wang, 2021). The efficacy of bulkSEQ has been firmly established, as evidenced by its successful application in various cancer studies. Examples of its usage include application in the study of glioblastoma multiforme, with bulkSEQ conducted as part of The Cancer Genome Atlas project, revealing recurrent mutations in genes such as TP53, PTEN, and EGFR (McLendon et al., 2008). In esophageal adenocarcinoma, bulkSEQ analysis has uncovered frequent mutations in genes such as APC, CDK2NA, and TP53, serving as biomarkers for prognosis and therapeutic targeting (Dulak et al., 2013). However, a stark limitation of bulkSEQ data is that it represents only an averaged genetic profile derived from a mixed population of cells from a tumour region, rather than capturing the full genetic heterogeneity present within each individual cell (Dong et al., 2020). Furthermore, tumour cells are typically interspersed with various non-tumour cells, such as endothelial cells, immune cells, stromal architecture cells, and remnants of normal host tissue (Lin et al., 2023). Even when the majority of cells are cancerous, the sampling process involved in bulkSEQ will result in a population of cells that exhibit varying genotypic and phenotypic properties (Schwede et al., 2020). The obtained genetic profile reflects a composite view of the tumour, rather than capturing the specific genetic characteristics of individual cell types within the tumour. This restriction hinders the resolution of tumour heterogeneity.

The introduction of single-cell sequencing (scSEQ) analyses has enabled researchers to study genetic diversity at a more granular level, focusing on individual cells. This capability allows for the identification of cellular diversity present within tumour samples. Single-cell genomics has provided insights into many areas of cancer research such as: heterogeneity in tumour composition and characteristics across different locations (spatial) and over time (temporal) (Loeffler-Wirth et al., 2018; Massalha et al., 2020; Nerurkar et al., 2020), metastasis (Han et al., 2022; T. Liu et al., 2022; Quinn et al., 2021), therapeutic resistance mechanisms (Aissa et al., 2021; Kashima et al., 2021). ScSEQ generally begins with the isolation of individual cells, whereby techniques such as fluorescence-activated cell sorting or microfluidic devices are used to extract single cells from a sample (Zhou et al., 2021). Once isolated, the genetic material within each cell undergoes amplification to generate a sufficient amount for downstream analysis. Amplification techniques, such as polymerase chain reaction or multiple displacement amplification are used to replicate the DNA present in the cell (Jovic et al., 2022). Amplification allows researchers to generate multiple copies of the genetic material from a single cell, thereby increasing the amount of material available for sequencing. Following amplification, the genetic material from each cell is then sequenced using NGS technologies to generate a profile of the genetic information within each individual cell (Zhou et al., 2021). However, while scSEQ offers enhanced resolution, it also presents certain challenges. These include allelic dropout (ADO), where one allele of a gene may not be detected during sequencing, leading to an incomplete representation of genetic information (Shestak et al., 2021). Additionally, errors can occur during amplification, where certain regions of the genome are preferentially amplified resulting in uneven coverage (Gawad et al., 2016). Other forms of sequencing errors or biases can occur, which may distort the interpretation of results (Mitchell et al., 2020). Despite advancements in scSEQ procedures, these causes of concern persist and continue to pose significant challenges in the field.

1.1.3 Cancer under an Evolutionary Model

A fundamental understanding of cancer lies in recognising it as an evolutionary process. Tumour cells will undergo selective pressure, a process reminiscent of Darwinian natural selection (Hanahan & Weinberg, 2000). This means that each tumour cell operates independently, undergoing mutations and adaptations over time, similar to individual organisms in a population evolving to better survive and reproduce in their environment. Here, the term selective pressure refers to the influence of factors within the body that favour the survival and proliferation of certain tumour cell variants over others, leading to the expansion of those variants that are best suited to the conditions within the tumour microenvironment (Greaves & Maley, 2012).

This conceptualisation of evolution in cancer has long been rooted in Nowell's clonal evolution theory, which proposed tumourigenesis as a genealogical model (Nowell, 1976). This model proposes that a tumour encounters selection pressures in an environment characterised by limited nutrition and metabolites, as well as competition between cells with higher reproductive rates and those that can escape immune surveillance. In this context, cells demonstrating heightened proliferation and enhanced survival are favoured (Merlo et al., 2006). Nowell's theory posits that cells, have the inherent capacity to acquire somatic mutations over time, particularly in genes pivotal to processes such as growth and division. This theory describes the initiation of cancer as when a single cell acquires a mutation that imparts a growth advantage. This serves as a catalyst for a process known as clonal expansion, where the mutated cell proliferates at an accelerated rate compared to its neighbouring cells. This acquisition can have a cascading effect, whereby the dynamic process of selection and adaptation leads to the development of subpopulations of cells, consisting of distinct genetic and phenotypic variations that can survive and co-exist (Bailey et al., 2020; Dagogo-Jack & Shaw, 2018).

Various models exist regarding the propagation of mutations within a tumour, which may occur in different tumours or at different stages of a tumour's progression. Traditionally, cancer development and mutation propagation were considered to involve a step-wise accumulation of genetic alterations, notably through the work of Fearon and Vogelstein (Fearon & Vogelstein, 1990) who proposed a theoretical framework for the progression of colorectal cancer. This is recognised as the linear evolution model. This model involves an initially monoclonal population of cells that individually accumulate genomic alterations gradually over time at a consistent evolutionary tempo. Cells undergo positive selection and proliferate, resulting in a cell population better equipped to survive against selection pressures compared to normal cells. This process leads to the development of highly homogeneous populations of cancer cells, with traces of unfavourable subclonal cells over time (de Bruin et al., 2013).

Advancements in NGS have led to the delineation of several other models of cancer evolution, indicating a spectrum of possible evolutionary trajectories (Vendramin et al., 2021). For instance, the divergence of subclones from a common ancestor and their independent evolution align with the branched evolution model (Greaves & Maley, 2012). In contrast to linear evolution, where genetic diversity is constrained, the branched evolution model, better accounts for the heterogeneous nature observed in tumour evolution. In this model, the tumour evolves through the parallel accumulation of diverse genetic alterations, resulting in the emergence of multiple subclones with distinct genomic profiles. This simultaneous expansion results in a diverse mixture of cell populations with distinct genetic profiles coexisting within the tumour mass.

In certain cancers, a model known as punctuated evolution has been proposed (Gould & Eldredge, 1977), whereby a substantial number of mutations occur through short bursts of time, particularly during the earliest stages of tumour progression. This model is defined by three key properties: stasis, marked by stable clonal expansions; evolution occurring in brief bursts; and a lack of persistent gradual intermediates during the evolutionary process (Cross et al., 2016). Punctuated evolution proposes that cells possess inherent or pre-existing traits or genetic alterations that predispose them to certain behaviours or phenotypes upon tumour initiation leading to the development of hallmark traits such as invasiveness, metastasis, and treatment resistance (Davis et al., 2017).

The timing of genomic changes and the role of selection may vary between different cases, or even within the same tumour during different stages of its development. As a result, different models may apply at various time-points of a tumour's development (Foo et al., 2011; Williams et al., 2016).

1.1.4 Intratumoural Heterogeneity and Metastasis

A subset of tumour heterogeneity is termed intratumoural heterogeneity, which pertains to the diverse tumour cell populations within the same tumour, as well as variations within and between different metastatic lesions of the same individual (Stanta & Bonin, 2018). Intratumoural heterogeneity significantly contributes to metastatic processes by fostering the emergence of advantageous mutations that support the acquisition and maintenance of hallmark traits of cancer (Hanahan & Weinberg, 2000). Importantly, it is also the inherent subclonality within tumours that holds profound implications for understanding the mapping of evolution between metastases (Hong et al., 2015).

Traditionally, metastasis was thought to occur sequentially, with cancer cells from the primary tumour first invading nearby tissues, then entering blood or lymphatic vessels, and finally colonising distant organs (Cheung & Ewald, 2016). However, recent research has shown that metastatic progression can occur through various mechanisms, including the concept of metastatic seeding, which suggests that metastatic sites can be seeded not only from the primary tumour but also from other metastatic sites (Gundem et al., 2015). This means that cancer cells from an existing metastasis can disseminate through the bloodstream or lymphatic system to colonise new sites in the body.

Metastatic seeding can be classified as monoclonal, originating from a single subclone, or polyclonal, arising from multiple subclones. While monoclonal seeding is by definition monophyletic, meaning it originates from a single ancestral cell lineage, polyclonal seeding can result from both mono- and polyphyletic seeding and leads to intermetastatic heterogeneity, wherein different metastatic sites within the same individual exhibit variations in their genetic composition. (Gui & Bivona, 2022). When metastatic seeding occurs, each newly formed metastatic site may contain a distinct subset of the genetic diversity present in the original tumour. While the genetic makeup of cells within a metastatic site accurately mirrors the genetic composition of that specific site, it may not encompass the entire range of genetic diversity found in the original tumour that seeded it. In essence, metastatic sites may exhibit variations in genetic diversity compared to the primary tumour, potentially leading to differences in genetic composition among metastatic sites within the same individual.

Therefore, analysing the evolutionary histories of a limited number of cells from each metastasis yields only partial insights into the genetic diversity within each site. Moreover, doing so for two metastatic sites does not ensure an accurate mapping of the subclonal cells that seeded from one site to the other. When analysing single cells from metastatic tumours, uncertainty arises regarding whether a cell accurately reflects its originating metastatic site or exhibits similarities with cells from other metastases due to their shared evolutionary history. This uncertainty is particularly pronounced in cases of polyclonal seeding, where cells from distinct subclones can populate multiple metastatic sites.

1.1.5 Implementation of Phylogenetics

Evolutionary relationships are primarily understood through the concept of coalescence (Kingman, 1982), where genetic lineages within a population are traced backward in time until they converge at a common ancestor. This mathematical framework is crucial for comprehending genealogical relationships among individuals within a population and the resulting patterns of genetic variation shaped by evolutionary processes.

Phylogenetics employs coalescent theory to reconstruct the evolutionary history of species or groups of organisms. By analysing genetic data from various species or populations, phylogenetic methods infer coalescent events occurring over evolutionary timescales (L. Liu et al., 2009). Represented as a diagram with nodes (points of divergence) and branches (connections between nodes), a phylogeny illustrates ancestral lineages and their descendants. Each branch represents a lineage's evolution over time. Closely positioned branches denote closely related organisms, while greater distances reflect more distant relationships. At the tips of the branches are the sampled groups or organisms of interest.

Given that cancer represents an evolutionary process driven by selection over time, cancer progression can be effectively traced through phylogenetic analyses. By detecting genetic mutations and alterations in samples across sequences, computational phylogenetic algorithms have been designed to reconstruct the evolutionary trajectories of cancer based on similarities and differences among genetic profiles (Schwartz & Schäffer, 2017).

1.1.6 The Problem of Incomplete Lineage Sorting

The coalescence and phylogenetics offer a framework for understanding the origin and diversification of genetic variation within populations over time. One challenge in evolutionary analysis lies in comparing species trees and gene trees as they often do not align. Species trees outline broader evolutionary relationships among different species, reflecting speciation events and divergence through evolutionary processes. In contrast, gene trees depict the relationships of gene sequences across species (Szöllősi et al., 2015). Ideally, species trees and gene trees should align, reflecting the evolutionary history of genes mirroring that of species. However, incongruence frequently occurs due to a phenomenon known as incomplete lineage sorting (ILS) (Maddison & Knowles, 2006).

ILS occurs when extant lineages in different groups share common ancestors, stemming from ancestral populations where multiple lineages coexisted prior to speciation. This persistence of ancestral polymorphism leads to incongruence between gene and species trees (Feng et al., 2022). ILS arises because gene lineages can coalesce at different points in the evolutionary history of a population. This means that genetic lineages from the ancestral population may not necessarily sort neatly into the descendant populations according to the branching pattern of the species tree. Instead, genetic lineages from different populations may coalesce back to a common ancestor within the ancestral population, leading to a shared genetic heritage between descendant populations.

Various factors influence the occurrence of incomplete lineage sorting. For instance, the effective population size, defined as the size of an idealised population that would experience the same amount of genetic drift as the actual population under consideration, plays a significant role. A higher effective population size often leads to more opportunities for incomplete lineage sorting (Hobolth et al., 2011). This is because larger population sizes result in longer coalescent times, which means that ancestral lineages persist for extended periods before merging into a common ancestor. During this extended duration, genetic diversity accumulates within the population, including ancestral polymorphisms that may persist across descendant populations. Moreover, the stochasticity inherent in longer coalescent times increases the chance of genetic drift and mutation events, further contributing to the retention of ancestral polymorphisms and the likelihood of incomplete lineage sorting.

As time since speciation increases, there is less chance of observing incomplete lineage sorting because genetic lineages have had more opportunity to fully differentiate into distinct species. This reduced persistence of ancestral polymorphisms and clearer species delineation decreases the likelihood of observing incongruent genealogies among genetic loci. Importantly, longer coalescent times can also increase with greater time since speciation. While longer coalescent times provide more opportunities for genetic lineages to coalesce and merge, potentially leading to incomplete lineage sorting, the overall impact of increased genetic divergence over time will generally outweigh this effect.

In the context of cancer research, ILS can be considered analogous to the divergence observed in metastatic seeding. In ILS, as time progresses, genetic lineages fail to merge before speciation events, resulting in divergence within species. Similarly, the process of metastatic seeding, which can occur in a polyclonal manner, leads to divergence in tumour lineages, as the genetic makeup of metastatic sites diverges from that of the tumour that seeded it, due to the presence of cells from different subclones. Just as ILS contributes to evolutionary divergence among species, metastatic seeding contributes to the genetic heterogeneity observed among tumours, highlighting parallels between evolutionary processes and cancer progression.

1.2 Motivation

Our motivation lies in addressing a researcher's choice in sequencing strategies when seeking to analyse metastatic progression. To be more specific, we consider a scenario where a researcher aims to map the evolutionary history among several tumours/metastatic sites within a patient. One approach they could consider is obtaining scSEQ data, selecting individual cells from each metastatic site. However, to depict the evolutionary history of multiple sites, rather than just the cells themselves, these cells must first be appropriately grouped. Alternatively, they could opt for bulkSEQ, where the composition of cells within a resection represents a cell admixture. Both sequencing methods exhibit particular use cases, each characterised by advantages and limitations (Table 1.1).

	Advantages	**Disadvantages**
BulkSEQ	Higher throughput	Masked cellular heterogeneity
	Cost-effective	May miss identification of rare cell subpopulations
	Less technical artifacts	Cannot accurately identify all specific cell types within a sample
	Less complex pipelines and resources	Measures average expression across a population of cells
scSEQ	High resolution of cellular heterogeneity	Data and time intensive
	Can identify rare cell subpopulations	High cost
	Can identify and classify cell types	Increased technical variability
	Measures the gene expression of individual cells	Less established technology

Table 1.1: The advantages and disadvantages of bulkSEQ and scSEQ are well-established at their respective scales of analysis: bulkSEQ in higher-level population dynamics and the latter on individual cellular interactions.

From these considerations, several questions emerge:

- Perhaps the researcher already possesses scSEQ data and is now seeking the optimal method to infer a multi-tumour phylogeny. Could applying a novel method using consensus sequences to pool data originally derived from scSEQ be more accurate in inferring the phylogeny versus just using bulkSEQ?

- Conversely, is the researcher interested in acquiring sequencing data to infer this evolutionary history, requiring determination of whether the investment in scSEQ data acquisition is worthwhile, or is bulk sequencing sufficient for their needs?

With evolutionary history in mind, our primary focus is on recovering the multi-tumour phylogeny, which drives our preference for pooled scSEQ data. This choice is motivated by concerns about potential interleaving that could arise if we were to rely solely on a phylogeny that displays the evolutionary history of individual cells as a fixed template for the evolutionary history of its higher population counterpart. In other words, the genealogy between individual cells a researcher samples may not correctly represent the genealogy of their corresponding metastatic sites. The individual cells are unlikely to cluster or evolve in perfect alignment with the progression depicted by a multi-tumour phylogeny. When we simulate cancer cells diverging from the primary tumour/metastatic sites to several other metastatic sites, we understand that the composition of cells in each tumour may be highly mixed. These tumours would have seeded from a cell/mixture of cells originating from an already highly heterogeneous environment. Understanding this, we would expect that the cell-cell history among multiple metastatic sites will often show extensive interleaving that does not align accurately with the true multi-tumour tree.

To address the questions stated above, we can simulate this scenario using computational methods and utilise the generated data to compare the reconstruction capabilities of various sequencing methods in delineating the evolutionary history of multi-tumour evolution within a patient.

1.3 Problem Statement

We aim to investigate the impact of different sequencing analysis methods on phylogenetic accuracy during multi-tumour tree reconstruction. Simulations offer a means to validate our methods. Here, phylogenetic accuracy refers to the degree of similarity between a reconstructed phylogenetic tree and a simulated true tree.

The main research aims we explore are:

1. Development of a simulation strategy that accurately represents the multi-tumour tree structure, effectively capturing the diverse parameters inherent in cancer single-cell data.

2. To assess the effectiveness of pooled consensus sequences from single-cell data as an alternative to bulk sequencing analysis, we aim to demonstrate their potential for achieving enhanced accuracy over bulk sequencing data in reconstructing the evolutionary histories of multiple metastatic sites within a patient.

3. To further evaluate the phylogenetic accuracy of multi-tumour tree estimation, we compile existing data by sets of replicates, to allow for performance comparisons between compiling directly between scSEQ data, which incorporates individual cell data, against methods involving pooled data. By doing so, we aim to ascertain the necessity of allocating resources to scSEQ data, or if the information obtained from reconstructing bulkSEQ data suffices.

1.4 Objectives

To meet these goals, we implement a stratified sampling method, wherein we vary the values of various biological and technical parameters associated with tumour evolution, executed across two simulation software programs. The first software tool simulates the true multi-tumour tree and its corresponding true cell-cell trees. Each cell on the tips of the cell-cell trees will be labelled based on the metastatic site from which they were sampled from. The second software tool will generate single-cell sequences based on the cell-cell tree files. By performing phylogenetic reconstruction on the generated sequences, we aim to determine whether scSEQ or bulkSEQ is more effective in estimating multi-tumour trees. This requires grouping the scSEQ data for two purposes: firstly, to ensure that the scSEQ data is aggregated at a higher hierarchical level, facilitating analysis across multiple tumours rather than their individual cells; and secondly, to create a pseudobulk dataset that will proxy as bulkSEQ data.

Once tree reconstruction and comparison of our initial pooling methods has been completed, we can further expand the analysis by comparing three different approaches that leverage the replicates of our reconstructed tree files: species tree estimation, supertree construction, and reconstruction of concatenated replicate sequences. The significance of this expanded comparison lies in directly evaluating the effectiveness of using reconstructed trees from scSEQ data (cell-cell level) to estimate multi-tumour data, against that of existing multi-tumour level data.

Our research will progress in the following ways:

1. Develop a pipeline that integrates two simulation software tools, incorporating parameter values generated by our sampling dataframe into each sample. This pipeline aims to accurately simulate true multi-tumour/cell-cell trees and their corresponding scSEQ data.

2. Evaluate whether obtaining scSEQ followed by building haplotype consensus sequences (H-CS) provides a more accurate approach than generating VCF pseudobulk (VCF-PB) data for reconstructing multi-tumour evolutionary histories.

3. Perform multi-tumour tree estimation using three different methods, each leveraging replicates of a sample: Using a species tree estimation software with reconstructed cell-cell tree files as input, generating a supertree from reconstructed multi-tumour tree files, and concatenating grouped sequences to reconstruct a multi-tumour tree.

1.5 Overview of Research

All computational experiments are performed under the NeSI (New Zealand eScience Infrastructure) environment. The execution of experiments involves utilising a combination of software tools and programming languages, primarily R, Linux and Python, along with the command-line interface for script execution. The sets of scripts utilised for our simulation runs, along with an updated final script set, are available at: https://github.com/fu-jono-283/thesis.git

In total, we record 3 simulations, each varying the range of a specific parameter, metastatic rate, which we define as the rate at which metastatic sites are produced. Our simulations consist of 200 distinct samples, each uniquely characterised by parameter values determined by a dataframe generated by latin hypercube sampling (LHS) (McKay et al., 1979). The simulation process results in the generation of 200 multi-tumour trees, with each multi-tumour tree consisting of 8 distinct cell-cell tree replicates. This totals to 1600 individual data points/iterations generated within the LHS dataframe. The term ‘multi-tumour tree’ is used to define a phylogenetic tree that represents a genealogy of multiple metastatic sites within an individual patient. The ‘cell-cell tree’ is it’s finer-scale counterpart, that represents the genealogy of the individual cells sampled from each metastatic site.

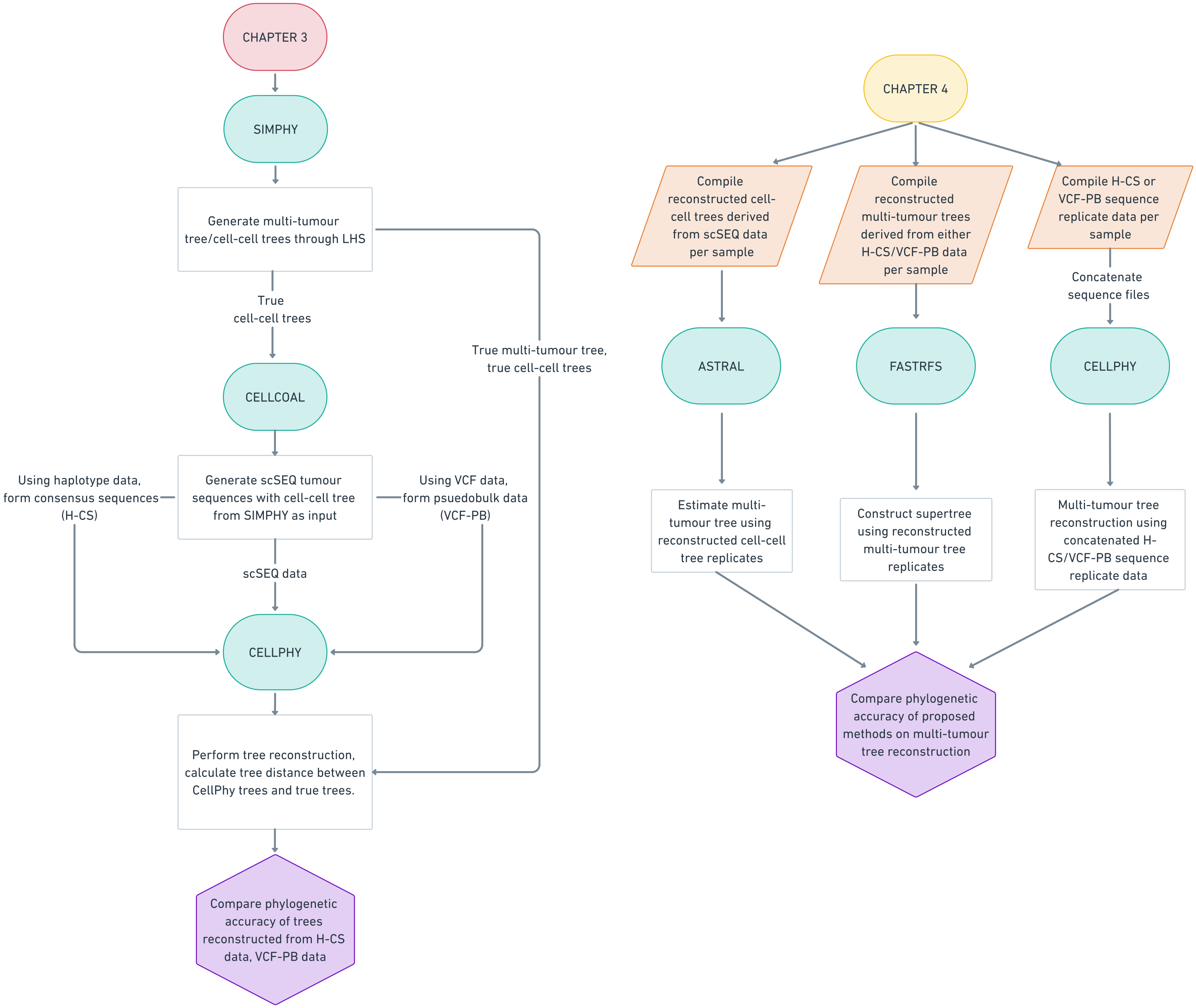

CHAPTER 3
SIMPHY
Generate multi-tumour tree/cell-cell trees through LHS
True cell-cell trees
True multi-tumour tree, true cell-cell trees
CELLCOAL
Using haplotype data, form consensus sequences (H-CS)
Generate scSEQ tumour sequences with cell-cell tree from SIMPHY as input
Using VCF data, form psuedobulk data (VCF-PB)
scSEQ data
CELLPHY
Perform tree reconstruction, calculate tree distance between CellPhy trees and true trees.
Compare phylogenetic accuracy of trees reconstructed from H-CS data, VCF-PB data
CHAPTER 4
Compile reconstructed cell-cell trees derived from scSEQ data per sample
Compile reconstructed multi-tumour trees derived from either H-CS/VCF-PB data per sample
Compile H-CS or VCF-PB sequence replicate data per sample
Concatenate sequence files
ASTRAL
FASTRFS
CELLPHY
Estimate multi-tumour tree using reconstructed cell-cell tree replicates
Construct supertree using reconstructed multi-tumour tree replicates
Multi-tumour tree reconstruction using concatenated H-CS/VCF-PB sequence replicate data
Compare phylogenetic accuracy of proposed methods on multi-tumour tree reconstruction

Figure 1.1: The Chapter 3 pipeline processes simulated scSEQ data to pool cells based on the metastatic sites each cell is sampled from. The Chapter 4 pipeline explores various methods for reconstructing the multi-tumour tree using compiled replicate information derived from data generated in Chapter 3.

We treat the terms ‘tumour’ and ‘metastatic site’ interchangeably throughout our project, allowing us to focus our analysis without the constraint of distinguishing a primary tumour. Our analysis prioritises assessing the similarity of branching patterns between reconstructed trees and the true evolutionary history rather than focusing on the precise chronological sequence of events. This perspective directs our choice to implement the generalised Robinson-Foulds distance (gRF) (Smith, 2020) as a metric for comparison. The gRF compares two phylogenetic trees by evaluating the agreement or disagreement between their internal branches. It accomplishes this by comparing shared bipartitions, which are the divisions of the leaves (for multi-tumour scaled trees, these leaves represent metastatic sites) into groups shared between the two trees. A bipartition is considered shared if it exists in both trees, meaning that the same set of leaves is grouped together in both trees. It then assigns similarity scores to each pair based on their degree of resemblance. An optimal matching is then determined to maximize the total similarity score across all pairs of splits. The gRF distance is computed as the dissimilarity between the trees based on the total similarity score of the optimal matching. By focusing on the agreement or disagreement of bipartitions rather than specific branch lengths or temporal information, the gRF metric provides a robust measure of tree similarity that is independent of the timing of evolutionary events.

In the Methods section for Chapter 3 and 4, where we describe how the generated LHS dataframe is updated with gRF scores, we retain the original definition of an optimal tree score, entering the values as 0, which aligns with the definition of the gRF distance. However, for subsequent statistical analyses and results, we normalise this score by adjusting it such that a score of 1 reflects perfect reconstruction accuracy. This adjustment facilitates ease of interpretation, where a higher score indicates better phylogenetic accuracy.

Replicates are an important component of this project. In Chapter 3, they allow for the exploration of outcome variability and assists with establishing statistical significance in our findings. In Chapter 4, the simulated data generated in Chapter 3 is compiled in sets of replicates to allow for the use of species estimation methods, supertree construction and phylogenetic reconstruction of concatenated sequences. Note that although parameter values remain consistent across replicates within a sample, each replicate is depicted by a unique cell-cell tree. These distinct trees correspond to the same multi-tumour tree initially generated by our genealogy simulator at the beginning of the pipeline. To provide a biological context, considering that we have conceptualised each sample in our dataframe as representing the parameter values of a specific cell, to further this, each replicate would then resemble a distinct, unlinked genomic region within that cell.

Finally, note that the utilisation of a pseudobulk dataset serves as a pragmatic solution to navigate around the logistical challenges associated with acquiring both bulkSEQ and scSEQ data from identical cellular samples. In practical scenarios, acquiring bulkSEQ data alongside scSEQ data from the same source presents significant challenges due to the constraints of available resources and technology, making it difficult to conduct both sequencing methodologies on the same set of cells. Furthermore, the processes of isolation, lysing and extraction required in scSEQ procedures would most likely preclude subsequent bulk sequencing methods to be applicable on the same cells, thus exacerbating the complexity of obtaining concurrent bulkSEQ and scSEQ data. In lieu of real bulkSEQ data, we leverage the scSEQ data generated to construct a pseudobulk dataset, effectively simulating bulkSEQ characteristics for comparative analyses.

1.6 Structure of Thesis

Chapter 2: Simulation Methodology
We present our chosen sampling method, LHS as the approach for generating our samples. Additionally, we introduce the five main software tools integral to our project: SimPhy (Mallo et al., 2016), CellCoal (Posada, 2020), CellPhy (Kozlov et al., 2022), ASTRAL (Zhang et al., 2018) and FastRFS (Vachaspati & Warnow, 2017).

Chapter 3: Assessing the Reconstruction Accuracy of Consensus Sequences
We discuss the computational processes by which we generate our dataframe using LHS, incorporating different parameter values for each sample, and how these values are utilised in our SimPhy and CellCoal simulations. Using SimPhy, within each sample, we generate a true multi-tumour tree along with its eight corresponding cell-cell tree replicates. For each cell-cell tree replicate, we simulate their single-cell tumour sequences using each tree file as input in the USERTREE parameter of CellCoal. CellCoal presents the resulting sequences as both a haplotype file and VCF file. For our project, the haplotype file is utilised in two ways. Firstly, the haplotype file can undergo

phylogenetic reconstruction through CellPhy in its original state. This will be an assessment of the accuracy of our simulated scSEQ data to reconstruct their true cell-cell tree. The haplotype file can also be processed prior to reconstruction, constructing a consensus sequence by pooling cells based on the metastatic site they have been sampled from. The VCF file can be transformed into a pseudobulk representation through weighted processing of the Phred-scaled likelihood score. After the grouped data is processed through CellPhy, the resulting reconstructed trees are compared to the true multi-tumour tree to assess phylogenetic accuracy. Additionally, parameters are assessed to determine their impact on phylogenetic accuracy.

Chapter 4: Extending Tree Reconstruction via Replicate Data Integration

In the previous chapter, we focus on comparing the accuracy of multi-tumour tree reconstruction exclusively between consensus sequences obtained from pooled individual cells and bulk sequencing data. However, to incorporate reconstruction derived directly from scSEQ data into this comparison, we must compile and utilise the existing data by sets of replicates per sample. Specifically, we can leverage species-tree estimation methods using the software tool ASTRAL, which conventionally utilises multi-locus data from multiple gene tree lineages to infer a species tree. However, in our approach, we adapt ASTRAL to compile the replicates of our cell-cell trees. Compiling our data in sets based on replicates allows us to approach the comparison of phylogenetic accuracy of scSEQ/bulkSEQ on a multi-tumour scale in several ways. To ensure a fair comparison with H-CS and VCF-PB, we must similarly apply the process of compiling datasets based on sets of replicates to these methods. We utilise the software tool FastRFS to construct supertrees, leveraging firstly the reconstructed H-CS tree files and then the VCF-PB tree files. Finally, we explore the potential of concatenating sequence files, again utilising both the H-CS and VCF-PB tree files, implementing CellPhy for phylogenetic reconstruction.

Chapter 5: General Discussion

We conclude this thesis by discussing our findings, acknowledging certain limitations of this study and present potential directions for future research.

2 Simulation Overview

In this chapter, we provide an overview of the main tools utilised in our project: This includes LHS implementation and the simulation software incorporated into our pipeline.

2.1 Latin Hypercube Sampling

Generating a dataframe in LHS, we introduce variability across multiple parameters within our simulations. These biological/technical parameters (Table 2.1) define a range of possible value combinations which can be represented as interconnected dimensions that form a complex high-dimensional grid. This is referred to as the sample space.

Parameter	Program	Description
Number of Tumours	SimPhy	Dictates the quantity of tumours simulated
Exponential Growth Rate	CellCoal	Rate at which the number of tumour cells increases exponentially per generation
Cells per Tumour	SimPhy	Number of cells sampled from each tumour
Effective Population Size	CellCoal	The number of cells affecting genetic variation and evolutionary processes within the simulated tumour population.
Metastatic Rate	SimPhy	The rate at which new tumours are formed per generation
Mutation Rate	SimPhy, CellCoal	The rate of how frequently mutations occur per site per generation
Number of Sites	CellCoal	The total length in bases, within the genomic sequences of our simulation
Allelic Dropout	CellCoal	A fixed rate probability that represents the failure to amplify one of the two alleles present, per genotype and cell
Amplification Error	CellCoal	Expected Beta-binomial probability (mean) of introducing an amplification error per read per site
Sequencing Error	CellCoal	A fixed rate probability that represents additional sequencing errors. Assumed constant across all sites

Table 2.1: Our choice of biological and technical parameters manipulated in our simulations is guided by their significance in effecting evolutionary patterns and/or their relevance to errors encountered during the sequencing of cancer data.

Each parameter contributes an axis to the sample space, collectively creating a grid structure (Fig. 2.1). When selecting for a suitable sampling method, we considered that the introduction of more parameters would lead to more interdependencies within the analysis and an expansion in complexity of this grid. Opting for an exhaustive grid search whereby we evaluate all conceivable combinations within our specified parameter ranges, would result in an excessive volume of data, imposing constraints on processing time. On the other hand, traditional random sampling would not be able to guarantee the generation of a data point from any predefined subset of the sample space. This introduces the potential for both oversampling or undersampling parts of the grid, wherein certain areas may be excessively emphasised or inadequately represented in the resulting data.

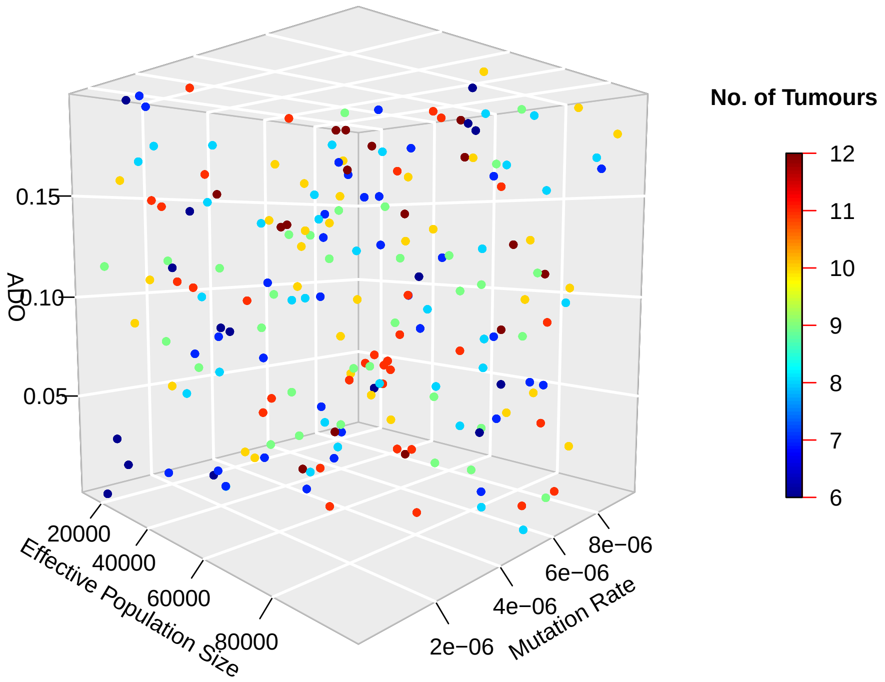


Figure 2.1: This 4D plot visualises the dispersion naturally observed in LHS throughout our first simulated run, comprising of 200 data points distributed across four parameters: allelic dropout, effective population size, mutation rate, and the number of tumours sampled.

LHS is integral to our sampling approach as it allows us to specify the desired number of samples and distributes the simulation evenly across each dimension, ensuring equal coverage of the entire parameter space. This capability enables us to thoroughly explore a diverse range of parameter combinations while maintaining consistency in sample size across different settings, thereby avoiding oversight of critical regions. It presents a structured approach to randomisation that can alleviate the issues associated with traditional random sampling. It does so by partitioning the sample space into intervals, with each interval contributing one data point. Within this context, granularity is a term that refers to the level of fineness by which the intervals are defined. Therefore, complexity of the sample space is dependent on our preference of granularity, determining how small or large we want each interval to be. Our ability to manage this complexity, whether by adjusting granularity or incorporating additional parameters, underscores the value of LHS in facilitating an efficient selection of samples.

2.2 SimPhy

Comparing the evolutionary history among multiple metastatic sites within a patient to those of their individual tumour cell counterparts presents a challenge. An option we propose is to model the simulation of scSEQ data with assigned tumour lineages for each cell. This requires a structured population that models the genealogy of each individual cell. With regard to studies of metastatic progression, researchers typically focus on inferring the evolutionary trajectory at a higher level, perhaps analysing the evolutionary progression of different tumours within the same individual. Or, at a finer resolution, they may examine the heterogeneity within individual tumour samples. They may attempt to do so by using either bulkSEQ data or scSEQ data. However, it is difficult to consider a scenario where one has access to both sets of data simultaneously. We understand that for bulkSEQ analysis, samples are often obtained from tumour resections, where a portion of a tumour mass is subjected to analysis. Note that this process does not preserve the cellular context of the individual cells within the tumour. On the other hand, obtaining scSEQ data requires the isolation of single cells, which is often achieved through techniques like microfluidics or droplet-based methods (Zhou et al., 2021). These approaches enable the capture of individual cells and their genetic information. Isolating the exact same single cells that were present in the bulkSEQ sample is technically challenging and realistically not feasible. Therefore, if our aim is to compare multi-tumour evolution reconstruction using both scSEQ and bulkSEQ datasets from the same source, we must utilise unique simulation approaches that allow us to construct structured populations representing the genealogies of multiple metastatic sites, as well as the genealogies of the corresponding cells sampled from each metastatic site.

We explore the synergistic use of two existing simulation software tools. The first, SimPhy (Mallo et al., 2016), is responsible for generating the genealogies of both the true multi-tumour tree and its corresponding true cell-cell trees for each iteration. Subsequently, we utilise the genealogies generated through SimPhy by inputting the cell-cell tree files through the UserTree parameter in our second tool, CellCoal, which simulates our single-cell sequences based on the provided tree files. This two-step process enables the generation of our structured cell populations, and allows us to simulate scSEQ data for our single-cells with established tumour lineages.

SimPhy is a computational software tool used for simulating the evolutionary processes of multiple gene families. It provides users with the ability to build phylogenetic trees, with the flexibility to adjust parameter values governing the effective population size, tree-wide substitution rates, species-birth rate and other factors influencing genetic variation. In its conventional usage, SimPhy finds application in evolutionary biology, where it serves as a tool for investigating species relationships and genetic evolution. Therefore, its parameter terminology, includes terms such as ‘tree-wide substitution rates’, representing the frequency of nucleotide substitutions occurring across the simulated phylogenetic tree, and ‘species-birth rate’, defined as the rate at which new species are generated in the simulated phylogenetic tree through speciation events. However, in our adaptation of SimPhy for the simulation of tumour sequences, we modify these terms to fit the context of our project. For instance, the parameter in SimPhy that denotes ‘tree-wide substitution rates’ will be interpreted as ‘mutation rate’ in our simulations, reflecting the frequency of mutations occurring in the simulated DNA sequence of a cell per unit of simulated time. Similarly, ‘species-birth rate’ is redefined as ‘metastatic rate’, representing the rate at which new tumours are generated in the simulated phylogenetic tree. We discern this difference in terminology to align with our adaptation of the SimPhy software.

In order to simulate the trajectories of multiple gene families, SimPhy relies on the Multispecies

Coalescent (MSC) model, a mathematical framework commonly used in phylogenetics to depict the genealogical relationships among individuals from multiple species. This model accounts for the shared ancestry of individuals within and between species over time, taking into consideration factors such as changes in population size, migration, and speciation events. SimPhy also accounts for various population-level phenomena such as ILS.

We use SimPhy to facilitate the comparison of pooled scSEQ data, either through the use of haplotype consensus sequences, or a processed VCF file that mimics bulkSEQ. The data that both methods use originate from an individual cell scale, and are compared against a true multi-tumour tree. SimPhy allows for this implementing a nested tree hierarchy derived from the MSC model that captures evolution across three levels: species trees, locus trees, and gene trees. In population genetics, research revolving around discordance of genealogies often centres on the relationship between species trees and gene trees, which represent different aspects of evolutionary history. In the context of our project, the relationship between species trees and gene trees can be analogous to the relationship between a multi-tumour tree and a cell-cell tree. In both cases, there is a hierarchical structure where higher-level entities (species or tumours) are composed of lower-level entities (genes or individual cells). Species trees represent the evolutionary history of species composed of multiple individuals, similar to how a multi-tumour tree represents the evolutionary history of multiple tumours. Likewise, gene trees depict the evolutionary relationships among individual genes within populations, akin to how a cell-cell tree represents the evolutionary relationships among individual cells within each metastatic site.

Therefore, in our tumour sequence simulations, SimPhy's hierarchical phylogenetic tree model will be implemented as follows:

- Species tree (multi-tumour tree) - At the apex is the species tree, a graphical representation depicting the evolutionary history of sampled tumours within a patient. This tree serves as a framework to illustrate the tumour-tumour relationships. Each leaf on the multi-tumour tree represents a distinct tumour sample from the primary tumour and different metastatic sites within the same individual. Nodes indicate the divergence of distinct tumour lineages. Branches provide a visual representation of the evolutionary trajectories of tumours, illustrating how metastases from different sites have evolved over time within a patient.

- Locus Tree - Although the locus tree is not explicitly used in our analyses, it serves as a default intermediary within the three-tree model, capturing the evolutionary history of genetic loci (regions of the genome) within a species.
- Gene tree (cell-cell tree) - At the lowermost level, within each locus of the locus tree, there exists a corresponding representation for individual tumour cells, analogous to the gene tree. This tree is built based on scSEQ data, capturing the evolutionary history of sampled cells. This cell-cell tree serves as a depiction of the evolutionary history of sampled tumour cell populations. Nodes indicate coalescent events.

The simulated tree files generated by SimPhy provide a ground truth against which reconstructed trees from phylogenetic inference methods can be evaluated. We can leverage these simulated trees to assess the accuracy and performance of different phylogenetic inference methods by comparing them to the true trees. These files also facilitate part of our evaluation of how various biological and technical factors may influence the phylogenetic accuracy of our methods. Furthermore, the cell-cell tree files generated by SimPhy will allocate a distinct label to each cell, denoting its association with a particular metastatic site generated within the multi-tumour tree. The cell-cell tree files are imported into CellCoal, where sequences are generated to align with the evolutionary relationships delineated by the tree file.

2.3 CellCoal

CellCoal (Posada, 2020) serves as the next tool in our simulation pipeline. Note that while SimPhy, our genealogy simulator, has the capability to generate sequence data corresponding to the genealogies it simulates, we have chosen to utilise CellCoal for constructing our single-cell sequences. This decision stems from CellCoal's specialisation in simulating somatic single-cell sequencing genotypes derived from cell populations. SimPhy, on the other hand, primarily functions to generate genealogies by incorporating the MSC model, which accounts for phenomena such as incomplete lineage sorting.

By itself, CellCoal operates with unstructured populations. Therefore, we first utilise SimPhy to generate true multi-tumour trees and their cell-cell tree counterparts. By importing these files into CellCoal as a USERTREE parameter, we can simulate sequences aligned with the evolution of cells, while incorporating their metastatic site affiliations. This enables the generation of sequences that reflect the initial cell-cell tree, with each cell consisting of a metastatic site label. As a result, while the sampled cells are structured according to their evolutionary relationships, the inclusion of metastatic site labels allows for the visualisation of how the cells may interleave, depicting divergence away from the initially sampled site.

CellCoal begins by generating a genealogy for the sampled cells, implementing the neutral coalescent. This model traces the ancestral relationships among the sampled cells by working backward in time. The genealogy is generated with the underlying assumption that the sampled cells are representative of a larger population that may have remained constant or fluctuated over time.

Additionally, it incorporates a tailored parameterization of the coalescent model, designed specifically for cancer cell samples. This parameterization accounts for key characteristics of cancer cell populations, including the presence of SNVs, genetic heterogeneity, and clonal expansion, enabling more precise simulations of cancer evolution.

Due to the nature of the program's simulation pipeline, two additional branches - a root branch and an outgroup branch - are automatically added to our genealogies and corresponding sequence data. These branches connect the most recent common ancestor within our cell-cell trees to the outgroup, resembling a normal somatic cell. As mentioned earlier, our analysis employs the gRF metric, chosen for its ability to assess phylogenetic tree similarity without the need for rooting, as we are not concerned with temporal directionality. While the root branch traditionally imparts directionality to the phylogeny, we are solely interested in tree similarity irrespective of the specific order of evolutionary events. We do not consider the notion of a primary tumour; instead, each terminal node in the generated multi-tumour tree is interpreted as a metastatic site. Consequently, we manually exclude root sequences from the resulting sequence files to align with our analytical focus.

CellCoal allows for the specification of a mutation model to estimate the rates at which nucleotide substitutions occur, determining the likelihood of one nucleotide replacing another. We have opted for the general-time-reversible (GTR) model due to its capability to accommodate rate heterogeneity among different nucleotide substitutions and variation in nucleotide frequencies. To incorporate rate heterogeneity, we integrate values extracted from a cosmic mutational signature (Alexandrov et al., 2020) into the substitution matrix. Obtained from a CellCoal example, these values were carefully chosen for their compatibility with the GTR model in our simulations. The GTR model enables variation in nucleotide frequencies across sites in a sequence, reflecting the composition of nucleotides at individual sites. We set uniform frequencies across all nucleotide bases, at 0.25.

From CellCoal, the two obtained file formats we utilise within our project are as follows:

1. Full Haplotype File - This file is formatted in PHYlogeny inference package (PHYLIP) file format. Sequences are arranged in a tabular structure, with each row representing a sequence and columns indicating nucleotides at particular positions in the alignment.

2. VCF File - This file provides information about genetic variants and their attributes across the sampled cells. Each entry in the VCF file represents a specific variant, such as SNVs, along with additional details such as genomic position, reference allele, alternative allele, quality scores, and genotype information for each cell.

2.4 CellPhy

CellPhy (Kozlov et al., 2022), our preferred tool for phylogenetic inference, is built upon RAxML-NG (Stamatakis, 2014), utilising its framework for maximum likelihood-based phylogenetic tree estimation. CellPhy effectively manages the output files generated by CellCoal. By leveraging the optimisation routines and tree search strategies of RAxML-NG, CellPhy is able to estimate maximum-likelihood values for model parameters to reconstruct phylogenies from SNVs. It also incorporates an error model to account for technical artefacts - amplification error and allelic dropout. The program expands upon the GTR model to infer tree topology, extending it from the traditional 4-state model to a 16-state model. This extension involves focusing on diploid genotypes rather than haploid genotypes, where each genetic locus possesses two copies of the genome.

By doing so, the model is enhanced to to accommodate the possible combinations of alleles at each site, resulting in a significantly larger state space. This model is renamed as GT16.

The set of all possible transitions between pairs of nucleotide states is as shown:

$$\Gamma = \{A|A, A|C, A|G, A|T, C|A, C|C, C|G, C|T, G|A, G|C, G|G, G|T, T|A, T|C, T|G, T|T\}$$

When provided with a set of observed single-cell SNV genotypes, either in PHYLIP format or as a standard VCF file, CellPhy implements its error and genotype models to calculate the likelihood of a given phylogenetic tree. This likelihood is computed as the product of independent probabilities across SNVs, using the Felsenstein pruning algorithm (Felsenstein, 1973). This algorithm is a fundamental method utilised in phylogenetic inference that evaluates the likelihood of a phylogenetic tree by considering the probability of observing the data at each node in the tree and propagating this information upwards through the tree structure.

There are two different models we utilise from CellPhy:

1. CellPhy-ML - This mode of CellPhy takes a genotype matrix as input and infers phylogenies under their genotype error model. We use this mode to reconstruct trees using our haplotype files as input.

2. CellPhy-GL - This mode of CellPhy utilises genotype likelihoods. It requires a VCF file with an appropriate PL field, which stores the Phred-scaled genotype likelihoods. These represent the likelihoods of each possible genotype given the observed data and therefore does not require the implementation of an error model. We utilise this to reconstruct trees using our VCF files as input.

The reconstructed trees obtained from CellPhy will be compared with the true multi-tumour trees generated by SimPhy using the generalised Robinson-Foulds distance metric. This comparison allows us to assess the phylogenetic accuracy of our various methods for obtaining multi-tumour sequence data from initial scSEQ data, either by constructing consensus sequences using haplotype data or generating a pseudobulk dataset using VCF data.

2.5 ASTRAL

Species tree inference, in the context of our project, involves deducing the evolutionary relationships among multiple metastatic sites using molecular data. Instead of species in the traditional sense, we are analysing genetic information from individual cells sampled from different tumours. In our project, each sample, characterised by unique parameter values, will consist of 8 replicates. Therefore, a sample will consist of 8 true individual cell-cell trees and 1 true multi-tumour tree. We treat these replicates as distinct, unlinked genomic regions.

ASTRAL (Zhang et al., 2018) is a widely established tool implemented for estimating species trees by leveraging multiple gene lineages. It incorporates algorithms to reconcile conflicting gene trees, estimating the most accurate species tree feasible from the available genetic data. Notably, ASTRAL operates under the multispecies coalescent, addressing complexities like ILS. Our objective here is to utilise ASTRAL to estimate the true multi-tumour tree by compiling our cell-cell tree replicates (scSEQ) as input.

ASTRAL uses dynamic programming to search for the tree that shares the maximum number of quartet topologies (the relationship between four taxa in a tree) with input gene trees. Essentially, a higher number of quartet topologies within the resulting tree suggests a better fit to the underlying evolutionary relationships among the sampled taxa. Therefore, the presence of more quartet topologies in the inferred tree is indicative of a higher likelihood of accurately capturing the true relationships among the taxa.

2.6 FastRFS

Supertree construction involves deriving a tree that encapsulates the evolutionary histories among a group of taxa, often species. One approach to sourcing these trees is by utilising sets of reconstructed species tree replicate files, or, in our case, reconstructed multi-tumour tree replicate files. Each file represents an independent estimate of the multi-tumour tree, offering multiple perspectives on the evolutionary histories among the metastatic sites. The consistency observed across these replicates can enhance confidence in the inferred relationships and mitigate the impact of stochastic variation.

Reconciling the information from independent estimates to construct a cohesive supertree presents a significant computational challenge. This task is formally known as the Robinson-Foulds Supertree problem. The objective is to find a supertree that accurately represents the consensus or majority signal among the source trees while minimising topological dissimilarity, quantified by the Robinson-Foulds (RF) distance. The RF distance is a metric that assesses topological dissimilarity by counting the number of bipartitions (i.e., splits of taxa into two disjoint subsets) present in one tree but not the other. Minimising this distance ensures that the resulting supertree closely aligns with the relationships inferred from the source trees.

We incorporate the use of FastRFS (Vachaspati & Warnow, 2017), a fast supertree method that finds an exact solution to the Robinson-Foulds Supertree problem within a constrained search space.

3 Assessing Reconstruction Accuracy of Consensus Sequences

3.1 Introduction

In this chapter, our primary objective is to assess whether consolidating single-cell sequences into consensus sequences yields improved performance in reconstructing the true tree compared to utilising phylogenetic reconstruction with a pseudobulk dataset sourced from the same data. We also assess whether scSEQ, in its own right, accurately reconstructs the phylogenetic relationships of the true cell-cell tree. We achieve this by evaluating tree similarity using the gRF distance, comparing the reconstructed tree of each method to the true trees initially generated in SimPhy.

We outline the methodology used to generate our data points within the LHS framework. We describe the process by which we use SimPhy to generate our initial true multi-tumour/cell-cell trees, and CellCoal to generate our tumour sequences. Additionally, we describe the process of manipulating the CellCoal output so that the scSEQ information is suitable for multi-tumour level analysis. More specifically, we explain the process of constructing consensus sequences from the haplotype data, as well as the construction of a pseudobulk dataset through a pooling method applied to the VCF data. Subsequently, reconstruction is performed through CellPhy, for the original single cell sequence files (scSEQ), the haplotype consensus sequences (H-CS), and the VCF pseudobulk (VCF-PB) dataset.

A researcher interested in studying the evolutionary histories of multiple metastatic sites within a patient will have the option of choosing between bulkSEQ data or scSEQ data. Although bulkSEQ presents advantages in computational efficiency, it tends to produce simplified representations of evolutionary relationships, leading to the possibility of inaccuracies in tumour phylogenies that deviate from true phylogenetic relationships. It provides an aggregated perspective that is vulnerable to misinterpretation owing to the incomplete representation of diverse subclones within the resected sample, thus failing to comprehensively capture genetic diversity. Conversely, while scSEQ has often demonstrated superior accuracy and precision in genomic analyses, primarily due to its capability to capture relationships between individual cells, it comes with significant resource constraints and the potential for errors throughout the sequencing process.

In this scenario, one concern regarding scSEQ is that the selection of sampled cells from each metastatic site may represent different subclones within the tumour due to intratumour heterogeneity. When examining the evolutionary relationships across multiple metastatic sites, it is probable that relying solely on the relationships of individual cells to portray the evolutionary histories of their higher-scale metastatic sites will lead to discordance. This discordance arises because individual cells may represent specific subclones within the tumour, each with its own evolutionary trajectory. As a result, the genetic relationships inferred from individual cells may not fully capture the evolutionary dynamics of the entire metastatic site. This is akin to ILS, where genetic lineages from ancestral populations fail to coalesce before speciation events. Our approach to overcome this discordance is to utilise consensus sequences, defined as sequences derived by bases that represent the most common nucleotide or amino acid at each position in an alignment of multiple related sequences (Schneider, 2002). To construct consensus sequences for this project, cells are pooled based on the metastatic site from which each cell is sampled from.

However, there has been limited attention given to the potential of leveraging consensus sequences as a method for aggregating cells derived from scSEQ data, particularly in facilitating cancer-related phylogenetic reconstruction. Currently, data for such an approach is absent. In phylogenetics, consensus sequences have been extensively utilised to summarise the genetic variation present in a group of related organisms, proving useful when analysing large datasets or when dealing with sequences that contain gaps or ambiguous nucleotide bases. For example, in virus transmission studies, consensus sequences have been used to show the genetic diversity and evolutionary dynamics, mapping transmission patterns and evolutionary pathways (Bouquet et al., 2012; Olvera et al., 2020).

To evaluate the effectiveness of this approach, we simulate true genealogies representing multiple metastatic sites within a patient, as well as the underlying genealogy representing the individual cells sampled from these sites. Sequence data is generated based on these genealogies. We test both methods using the same sequence data: pooling individual cells to construct consensus sequences and creating a pseudobulk dataset to mimic bulkSEQ data. Subsequently, we compare the reconstructed phylogenetic trees from both methods against the true multi-tumour tree to assess the extent to which each method accurately represents the true metastatic relationships. In doing so, we can determine whether the use of consensus sequences is a feasible method when one has scSEQ data, and whether it can outperform bulkSEQ in phylogenetic reconstruction. We also explore how variations in biological and technical parameters across a diverse sample space impact the reliability of our methods.

By analysing overlapping distributions for phylogenetic accuracy scores between H-CS and VCF-PB data, we determine that the VCF-PB method consistently outperforms the H-CS method across various scenarios, underscoring its effectiveness for diverse applications. Given its representation of bulkSEQ data, researchers aiming to reconstruct multi-tumour evolution could confidently rely on the VCF-PB dataset as a reliable option for tracing multi-tumour evolution, alleviating the need for the complex and resource-intensive process of acquiring scSEQ data.

3.2 Methods

3.2.1 Pipeline

In Chapter 3 of the pipeline, a R script denoted as '3_main' controls the execution of numerous other scripts, nested within one another, and encapsulates the design of our LHS dataframe. It oversees the generation of multi-tumour/cell-cell genealogies from SimPhy, the production of tumour sequences from CellCoal, the implementation of pooling methods, phylogenetic reconstruction via CellPhy, and the computation of our tree scores.

We select for simulation software tools SimPhy and CellCoal for their ability to simulate genealogies at both the multi-tumour and cell-cell levels, along with their corresponding sequences. These tools also enable us to adjust relevant parameters during the sampling process, facilitating the coverage of an appropriate sample space and allowing us to observe the impact of parameters throughout the simulation. Utilising LHS, we construct a dataframe, systematically varying these parameters in each sample.

This dataframe is applied to SimPhy, our MSC simulator. Prior to execution of our simulation, we assign each cell a label corresponding to the metastatic site from which it was sampled, within the SimPhy configuration file. For the tips of the multi-tumour tree, we label them numerically, where each number corresponds to a specific metastatic site. Similarly, the labels assigned to the tips of the cell-cell trees follow the format '1_0_3', where the first number denotes the metastatic site (metastatic site 1), and the third number identifies the cell (cell 3). This enables us to establish the true genealogy, where cells are labelled based on the metastatic site they originate from before simulating the sequences representing these cells.

By inputting the cell-cell tree files into our designated scSEQ simulator, CellCoal, we can generate tumour sequences, whereby each cell sequenced is labelled with the metastatic site from which they were sampled. Pooling methods are then applied to the tumour sequences.

We then proceed with the phylogenetic tree reconstruction process using CellPhy. We implement the gRF distance to quantify the dissimilarity between two phylogenetic trees, allowing us to compare our reconstructed trees against the true multi-tumour tree. We conduct statistical analyses to model the impact of our parameters across the sample space. We observe the distribution of phylogenetic accuracy for our proposed sequencing methods.

3.2.2 LHS Dataframe Generation

We defined ten parameters for our simulations, along with their respective lower and upper bounds (Table 3.1).

Parameter	Lower Bound	Upper Bound
Number of Tumours	6	12
Cells Per Tumour	4	8
Effective Population Size	10000	100000
Metastatic Rate	0.0000001	0.0001
Mutation Rate	0.00000001	0.00001
Number of Sites	10000	100000
Exponential Growth Rate	0.00001	0.001
Allelic Dropout	0	0.2
Amplification Error	0.001	0.1
Sequencing Errors	0	0.05

Table 3.1: The lower and upper bounds of the parameters varied in our simulations

As these parameters will be inserted into our simulation tools, parameter mapping is expressed for SimPhy and CellCoal, defining a correspondence between parameter identifiers and their respective format strings. The format of these placeholders align with the specifications provided by the developers of each software in their configuration files.

The parameter mapping is shown here:

$$\text{SimPhy Parameters} = \begin{cases} \text{Number of Taxa} & = -sl \text{ f:\%d}, \\ \text{Individuals per Taxa} & = -si \text{ f:\%d}, \\ \text{Effective Population Size} & = -sp \text{ f:\%d}, \\ \text{Metastatic Rate} & = -sb \text{ f:\%f}, \\ \text{Mutation Rate} & = -su \text{ f:\%f} \end{cases}$$

$$\text{CellCoal Parameters} = \begin{cases} \text{Number of Cells} & = s\%d, \\ \text{Number of Sites} & = l\%d, \\ \text{Exponential Growth Rate} & = g\%.4e, \\ \text{ADO} & = D\%.f, \\ \text{Amplification Error} & = A\%.f, \\ \text{Sequencing Errors} & = E\%.f \end{cases}$$

These format strings guide the representation of the parameters when incorporated into the configuration files of our simulation tools, allowing us to substitute the values from the LHS dataframe for each iteration.

The LHS dataframe is generated using the R package ‘lhs’. The ‘randomLHS’ function is applied to design our LHS dataframe. This automates the process of dividing the range of each parameter into equally probable intervals. It then randomly samples one value from each interval for each parameter, subject to the constraint that each value of a parameter occurs only once. This process ensures that the resulting sample covers the entire range of each parameter while avoiding clustering in any particular region of the parameter space.

We specify s, which denotes the number of LHS samples to generate. We also specify m, which denotes the number of parameters being varied per sample. In our case, we set $s = 200, m = 10$. Therefore, the LHS dataframe is structured as a matrix with dimensions $s \times m$. In this matrix, each column corresponds to a parameter, while each row signifies a unique sample.

We define a function to assess whether the resulting matrix satisfies the properties of LHS. Given a matrix X with s rows and m columns, denoted as $X = (x_{im})_{s\times m}$, for each column m of X: the function checks the following conditions:

- Ensure that the minimum value of the column x_{ij} is greater than 0, i.e., $\min(X_{:,j}) > 0$
- Ensure that the maximum value of the column x_{ij} is less than 1, i.e., $\max(X_{:,j}) < 1$
- Ensure that there are no missing values in the column x_{ij}

Furthermore, for each column j of X, it verifies that each unique value occurs exactly once. A scaling function is then applied to each sample, multiplying each generated parameter value by its corresponding range and then adding the minimum value of the range. Since LHS generates samples from a normalised space ranging from 0 to 1, scaling ensures that the parameter values align with their specified ranges.

3.2.3 True Multi-tumour/Cell-cell Tree Simulation

To utilise SimPhy for generating tree files, we iterate through a loop over each row of our LHS dataframe, extracting the parameter values for each row. Each iteration of our loop substitutes these parameter values into a designated SimPhy configuration file using the `gsub` function, adjusting the parameters accordingly.

Note that within the SimPhy configuration file, we have specified a value of 8 for the argument `RL`, which denotes the number of locus trees per species tree. As mentioned previously, the locus tree represents the genealogies of genomic loci within a species. In this case, the 8 locus tree generated per sample by the argument `RL` can be viewed as cell-cell tree replicates. We specifically want to build our replicates at this particular point of our simulation process rather than during the initial generation of our LHS dataframe as this ensures that each replicate cell-cell tree of a sample is distinct, while corresponding to the same multi-tumour tree, portraying the genealogies of various genomic regions of a tumour cell.

To uniquely identify the output folder of each iteration, we use the `paste` function to create a label for our files, combining the string `_pop` with the effective population size value, followed by `_sites` and the number of sites value, unique to each iteration. This label serves as an identifier throughout our project to distinguish each iterative result.

3.2.4 Tumour Sequence Simulation

The script transitions to the execution of CellCoal, as it generates scSEQ sequences based on the provided cell-cell tree files from SimPhy. This is achieved by substituting the corresponding tree file as a USERTREE in the CellCoal simulation. For this portion of 3_main, an external SLURM script is executed, which in turn executes a job array, with each job aligned to an iteration from the dataframe. Each job executes an external R script dedicated to the CellCoal simulation. This script can be referred to as `cellcoal.R`. This SLURM script is required so as to accommodate for substantial memory requirements of the CellCoal simulation. Without an allocation of approximately 150GB per job, CellCoal risks premature termination during the process of generating the VCF file output. To our benefit, the simultaneous execution of a collection of jobs divides the time required to complete this section of our workflow.

We leverage the SLURM array task ID to access specific rows of data from our dataframe. When a job is submitted, SLURM assigns it a unique ID, which is then passed as an argument to `cellcoal.R`. This R script converts the task ID into an array index, enabling us to retrieve the corresponding parameter values from our dataset. It begins by reading the LHS dataframe, identifying available tasks, and then extracting parameter values for each iteration. These values are updated within a configuration file using the `gsub` function. Additionally, the script locates the cell-cell tree file associated with each iteration in the SimPhy directory, incorporating it into the USERTREE argument.

Recall that CellCoal automatically includes outgroup sequence data in any simulated dataset, effectively introducing an additional population that can represent a primary tumour/healthy cells from which all other populations are derived from. To omit this, `cellcoal.R` locates the haplotype file within the output folder of each iteration. It identifies the outgroup cell within the file and removes all lines of data following it. Additionally, the script updates the first line of the file to reflect the correct number of metastatic sites.

3.2.5 Haplotype Consensus Sequence Generation

`cellcoal.R` concludes by constructing the haplotype consensus sequences. It does so by initiating an external Python script designed to generate our H-CS datasets in FASTA format from the resulting haplotype file. This Python script pools sequences into alignments based on the metastatic site to which each cell within the haplotype was sampled from. It achieves this by extracting the first number from each cell label. It then analyses the occurrence of nucleotide bases within these pooled alignments, taking into account the ambiguous bases represented by the IUPAC code. This includes the four nucleotide bases comprising DNA, as well as codes indicating situations where more than one nucleotide could be present at a specific position in the sequence alignment.

The nucleotide bases 'A', 'T', 'G', and 'C' constitute the four fundamental components of DNA:

A :	Adenine	T :	Thymine	G :	Guanine	C :	Cytosine

The ambiguous bases represent possible combinations of these four nucleotides:

M:	$\{A,\ C\}$	*B:*	$\{C,\ G,\ T\}$
W:	$\{A,\ T\}$	*D:*	$\{A,\ G,\ T\}$

R:	$\{A,\ G\}$	*V:*	$\{A,\ C,\ G\}$
Y:	$\{C,\ T\}$	*H:*	$\{A,\ C,\ T\}$
S:	$\{G,\ C\}$	*N:*	$\{A,\ T,\ G,\ C\}$
K:	$\{G,\ T\}$		

Note that in instances when the script encounters a situation where it would typically designate a column with base codes ‘B’, ‘D’, ‘V’, or ‘H’, we instead implement ‘N’ to signify ambiguity within the sequence. This approach is adopted because, by design, the reconstruction tool CellPhy does not interpret ambiguity bases ‘B’, ‘D’, ‘V’ or ‘H’. Hence, we adopt this substitution to ensure that consensus sequences can still be generated while maintaining compatibility with CellPhy.

$$N :\{C, G, T\}$$

$$N :\{A, G, T\}$$

$$N :\{A, C, G\}$$

$$N :\{A, C, T\}$$

Subsequently, the script applies a specified cutoff threshold to determine which bases are considered for consensus determination.

```
def consen(alig, ids):
    s_bases = (’A’, ’C’, ’G’, ’T’)
    d_bases = (’R’, ’Y’, ’S’, ’W’, ’K’, ’M’, ’N’)
    bases = s_bases + d_bases
    con_seq = ""
    cutoff = 0.35
    for i in range(alig.get_alignment_length()):
        col = alig[:, i]
        col = col.upper()
        base_count = {b: col.count(b) for b in bases}
        genotypes = sorted(base_count.items(), key=lambda x: -x[1])
        # genotypes = [b for b in genotypes if b[1] > 0]
        genotypes = [b for b in genotypes if b[1] >= len(col) * cutoff]
        # genotypes = [e for e in genotypes if not (genotypes[0][1] !=  e[1]
        and e[1] <= 1)]
        if len(genotypes) == 0:
            con_seq += ’N’
        else:
            if len(genotypes) <= 1:
                con_seq += genotypes[0][0]
            else:
                amb_base = [IUPAC[c[0]] for c in genotypes]
                amb_base_list = []
                for d in amb_base:
                    amb_base_list.extend(list(d))
                amb_base_set = sorted(set(amb_base_list))
                con_seq += consensus_iupac[tuple(amb_base_set)]

    con_rec = SeqRecord(Seq(con_seq), id=ids, description=’’)
    return con_rec
```

This threshold imposes three conditions:

- If all observed bases have frequencies below the cutoff threshold, then the consensus base is assigned as ‘N’.
- If there is only one observed base with a frequency at or above the cutoff threshold, then that base becomes the consensus.
- If there are multiple observed bases with frequencies at or above the cutoff threshold, then the consensus is determined based on utilising ambiguity bases.

The cutoff threshold value is user-adjustable, and therefore imposes flexibility for consensus determination.

3.2.6 VCF Pseudobulk Data Generation

VCF-PB files undergo external processing via a Python script. Using VCFtools, a command-line program for processing VCF files, the individual labelled as ‘outgcell’ in each VCF file – automatically inserted by CellCoal – is removed from a modified VCF file using the `--remove-indv` argument and entering ‘outgcell’. The `--recode` option instructs VCFtools to create a new VCF file with the filtered data. The `--recode-INFO-all` option instructs `vcftools` to include all INFO fields in the output VCF file. The INFO fields in VCF files contain additional information associated with each variant call. Among the data included in these fields, the PL (Phred-scaled likelihood) score holds particular significance for our analysis as we leverage the PL score to generate our pseudobulk data.

Our Python script generates VCF-PB data by grouping cells based on the metastatic site from which they were sampled. This script operates similarly to our construction of H-CS data and identifies these groups based on the label of each cell, specifically by extracting the first number of the label. We utilise the PL scores extracted from the INFO fields of our VCF, which represent the likelihoods of different genotypes at a given variant site. These scores are logarithmically transformed probabilities, which means they provide a more compressed and standardised way to express the likelihood of an error in base calling. Each PL score in a VCF file corresponds to a specific genotype, and higher PL scores indicate higher confidence in the genotype call.

The need to pool our cells by metastatic site enables the application of a weighted PL formula to compute normalised genotype likelihoods for each genotype across all cells within the same group. This formula accounts for the likelihoods of different genotypes across all cells and computes a weighted normalisation, considering each cell’s contribution to the overall likelihood. These normalised PL values represent genotype likelihoods within each metastatic site, facilitating the identification of aggregated genotypes and enabling the construction of a VCF-PB dataset.

The formula for the weighted PL code:

```
def sum_lists(*args):
    return list(map(sum, zip(*args)))
```

```
def weighted_PL(PL_list):
    if len(PL_list) ==1:
        return PL_list[0]
    sublist_sums = [sum(xx for xx in sublist if xx!= -math.inf and xx
    is not None and xx!=-2147483648) for sublist in PL_list]
    total_sum = sum(1 / total for total in sublist_sums if total != 0)
    indiv_res = [[(1 / total) * value if total != 0 else 0 for value in sublist]
    for sublist, total in zip(PL_list, sublist_sums)]
    norm_res = [round(x/total_sum,2) for x in sum_lists(*indiv_res)]
    return norm_res
```

can be expressed as:

$$PL_{\text{weighted}}(\text{genotype}) = \sum_{i=1}^{N} \left(\frac{PL(\text{total})_i \times PL(\text{genotype})_i}{\sum_{j=1}^{N} PL(\text{total})_j} \right)$$

- $PL(\text{total})_i$: This refers to the total PL (Phred-scaled likelihood) score for the ith metastatic site. It represents the cumulative likelihood score considering all variants at the site across all cells within that metastatic site.

- $PL(\text{genotype})_i$: This denotes the PL score specifically assigned to a particular genotype at the ith metastatic site. It reflects the likelihood of observing the given genotype.

- $\sum_{i=1}^{N}$: This represents the summation operation performed over all N metastatic sites. The index i, is used to iterate through each individual metastatic site

- $\sum_{j=1}^{N} PL(\text{total})_j$: This expression represents the summation of total PL scores for all metastatic sites. It calculates the combined likelihood considering all variants across all metastatic sites.

This formula calculates the weighted normalisation of PL values for a specific genotype across all cells in the metastatic site, taking into account the contribution of each cell to the overall likelihood. The resulting weighted normalised PL values represents the genotype likelihoods within each metastatic site.

3.2.7 Calculating Tree Reconstruction Accuracy

The tree reconstruction workflow comprises three distinct scripts that execute sequentially, with each script invoking the next in the sequence. The first R script specifies CPU usage requirements for each iteration. The second is a SLURM script that executes the tree reconstruction software using CellPhy. Finally, the third script calculates the tree distance, measuring dissimilarity between the initial true tree and our reconstructed tree.

Descriptions of each script are listed in order below:

1. `pre-cellphy_cpu.R` - The initial R script serves as a preprocessing step in our workflow. The necessity of this R script arises from its capability to manually adjust the CPU allocation for the set of job submissions it initialises. We have constructed adjustable parameters to enable dynamic CPU usage based on extracted values from our data. Consequently, we have the capability to allocate either additional or fewer CPU resources as necessary to expedite processing for iterations that may vary in duration. The script iterates through the output folders of our CellCoal directory, performing a calculation of the total number of cells multiplied by the number of sites per cell for each iteration. It does so by extracting the number of cells from the first line of each haplotype file, which represents the total count of cells. Then, it parses the folder name to extract the number of sites, specified in the folder name format.

   ```
   result_variable <- num_sites * num_from_full_hap
   if (result_variable > 20000000) {
       cpus_per_task <- 20
       mem <- "20G"
   } else if (result_variable > 2500000) {
       cpus_per_task <- 12
       mem <- "10G"
   } else {
       cpus_per_task <- 2
   ```

```
    mem <- "4G"
}
```

Generally with RAxML-NG, the program of which CellPhy is an extension of, computational time increases as the number of patterns to analyse rises. Factors influencing pattern count include sequence diversity, the number of sequences, and alignment length. Here, we assume that the `result_variable` metric may generally reflect the computational intensity of each iteration, as it represents the total number of sites within the haplotype file subjected to analysis, and therefore higher values of this metric would require greater computational demands. Our approach prioritises the number of sequences and alignment length to ensure a practical run-time for our simulation, as these factors are rather straightforward to extract. We deemed it necessary to develop this pre-processing script after observing significant reduction in time to completion from increased CPU usage, up to a certain threshold. However, beyond this threshold, further increases in CPU usage paradoxically led to an increase in completion time rather than a decrease. This is likely attributed to factors such as resource contention, and overhead associated with excessive parallelization.

2. `cellphy_single.sh` - The intermediate SLURM script is primarily used to initiate the execution of CellPhy. The string of code for CellPhy execution specifies various parameters, including the substitution model that varies depending on the data type – `GT16+FO+E` is used for our haplotype files (such as our scSEQ or H-CS data) and `GT16+FO` is used for our VCF-PB data.

3. `cellphy.R` - Upon completing phylogenetic reconstruction, a final R script is executed by the preceding SLURM script, tasked with calculating the phylogenetic accuracy and updating the dataframe with tree scores for each iteration. This involves computing the gRF distance, wherein the script identifies the reconstructed tree from our results and pairs it with the corresponding true tree generated by SimPhy. To compute the tree distance, our script utilises the `TreeDistance` function from the R package ‘TreeDist’.

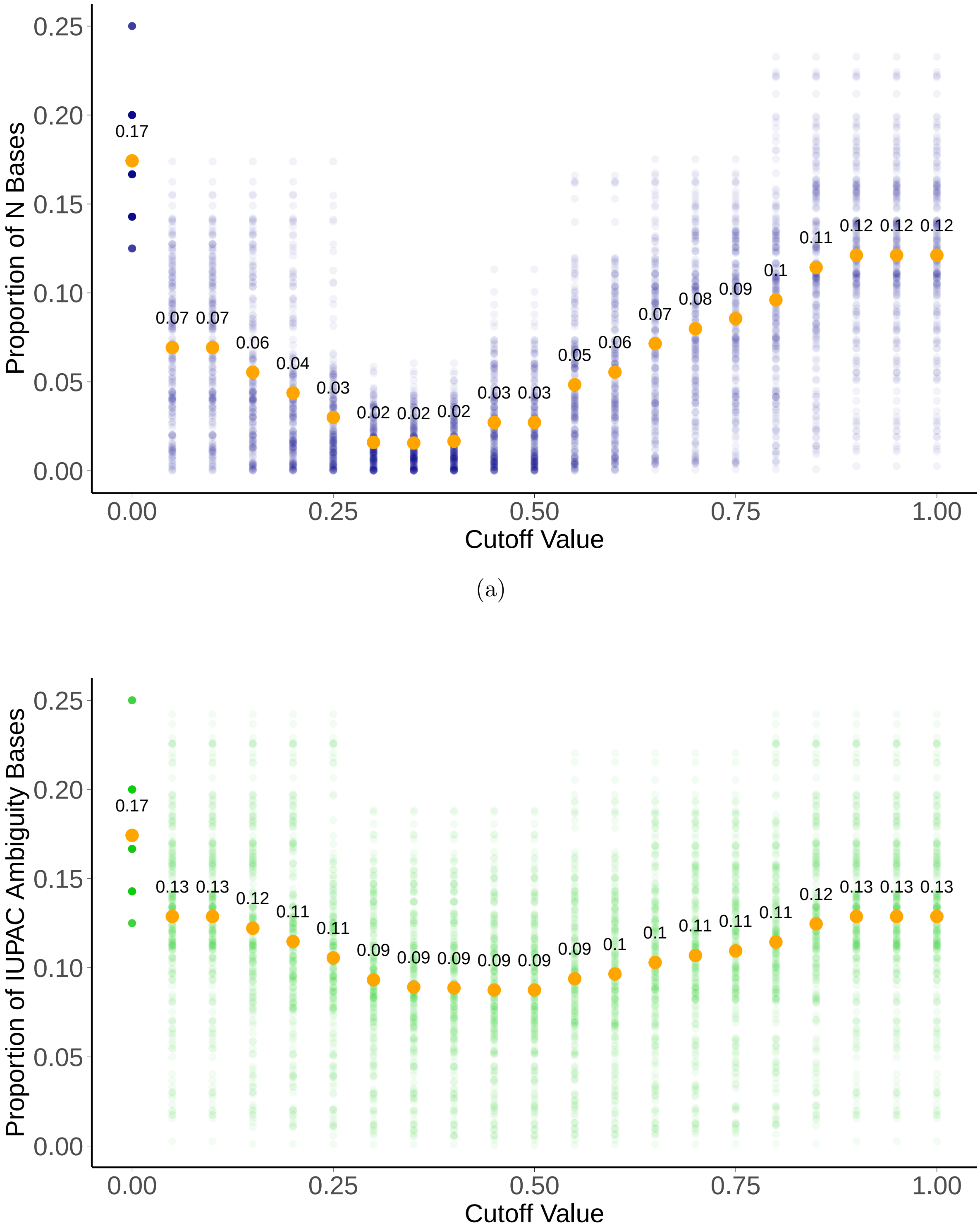
3.3 Results

3.3.1 Constructing Consensus Sequences

When constructing our consensus sequences, the cutoff threshold value is a parameter that can be adjusted to optimise sequence quality. In order to establish an optimal cutoff threshold value, we explore a range of cutoff threshold values from 0 to 1 at intervals of 0.05, measuring the values that result in consensus sequences with the lowest total number of ‘N’ bases. This choice is motivated by the need to mitigate uncertainty and increase the reliability of our consensus sequences. The strategy of reducing the count of ‘N’ bases to manage uncertainty is particularly crucial. This is because the four IUPAC ambiguity bases ‘B’, ‘D’, ‘V’, and ‘H’, typically indicating situations where three possible nucleotides may occur at a particular site, have also been programmed to be interpreted as ‘N’ bases.

While minimising the count of ‘N’ provides a straightforward method for identifying positions with unresolved ambiguity, we also consider the importance of assessing whether the threshold value that minimises ‘N’ count aligns with a reduction in overall IUPAC ambiguity bases. Therefore, we conduct a similar test to determine the cutoff threshold value that most effectively reduces the overall ambiguity count, as defined by a decrease in the proportion of all IUPAC ambiguity bases (‘M’, ‘W’, ‘R’, ‘Y’, ‘S’, ‘K’, and ‘N’).

As shown in Fig 3.1, our analyses involves a simple comparison against the proportion of bases, either ‘N’ or total IUPAC ambiguity bases, relative to the total bases within a consensus sequence, across varying cutoff threshold values. For each sample, we determine the frequency of ambiguity bases and total bases. Then, we calculate the ratio of ambiguity bases to total bases within each sample. Our analyses reveals that lowest mean proportions occurred within a cutoff value range of 0.3 to 0.4 for ‘N’ bases and a range of 0.3 to 0.55 for total ambiguity bases. We observe that the proportion remains relatively linear within these ranges. Consequently, for the construction of our H-CS dataset, we select a cutoff value that exhibits the lowest proportion for both ‘N’ bases and total ambiguity bases, which is determined to be 0.35.



(a)

(b)

Figure 3.1: Comparison of optimal cutoff threshold values for the lowest proportion of ‘N’ bases against the optimal cutoff threshold values for the lowest proportion of total IUPAC ambiguity bases.

3.3.2 Effect of Parameters on Phylogenetic Accuracy

Throughout this project, Generalised Linear Model (GLM) analyses are conducted to explore the relationship between the phylogenetic accuracy of the proposed sequencing methods. In this chapter, the methods of scSEQ, H-CS, and VCF-PB, are assessed against the ten parameters varied throughout our simulations. Note that the estimates of each GLM coefficient are not normalised, meaning they lack a consistent baseline for comparison. Therefore, the absolute values of the estimates do not represent any meaningful quantity but rather indicate whether a parameter has a positive or negative influence on phylogenetic accuracy. The significance level of each estimate indicates the strength of evidence for its effect on the outcome variable. We assess the significance and direction of each parameter's impact on phylogenetic accuracy while considering potential confounding factors.

We perform three simulations, with differing metastatic rate ranges, as shown in Table 3.2. This is because, in the analysis of our initial GLM analysis, shown in Table 3.3, we observe unexpected significance concerning the influence of metastatic rate and allelic dropout (ADO) on the phylogenetic accuracy of our scSEQ data. To investigate further, we generate scatter plots (Fig 3.2 and Fig 3.3) between these parameters and phylogenetic accuracy, which reveal an anomalous pattern associated with metastatic rates below a certain threshold. To confirm these anomalies, we conduct two additional simulations, exploring metastatic rates within two extra ranges while keeping the range of other parameter values fixed. In Fig 3.2, we confirm from our MR-Low simulation that low metastatic rates induced anomalous accuracy scores that are dispersed, diverging away from a predominant cluster. This is then rectified by maintaining the metastatic rate above this threshold.

Simulation	Lower Bound	Upper Bound
MR-Full	0.0000001	0.0001
MR-Low	0.0000001	0.000009
MR-High	0.00001	0.0001

Table 3.2: The lower and upper bounds for metastatic rate of our three simulations - MR-Full, MR-Low, and MR-High - each simulation consists of a distinct range of metastatic rates in which the LHS can operate.

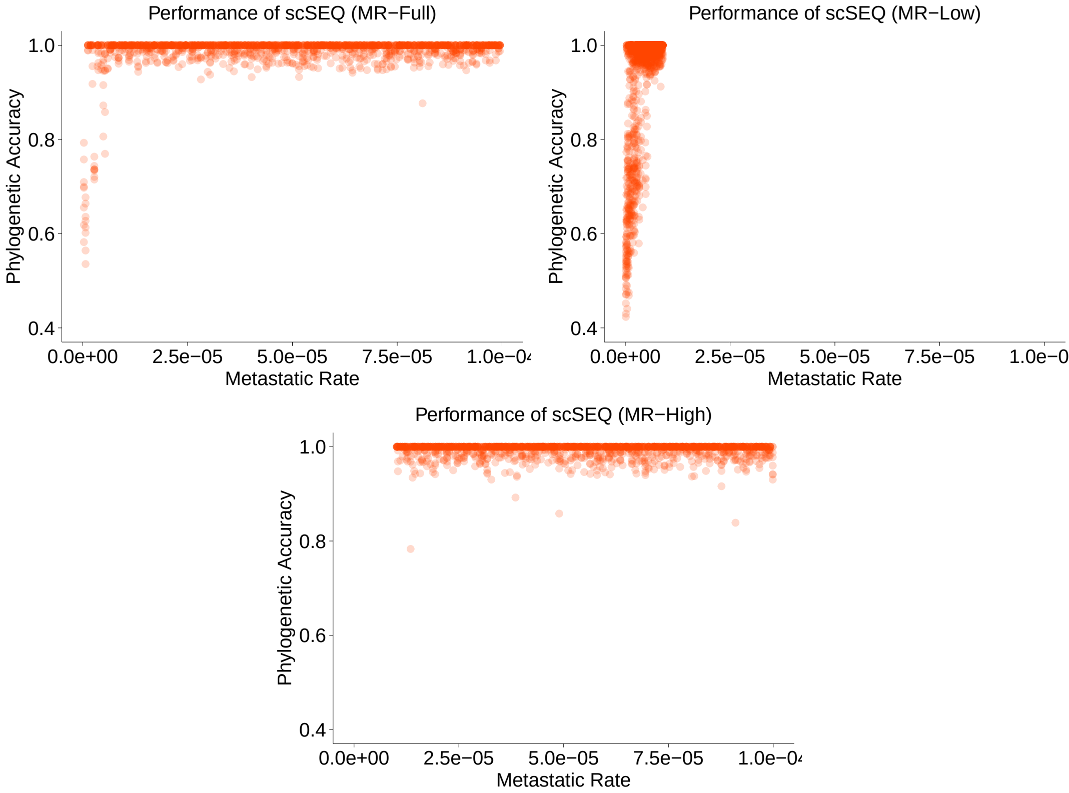


Figure 3.2: Plots between metastatic rate and phylogenetic accuracy of scSEQ at three different metastatic rate ranges.

In Table 3.3, we present the GLM analysis results for our initial simulation, MR-Full, with statistical summaries for each of our three methods: scSEQ, H-CS, and VCF-PB. Firstly, the intercept value demonstrates high significance ($p < 0.001$) for all methods, indicating that even in the absence of specific parameter variations, there exists a notable level of accuracy in phylogenetic reconstruction. This also suggests that, as expected, there are other factors excluded in our model that contributes to the phylogenetic accuracy of our methods.

When analysing phylogenetic accuracy of scSEQ within Chapter 3, we are specifically analysing each iteration's capability to accurately reconstruct the true cell-cell tree using scSEQ (leveraging the tree file obtained from SimPhy to build tumour sequences from CellCoal, directly reconstructing the haplotype sequences through CellPhy). Note that our assessment doesn't extend to the method's proficiency in reconstructing multi-tumour evolution.

In our GLM analysis of the scSEQ method, we observe several factors that significantly influence

Method	Parameter	Estimate	Std. Error	t value	p value	
scSEQ	Intercept	0.90687	0.00855	106.01789	0.00000	***
	Number of Tumours	0.00272	0.00057	4.73243	0.00000	***
	Exponential Growth Rate	7.80537	3.59364	2.17200	0.03000	*
	Cells per Tumour	0.00432	0.00083	5.23120	0.00000	***
	Effective Population Size	0.00000	0.00000	0.85715	0.39149	
	Metastatic Rate	278.86990	35.10006	7.94500	0.00000	***
	Mutation Rate	-298.22030	349.79750	-0.85255	0.39404	
	Number of Sites	0.00000	0.00000	5.76345	0.00000	***
	Allelic Dropout	0.06893	0.01770	3.89402	0.00010	***
	Amplification Error	0.04724	0.03655	1.29258	0.19634	
	Sequencing Errors	-0.37607	0.06913	-5.43982	0.00000	***
H-CS	Intercept	0.79045	0.03901	20.26416	0.00000	***
	Number of Tumours	-0.00202	0.00262	-0.76995	0.44144	
	Exponential Growth Rate	38.62828	16.38752	2.35718	0.01854	*
	Cells per Tumour	0.01080	0.00377	2.86474	0.00423	**
	Effective Population Size	-0.00000	0.00000	-12.80550	0.00000	***
	Metastatic Rate	-3381.33200	160.06160	-21.12519	0.00000	***
	Mutation Rate	4012.08800	1595.13000	2.51521	0.01199	*
	Number of Sites	0.00000	0.00000	4.26493	0.00002	***
	Allelic Dropout	-0.05129	0.08072	-0.63545	0.52522	
	Amplification Error	-0.19027	0.16668	-1.14157	0.25380	
	Sequencing Errors	0.33991	0.31526	1.07819	0.28111	
VCF-PB	Intercept	0.77907	0.03826	20.36455	0.00000	***
	Number of Tumours	0.00469	0.00257	1.82408	0.06833	
	Exponential Growth Rate	16.08758	16.07199	1.00097	0.31699	
	Cells per Tumour	0.01740	0.00370	4.70723	0.00000	***
	Effective Population Size	-0.00000	0.00000	-13.68756	0.00000	***
	Metastatic Rate	-2457.47600	156.97970	-15.65474	0.00000	***
	Mutation Rate	-3279.35200	1564.41600	-2.09621	0.03622	*
	Number of Sites	0.00000	0.00000	3.39077	0.00071	***
	Allelic Dropout	0.07499	0.07916	0.94730	0.34363	
	Amplification Error	-0.37552	0.16347	-2.29723	0.02174	*
	Sequencing Errors	0.33508	0.30919	1.08376	0.27863	

Table 3.3: For simulation MR-Full, a GLM analysis modelling the relationship between variation in biological/technical parameters and phylogenetic accuracy of reconstructed trees.

phylogenetic accuracy. Firstly, an increase in the number of tumours results in a significant improvement in accuracy ($p < 0.001$). The exponential growth rate shows marginal significance ($p = 0.03$), improving reconstruction accuracy. A higher number of cells per tumour sample also significantly contributes to enhanced accuracy ($p < 0.001$). Similarly, an increase in the number of sequencing sites significantly improves accuracy ($p < 0.001$). Reducing sequencing errors also shows to be a significant factor for improve accuracy ($p < 0.001$). These results are generally in line with expectations, as they all contribute to a general increase in genetic diversity and information, allowing for more robust phylogenetic reconstructions. However, for scSEQ in particular, it is also important to consider that they can also increase the complexity of phylogenetic analysis.

We do observe a perplexing finding: a significant correlation between a higher metastatic rate and an improvement in phylogenetic accuracy ($p < 0.001$). This seems counter-intuitive, as higher metastatic rates typically lead to shorter branch lengths in the evolutionary tree. This results in less time for mutations to accumulate and for cell lineages to diverge within metastatic sites. Moreover, increased migration to other sites due to higher metastatic rates leads to greater intermixing of cell lineages. This intermixing complicates the reconstruction of phylogenetic relationships. In general, one would expect a correlation between a lower metastatic rate and an improvement in phylogenetic accuracy, given the assumption that longer branch lengths allow for more informative genetic divergence.

Furthermore, a significant correlation is observed between a higher rate of allelic dropout and an improvement in phylogenetic accuracy ($p < 0.001$). This too contradicts conventional assumptions, as allelic dropout introduces ambiguity by causing alleles to disappear from the sequence data, which should obscure the underlying genetic information, leading to a decrease in phylogenetic accuracy. A plausible hypothesis may be that ADO may reduce the amount of ambiguity by effectively reducing the number of potential heterozygous mutations at specific sites. In other words, when ADO occurs, it may eliminate one of the alleles at a particular genomic position, thereby simplifying the interpretation of the sequence data by reducing the number of possible genetic variations that need to be considered. This reduction in ambiguity could potentially lead to improved accuracy in tree reconstruction, as there is less uncertainty in determining the nucleotide state at each position.

We speculate that the observed significance of ADO in improving reconstruction accuracy may actually be confounded by the low metastatic rate values in certain samples. This suggests that samples demonstrating poor phylogenetic accuracy due to a low metastatic rate may also have been assigned lower ADO values. Consequently, due in part to a small sample size, this alignment of low metastatic rate and low ADO could potentially skew the GLM analysis, displaying an anomalous association between increased ADO and improved accuracy.

Unlike scSEQ, H-CS, and VCF-PB methods are associated with phylogenetic reconstructions on a multi-tumour scale, having pooled cells based on the metastatic site from which they were sampled. For the H-CS analysis, we observe similar trends to scSEQ; the exponential growth rate shows significance ($p = 0.018$) on improving accuracy, as does the number of cells sampled per tumour, exhibiting a positive association ($p < 0.05$). The mutation rate ($p = 0.012$) and the number of sites ($p < 0.001$) also shows significance, contributing to improved phylogenetic accuracy.

In the VCF-PB analysis, we observe a highly significant relationship between the number of cells per tumour and an increase in phylogenetic accuracy ($p < 0.001$). The number of sites is also statistically significant ($p < 0.001$), the estimate indicating its positive impact on phylogenetic accuracy. Conversely, we find a significant negative relationship between amplification error and accuracy ($p = 0.02174$).

Despite it being marginal, there exists a statistically significant relationship between a decrease in mutation rate and an improvement in phylogenetic accuracy in the VCF-PB analyses. This may stem from a reduction in heterozygous bases, which tend to arise more frequently as mutation rates increase, complicating the interpretation of sequencing data. It is therefore essential to take into consideration the differing approaches of our two methods in handling ambiguity bases. In H-CS data, we have implemented a specific cutoff threshold value aimed at minimising the proportion of ambiguity bases. Contrastingly, for VCF-PB data, ambiguity is accounted for through the weighted PL formula, but there is no explicit control over its spectrum. This difference in methodology could explain the slight difference in effect of mutation rate on accuracy observed between the two methods.

Both H-CS and VCF-PB, exhibit a highly significant negative association between the effective population size and phylogenetic accuracy ($p < 0.001$). This can be expected, as a smaller effective population size suggests a quicker coalescent time within a tumour. Quicker coalescence means that cell lineages merge more rapidly, reducing the opportunity for interleaving to occur among subclonal cells. As a result, smaller effective population sizes tend to lead to more accurate phylogenetic reconstructions.

In both the GLM analysis of H-CS and VCF-PB data, there is a highly significant, negative association between metastatic rate and phylogenetic accuracy ($p < 0.001$). If we consider the assumption that the lower bound of our initial range of metastatic rates acts as a confounding factor in our analysis of scSEQ, it implies that the effect of a low metastatic rate within the range of the MR-Full simulation does not manifest in the GLM for H-CS and VCF-PB as it does for scSEQ. We posit that this inconsistency may stem from the scoring mechanism employed in H-CS and VCF-PB.

Recognising the anomalies revealed by our GLM analysis, we generated a scatter plot between metastatic rate and ADO against phylogenetic accuracy. In Figure 3.2, labelled as MR-Full, we examine the relationship between metastatic rate and phylogenetic accuracy observed in our initial simulation. Toward the lower boundary of the metastatic rate range, a discernible tail of outliers emerge, exhibiting scores that reach below 0.6. Upon examination, we determined that these data points correspond to the replicates of 5 samples consistently exhibiting lower scores, diverging from the main cluster. These samples are listed as follows:

```
_pop33305_sites19910
_pop40092_sites55025
_pop68658_sites21093
_pop83673_sites63544
_pop99248_sites42162
```

Among these samples, the highest metastatic rate value identified is 0.000005. We therefore conduct two additional simulations:

- MR-Low - This simulation aims to capture scenarios with lower metastatic rates. To ensure that it captures the observed rate of 0.000005 and below, the upper bound is set at 0.000009.

- MR-High - This simulation aims to capture scenarios with higher metastatic rates. To avoid overlap with the MR-Low simulation the lower bound is set at 0.00001.

In Fig 3.2, simulation MR-Low indicates a more pronounced dispersion of scores with a tail of data points reaching below 0.45. Despite this the majority of scores still clustered around a score of 0.9 and 1. In contrast, MR-High, lacks an obvious dispersion of scores. Instead, scores predominantly clustered between 0.9-1, with only one data point falling below 0.8. Here, we do not observe an obvious discernible effect of metastatic rate on phylogenetic accuracy.

In Fig 3.3, we generate scatter plots depicting the relationship between ADO and phylogenetic accuracy across all three simulations.

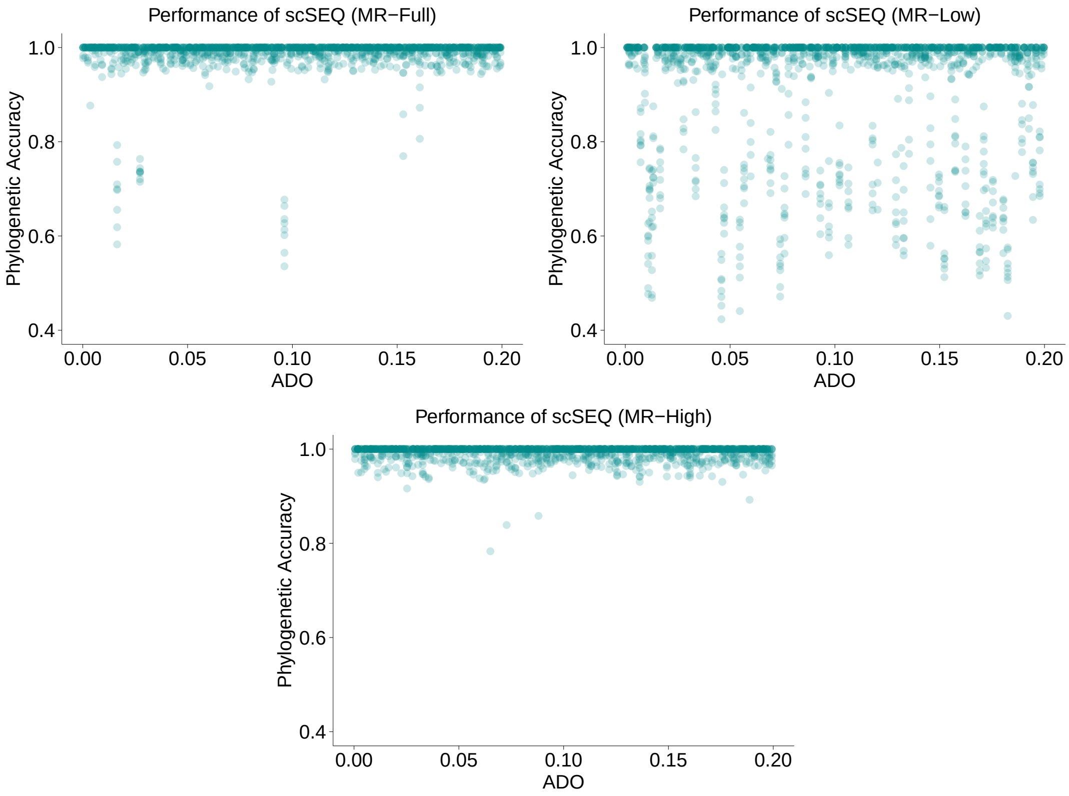


Figure 3.3: Plot between allelic dropout and phylogenetic accuracy of scSEQ at three different metastatic rate ranges.

In MR-Full, we observe that 2 sets of sample replicates exhibiting phylogenetic accuracy scores below 0.6, have values of ADO between 0 and 0.05. Another set of sample replicates assigned an ADO value between 0.09 and 0.1, have a set of scores that reach below 0.6. In MR-Low, the dispersion of scores from 0.4 to 1 appear relatively even. In MR-High, bar outliers that have an accuracy score above 0.75, all scores cluster near 0.9-1, and there is no apparent tail that might skew the GLM analysis.

We analyse the GLM analysis of simulation MR-Low, outlined in Table 3.4. In contrast to the results observed in MR-Full, for scSEQ, the coefficient for ADO demonstrates a negative influence in MR-Low. This observation suggests that in MR-Low, characterised by a more dispersed distribution of phylogenetic accuracy scores, there are no outliers assigned with low ADO values to skew the GLM analysis. This confirms our speculation, that the influence of ADO observed in MR-Full may have been confounded by low metastatic rate values.

For scSEQ, H-CS and VCF-PB, both the number of tumours ($p = 0.01365$, $p < 0.001$, $p < 0.001$) and cells per tumour ($p < 0.001$, $p < 0.001$, $p < 0.001$) demonstrate significant positive influences on accuracy. Exponential growth rate ($p < 0.001$, $p = 0.00101$, $p < 0.001$) exhibits significant influence, associated with enhancing phylogenetic accuracy. Furthermore, all three methods demonstrate a significant positive correlation between metastatic rate and phylogenetic accuracy ($p < 0.001$), while displaying a significant negative correlation with mutation rate and phylogenetic accuracy. When simulating a specified number of metastatic sites under a low metastatic rate, longer branches result, as metastases occur relatively infrequently. In MR-Low, we've constrained the metastatic rate to be extremely low. One would expect there to be little to no intermixing of subclonal cells among metastatic sites. Increasing the metastatic rate should lead to shorter branch lengths, potentially exacerbating the issue of heterogeneity. Finally, mutation rate appears to worsen accuracy, but mutation rate is generally expected to increase genetic information, theoretically enhancing reconstruction accuracy.

For both H-CS and VCF-PB, a consistent negative association with effective population size persists from the MR-Full to MR-Low simulations ($p < 0.001$, $p < 0.001$). However, scSEQ diverges from this trend, suggesting a significant positive influence of effective population size on phylogenetic accuracy. Increasing the effective population size typically extends coalescence time, which would therefore reduce the chance of incomplete lineage sorting. This anomalous effect may be masked in H-CS and VCF-PB as the sequences are pooled.

We acknowledge that the anomalous effects observed in this GLM analysis is likely due to the extremely low metastatic range we have restricted values to. This confounds the interpretation of

Method	Parameter	Estimate	Std. Error	t value	p value	
scSEQ	Intercept	0.73656	0.02367	31.11821	0.00000	***
	Number of Tumours	0.00341	0.00138	2.46915	0.01365	*
	Exponential Growth Rate	53.53794	8.52768	6.27814	0.00000	***
	Cells per Tumour	0.01055	0.00201	5.25155	0.00000	***
	Effective Population Size	0.00000	0.00000	4.49652	0.00001	***
	Metastatic Rate	29829.97133	932.14652	32.00138	0.00000	***
	Mutation Rate	-15681.23414	833.40242	-18.81592	0.00000	***
	Number of Sites	0.00000	0.00000	0.07619	0.93927	
	Allelic Dropout	-0.10314	0.04163	-2.47759	0.01333	*
	Amplification Error	0.08071	0.08337	0.96817	0.33310	
	Sequencing Errors	0.02864	0.16462	0.17398	0.86191	
H-CS	Intercept	0.62869	0.04530	13.87982	0.00000	***
	Number of Tumours	0.00951	0.00264	3.59749	0.00033	**
	Exponential Growth Rate	53.75077	16.31875	3.29381	0.00101	**
	Cells per Tumour	0.01756	0.00384	4.56664	0.00001	***
	Effective Population Size	-0.00000	0.00000	-3.45008	0.00058	***
	Metastatic Rate	21692.41601	1783.77607	12.16095	0.00000	***
	Mutation Rate	-19477.84129	1594.81719	-12.21321	0.00000	***
	Number of Sites	0.00000	0.00000	0.16172	0.87155	
	Allelic Dropout	-0.05523	0.07966	-0.69330	0.48822	
	Amplification Error	-0.07340	0.15953	-0.46011	0.64550	
	Sequencing Errors	-0.08750	0.31502	-0.27775	0.78124	
VCF-PB	Intercept	0.64503	0.04549	14.17953	0.00000	***
	Number of Tumours	0.00857	0.00266	3.22587	0.00128	**
	Exponential Growth Rate	64.60603	16.38898	3.94204	0.00008	***
	Cells per Tumour	0.01476	0.00386	3.82351	0.00014	***
	Effective Population Size	-0.00000	0.00000	-2.48942	0.01290	*
	Metastatic Rate	27376.55086	1791.45357	15.28175	0.00000	***
	Mutation Rate	-22671.67340	1601.68139	-14.15492	0.00000	***
	Number of Sites	-0.00000	0.00000	-0.54818	0.58365	
	Allelic Dropout	-0.06172	0.08000	-0.77143	0.44057	
	Amplification Error	-0.14843	0.16022	-0.92643	0.35436	
	Sequencing Errors	-0.01263	0.31638	-0.03992	0.96816	

Table 3.4: For simulation MR-Low, a GLM analysis modelling the relationship between variation in biological/technical parameters and phylogenetic accuracy of reconstructed trees.

effective population size, metastatic rate and mutation rate, and potentially skews the influence of additional parameters in unexpected directions.

In our final simulation, MR-High, the metastatic rate has been intentionally adjusted to a higher range, aiming to mitigate its confounding effect on the analysis. With this adjustment, we anticipate observing more expected results throughout the GLM analysis (Table 3.5) overall.

For scSEQ, two parameters show statistically significant effects on phylogenetic accuracy based on their p-values. The effective population size exhibits a negative association (p = 0.03773), implying that an increase in effective population size correlates with a decrease in phylogenetic accuracy. The number of sites demonstrates a positive association ($p < 0.001$), indicating that higher accuracy is associated with a greater number of sites.

For H-CS, several parameters exhibit statistically significant effects on phylogenetic accuracy based on their p-values. The number of simulated tumours demonstrates a negative association ($p < 0.001$), suggesting that an increase in the number of tumours correlates with a decrease in phylogenetic accuracy. Similarly, the effective population size displays a negative association ($p <$

0.001). The metastatic rate coefficient also shows a negative association ($p < 0.001$), implying that accuracy decreases as metastatic rate increases. Conversely, the mutation rate exhibits a positive association ($p < 0.001$) with increased phylogenetic accuracy. Intriguingly, the number of sites shows a significantly negative association (p = 0.00116).

For VCF-PB, several coefficients demonstrate statistically significant effects on phylogenetic accuracy. The number of tumours exhibits a negative association ($p < 0.001$), indicating that an increase in the number of tumours correlates with a decrease in accuracy. The exponential growth rate coefficient shows a marginally significant negative association (p = 0.02141), suggesting that higher values of this parameter are associated with reduced accuracy. The cells per tumour coefficient demonstrates a positive association ($p < 0.001$), indicating that a higher number of sampled cells are associated with increased accuracy. Similarly, the effective population size exhibits a negative association ($p < 0.001$). The metastatic rate coefficient also shows a negative association ($p < 0.001$), implying that accuracy decreases as metastatic rate increases.

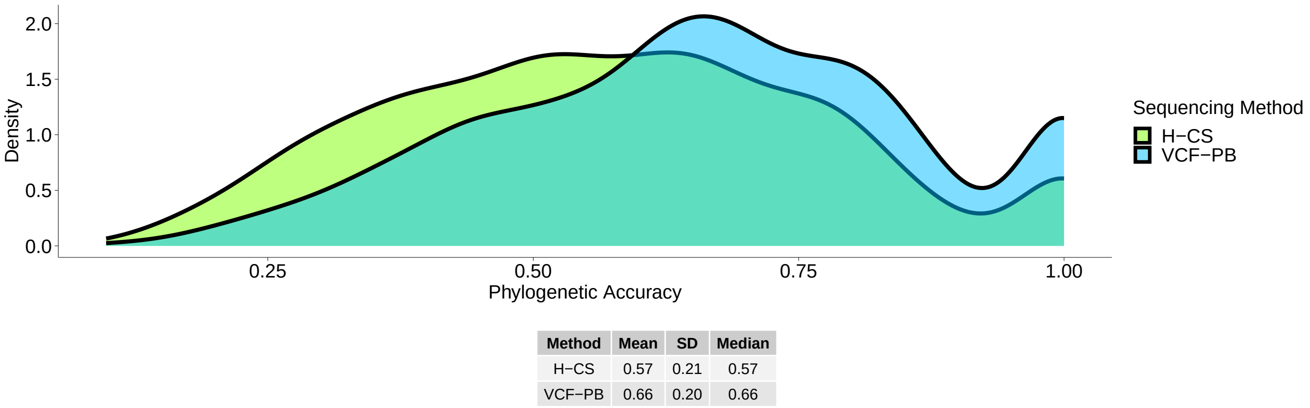
Method	Parameter	Estimate	Std. Error	t value	p value	
scSEQ	Intercept	0.99093	0.00324	305.90145	0.00000	***
	Number of Tumours	-0.00010	0.00022	-0.48396	0.62848	
	Exponential Growth Rate	-1.64127	1.33921	-1.22555	0.22055	
	Cells per Tumour	-0.00027	0.00032	-0.84214	0.39983	
	Effective Population Size	-0.00000	0.00000	-2.07955	0.03773	*
	Metastatic Rate	-1.01047	14.80343	-0.06826	0.94559	
	Mutation Rate	112.33652	135.09574	0.83153	0.40580	
	Number of Sites	0.00000	0.00000	6.78149	0.00000	***
	Allelic Dropout	0.00049	0.00669	0.07369	0.94126	
	Amplification Error	0.00349	0.01348	0.25912	0.79558	
	Sequencing Errors	0.02213	0.02660	0.83208	0.40549	
H-CS	Intercept	0.97942	0.03755	26.08410	0.00000	***
	Number of Tumours	-0.01310	0.00249	-5.25176	0.00000	***
	Exponential Growth Rate	-10.60742	15.52320	-0.68333	0.49450	
	Cells per Tumour	0.00512	0.00367	1.39499	0.16321	
	Effective Population Size	-0.00000	0.00000	-18.74405	0.00000	***
	Metastatic Rate	-2140.28596	171.59071	-12.47320	0.00000	***
	Mutation Rate	7154.90142	1565.93290	4.56910	0.00001	***
	Number of Sites	-0.00000	0.00000	-3.25345	0.00116	**
	Allelic Dropout	-0.07559	0.07752	-0.97516	0.32963	
	Amplification Error	0.17532	0.15627	1.12194	0.26206	
	Sequencing Errors	-0.94801	0.30828	-3.07512	0.00214	**
VCF-PB	Intercept	1.01497	0.03789	26.78616	0.00000	***
	Number of Tumours	-0.01344	0.00252	-5.34019	0.00000	***
	Exponential Growth Rate	-36.07609	15.66501	-2.30297	0.02141	*
	Cells per Tumour	0.01338	0.00370	3.61432	0.00031	***
	Effective Population Size	-0.00000	0.00000	-16.86703	0.00000	***
	Metastatic Rate	-1566.34012	173.15815	-9.04572	0.00000	***
	Mutation Rate	-1498.62551	1580.23731	-0.94835	0.34309	
	Number of Sites	-0.00000	0.00000	-0.76286	0.44566	
	Allelic Dropout	-0.14012	0.07822	-1.79120	0.07345	
	Amplification Error	0.08320	0.15769	0.52759	0.59786	
	Sequencing Errors	-0.93391	0.31110	-3.00196	0.00272	**

Table 3.5: For simulation MR-High, a GLM analysis modelling the relationship between variation in biological/technical parameters and phylogenetic accuracy of reconstructed trees.

3.3.3 Distribution of Phylogenetic Accuracy Score

After identifying anomalous effects resulting from our initial metastatic rate range in our analyses, we have decided to narrow our focus exclusively to simulation MR-High. Our GLM analyses suggest that this particular simulation, overall, exhibits the expected influences of the parameters under study.

In Fig 3.4, an overlapping distribution graph displays a comparison of the phylogenetic accuracy scores obtained from H-CS and VCF-PB data for simulation MR-High. Notable disparities emerge in the performance of the two methods. On average, VCF-PB exhibits higher accuracy scores, with a mean of 0.66 and a smaller standard deviation of 0.2, compared to H-CS with a mean of 0.57 and a slightly broader standard deviation of 0.21. This indicates that the utilisation of the VCF-PB method tends to yield more consistent and superior accuracy across the dataset.



Method	Mean	SD	Median
H−CS	0.57	0.21	0.57
VCF−PB	0.66	0.20	0.66

Figure 3.4: In simulation MR-High, an overlapping distribution graph comparing the phylogenetic accuracy of H-CS and VCF-PB methods at reconstructing the true multi-tumour tree.

Both methods present a bimodal distribution pattern, with peaks occurring in the range of 0.5-0.75, followed by a decline near 0.9, and a subsequent slight increase towards 1. However, the VCF-PB data demonstrates a denser concentration of scores above 0.5 compared to H-CS, implying a more consistent performance in the higher accuracy range. Overall, these findings suggest that VCF-PB outperforms H-CS in terms of phylogenetic accuracy, as evidenced by its higher central tendency and more consistent performance throughout the dataset. Note that the difference in performance between the two methods appears to be relatively minor.

3.4 Discussion

In Chapter 3, our main aim was to assess the reconstruction accuracy between methods that manipulated single-cell sequencing data (scSEQ) using haplotype consensus sequences (H-CS) and VCF pseudobulk (VCF-PB) data, a proxy of bulk sequencing (bulkSEQ). We observed if using consensus sequences was a viable, or even a better way of obtaining multi-tumour evolution compared to bulkSEQ. Additionally, we assessed the efficacy of scSEQ data on its own in reconstructing the true cell-cell trees. Throughout this process, we confirmed whether our model ran effectively by using a GLM analysis to assess the parameters we varied throughout our dataframe, and repeated our simulations until GLM analyses resulted as expected.

Firstly, we demonstrated that when constructing consensus sequences, adjusting the cutoff threshold value effectively reduced the proportion of total ambiguity bases. The rationale behind implementing an adjustable cutoff is to mitigate overall ambiguity. As several cell sequences are pooled together, there is opportunity for bases to introduce ambiguity by not determining a complete consensus. However, by selecting an appropriate threshold, we established a limit on which bases were assessed, essentially reducing ambiguity. This process ensured that only bases meeting the specified threshold were retained in the consensus sequences. We selected a cutoff threshold value that resulted in the lowest proportion of ambiguity bases throughout our data points. However, it is an assumption that reducing the maximum proportion of ambiguity significantly improves results. There is also the possibility that cutting off too much information could lead to loss of valuable data and result in an overly aggregated perspective of the original single-cell data. If we perceive VCF-PB as a proxy for bulk sequencing, it's worth noting that our results suggest that bulk sequencing performs better than this approach. This raises questions about the construction of the consensus sequence and our reliance on the cutoff threshold value. If not for time constraints, we would consider assessing the phylogenetic accuracy of a range of cutoff values through one sample.

We conducted GLM analyses at various levels of metastatic rate to assess the relationships between the parameters we varied and their impact on phylogenetic accuracy. Initially, we explored a specific range of metastatic rates, but certain parameters exhibited unexpected influence. Particularly, in the scSEQ method, we observed positive associations between metastatic rate and negative associations in allelic dropout (ADO) concerning phylogenetic accuracy improvement. This initial positive association with metastatic rate is intriguing because it would be assumed that low metastatic rates lead to longer branch lengths due to increased time for coalescent events. However, it appears we may have extended the range too far, resulting in long periods with virtually no significant events, causing anomalies in the results. This discrepancy doesn't align with typical sequencing scenarios. Also, ADO is known to introduce noise by confounding data, and its association with phylogenetic accuracy improvement is counter-intuitive. ADO is a known technical artefact that disrupts sequencing, resulting in a reduction of allele representation rather than improving accuracy (Lähnemann et al., 2021).

In our final simulation, MR-High, where the metastatic range is shifted to a higher range, we observed more satisfactory results. Consistently, we observed negative associations between metastatic rate and effective population size, as well as positive associations with mutation rate, regarding phylogenetic accuracy. These findings suggest that the simulation functioned in an appropriate sample space. There is greater confidence in the simulation's ability to effectively sample the parameter space without being adversely affected by the parameters themselves.

Based on the results from MR-High, our overlapping distribution graphs suggest that scSEQ alone is highly proficient in accurately reconstructing the true cell-cell tree. Additionally, VCF-PB outperforms H-CS. This implies two key points: Firstly, bulk sequencing, of which VCF-PB is a proxy, is superior to using consensus sequences derived from scSEQ data for multi-tumour reconstruction. However, it is important to note that VCF-PB is technically a derivative of scSEQ data. As a natural consequence, it is the only method available to mimic bulk data from the same sample of scSEQ. Therefore, VCF-PB represents a unique approach to manipulating scSEQ data and can be considered a better method than using consensus sequences for reconstruction of multi-tumour evolution, assuming one only has scSEQ data on hand.

4 Extending Tree Estimation via Replicate Data Integration

4.1 Introduction

In the preceding chapter, we conducted a direct comparison between the efficacy of two distinct methods that manipulated scSEQ data: phylogenetic reconstruction from H-CS against VCF-PB data, to recover the true multi-tumour tree. This comparison involved evaluating two forms of pooled scSEQ data. However, one may inquire: What if a researcher seeks to determine if scSEQ alone is sufficient, without the need to resort to consensus sequences, and how does it compare to bulk sequencing? Previously, the comparison between directly using individual scSEQ data against H-CS and VCF-PB data was not possible, as the reconstructed trees produced from unprocessed scSEQ data existed on a different scale to those of the latter. In this chapter, we extend the comparison of methods to include the direct assessment of phylogenetic inference using scSEQ data against those that have implemented H-CS and VCF-PB methods. We accomplish this by conducting phylogenetic reconstruction using tree estimation techniques, leveraging the reconstructed tree files generated in the previous chapter.

Recall that we have produced 8 cell-cell tree replicates for each multi-tumour tree. These replicates represent different genomic loci within the same tumour cell, each carrying a distinct genetic lineage. We can leverage various software tools, and by adapting species tree estimation methods, combine these lineages and estimate a tree that captures the evolutionary relationships on a multi-tumour level.

Species tree estimation is an extensively researched problem in phylogenetics, that has posed significant computational challenges. It encompasses a multitude of approaches constructed by different algorithms aimed at reconstructing evolutionary trees from analyses of biological datasets. It has also garnered significant interest in recent years, driven by the need to address heterogeneity caused by ILS observed within the MSC model. In our study, we are particularly interested in estimation methods capable of utilising compiled tree files as input data. An established approach for species tree estimation involves computing gene trees, representing trees on different genomic regions, and then integrating these trees into a species tree, often under the framework of the multispecies coalescent (MSC) model, which as we know, accounts for gene tree heterogeneity. ASTRAL is a widely utilised method for species tree estimation.

There is also supertree estimation, which involves assembling a single phylogenetic tree from multiple smaller trees, with supertree algorithms aiming to minimise discrepancies between them. FastRFS (Fast Robinson-Foulds Supertrees) is a software tool known for generating optimal supertrees.

Concatenation methods represent an alternative approach (Warnow, 2015), where sequences from multiple distinct genomic regions are combined into a single alignment. This concatenated alignment is then utilised to infer the species tree, operating under the assumption that all genes share the same evolutionary history.

Rather than compiling gene trees, or species trees, here we compile the replicates of either existing cell-cell trees, or multi-tumour trees as input so as to reconstruct a multi-tumour tree. Additionally, we can explore whether concatenating replicate H-CS/VCF-PB sequences offers an alternative approach to infer the multi-tumour tree from the H-CS or VCF-PB data. These data sources already exist on a multi-tumour level scale, providing an opportunity to investigate their effectiveness in reconstructing the multi-tumour tree directly.

4.2 Methods

4.2.1 Pipeline

For Chapter 4, we have a primary R script dedicated to compiling and processing our data. The primary R script for this chapter will be referred to as '4_main'. This script, similar to '3_main', controls the initiation of several other scripts, embedded within each other. It is responsible for extracting the necessary files, grouping them into condensed folders that now contain all replicate files per sample. 4_main is also responsible for formatting these files so that they are compatible with the methods we input them through, initiating the tools, ASTRAL, FastRFS, and CellPhy, calculating and updating a new dataframe with tree scores.

In order to reconstruct the evolutionary history of multiple metastatic sites we previously performed phylogenetic reconstruction on haplotype consensus sequences as well as VCF pseudobulk data. However, this analysis was conducted on each individual replicate. Our objective is to estimate a multi-tumour tree directly from scSEQ data. This can be done by compiling the existing data as sets of replicates. The distinction here is important. The trees reconstructed from H-CS or VCF-PB data cannot be directly compared to the tree resulting from multi-tumour tree estimation using scSEQ.

Although the methods of H-CS, VCF-PB, discussed in Chapter 3 are potential ways to reconstruct multi-tumour evolutionary histories, they differ fundamentally to those in Chapter 4 due to the fact that the former are based on individual replicate data, whereas the objective of estimation is to compile the replicate data from scSEQ. This is because when conducting individual replicate analysis (Chapter 3), each tree reflects the evolutionary history of a single genomic loci per cell sampled, capturing its specific lineage. In contrast, estimation from compiled replicate data aims to synthesise information across multiple genomic loci to infer a combined evolutionary history. Therefore, in Chapter 4, extending the analysis by applying the same replicate integration approach to H-CS and VCF-PB data is essential to ensure a fair comparison with scSEQ multi-tumour tree estimation.

We present three different multi-tumour tree estimation methods, facilitating direct comparisons across our three types of data:

1. We compile previously obtained reconstructed cell-cell tree files from scSEQ data, organising them based on their status as replicates of the same sample. These replicate files depict the genealogies of different genomic loci of individual cells sampled from multiple metastatic sites. We convert these individual tree files into a format compatible with ASTRAL. ASTRAL then analyses these replicate tree files to estimate a multi-tumour tree.

2. We compile the reconstructed multi-tumour tree files, previously obtained from either H-CS data or VCF-PB data based on their status as replicates of the same sample. This method is divided into two steps: first, we apply the tree estimator to H-CS data, and then to VCF-PB data. These replicate tree files depict the genealogies of multiple metastatic sites. They contain data originating from various genomic loci of each cell sampled from each metastatic site. Initially simulated from scSEQ data, the data has been pooled based

on the metastatic site from which each cell was sampled from. We convert these individual tree files into a format compatible with FastRFS. FastRFS integrates the information from our multi-tumour tree replicate files to construct a supertree, representing their combined evolutionary relationships.

3. The final method, is divided into two steps: the first is applied to H-CS, and then to VCF-PB. It involves concatenating sequence files based on their status as replicates of the same sample. For the haplotype files, this is performed manually through basic file manipulation. For the VCF files, we utilise VCFtools to concatenate the relevant data together. Subsequently, a file containing the compiled data of all concatenated sequences per sample undergoes tree reconstruction.

By utilising these methods and leveraging different data types, we aim to determine which approach is most effective in recovering the true multi-tumour evolutionary tree. Our rationale for this extended analysis lies in the unique advantage offered by the presence of multiple genomic loci, or ‘replicates’, for each sample. These replicates exhibit distinct genomic lineages, enabling us to implement them as a basis for conducting three distinct methods of tree reconstruction. Through this approach, we can directly compare the efficacy of estimating trees using individual cell-cell lineages against the manipulation of reconstructed trees obtained through H-CS and VCF-PB.

This evaluation enables us to assess the strengths and limitations of each method we propose up to this point in recovering the underlying evolutionary dynamics of multi-tumour evolution. We again explore how variations in biological and technical parameters across a diverse sample space impact the reliability of our methods.

4.2.2 Tree Estimation using scSEQ and ASTRAL

Our primary R script for Chapter 4, 4_main, first reads the existing LHS dataframe. It generates a new dataframe, condensing the original by filtering for a single replicate. This reduction condenses the dataset from 1600 individual data points to the core samples themselves, which are represented by just 8 replicates. As a result, the new dataframe comprises only 200 rows.

Subsequently, the script iterates through the output folders within the CellCoal directory, extracting the first portion of the iteration identifier that labels each output folder. This string is obtained by removing the numeric suffix `r[0-9]+`, which is used to represent the replicate number from the folder name. Following this, the script creates corresponding folders within a new directory allocated for ASTRAL output. Finally, for each output folder within the CellCoal directory, the script locates and lists all cell-cell tree files. These tree files are copied to their respective folder within the ASTRAL directory. Prior to being copied, each tree file is appended with the replicate label. Each sample folder embedded within the ASTRAL directory will contain 8 cell-cell tree files. In summary, the replicate cell-cell tree files of each sample are extracted from the CellCoal directory, compiled together in the output folder of a new ASTRAL directory.

At this stage, two external Python scripts are executed. One is responsible for constructing a mapping file for each sample. The other is responsible for compiling the replicate tree files of each sample into a single tree file.

ASTRAL necessitates a mapping file to be provided alongside the compiled tree files, ensuring conformity with the specified data formatting within the tree files. On the left, is the developer's guideline delineating the required mapping file format. Next to this, we present an example of our adjusted tree data in accordance with their guideline.

```
individual_A1 species_name_A     7_0_3 7
individual_A2 species_name_A     7_0_2 7
individual_B1 species_name_B     7_0_5 7
individual_B2 species_name_B     7_0_0 7
individual_B3 species_name_B     7_0_4 7
```

This is constructed using Python library ‘DendroPy’. The Python script traverses the ASTRAL directory and for each tree file located, parses the tree string based on commas assuming that each cell or metastatic site is separated by a comma. For each cell in the tree, the code splits the cell label at the colon to extract the identifier. It then processes this identifier to obtain the metastatic site label and cell identifier, which are subsequently used to create a unique individual key.

ASTRAL requires input trees to be compiled together into a single tree file. This file incorporates the information from multiple individual trees, enabling ASTRAL to analyse the evolutionary information from all input trees simultaneously. For each sample, this is done by traversing the ASTRAL directory, exploring each folder and identifying the tree files that adhere to the naming convention, commencing with `r` and concluding with `.raxml.bestTree`. The script proceeds to load each tree using DendroPy's `TreeList.get_from_path()` function, accumulating them into a single tree file.

The Python script initiates the execution of ASTRAL. Upon completion of the ASTRAL analysis, gRF scores are calculated between the estimated tree from ASTRAL and the true multi-tumour tree from SimPhy and updated into the new dataframe.

4.2.3 Supertree Construction and FastRFS

The approach we implement to manipulate our files in preparation for supertree construction resembles that from tree estimation using scSEQ and ASTRAL. Note that this method is initially applied to the reconstructed trees obtained from our H-CS data, followed by those derived from VCF-PB data. Since the data manipulation remains consistent across both sets of data, we describe the process once, explaining how both are utilised in unison.

Once again, the Python script traverses the output folders within the CellCoal directory, extracting the initial portion of the iteration identifier that identifies each output folder. Subsequently, new directories for FastRFS, allocated for H-CS and VCF-PB, are created. The script then identifies all multi-tumour tree files and transfers them to their corresponding folders within the FastRFS directories. Before the copying process, each tree file is appended with the replicate label. As a result, each sample folder within the FastRFS directory should contain 8 multi-tumour tree files. In summary, the replicate multi-tumour tree files of each sample are extracted from the CellCoal directory and compiled together in the output folder of a new FastRFS directory, with one directory dedicated to H-CS and the other to VCF-PB.

Unlike ASTRAL, FastRFS does not require a mapping file. However, it still requires that individual tree files are compiled into a single file. An external Python script that iterates over all samples within the respective FastRFS directories is executed to perform this task.

Following the merging process, FastRFS is executed to construct our supertrees. For each supertree constructed, the script calculates the gRF score between the FastRFS supertree derived from either H-CS/VCF-PB data and the true multi-tumour tree first generated by SimPhy. Scores are subsequently updated in the new dataframe.

4.2.4 Sequence Data Concatenation and CellPhy

Finally, we investigate the strategy of concatenating data separately for haplotype and VCF files within CellCoal. For each sample, we concatenate either the pooled sequences from H-CS data or VCF-PB data, compiling sequences based on the fact that each sequence represents a replicate of that sample.

We repeat the process of manipulating and organising files used in previous methods to condense the CellCoal directory. We extract the first portion of the iteration identifier, omitting the replicate label, to create new directories dedicated for each method. Each embedded folder contains only the first portion of the iteration identifier. By condensing 1600 folders into 200, the embedded folders within each new directory contains 8 replicate sequences whereby the cells have been pooled together based on the metastatic site from where each was sampled from.

At this stage, it is necessary to separate the concatenation methods applied for H-CS and VCF-PB data. Concatenation for H-CS involves a process where sequence files from each FASTA file are appended one after the other. However, VCF-PB file concatenation necessitates the use of VCFtools, implementing the `vcf-concat` function.

1. The process of concatenating H-CS sequences into a single file involves reading each FASTA file within a folder, then parsing its contents line by line. As the script encounters header lines (denoted by ">" at the beginning), it captures the header information and initialises an empty string to store the associated sequence. Subsequent lines are appended to this string until a new header line is encountered, indicating the start of a new sequence. At this point, the script pairs the header with its corresponding sequence and stores this pair in a list. Once all replicate sequences within the folder are processed, the script iterates over the collected pairs and writes them to a new FASTA file, where each pair (header and sequence) occupies a single block of text, adhering to the FASTA file format.

2. Concatenating the VCF-PB files involves the use of the `vcf-concat` function. The command takes multiple input VCF files as arguments and merges them into a single VCF file, appending the contents of each input file to the end of the previous one, thereby preserving the variant records from each input file. Note that this does not perform merging of variant

records. In our script, `vcf-concat` is used within a loop to concatenate multiple VCF files found within each sample folder, creating a single concatenated VCF file for each sample.

As the concatenation method requires the use of CellPhy to undergo phylogenetic reconstruction, we implement the use of our three-tiered script structure from Chapter 3 so as to allow for adjustment of CPU usage. After phylogenetic trees are obtained, the gRF scores for comparison against the true multi-tumour tree are obtained and updated to the dataframe.

4.3 Results

4.3.1 Extended: Effect of Parameters on Phylogenetic Accuracy

Recall that in the previous chapter, we determined that the final simulation, MR-High, where metastatic rate is set higher than prior simulations returned meaningful results. The GLM analyses did not show confounding parameter influences. Because of this, the results presented here will focus only on the data simulated within MR-High.

In our extended analysis, we examine the outcomes obtained using different methodologies. Specifically, we use ASTRAL to reconstruct trees from scSEQ data, while FastRFS is utilised for reconstructing trees from both H-CS and VCF-PB data. Moreover, concatenation is employed to reconstruct trees from both H-CS and VCF-PB datasets. We conduct a GLM analysis on these methods, as shown in Table 4.1.

The significance of intercepts across all five analyses suggests that a baseline level of accuracy can be observed even without the inclusion of the parameters assessed. This underscores the inherent effectiveness of these methods in tree reconstruction tasks.

For ASTRAL, the influence of effective population size is statistically significant ($p < 0.001$), with a negative impact on phylogenetic accuracy. As the effective population size decreases, there is an increase in phylogenetic accuracy. As we have observed in previous simulations, this is an expected outcome, due to the fact that a smaller effective population size leads to quicker coalescence times. The influence of metastatic rate is also highly significant ($p = 0.00511$), again with a negative impact on phylogenetic accuracy. When we have low metastatic rates, this means that the generation of a tumour occurs at a slower rate within the simulation. This leads to longer

Method	Parameter	Estimate	Std. Error	t value	p-value	
scSEQ-ASTRAL	Intercept	1.04725	0.07920	13.22314	0.00000	***
	Number of Tumours	-0.01002	0.00526	-1.90339	0.05851	
	Exponential Growth Rate	29.06761	32.74186	0.88778	0.37579	
	Cells per Tumour	0.00236	0.00774	0.30446	0.76112	
	Effective Population Size	-0.00000	0.00000	-3.89443	0.00014	***
	Metastatic Rate	-1025.32543	361.92260	-2.83300	0.00511	**
	Mutation Rate	4852.02131	3302.89731	1.46902	0.14349	
	Number of Sites	-0.00000	0.00000	-1.25798	0.20995	
	Allelic Dropout	0.08903	0.16350	0.54456	0.58670	
	Amplification Error	0.03367	0.32960	0.10215	0.91875	
	Sequencing Errors	-1.05819	0.65024	-1.62739	0.10532	
H-CS-CONCAT	Intercept	0.84249	0.10294	8.18386	0.00000	***
	Number of Tumours	-0.00110	0.00684	-0.16034	0.87278	
	Exponential Growth Rate	-22.88528	42.55910	-0.53773	0.59140	
	Cells per Tumour	0.01003	0.01006	0.99763	0.31973	
	Effective Population Size	-0.00000	0.00000	-5.60188	0.00000	***
	Metastatic Rate	-2044.49322	470.44067	-4.34591	0.00002	***
	Mutation Rate	14266.65993	4293.23077	3.32306	0.00107	**
	Number of Sites	-0.00000	0.00000	-0.23106	0.81752	
	Allelic Dropout	0.18149	0.21252	0.85396	0.39421	
	Amplification Error	0.62868	0.42842	1.46741	0.14393	
	Sequencing Errors	-0.75393	0.84520	-0.89201	0.37352	
H-CS-FastRFS	Intercept	1.03720	0.10324	10.04630	0.00000	***
	Number of Tumours	-0.01869	0.00686	-2.72552	0.00702	**
	Exponential Growth Rate	-3.48306	42.68191	-0.08161	0.93505	
	Cells per Tumour	0.01809	0.01009	1.79342	0.07450	
	Effective Population Size	-0.00000	0.00000	-6.28287	0.00000	***
	Metastatic Rate	-1542.84547	471.79818	-3.27014	0.00128	**
	Mutation Rate	6943.58629	4305.61931	1.61268	0.10848	
	Number of Sites	-0.00000	0.00000	-1.31985	0.18848	
	Allelic Dropout	-0.02220	0.21314	-0.10415	0.91716	
	Amplification Error	0.58732	0.42966	1.36694	0.17327	
	Sequencing Errors	-0.37571	0.84764	-0.44325	0.65809	
VCF-PB-CONCAT	Intercept	1.05201	0.08241	12.76513	0.00000	***
	Number of Tumours	-0.01356	0.00548	-2.47690	0.01413	*
	Exponential Growth Rate	7.36100	34.07090	0.21605	0.82918	
	Cells per Tumour	0.00747	0.00805	0.92785	0.35467	
	Effective Population Size	-0.00000	0.00000	-3.48237	0.00062	***
	Metastatic Rate	-1019.88673	376.61362	-2.70805	0.00739	**
	Mutation Rate	6028.76041	3436.96728	1.75409	0.08104	
	Number of Sites	-0.00000	0.00000	-1.22667	0.22147	
	Allelic Dropout	-0.04343	0.17014	-0.25526	0.79880	
	Amplification Error	0.43834	0.34298	1.27804	0.20280	
	Sequencing Errors	-1.28508	0.67663	-1.89923	0.05906	
VCF-PB-FastRFS	Intercept	0.96552	0.08630	11.18828	0.00000	***
	Number of Tumours	-0.00746	0.00573	-1.30102	0.19484	
	Exponential Growth Rate	25.63038	35.67674	0.71841	0.47339	
	Cells per Tumour	0.01012	0.00843	1.20004	0.23163	
	Effective Population Size	-0.00000	0.00000	-5.21679	0.00000	***
	Metastatic Rate	-176.10749	394.36425	-0.44656	0.65570	
	Mutation Rate	-1667.64233	3598.95904	-0.46337	0.64363	
	Number of Sites	-0.00000	0.00000	-0.14145	0.88766	
	Allelic Dropout	-0.01789	0.17815	-0.10041	0.92013	
	Amplification Error	0.00422	0.35914	0.01176	0.99063	
	Sequencing Errors	-0.12469	0.70852	-0.17599	0.86049	

Table 4.1: For simulation MR-High, the impact of variation in biological/technical parameters on the accuracy of trees reconstructed using ASTRAL, FastRFS and concatenated sequences.

branch lengths between each metastatic event, which provides more time for subclonal cell lineages to coalesce within its population.

For the H-CS concatenated sequences, we observe a significant negative association with effective population size (p-value < 0.001), again implying that as effective population size decreases, the phylogenetic accuracy increases. A significant negative association with metastatic rate is also observed (p-value < 0.001), which reaffirms that lower metastatic rates lead to higher phylogenetic accuracy. Furthermore, mutation rate is statistically significant ($p = 0.00107$), shown to have a positive association with accuracy. This may be because the higher mutation rates increase genetic diversity within each metastatic site, providing more genetic markers for analysis. This can enhance the resolution of inference methods, allowing them to better distinguish the true evolutionary histories between closely related lineages, despite the potential noise introduced by incomplete lineage sorting.

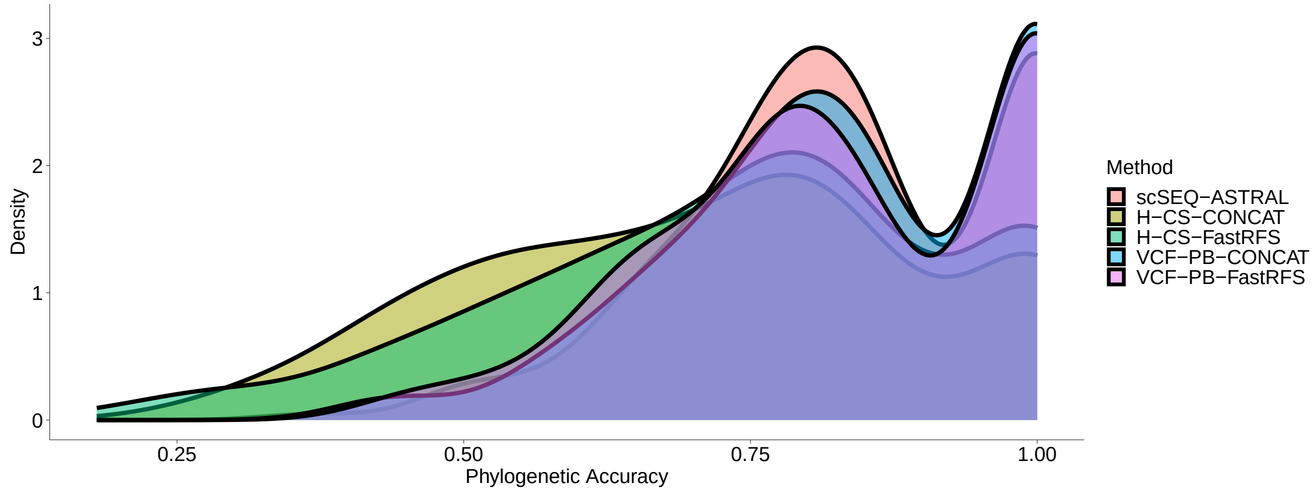
For H-CS FastRFS, both effective population size and metastatic rate exhibit statistically significant impacts on phylogenetic accuracy ($p < 0.001$, $p = 0.00128$). Interestingly, there is a negative association between the number of tumours and phylogenetic accuracy. An increase in the number of populations studied in phylogenetic trees has shown to improve accuracy. However, this increase in the number of populations may also lead to a higher likelihood of incomplete lineage sorting due to the need to accommodate multiple independent evolutionary histories.

For both VCF-PB concatenated sequences and VCF-PB FastRFS, there are significant negative associations with effective population size on phylogenetic accuracy ($p < 0.001$ for both), and for VCF-PB concatenated sequences, there is also a significant negative association with metastatic rate ($p = 0.00739$).

4.3.2 Extended: Distribution of Phylogenetic Accuracy

Upon examining the overlapping distribution graph (Fig 4.1), it is evident that all methods demonstrate varying levels of performance in tree reconstruction. For ASTRAL, the mean accuracy score is 0.84, with a standard deviation of 0.14 and a median of 0.83. H-CS concatenation shows a mean accuracy of 0.72, slightly lower than ASTRAL, with a standard deviation of 0.2 and a median of 0.74. H-CS FastRFS yields a mean accuracy of 0.75, with a similar standard deviation of 0.2

and a median of 0.77. Interestingly, both VCF-PB concatenation and VCF-PB FastRFS exhibit near-identical mean and median accuracy scores of 0.84, 0.83 and 0.83, 0.83, respectively, suggesting comparable performance within the VCF-PB dataset. Interestingly, we identify a bimodal distribution pattern across all methods, characterised by a notable decline in density around an accuracy score of 0.825, similar to the curve depicted in the overlapping distribution graph (Fig 3.4) in Chapter 3.



Method	Mean	SD	Median
scSEQ−ASTRAL	0.84	0.14	0.83
H−CS−CONCAT	0.72	0.20	0.74
H−CS−FastRFS	0.75	0.20	0.77
VCF−PB−CONCAT	0.84	0.15	0.83
VCF−PB−FASTRFS	0.83	0.15	0.83

Figure 4.1: In simulation MR-High, an overlapping distribution graph comparing the phylogenetic accuracy of replicate data integration methods at reconstructing the true multi-tumour tree.

With regard to density distribution, the scores for ASTRAL and VCF-PB methods are predominantly concentrated in the higher range (0.7 and above), indicating superior performance compared to H-CS, which tends to cluster in the lower score region. This disparity suggests that scSEQ-Astral and VCF-PB exhibit greater accuracy and consistency in reconstructing phylogenetic trees, while H-CS shows comparatively lower performance across the evaluated metrics. Visually, we observe that ASTRAL has a slightly larger density of scores between 0.75-8, in comparison to the VCF-PB methods, while the VCF-PB methods have a slightly higher density near a perfect accuracy score of 1. While the difference is relatively small, we also observe that concatenation of the VCF-PB data shows to perform slightly better throughout the density distribution, which is also evidence in the mean scores.

4.4 Discussion

This chapter served as an extension of the preliminary analysis conducted in Chapter 3. In Chapter 3, comparisons were limited to between H-CS and VCF-PB due to the nature of the analysis, which relied on manipulating single-cell sequencing data by pooling cells together. In Chapter 4, we explored additional approaches beyond the use of H-CS and VCF-PB. While the inclusion of replicates in our simulations enhances the statistical significance of the data, they also allow us to compile the information generated from each replicate of a sample, using these compiled sets to reconstruct trees.

We represented scSEQ replicates as the unlinked genomic loci of a cell. With this approach, we have an additional way of utilising single-cell sequencing data directly, known as species estimation. This method involved providing an estimation tool with multiple lineages, and enables the inference of a tree while considering factors such as ILS.

To compare this approach against H-CS and VCF-PB, we must also utilise replicates for these methods. For each sample that consists of reconstructed H-CS tree or VCF-PB tree, we created sets comprising their respective replicate tree files. For H-CS and VCF-PB, we have two options for comparison: we can either construct a supertree using the trees themselves, or perform phylogenetic reconstruction on concatenated sequences.

Here, the prioritised data in our results were derived from the MR-High simulation, where we first observed favourable associations between our parameters and phylogenetic accuracy. We performed a final GLM analysis comparing methods that integrate the use of replicate data. We noted consistent trends, wherein improved phylogenetic accuracy was statistically linked to negative associations with metastatic rate and effective population size, and conversely, positive associations with mutation rate. Note that fewer parameter associations reached statistical significance. This could be attributed to the condensation of data, resulting from the use of replicates. Previously, our analysis comprised of 1600 data points, whereas now there are 200 available data points.

Our overlapping distribution graph reveals that both scSEQ and VCF-PB methods exhibited similar and optimal performance. This highlights two significant points: Firstly, if only bulk sequencing data is available, our results suggest that bulk sequencing data alone may be just as effective in reconstructing multi-tumour evolution as single-cell estimation methods or the use of haplotype consensus sequences. Secondly, for those utilising single-cell sequencing replicate data, the most beneficial approach involves leveraging either species estimation methods, or pooling VCF data, then compiling replicate data through concatenation or construction of supertrees.

5 General Discussion

In this project, we developed a strategy that combined two simulation tools using a stratified sampling approach to accurately capture the sample space inherent in single-cell data. Initially, we aimed to assess the efficacy of pooling consensus sequences to determine if it could outperform bulk sequencing in achieving reconstruction accuracy. Subsequently, our objective shifted to investigating whether the utilisation of replicates could facilitate a comparison involving single-cell sequencing (scSEQ) data, alongside the construction of haplotype consensus sequences (H-CS), and the use of a VCF pseudobulk (VCF-PB) dataset, regarding their ability to reconstruct multi-tumour evolution.

5.1 Common Trends in Parameter Analyses

Generalised linear model (GLM) analyses have proven to be effective in assessing the impact and significance of variables, providing a common method for evaluating their effects. In Chapter 3, our initial observations revealed several anomalies, likely attributed to extensively low metastatic rates. However, upon shifting to a more appropriate range, such as the simulation MR-Full, we found that parameters had influenced phylogenetic accuracy in a more expected manner. Notably, in MR-Full, when the metastatic rate is relatively high, indicative of quicker progression, scSEQ tended to accurately reconstruct the true cell-cell tree, but was prone to incomplete lineage sorting (ILS) at higher rates. We also note that scSEQ's performance improved as the effective population size decreased. Similarly, these trends observed in scSEQ were also observed in H-CS and VCF-PB. Furthermore, they were consistent across all other methods explored in Chapter 4. We consistently observed that a decrease in metastatic rate improves phylogenetic accuracy, while a decrease in effective population size also enhances the performance of all methods. This emphasises the importance of considering these factors from an evolutionary perspective, particularly in the context of somatic clonal evolution. When metastatic progression is low, it is more likely that researchers will be able to recover the true evolutionary histories, as subclonal cells have had less chance to intermix between tumours. Similarly, a lower effective population size enhances the accuracy of evolutionary reconstructions.

5.2 Phylogenetic Accuracy of Consensus Sequence Tree Reconstruction

We find that the utilisation of H-CS data, derived from scSEQ, does not outperform a pseudobulk dataset in reconstructing multi-tumour evolution. There are several potential reasons for this outcome. Firstly, our selection of the cutoff threshold value for consensus sequence generation may be flawed, limiting the performance of consensus sequences. Additionally, VCF-PB, which serves as a proxy for bulk sequencing, essentially aggregates scSEQ data. If we consider VCF-PB as another method to manipulate scSEQ data, it can therefore be suggested that when presented with scSEQ data, constructing VCF-PB data is a preferable method of pooling, as it exhibited better performance scores than H-CS. Overall, if a researcher has scSEQ data on hand, to perform multi-tumour reconstruction, they should utilise pooling cells via the VCF-PB method. If a researcher possesses bulk sequencing data, it is reasonable to assume that their multi-tumour reconstruction performance would surpass that of pooling using consensus sequences derived by scSEQ.

5.3 Methods using VCF-PB Rivals Accuracy of scSEQ data

If multiple unlinked genomic regions of the same cell are sampled, estimation methods can be performed on existing scSEQ data. The replicate data of H-CS and VCF-PB can also be utilised to reconstruct trees via the generation of supertrees or through phylogenetic reconstruction of their concatenated sequences. When comparing these methods (scSEQ using ASTRAL, H-CS via FastRFS/concatenation, VCF-PB via FastRFS/concatenation), we find that compiling scSEQ cell-cell trees through ASTRAL to estimate trees yields similar performance statistics to constructing FastRFS supertrees using existing VCF-PB multi-tumour tree files or concatenating the pooled sequences of VCF-PB. Furthermore, out of all available methods, these three demonstrated the highest performance. We demonstrate that using methods that take advantage of multiple unlinked genomic data is a viable way to estimate multi-tumour evolution. It is likely that increasing the number of tree files used to reconstruct the tree will lead to more accurate trees. Overall, if a researcher is able to obtain replicate data of scSEQ, they can directly reconstruct their individual tree lineages, and compile the data to estimate multi-tumour evolution effectively.

5.4 Limitations

One of the primary limitations of our study relates to the total sample size of our simulations. In Chapter 3, the low sample size may have contributed to confounding effects in our GLM analyses such as notable association between ADO and phylogenetic accuracy, which we attribute to the coincidental alignment of low ADO with other outliers. With a larger sample size, the dispersion of data points across different levels of ADO would likely have been more pronounced, potentially mitigating such chance associations. In Chapter 4, we observed that only a limited number of parameters exhibited statistical significance. This reduced significance likely arose from compiling our data based on sets of replicates. We initially had 1600 data points in Chapter 3 but consolidated them to 200 data points. This decrease may have reduced overall significance. Furthermore, increasing the number of replicates per sample would likely have improved the accuracy of our tree reconstruction methods in Chapter 4. This improvement would stem from the nature of the tools we utilised, ASTRAL, FastRFS and concatenation, which rely on multiple distinct lineages (or their sequence information) of the same population/individual for accurate estimation. Methods such as species estimation and supertree construction typically utilise numerous tree files, often numbering in the hundreds, to estimate accurate phylogenetic trees. However, increasing the sample size posed challenges for our project due to computational and time constraints. Specifically, iterations with large numbers of sites or total cells resulted in large VCF files, often several gigabytes in size. Larger simulations were occupying several terabytes of space. Additionally, we had to account for the computational resources required to run CellPhy efficiently, which increased with larger datasets. These factors limited our ability to expand the sample size beyond what was feasible within our limited data storage capacity and time constraints. We had hoped to gain deeper insight, expecting a greater number of parameters to indicate how the accuracy scores were influenced. Once we identify strategies for optimising the entire simulation process, including the appropriate

CPU allocation and logging run-times, implementing fully automated runs would provide a clearer path forward. This automation could facilitate scaling up the simulation size to improve statistical power effectively.

Our second limitation concerns the inclusion of ten parameters in our analyses, prompting a reflection on the appropriateness of this number. Firstly, the selection of these specific parameters aligns with the expected characteristics of a simulated system that models multiple populations among cancer cells. For instance, the incorporation of exponential growth rate and considerations for sequencing errors enhances the model's realism, addressing common phenomena observed in tumour studies. We intentionally omitted the variation of parameters to simplify the model. For example, haploid coverage (level of sequencing coverage assigned to haploid regions) was fixed at a value of 0.5, instead of being treated as variables. This decision aimed to simplify our model of cancer by reducing the number of variable parameters. However, by assigning fixed values to parameters such as haploid coverage we may have unintentionally increased complexity. These parameters could have been omitted entirely, as they may not have been necessary for our analysis. Upon reflection, it becomes apparent that further streamlining complexity could be advantageous. For instance, keeping the exponential growth rate constant rather than making it a variable could have simplified the model without sacrificing accuracy. Additionally, focusing on simulating scenarios that align with realistic expectations is crucial. For instance, setting the metastatic rate at a higher level from the beginning would have better reflected real-world conditions and could have mitigated confounding factors in our simulations, reducing the need for repeated simulation runs.

An inherent limitation of our model, as previously described, is the challenge of obtaining bulk and single-cell data from the same source. Practically, this is unattainable unless a restricted space is simulated, generating the single cells sampled alongside other cells within the mix. This approach could provide a more accurate comparison between bulk and single-cell sequencing data. Consequently, the utilisation of VCF-PB experiments serves two roles: it can represent bulk sequencing data or it can be used as a method to modify single-cell sequencing data for scaling up to multi-tumour evolutionary simulations. This makes it difficult to discern whether VCF-PB represents bulk-sequencing well, or if it is just an extended method of utilising scSEQ to gather multi-tumour histories.

The construction of consensus sequences in our study raises considerations and potential areas for improvement. One notable aspect is the handling of ambiguous bases, particularly when dealing with bases containing 'B', 'D','V', 'H' ambiguity codes, which are instead represented as 'N' to accommodate for CellPhy's GT16 model. This conversion to 'N' may have influenced our results and introduces a level of uncertainty. Exploring alternative approaches to handle ambiguous bases could offer more accurate consensus sequences. The selection of a cutoff threshold value for consensus sequence generation is another critical factor. While setting a cutoff threshold can effectively filter out noise from low-frequency mutations, it also runs the risk of aggregating single-cell data to the extent that its original granularity and usefulness are compromised. Perhaps one could perform phylogenetic reconstruction on several consensus sequences of the sample sample, where the consensus sequences vary on cutoff threshold value, to determine how it effects scores, and how effectively.

CellCoal, a coalescent simulator, typically assumes neutral selection, meaning it assumes that evolutionary changes occur randomly and are not influenced by natural selection. This can significantly impact the simulation outcomes. While neutral selection assumes that genetic variations do not confer a selective advantage or disadvantage, incorporating selection pressure can alter the dynamics of the simulated populations. Selection acting similarly to lowering the effective population size can potentially lead to more accurate scSEQ trees and reduce the occurrence of ILS. However, simulating forward selection can be challenging. Firstly, accurately modelling the complex interplay between genetic variation and selection pressure requires comprehensive understanding and precise parameterization of clonal somatic evolution and the inherent errors that occur in scSEQ. Additionally, simulating forward selection may introduce computational complexities, as it involves predicting the future trajectory of genetic variations under evolving selection pressures.

5.5 Future Experiments

Expanding the sample size of our simulations presents a clear opportunity for improvement in our simulations. By increasing the number of replicates for estimation methods, we can better align with the requirements of tools like ASTRAL and FastRFS, which generally benefit from larger sets of lineage tree files. Additionally, exploring alternative phylogenetic inference methods, such as the use of Bayesian approaches like StarBEAST (Douglas et al., 2022), is worth considering if computational resources allow.

We could simplify the model by reducing the number of parameters while still maintaining a diverse sample space. This would provide more statistical power to detect influences. Additional statistical tests could be performed to determine the magnitude of effects of parameters, rather than solely focusing on significance. For example, using z-statistics to standardise the coefficient estimates would allow for this comparison.

Exploring the impact of consensus sequence construction methods on phylogenetic accuracy could provide insights into potential improvements. One approach could involve comparing phylogenetic accuracy across different consensus cutoff threshold values. By varying the cutoff threshold and analysing the resulting consensus sequences, we can assess how different thresholds affect phylogenetic accuracy. Additionally, investigating alternative consensus sequence construction methods or refining existing approaches may offer further opportunities for improvement.

References

Aissa, A. F., Islam, A. B. M. M. K., Ariss, M. M., Go, C. C., Rader, A. E., Conrardy, R. D., Gajda, A. M., Rubio-Perez, C., Valyi-Nagy, K., Pasquinelli, M., Feldman, L. E., Green, S. J., Lopez-Bigas, N., Frolov, M. V., & Benevolenskaya, E. V. (2021). Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature Communications*, *12*(1), 1628. https://doi.org/10.1038/s41467-021-21884-z

Aktipis, C. A., Boddy, A. M., Jansen, G., Hibner, U., Hochberg, M. E., Maley, C. C., & Wilkinson, G. S. (2015). Cancer across the tree of life: Cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1673), 20140219. https://doi.org/10.1098/rstb.2014.0219

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Getz, G., . . . Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alves, J. M., Prieto, T., & Posada, D. (2017). Multiregional Tumor Trees Are Not Phylogenies. *Trends in Cancer*, *3*(8), 546–550. https://doi.org/10.1016/j.trecan.2017.06.004

Bailey, C., Shoura, M. J., Mischel, P. S., & Swanton, C. (2020). Extrachromosomal DNA—relieving heredity constraints, accelerating tumour evolution. *Annals of Oncology*, *31*(7), 884–893. https://doi.org/10.1016/j.annonc.2020.03.303

Boddy, A. M. (2023). The need for evolutionary theory in cancer research. *European Journal of Epidemiology*, *38*(12), 1259–1264. https://doi.org/10.1007/s10654-022-00936-8

Bouquet, J., Cheval, J., Rogée, S., Pavio, N., & Eloit, M. (2012). Identical Consensus Sequence and Conserved Genomic Polymorphism of Hepatitis E Virus during Controlled Interspecies Transmission [Publisher: American Society for Microbiology]. *Journal of Virology*, *86*(11), 6238–6245. https://doi.org/10.1128/jvi.06843-11

Bray, F., Laversanne, M., Weiderpass, E., & Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, *127*(16), 3029–3030. https://doi.org/10.1002/cncr.33587

Cheung, K. J., & Ewald, A. J. (2016). A collective route to metastasis: Seeding by tumor cell clusters. *Science (New York, N.Y.)*, *352*(6282), 167–169. https://doi.org/10.1126/science.aaf6546

Cross, W. C., Graham, T. A., & Wright, N. A. (2016). New paradigms in clonal evolution: Punctuated equilibrium in cancer. *The Journal of Pathology*, *240*(2), 126–136. https://doi.org/10.1002/path.4757

Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, *15*(2), 81–94. https://doi.org/10.1038/nrclinonc.2017.166

Davis, A., Gao, R., & Navin, N. (2017). Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et biophysica acta*, *1867*(2), 151–161. https://doi.org/10.1016/j.bbcan.2017.01.003

de Bruin, E. C., Taylor, T. B., & Swanton, C. (2013). Intra-tumor heterogeneity: Lessons from microbial evolution and clinical implications. *Genome Medicine*, *5*(11), 101. https://doi.org/10.1186/gm505

Dimitrakopoulos, C. M., & Beerenwinkel, N. (2017). Computational approaches for the identification of cancer genes and pathways. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, *9*(1), e1364. https://doi.org/10.1002/wsbm.1364

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., & Jiang, Y. (2020). SCDC: Bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, *22*(1), 416–427. https://doi.org/10.1093/bib/bbz166

Douglas, J., Jiménez-Silva, C. L., & Bouckaert, R. (2022). StarBeast3: Adaptive Parallelized Bayesian Inference under the Multispecies Coalescent. *Systematic Biology*, *71*(4), 901–916. https://doi.org/10.1093/sysbio/syac010

Dulak, A. M., Stojanov, P., Peng, S., Lawrence, M. S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S. E., Shefler, E., McKenna, A., Cibulskis, K., Sivachenko, A., Carter, S. L., Saksena, G., Voet, D., Ramos, A. H., Auclair, D., Thompson, K., … Bass, A. J. (2013). Exome and whole genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature genetics*, *45*(5), 478–486. https://doi.org/10.1038/ng.2591

Fearon, E. R., & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, *61*(5), 759–767. https://doi.org/10.1016/0092-8674(90)90186-i

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Biology*, *22*(3), 240–249. https://doi.org/10.1093/sysbio/22.3.240

Feng, S., Bai, M., Rivas-González, I., Li, C., Liu, S., Tong, Y., Yang, H., Chen, G., Xie, D., Sears, K. E., Franco, L. M., Gaitan-Espitia, J. D., Nespolo, R. F., Johnson, W. E., Yang, H., Brandies, P. A., Hogg, C. J., Belov, K., Renfree, M. B., … Zhang, G. (2022). Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell*, *185*(10), 1646–1660.e18. https://doi.org/10.1016/j.cell.2022.03.034

Foo, J., Leder, K., & Michor, F. (2011). Stochastic dynamics of cancer initiation. *Physical biology*, *8*(1), 015002. https://doi.org/10.1088/1478-3975/8/1/015002

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics*, *17*(3), 175–188. https://doi.org/10.1038/nrg.2015.16

Gould, S. J., & Eldredge, N. (1977). Punctuated Equilibria: The Tempo and Mode of Evolution Reconsidered. *Paleobiology*, *3*(2), 115–151. Retrieved March 15, 2024, from https://www.jstor.org/stable/2400177

Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, *481*(7381), 306–313. https://doi.org/10.1038/nature10762

Gui, P., & Bivona, T. G. (2022). Evolution of metastasis: New tools and insights. *Trends in Cancer*, *8*(2), 98–109. https://doi.org/10.1016/j.trecan.2021.11.002

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M. C., Papaemmanuil, E., Brewer, D. S., Kallio, H. M. L., Högnäs, G., Annala, M., Kivinummi, K., Goody, V., Latimer, C., O'Meara, S., Dawson, K. J., Isaacs, W., Emmert-Buck, M. R., Nykter, M., Foster, C., . . . Bova, G. S. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature*, *520*(7547), 353–357. https://doi.org/10.1038/nature14347

Han, Y., Wang, D., Peng, L., Huang, T., He, X., Wang, J., & Ou, C. (2022). Single-cell sequencing: A promising approach for uncovering the mechanisms of tumor metastasis. *Journal of Hematology & Oncology*, *15*(1), 59. https://doi.org/10.1186/s13045-022-01280-w

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., & Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, *21*(3), 349–356. https://doi.org/10.1101/gr.114751.110

Hong, M. K., Macintyre, G., Wedge, D. C., Van Loo, P., Patel, K., Lunke, S., Alexandrov, L. B., Sloggett, C., Cmero, M., Marass, F., Tsui, D., Mangiola, S., Lonie, A., Naeem, H., Sapre, N., Phal, P. M., Kurganovs, N., Chin, X., Kerger, M., . . . Hovens, C. M. (2015). Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nature Communications*, *6*, 6605. https://doi.org/10.1038/ncomms7605

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, *12*(3), e694. https://doi.org/10.1002/ctm2.694

Kashima, Y., Shibahara, D., Suzuki, A., Muto, K., Kobayashi, I. S., Plotnick, D., Udagawa, H., Izumi, H., Shibata, Y., Tanaka, K., Fujii, M., Ohashi, A., Seki, M., Goto, K., Tsuchihara, K., Suzuki, Y., & Kobayashi, S. S. (2021). Single-Cell Analyses Reveal Diverse Mechanisms of Resistance to EGFR Tyrosine Kinase Inhibitors in Lung Cancer. *Cancer Research*, *81*(18), 4835–4848. https://doi.org/10.1158/0008-5472.CAN-20-2811

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, *13*(3), 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kozlov, A., Alves, J. M., Stamatakis, A., & Posada, D. (2022). CellPhy: Accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *Genome Biology*, *23*(1), 37. https://doi.org/10.1186/s13059-021-02583-w

Kyrochristos, I. D., Ziogas, D. E., Goussia, A., Glantzounis, G. K., & Roukos, D. H. (2019). Bulk and Single-Cell Next-Generation Sequencing: Individualizing Treatment for Colorectal Cancer. *Cancers*, *11*(11), 1809. https://doi.org/10.3390/cancers11111809

Lähnemann, D., Köster, J., Fischer, U., Borkhardt, A., McHardy, A. C., & Schönhuth, A. (2021). Accurate and scalable variant calling from single cell DNA sequencing data with ProSolo. *Nature Communications*, *12*, 6744. https://doi.org/10.1038/s41467-021-26938-w

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., ... Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31. https://doi.org/10.1186/s13059-020-1926-6

Li, X., & Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science*, *13*(1), 1–6. https://doi.org/10.1038/s41368-021-00146-0

Lin, J.-R., Wang, S., Coy, S., Chen, Y.-A., Yapp, C., Tyler, M., Nariya, M. K., Heiser, C. N., Lau, K. S., Santagata, S., & Sorger, P. K. (2023). Multiplexed 3D atlas of state transitions and immune interaction in colorectal cancer. *Cell*, *186*(2), 363–381.e19. https://doi.org/10.1016/j.cell.2022.12.028

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., & Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, *53*(1), 320–328. https://doi.org/10.1016/j.ympev.2009.05.033

Liu, T., Liu, C., Yan, M., Zhang, L., Zhang, J., Xiao, M., Li, Z., Wei, X., & Zhang, H. (2022). Single cell profiling of primary and paired metastatic lymph node tumors in breast cancer patients. *Nature Communications*, *13*(1), 6823. https://doi.org/10.1038/s41467-022-34581-2

Loeffler-Wirth, H., Binder, H., Willscher, E., Gerber, T., & Kunz, M. (2018). Pseudotime Dynamics in Melanoma Single-Cell Transcriptomes Reveals Different Mechanisms of Tumor Progression. *Biology*, *7*(2), 23. https://doi.org/10.3390/biology7020023

Luzzatto, L. (2011). Somatic mutations in cancer development. *Environmental Health: A Global Access Science Source*, *10 Suppl 1*(Suppl 1), S12. https://doi.org/10.1186/1476-069X-10-S1-S12

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30. https://doi.org/10.1080/10635150500354928

Mallo, D., De Oliveira Martins, L., & Posada, D. (2016). SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology*, *65*(2), 334–344. https://doi.org/10.1093/sysbio/syv082

Martin, T. A., Ye, L., Sanders, A. J., Lane, J., & Jiang, W. G. (2013). Cancer Invasion and Metastasis: Molecular and Cellular Perspective. In *Madame Curie Bioscience Database [Internet]*. Landes Bioscience. Retrieved March 29, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK164700/

Massalha, H., Bahar Halpern, K., Abu-Gazala, S., Jana, T., Massasa, E. E., Moor, A. E., Buchauer, L., Rozenberg, M., Pikarsky, E., Amit, I., Zamir, G., & Itzkovitz, S. (2020). A single cell atlas of the human liver tumor microenvironment. *Molecular Systems Biology*, *16*(12), e9682. https://doi.org/10.15252/msb.20209682

McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019). Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine*, *9*(11), a036798. https://doi.org/10.1101/cshperspect.a036798

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code [Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality]]. *Technometrics*, *21*(2), 239–245. https://doi.org/10.2307/1268522

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., M. Mastrogianakis, G., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., Alfred Yung, W. K., Bogler, O., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., Sabo, A., ... National Human Genome Research Institute. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. https://doi.org/10.1038/nature07385

Merlo, L. M. F., Pepper, J. W., Reid, B. J., & Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, *6*(12), 924–935. https://doi.org/10.1038/nrc2013

Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, *11*(10), 685–696. https://doi.org/10.1038/nrg2841

Mitchell, K., Brito, J. J., Mandric, I., Wu, Q., Knyazev, S., Chang, S., Martin, L. S., Karlsberg, A., Gerasimov, E., Littman, R., Hill, B. L., Wu, N. C., Yang, H. T., Hsieh, K., Chen, L., Littman, E., Shabani, T., Enik, G., Yao, D., ... Mangul, S. (2020). Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biology*, *21*(1), 71. https://doi.org/10.1186/s13059-020-01988-3

Nerurkar, S. N., Goh, D., Cheung, C. C. L., Nga, P. Q. Y., Lim, J. C. T., & Yeong, J. P. S. (2020). Transcriptional Spatial Profiling of Cancer Tissues in the Era of Immunotherapy: The Potential and Promise. *Cancers*, *12*(9), 2572. https://doi.org/10.3390/cancers12092572

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science (New York, N.Y.)*, *194*(4260), 23–28. https://doi.org/10.1126/science.959840

Olvera, A., Noguera-Julian, M., Kilpelainen, A., Romero-Martín, L., Prado, J. G., & Brander, C. (2020). SARS-CoV-2 Consensus-Sequence and Matching Overlapping Peptides Design for COVID19 Immune Studies and Vaccine Development. *Vaccines*, *8*(3), 444. https://doi.org/10.3390/vaccines8030444

Posada, D. (2020). CellCoal: Coalescent Simulation of Single-Cell Sequencing Samples. *Molecular Biology and Evolution*, *37*(5), 1535–1542. https://doi.org/10.1093/molbev/msaa025

Quinn, J. J., Jones, M. G., Okimoto, R. A., Nanjo, S., Chan, M. M., Yosef, N., Bivona, T. G., & Weissman, J. S. (2021). Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science*, *371*(6532), eabc1944. https://doi.org/10.1126/science.abc1944

Raphael, B. J., Dobson, J. R., Oesper, L., & Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: Computational approaches to enable precision medicine. *Genome Medicine*, *6*(1), 5. https://doi.org/10.1186/gm524

Rogiers, A., Lobon, I., Spain, L., & Turajlic, S. (2022). The Genetic Evolution of Metastasis. *Cancer Research*, *82*(10), 1849–1857. https://doi.org/10.1158/0008-5472.CAN-21-3863

Sarkar, S., Horn, G., Moulton, K., Oza, A., Byler, S., Kokolus, S., & Longacre, M. (2013). Cancer Development, Progression, and Therapy: An Epigenetic Overview. *International Journal of Molecular Sciences*, *14*(10), 21087–21113. https://doi.org/10.3390/ijms141021087

Schneider, T. D. (2002). Consensus Sequence Zen. *Applied bioinformatics*, *1*(3), 111–119. Retrieved March 31, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1852464/

Schwartz, R., & Schäffer, A. A. (2017). The evolution of tumour phylogenetics: Principles and practice. *Nature reviews. Genetics*, *18*(4), 213–229. https://doi.org/10.1038/nrg.2016.170

Schwede, M., Waldron, L., Mok, S. C., Wei, W., Basunia, A., Merritt, M. A., Mitsiades, C. S., Parmigiani, G., Harrington, D., Quackenbush, J., Birrer, M. J., & Culhane, A. C. (2020). The impact of stroma admixture on molecular subtypes and prognostic gene signatures in serous ovarian cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *29*(2), 509–519. https://doi.org/10.1158/1055-9965.EPI-18-1359

Shestak, A. G., Bukaeva, A. A., Saber, S., & Zaklyazminskaya, E. V. (2021). Allelic Dropout Is a Common Phenomenon That Reduces the Diagnostic Yield of PCR-Based Sequencing of Targeted Gene Panels. *Frontiers in Genetics*, *12*, 620337. https://doi.org/10.3389/fgene.2021.620337

Siegel, R. L., Jemal, A., Wender, R. C., Gansler, T., Ma, J., & Brawley, O. W. (2018). An assessment of progress in cancer control. *CA: A Cancer Journal for Clinicians*, *68*(5), 329–339. https://doi.org/10.3322/caac.21460

Smith, M. R. (2020). Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics*, *36*(20), 5007–5013. https://doi.org/10.1093/bioinformatics/btaa614

Soerjomataram, I., & Bray, F. (2021). Planning for tomorrow: Global cancer incidence and the role of prevention 2020-2070. *Nature Reviews. Clinical Oncology*, *18*(10), 663–672. https://doi.org/10.1038/s41571-021-00514-z

Somarelli, J. A., Ware, K. E., Kostadinov, R., Robinson, J. M., Amri, H., Abu-Asab, M., Fourie, N., Diogo, R., Swofford, D., & Townsend, J. P. (2017). PhyloOncology: Understanding cancer through phylogenetic analysis. *Biochimica et biophysica acta. Reviews on cancer*, *1867*(2), 101–108. https://doi.org/10.1016/j.bbcan.2016.10.006

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanta, G., & Bonin, S. (2018). Overview on Clinical Relevance of Intra-Tumor Heterogeneity. *Frontiers in Medicine*, *5*. https://doi.org/10.3389/fmed.2018.00085

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724. https://doi.org/10.1038/nature07943

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249.

Szöllősi, G. J., Tannier, E., Daubin, V., & Boussau, B. (2015). The Inference of Gene Trees with Species Trees. *Systematic Biology*, *64*(1), e42–e62. https://doi.org/10.1093/sysbio/syu048

Vachaspati, P., & Warnow, T. (2017). FastRFS: Fast and accurate Robinson-Foulds Supertrees using constrained exact optimization. *Bioinformatics*, *33*(5), 631–639. https://doi.org/10.1093/bioinformatics/btw600

Vendramin, R., Litchfield, K., & Swanton, C. (2021). Cancer evolution: Darwin and beyond. *The EMBO Journal*, *40*(18), e108389. https://doi.org/10.15252/embj.2021108389

Vincze, O., Colchero, F., Lemaître, J.-F., Conde, D. A., Pavard, S., Bieuville, M., Urrutia, A. O., Ujvari, B., Boddy, A. M., Maley, C. C., Thomas, F., & Giraudeau, M. (2022). Cancer risk across mammals. *Nature*, *601*(7892), 263–267. https://doi.org/10.1038/s41586-021-04224-5

Wang, Q. (2016). Cancer predisposition genes: Molecular mechanisms and clinical impact on personalized cancer care: Examples of Lynch and HBOC syndromes. *Acta Pharmacologica Sinica*, *37*(2), 143–149. https://doi.org/10.1038/aps.2015.89

Warnow, T. (2015). Concatenation Analyses in the Presence of Incomplete Lineage Sorting. *PLoS Currents*, *7*, ecurrents.currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7. https://doi.org/10.1371/currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7

Werner, B., Case, J., Williams, M. J., Chkhaidze, K., Temko, D., Fernández-Mateos, J., Cresswell, G. D., Nichol, D., Cross, W., Spiteri, I., Huang, W., Tomlinson, I. P. M., Barnes, C. P., Graham, T. A., & Sottoriva, A. (2020). Measuring single cell divisions in human tissues from multi-region sequencing data. *Nature Communications*, *11*(1), 1035. https://doi.org/10.1038/s41467-020-14844-6

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., & Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nature genetics*, *48*(3), 238–244. https://doi.org/10.1038/ng.3489

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhou, W.-m., Yan, Y.-y., Guo, Q.-r., Ji, H., Wang, H., Xu, T.-t., Makabel, B., Pilarsky, C., He, G., Yu, X.-y., & Zhang, J.-y. (2021). Microfluidics applications for high-throughput single cell sequencing. *Journal of Nanobiotechnology*, *19*(1), 312. https://doi.org/10.1186/s12951-021-01045-6

Zhu, L., Jiang, M., Wang, H., Sun, H., Zhu, J., Zhao, W., Fang, Q., Yu, J., Chen, P., Wu, S., Zheng, Z., & He, Y. (2021). A narrative review of tumor heterogeneity and challenges to tumor drug therapy. *Annals of Translational Medicine*, *9*(16), 1351. https://doi.org/10.21037/atm-21-1948