Implementing tumour growth models in BEAST2

Author: Yuan Xu

(Student ID: 291641167, UPI: yxu927, Current Programme: Master of Science)

Supervisor(s): Alexei J. Drummond, Kylie Chen, Jonathan Klawitter

Thomas Building, School of Biological Sciences, University of Auckland, Auckland, New Zealand.

Key Words: Bayesian phylogenetics, coalescent, growth model, single-cell, tumours

Word Count: 4572 words

21 Pages

Abstract

Evolutionary analysis based on single-cell sequencing data will allow us to explore cell population demographics and tumour growth rates. Specific models are required to make demographic inferences about cell populations. Models for cancer cell growth, such as the Gompertzian growth, can be used for modeling growth beyond the early exponential phase. However, the Gompertz growth model has not yet been fully integrated into a Bayesian phylogenetics framework.

In this project, we aim to implement a Gompertz model for inference and a forward sampler for simulation. We will implement a Gompertzian model to be used with the coalescent model for tumor growth. This model will be integrated into the Bayesian phylogenetics software BEAST2. We hope this can improve estimates of tumor age and growth rates of tumour cells. Using both simulated data for validation and real single-cell cancer data for evaluation, we will assess our model's performance using priors from empirical studies. Future applications of this model could include inferring cancer cell driver mutations based on Bayesian phylogenetic models. The methods of our study can be potentially applied to estimate cell evolution histories, population growth dynamics, and identify parameters of clinical significance in cancers.

1. Background

Tumour biology has many complexities, therefore modelling these processes accurately is an important step to understanding tumour evolution and growth. Mathematical models assume cells proliferate at a constant rate. whereas the Mendelsohn model adjusts the growth to be proportional to the surface area of the tumour. For instance, the exponential model assumes cells proliferate at a constant rate, whereas the Mendelsohn model adjusts the growth to be proportional to the surface area of the tumour. The Gompertz model has an S-shaped curve that allows the growth rate to decrease after reaching a certain capacity (Heesterman et al., 2019). Using statistical phylogenetics methods, we can estimate phylogenetic trees that inform us about ancestral relationships, as well as evolutionary parameters. Using sequencing information, we can gain an understanding of evolutionary patterns and ancestral relationships. Bayesian methods such as BEAST2 (Bouckaert et al., 2019), provide a framework for incorporating prior knowledge in the analysis. Coalescent models provide a natural way to incorporate growth dynamics into the tree prior for Bayesian inference.

1.1 Biology of Tumours

Cancer originates from the uncontrolled proliferation of a single cell, which accumulates genetic mutations and distinct biomarkers over time (D. Hanahan & Weinberg, 2000). This leads to the formation of a clonal family of cells with similar genetic characteristics (Nowell, 1976). During the clonal evolution process, somatic mutations, genetic drift, natural selection, and the effects of spatial heterogeneity, give rise to cell populations carrying various genotypes, referred to as subclones(Tarabichi et al., 2021). Increasingly, scientific literature emphasizes that the tumour microenvironment, especially the intercellular communication between tumour cells and the normal cells surrounding them, plays an important role in the development and progression of tumours (Apte et al., 2006; Douglas Hanahan & Weinberg, 2011; Junttila & de Sauvage, 2013).

The complexities in the tumour microenvironment are challenging to describe and simulate in vivo or in vitro methods (Wu & Swartz, 2014). The perspectives on tumour biology are multi-layered and complex (Casas-Selves & Degregori, 2011; Gatenby & Gillies, 2008).

1.2 Mathematical Models of Tumour Growth

With the advancement of personalized cancer treatment, mathematical modeling offers the possibility to quantitatively describe pathophysiological properties of tumours (Agur & Vuk-Pavlovic, 2012). It can assist in understanding the progression of tumours, such as predicting the efficacy of new therapies, to guide the optimization of cancer treatment plans and the development of personalized dosing strategies(Agur, Elishmereni, & Kheifetz, 2014; Agur, Halevi-Tobias, Kogan, & Shlagman, 2016).

Over the past several decades, researchers have proposed numerous ordinary differential equation (ODE) models suitable for tumour growth dynamics (Gerlee, 2013; Wodarz & Komarova, 2009). These models are frequently used to predict patterns of tumour growth (Ghaffari Laleh et al., 2022). However, the basis for choosing a growth model is often for convenience of mathematical

analysis rather than biological accuracy (Gerlee, 2013). Multiple efforts have been made to identify the optimal ODE growth model by fitting tumour growth data(Benzekry et al., 2014; Sarapata & de Pillis, 2014; Vaidya & Alexandro, 1982). However, the results have been inconsistent, indicating that the choice of model is largely influenced by the type of tumour (Benzekry et al., 2014; Sarapata & de Pillis, 2014). Without clear guidance, choosing an inappropriate growth model can pose challenges (Gerlee, 2013), potentially impacting the prediction of treatment outcomes (Wodarz, 2013; Wodarz & Komarova, 2009). Here, we present some commonly used ODE growth models to describe tumour dynamics. The models predict the growth of a tumour by describing the change in tumour volume, V, over time. The model equations used are displayed in Table 1, with a, b, and c as model parameters.

Model	Equation
Exponential	$V = aV$
Mendelsohn	$V = aV^b$
Logistic	$V = aV(1 - \frac{V}{b})$
Linear	$V = \frac{aV}{(V + b)}$
Surface	$V = \frac{aV}{(V + b)^{\frac{1}{3}}}$
Gompertz	$V = aV \ln\frac{b}{(V + c)}$

Table 1: ODE models of tumour growth

Exponential: During the early stages of tumour growth, cell division follows a consistent pattern, producing two offspring cells with each division (Shahriyari & Komarova, 2013), leading to an exponential growth trajectory(Collins, Loeffler, & Tivey, 1956). This model suggests that the growth rate is directly proportional to the total number of cells, with the proportionality constant 'a' characterizing the rate of growth. While this model has been widely adopted for its accuracy in depicting early tumour growth dynamics (Rodriguez-Brenes, Komarova, & Wodarz, 2013; Summers, 1966), it fails to describe the later stages of tumour progression, especially when angiogenesis and nutritional constraints come into play (Benzekry et al., 2014; Gerlee, 2013).

Mendelsohn: Mendelsohn extended the exponential growth model (Mendelsohn, 1963) such that growth rate is directly proportional to a power of the cell count.

Logistic: The Logistic model, also referred to as the Pearl-Verhulst model, was formulated by Pierre Francois Verhulst in 1838 (Verhulst, 1838). Here, the population growth is constrained by a carrying capacity 'b'. As the population size increases, the growth rate linearly declines until it reaches a plateau hitting the carrying capacity.

Linear: The linear model describes an initial exponential growth trend, which eventually transitions into a steady growth phase. In this model formulation, ‘a/b’ is the initial exponential

growth rate, ‘a’ is the subsequent steady growth. This model was employed in earlier research to analyze the growth dynamics of cancer cell colonies (Dethlefsen, Prewitt, & Mendelsohn, 1968).

Surface: This model operates on the premise that cell proliferation is restricted to the periphery, implying only the outermost layer of the tumour undergoes division, whereas the internal cells of a solid tumour remain dormant (Patt & Blackford, 1954).

Gompertz: Benjamin Gompertz initially introduced the Gompertz model in 1825 (Gompertz, 1825). Early research indicated that transplanted or primary tumours have initial growth phases that can be described by the exponential model (Benzekry et al., 2014). However, in the later phases, as angiogenesis and nutrient depletion take hold, there's a marked deceleration in the growth rate (Douglas Hanahan & Weinberg, 2011). This slowdown can be described by the Gompertz equation, an S-shaped curve that is asymmetric relative to the inflection point (Yin, Goudriaan, Lantinga, Vos, & Spiertz, 2003). This growth can be interpreted as exponential process which is constrained by an exponential retardation as time progresses (Sullivan & Salmon, 1972).

1.3 Statistical Phylogenetics

Phylogenetics explores the evolution and ancestral history of species through the analysis of genomic sequences (Penny, 2004). By using phylogenetic tools to reconstruct phylogenetic trees from sequencing data, we can construct the evolutionary lineage of tumours, enabling a deeper understanding of their origins and evolution (Schwartz & Schaffer, 2017). These analytical methods primarily rely on genomic sequences such as mutations. These methods are essential for identifying key driver mutations (Leung et al., 2017), pathways of tumour evolution (Leung et al., 2017; Williams, Cook, & Smerdon, 2022), and inferring the timing of metastases (Hong, Shpak, & Townsend, 2015). Popular phylogenetic tools mainly utilize Bayesian algorithms and require the integration of various evolutionary models such substitution models, molecular clocks, and tree priors (dos Reis, Donoghue, & Yang, 2016). Next, we provide a summary of the Bayesian phylogenetic approach.

1.4 Bayesian Methods

Given the sequence data D and a model characterized by parameters θ, Bayes' theorem states:

$$P(\theta|D) = \frac{P(\theta) \cdot P(D|\theta)}{P(D)}.$$

The prior distribution P(θ) reflects our initial knowledge of the model parameters based on prior research and experience, such as substitution model parameters, molecular clock model parameters, and evolutionary trees. Under the given model parameter θ, the likelihood function P(D | θ) describes the probability of observing the data D. The likelihood value of the phylogenetic model can be calculated using Felsenstein's pruning algorithm (Felsenstein, 1981). The posterior distribution P(θ | D) updates our knowledge of the model parameter θ after observing the data and is proportional to the prior distribution P(θ) and likelihood function

$P(D \mid \theta)$. To sample the posterior distribution, we can adopt sampling techniques such as Markov Chain Monte Carlo (MCMC) (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Yang & Rannala, 1997). Software supporting Bayesian phylogenetic inference includes RevBayes (Hohna et al., 2016), MrBayes (F. Ronquist & Huelsenbeck, 2003; Fredrik Ronquist et al., 2012), BEAST1 (Drummond & Rambaut, 2007) and BEAST2 (Bouckaert et al., 2019).

1.4.1 Substitution Models

Substitution models use continuous-time discrete-state Markov chains to model the evolutionary change of homologous sites over time. The level of detail modelled can be nucleotides, amino acids, or more abstract discrete features like absence/presence (0/1) of morphological features. Based on character state transition probabilities, we can infer the likelihood of the phylogenetic tree under multiple sequence alignments (Felsenstein, 1981). DNA substitution models are established based on the equilibrium frequencies of nucleotides and instantaneous changing rates. The Jukes-Cantor model(Jukes & Cantor, 1969) is considered a benchmark model characterized by its identical frequencies and rates. Later models, such as K80 (Kimura, 1981) and HKY (Hasegawa, Kishino, & Yano, 1985), differentiate between rates for within and between purines and pyrimidines nucleotide changes. The GTR model (Tavaré, 1986) is viewed as the most generalized time-reversible model, with both its frequencies and rates set as free parameters.

1.4.2 Molecular Clocks

To understand the dynamics of evolution, we often rely on a molecular clock for the character state changes (Zuckerkandl, 1962; Zuckerkandl & Pauling, 1965). The model proposed by Zuckerkandl in 1962 use a constant molecular clock (Zuckerkandl, 1962), where the rate of evolution was considered constant. However, in many empirical observations, this assumption of a constant evolutionary rate does not always hold true (Cutler, 2000; Gillooly, Allen, West, & Brown, 2005; Kishino & Hasegawa, 1989). In particular for the tumour environment, where heterogeneity in proliferation and mutation rates are often observed (Douglas Hanahan & Weinberg, 2011). During the development of tumours, some studies suggest that the mutation rate gradually increases(Douglas Hanahan & Weinberg, 2011). To accommodate rate variation, researchers have introduced models such as the local clock (Drummond & Suchard, 2010) and the relaxed clock (Aris-Brosou & Yang, 2002; Drummond, Ho, Phillips, & Rambaut, 2006).

1.4.3 Tree Priors

Tree priors express how information is passed from the current generation to the next, relying on population genetics or birth-death branching models. In these model-based priors, organisms follow processes like pure birth or Yule model (Yule, 1925), birth-death, or coalescent to describe the generation of offspring (Fisher, 1958; Kingman, 1982; Wright, 1931). Traditional population genetics models assume a population size that is constant and remains stable over time (Kingman, 1982). However, there are scenarios where the population size can vary, such as tumour growth often displaying an S-shaped Gompertzian trajectory (Comen, Norton, & Massague, 2011; Norton,

1988; Norton & Massagué, 2006) and epidemics that show exponential growth in their initial stages (Gire et al., 2014). To accommodate changes in population size, researchers developed models like the Bayesian skyline coalescent (Drummond, Rambaut, Shapiro, & Pybus, 2005).

1.5 Coalescent Inference

In recent years, with the growth of computational processing capabilities and the availability of gene sequence data, have led to increasing efforts to infer evolutionary histories. Most current methods are based on coalescent theory, which is a stochastic model that describes how population genetic dynamics affect the lineage structure of sampled gene sequences. Inference methods of coalescent theory allow for direct estimation of various population genetic parameters from gene sequence data. This includes recombination (Fearnhead & Donnelly, 2001; Griffiths & Marjoram, 1996; Kuhner, Yamato, & Felsenstein, 2000), population differentiation (Bahlo & Griffiths, 2000; Beerli & Felsenstein, 2001), and changing population sizes (Beaumont, 1999; Drummond, Nicholls, Rodrigo, & Solomon, 2002; Kuhner, Yamato, & Felsenstein, 1998).

The variable population coalescent models designed by Griffiths and Tavare (Griffiths & Tavare, 1994) and Donnelly and Tavare (Donnelly & Tavare, 1995) allow for retracing the historical dynamics of populations based on existing gene sequences. This model was further extended by Rodrigo and Felsenstein (Rodrigo & Felsenstein, 1999), allowing analysis of sequences obtained at different time points, such as rapidly evolving pathogens or ancient DNA sequences.

Utilizing genetic data to infer the history of populations provides valuable tools for hypothesis testing in multiple subfields, including anthropology (Stephens et al., 1998), epidemiology (Joy et al., 2003; Pybus, Charleston, Gupta, Rambaut, Holmes, & Harvey, 2001), conservation biology (Roman & Palumbi, 2003), and ecology (Flanagan et al., 2004; Storz, Beaumont, & Alberts, 2002).

Population genetic modeling of population growth is crucial for many phylogenetic problems such as the inference of virus evolution (Moya, Holmes, & González-Candelas, 2004). The coalescent, as a stochastic process, presents us with the distribution of approximate ancestral histories (Spence, Steinrücken, Terhorst, & Song, 2018). Most coalescent-based estimation methods focus on a single genealogy, usually obtained through standard phylogenetic methods (Liu, Yu, Kubatko, Pearl, & Edwards, 2009). However, these reconstructed genealogies often come with significant uncertainty (Riester, Stadler, & Klemm, 2009). We use Monte Carlo methods to sample parameters of the coalescent model given these uncertainties (Jäckel, 2002).

Population models such as the coalescent, provide a way to describe how the effective population size evolves over time. Each model has population parameters. Commonly used population models include: constant size (single parameter), exponential growth (a fixed growth rate over time; two parameters), and logistic growth (a declining growth rate over time; three parameters) (Drummond et al., 2005). Using Bayesian analysis to estimate these population parameters can model the historical changes of populations (Drummond et al., 2002; Kuhner, Yamato, & Felsenstein, 1998; Pybus, Rambaut, & Harvey, 2000).

1.6 BEAST2

The BEAST2 software package provides a unified Bayesian statistical framework for parameter estimation and hypothesis testing of evolutionary models for molecular sequence data(Bouckaert et al., 2019). This approach combines prior knowledge with new data to perform Bayesian inference of evolutionary trees. The models applied encompass phylogenetics, coalescent-based population genetics, multi-locus demographic models, phylogeography, divergence time estimation, gene and species tree inference (Thomé & Carstens, 2016). The BEAST2 software uses Markov Chain Monte Carlo (MCMC) sampling to perform the inference (Bouckaert et al., 2019). BEAST2 offers many sequence evolution models and tree-based models for within-species and between-species sequence data (Ogilvie, Bouckaert, & Drummond, 2017). Here, we achieve evolutionary inference by specifying evolutionary models (substitution models, demographic models, tree prior models, clock models and optionally insertion-deletion models). Based on these methods, we can conduct evolutionary analyses based on flexible models tailored for various scenarios.

Figure 1 shows the Bayesian model used in the BEAST2 software, including the evolutionary model and model parameters. Here, we assume the data (D) evolves according to a continuous time Markov chain on a phylogeny (PhyloCTMC). Components of the model include the substitution model, the clock model and the tree prior. For example, the HKY substitution model has an instantaneous rate matrix "Q" with parameters for the base frequencies (π) and kappa (κ). The tree prior model gives the probability of the time tree (T), and the clock rate is defined by the clock model with a rate (r). The clock model can be a strict or relaxed clock.

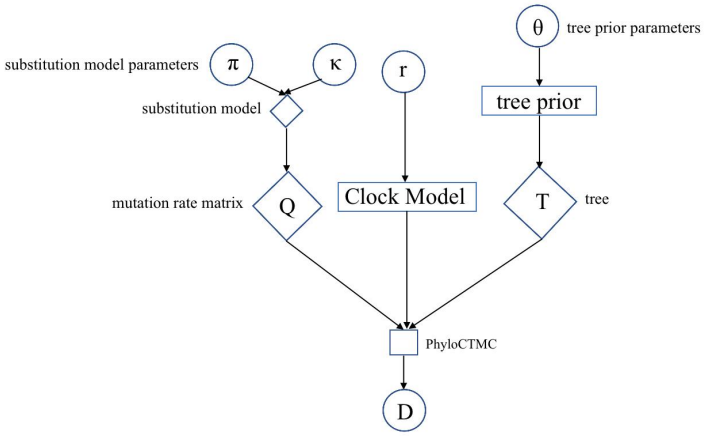

Figure 1: A complete phylogenetic model, including the data (D) generated by the phylogenetic continuous time Markov chain, the relevant parameters of the substitution model (such as π and κ of the HKY model), the clock model related to the branching rate, and the tree determined by the tree prior.

2. Research Objectives

The research in this proposed study will build a joint inference of cancer-specific population growth dynamics and phylogenetics in a fully probabilistic framework for tumour cell evolution analysis of single-cell DNA sequencing data (scDNA-seq). Specifically, we want to address the following questions:

How can we develop a Bayesian phylogenetic model to simulate the evolution of tumours growth more accurately? Can key features of the nature and timing of tumour evolution be inferred by applying the newly implemented growth model to real tumour data?

In order to achieve quantitative estimates of timing, growth rates, and lineage divergence times for cancer data, it is necessary to develop models for demographic inference specific to cell populations based on single-cell sequencing data. Traditional population models, such as birth-death and pure birth models (Yule models), do not match the growth pattern of cancer cells, so we believe a coalescent-based model is more appropriate. The main research is to develop cellular growth models of cells for a single patient. However, general population growth models such as the Skyline models can be biased depending on the way the data was pre-processed. Models that more realistically describe cancer cell growth, such as Gompertzian growth, are not yet fully integrated into a full Bayesian framework.

2.1 Aim and Objective

In this study, I will combine coalescent theory and density dependent cell growth dynamics to develop a cell tree prior based on Gompertz growth for solid tumours. The above model will be implemented in the Bayesian phylogeny software BEAST 2.

AIM1: Develop a Gompertz cell growth model for somatic cell evolution in BEAST 2.

AIM2: Evaluate the Gompertz growth model and coalescent tree prior model using simulated single-cell DNA data.

AIM3: Demonstrate the Gompertz coalescent model on real single-cell DNA datasets from cancer patients.

3. Research Design

3.1 Develop a Gompertz cell growth model for somatic cell evolution in BEAST2

3.1.1 Gompertz function and Integral for Gompertz

I will apply the Gompertz equation for modeling tumour growth. As mentioned earlier, as the tumour grows larger the growth rate should plateau (become stable). There are fewer cells at the boundaries (proliferative zone) compared to the larger mass of cells in the center of the tumour. Late-stage tumour growth is usually not exponential but is well-fitted by the Gompertz model which has a decay in growth rate at later stages.

In 1988, Larry Norton proposed the Gompertz equation for breast cancer growth:

$$N(t) = N(0) \bullet \exp\ (k \bullet (1 - \exp\ (-bt)))$$

where N(t) is the population function over time, which gives us the population at time t. This is a function of N(0) which is the initial population size, b is the initial growth rate of tumour growth, N(∞) is the limiting population size or carrying capacity and k= log[N(∞)/N(0)] is a ratio.

Using the Griffiths and Tavare equation, we compute the coalescent density for time intervals by integrating over 1/N(t) (Griffiths & Tavare, 1994). By iterating across all time intervals, we can calculate the density of the entire tree. This provides a generalized coalescent framework for varying population sizes, N=N(t) (Griffiths & Tavare, 1994). They showed that the coalescent density for the first coalescence event being at time t in the past given n lineages is:

$$f(t) = \frac{1}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(x)} dx \right)$$

and here in my research, the N(t) I use is the Gompertz function.

Here, I will use Java code to integrate the Gompertz model and coalescent model. I will implement a GompertzFunction class which implements the PopulationFunction interface. The PopulationFunction provides a blueprint for population functions and requires classes to implement several key methods. Fig. 1 shows the main implementation steps.

The three model parameters b, N0, N_inf to the GompertzFunction are sampled during MCMC. The GompertzFunction can then be passed in as the PopulationModel input to the Coalescent class. We will describe in detail the structure of our GompertzFunction class below.

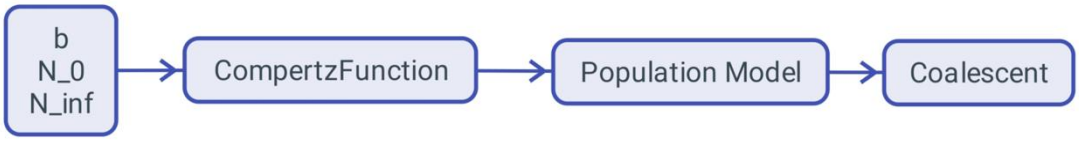

Figure 2: Diagram of model parameters for the Gompertz function, and how the GompertzFunction is used as an

input Population Model to the Coalescent model.

3.1.2 Implementation Details

A population function is defined to represent the Gompertz coalescent model. This can be achieved by implementing the Population Function interface, which is to describe the size and change of a population. I will have three input parameters to the GompertzFunction class: N0, Ninf, b. These model parameters can be sampled during the MCMC. The implementation is achieved by building a custom population function for the Gompertz growth model.

```
double getPopSize(double t)
```

The getPopSize(double t) method above is used to calculate the population size at a given time "t" using the Gompertz function. This method receives a time parameter t and returns the value of the population function N(t) at time t. This method gives us a convenient way to query the population size at a specific point in time.

```
double getIntensity(double t)
double getInverseIntensity(double x)
```

The methods getIntensity(double t) and getInverseIntensity(double x) can either use the analytical form, or be calculated numerically using numeric integration methods.

The getIntensity(double t) method contains parameter ‘t’, return value of inverse demographic intensity function. It calculates the demographic intensity by integrating 1/N(s) from 0 to t.

On the other hand, getInverseIntensity(double x) method takes in an input parameter x, parameter ‘x’ used to represent the coalescent intensity (x = integral 1/N(s) ds from 0 to t) and returns the inverse demographic intensity function. This can be used for simulating coalescent intervals when sampling from the prior.

```
double getIntergral(double start, double finish)
```

The getIntegral(double start, double finish) method performs integration of 1/N(t) within the specified time range (between start and finish) and returns the integral value. This is used to calculate the coalescent density for a time interval, and can also be used to calculate the coalescent density over the whole tree.

3.1.3 Implementing the Gompertz Coalescent model

This part we describe the link between gene lineages and the Gompertz coalescent model. All descendent cells that develop from the same ancestral cell have a common ancestor that can be traced back to their origin. The time that two samples coalesce to a common ancestor is called a coalescent time. For different population sizes, the probability of coalescent events will be different. The generalization of the coalescent for the case where the population size changes over time N = N(t) is given by Griffiths and Tavare (1994). When we consider a population whose population changes over time according to some growth model (such as the Gompertz function), we need to calculate the population size at a specific time t. To perform this calculation, we have to integrate the population function N(t).

We use the 'PopulationFunction' interface along with the 'Coalescent' class, to calculate the probability of a tree by using a specific population size function.

Coalescent class. The main function of this class is to calculate the probability of a tree based on the population size function using the 'PopulationFunction'. This class has a 'popSizeInput' parameter and requires an implementation of the 'PopulationFunction' interface for the population model.

Integration and application. The 'PopulationFunction' interface provides a common framework for different population models, and the 'Coalescent' class can utilize these models to simulate and analyze the probability distribution of evolutionary trees. By applying specific population models to the 'Coalescent' class, we can observe and analyze the impact of different population dynamics and model parameters on evolutionary inference.

3.2 Evaluate the Gompertz growth model and coalescent tree prior model using simulated single-cell DNA data

3.2.1 Simulating under the Gompertz Coalescent model

To perform simulation under the coalescent, we need to take the inverse. The simulation model will be embedded in LPhyStudio (Drummond, Chen, Mendes, & Xie, 2023). In order to verify the performance of our model, we will evaluate the performance of the Gompertz coalescent model by simulating single-cell sequencing DNA data using the LPhy simulation framework. This framework allows integration with phylogenetic models including substitution models, population dynamics models, and error models. We will simulate the phylogenetic tree based on coalescent times using the Gompertz growth function. At the root of the tree, the equilibrium frequencies of GT16 substitution model (Kozlov, Alves, Stamatakis, & Posada, 2022) will be used to simulate the ancestral genotypes. Next, the evolution process along the tree branches is simulated using the GT16 substitution model to obtain the terminal sequences (leaves of the tree). These terminal sequences are our real genotypes from our simulation. Since we are simulating single-cell DNA data, we also consider errors (specifically allelic dropout, and sequencing/amplification errors). We will add errors to the diploid genotypes at the leaves.

I will extend the LPhy and LPhyStudio software to implement a simulator for the Gompertz coalescent model. The benefits of using the LPhy framework is that it provides a user-friendly graphical interface for specification of phylogenetic models, visualization of simulation data, and generation of textual descriptions and graphical diagrams. Users can intuitively define and edit phylogenetic models through LPhyStudio, and generate corresponding model specifications using the LPhy language. Figure 2 shows a screenshot of a set of models developed for single-cell sequencing DNA data using the Phylonco package (Chen, Moravec, Gavryushkin, Welch, & Drummond, 2022) in LPhyStudio: the GT16 substitution model and the GT 16 error model to illustrate the functions implemented by the Lphy software. Here, using LPhy scripts, LPhyStudio provides graphical views and text descriptions of the corresponding model, while displaying data results simulated based on the model. This graphical modeling method can construct new models from existing ones by altering the probability distribution assigned to random variables, or by changing the functions assigned to deterministic variables, or by introducing new relationships between variables. Our simulation will use a similar LPhy script, but with modifications to the

coalescent function, and with the addition of a Gompertz function and priors.

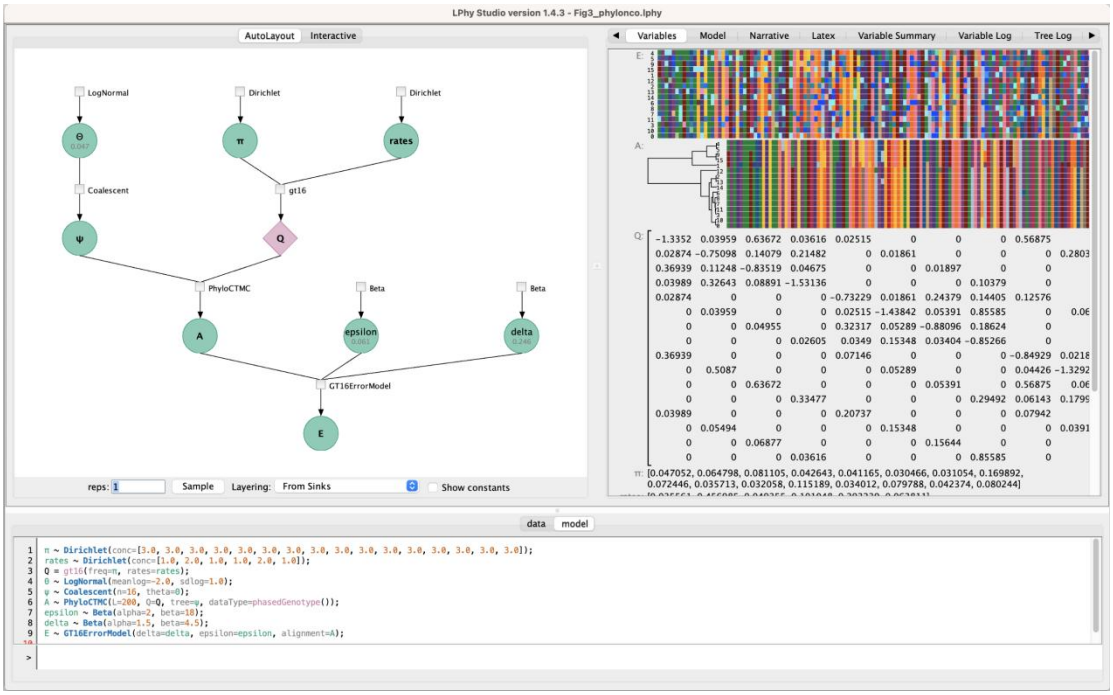

Figure 3: Screenshot of applying the GT16 model and the GT16 error model in LPhyStudio. The left side shows a probabilistic graphical model composed of different nodes, and the right side shows the text narrative format corresponding to the data and phylogenetic model, the phylogenetic tree based on model simulation and the corresponding diploid nucleotide genotypes. Diamond-shaped nodes are represented as deterministic nodes (such as the transfer "Q" matrix of CTMC), and circular nodes are represented as random nodes (such as the "θ" of the group size).

Starting with the LPhy script shown in the picture above, which displays the above GT 16 substitution model as well as the GT 16 error model. The GT16 error model simulates single-cell diploid nucleotides considering potential errors that may occur in single-cell sequencing data. The script defines 16 sequences, among which the "A" sequence is generated according to the GT16 substitution model, and its rate and genotype frequency are sampled from the Dirichlet distribution. The tree is simulated according to a constant coalescent model with theta following a Lognormal distribution. The process of generating the observed noisy sequence "E" involves applying sequencing/amplification error (epsilon) and allele loss error (delta) derived from a beta distribution to sequence "A". In the Dirichlet generators, the parameter "conc" represents the concentration parameter.

Incorporating the Gompertz coalescent into our simulation framework, I will define specific priors on the model parameters. These priors will be chosen based on empirical studies and previous papers to provide realistic evolutionary scenarios. The parameters b, N0 and Ninf will be drawn from some distributions. For example, we could use a Normal distribution for the growth rate parameter b~Normal(5,1). Parameters N0 and Ninf will be specified following Lognormal distributions. The Gompertz growth model can be defined as growthModel~Gompertz(b, N0, Ninf). Finally, the coalescent model for 16 sequences can follow the form Coalescent(n=16, growthModel).

3.2.2 Model validation using simulations

The tip sequences and prior information will be input into the Bayesian framework BEAST2 to perform phylogenetic inference. The inferences obtained by BEAST2 will provide us with estimates of the tree and model parameters. We will compare the 95% Highest Posterior Density interval (HPD) of our estimates with the true parameters from the simulated process. To compare the properties of phylogenetic trees, we will use some common metrics, including tree length, tree height, and topological similarity. For tree topology we will use metrics such as Robinson-Foulds (RF) or Subtree Prune and Regraft (SPR) to evaluate topological differences between the simulated and true tree (Robinson & Foulds, 1981; Swofford, 1996). Here we will perform a well-calibrated study (Dawid, 1982). We expect that the 95% interval should cover the true value 95% of the time. This means that when we perform 100 simulations and use the same prior for inference, we expect that the true value will be within the 95% interval 95 times out of 100. If our implementation passes this test, we can say that it is "well calibrated".

3.2.3 Comparison of Gompertz with Exponential and logistic functions

Next, we want to look at the bias introduced when simulating under the Gompertz coalescent, and performing inference using other population functions (Exponential or Logistic functions).

First, we simulate datasets under the Gompertz coalescent. Then, we perform inference under three scenarios:

Gompertz Function: As our control, we expect the Gompertz function to be the most accurate since the data originates from a Gompertz coalescent.

Exponential Function: Here, we anticipate deviations. Given that the Exponential model entails a constant rate of growth, we will assess the bias in the estimated growth curve.

Logistic Function: With its characteristic S-shape, the Logistic model presents a bounded growth representation. This analysis will reveal whether such a model can accurately capture or misrepresents the demographics of populations simulated under Gompertz coalescent.

We will compare the estimate population functions with the true Gompertz function we used for simulation.

3.3 Evaluation on real single-cell DNA cancer data

For real data, we will use single-cell colorectal cancer data from seven patients produced by David Posada's lab (with 24 - 79 cells, 13,000 - 55,000 single nucleotide variants). Drawing from previous work, we will select relevant priors from empirical studies.

We will use the GT16 substitution model and error model, with the Gompertz coalescent model to perform inference on the real data. Our priors will be chosen with guidance from our collaborators and previous literature.

4 Implications and Challenges

The Gompertz coalescent model moves us one step forward in accurately representing cell growth dynamics for tumour growth. This model should better capture the cell growth dynamics of tumour cells. One limitation is that the current formulation is unable to account for different growth rates in different groups of cells (e.g., population structure, metastatic groups). We believe a multi-state or structured coalescent model would be required to achieve this. A key challenge in our work is the selection of priors based on tumour biology or medical knowledge. We believe the choice of priors may have significant effects on the population and phylogenetic inference such as the branch lengths. Our work is significant in that it will be embedded in a rich and fully Bayesian phylogenetics framework BEAST2 and will enable analyses to have models better tailored towards cancer phylogenetics.

References

Agur, Z., Elishmereni, M., & Kheifetz, Y. (2014). Personalizing oncology treatments by predicting drug efficacy, side-effects, and improved therapy: mathematics, statistics, and their integration. *Wiley Interdiscip Rev Syst Biol Med, 6*(3), 239-253. doi:10.1002/wsbm.1263

Agur, Z., Halevi-Tobias, K., Kogan, Y., & Shlagman, O. (2016). Employing dynamical computational models for personalizing cancer immunotherapy. *Expert Opin Biol Ther, 16*(11), 1373-1385. doi:10.1080/14712598.2016.1223622

Agur, Z., & Vuk-Pavlovic, S. (2012). Mathematical modeling in immunotherapy of cancer: personalizing clinical trials. *Mol Ther, 20*(1), 1-2. doi:10.1038/mt.2011.272

Apte, R. N., Krelin, Y., Song, X., Dotan, S., Recih, E., Elkabets, M., . . . Voronov, E. (2006). Effects of micro-environment- and malignant cell-derived interleukin-1 in carcinogenesis, tumour invasiveness and tumour-host interactions. *Eur J Cancer, 42*(6), 751-759. doi:10.1016/j.ejca.2006.01.010

Aris-Brosou, S., & Yang, Z. (2002). Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Systematic Biology, 51*(5), 703-714.

Bahlo, M., & Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theoretical population biology, 57*(2), 79-95.

Beaumont, M. A. (1999). Detecting population expansion and decline using microsatellites. *Genetics, 153*(4), 2013-2029.

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences, 98*(8), 4563-4568.

Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J. M., Hlatky, L., & Hahnfeldt, P. (2014). Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput Biol, 10*(8), e1003800. doi:10.1371/journal.pcbi.1003800

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., . . . Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol, 15*(4), e1006650. doi:10.1371/journal.pcbi.1006650

Casas-Selves, M., & Degregori, J. (2011). How cancer shapes evolution, and how evolution shapes cancer. *Evolution (N Y), 4*(4), 624-634. doi:10.1007/s12052-011-0373-y

Chen, K., Moravec, J. C., Gavryushkin, A., Welch, D., & Drummond, A. J. (2022). Accounting for errors in data improves divergence time estimates in single-cell cancer evolution. *Molecular biology and evolution, 39*(8), msac143.

Collins, V. P., Loeffler, R. K., & Tivey, H. (1956). Observations on growth rates of human tumors. *Am J Roentgenol Radium Ther Nucl Med, 76*(5), 988-1000.

Comen, E., Norton, L., & Massague, J. (2011). Clinical implications of cancer self-seeding. *Nature reviews Clinical oncology, 8*(6), 369-377.

Cutler, D. J. (2000). Understanding the overdispersed molecular clock. *Genetics, 154*(3), 1403-1417.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association, 77*(379), 605-610.

Dethlefsen, L. A., Prewitt, J. M., & Mendelsohn, M. L. (1968). Analysis of tumor growth curves. *J Natl Cancer Inst, 40*(2), 389-405. doi:10.1093/jnci/40.2.389

Donnelly, P., & Tavare, S. (1995). Coalescents and genealogical structure under neutrality. *Annual review of genetics, 29*(1), 401-421.

dos Reis, M., Donoghue, P. C., & Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nat Rev Genet, 17*(2), 71-80. doi:10.1038/nrg.2015.8

Drummond, A. J., Chen, K., Mendes, F. K., & Xie, D. (2023). LinguaPhylo: a probabilistic model specification language for reproducible phylogenetic analyses. *PLOS Computational Biology, 19*(7), e1011226.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS biology, 4*(5), e88.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics, 161*(3), 1307-1320.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology, 7*(1), 1-8.

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution, 22*(5), 1185-1192.

Drummond, A. J., & Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC biology, 8*(1), 1-12.

Fearnhead, P., & Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics, 159*(3), 1299-1318.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol, 17*(6), 368-376. doi:10.1007/bf01734359

Fisher, R. A. (1958). The genetical theory of natural selection Second edition. In: Dover, New York, New York, USA.

Flanagan, N., Tobler, A., Davison, A., Pybus, O., Kapan, D., Planas, S., . . . McMillan, W. O. (2004). Historical demography of Müllerian mimicry in the neotropical Heliconius butterflies. *Proceedings of the National Academy of Sciences, 101*(26), 9704-9709.

Gatenby, R. A., & Gillies, R. J. (2008). A microenvironmental model of carcinogenesis. *Nat Rev Cancer, 8*(1), 56-61. doi:10.1038/nrc2255

Gerlee, P. (2013). The model muddle: in search of tumor growth laws. *Cancer Res, 73*(8), 2407-2411. doi:10.1158/0008-5472.Can-12-4355

Ghaffari Laleh, N., Loeffler, C. M. L., Grajek, J., Stankova, K., Pearson, A. T., Muti, H. S., . . . Kather, J. N. (2022). Classical mathematical models for prediction of response to chemotherapy and immunotherapy. *PLoS Comput Biol, 18*(2), e1009822. doi:10.1371/journal.pcbi.1009822

Gillooly, J. F., Allen, A. P., West, G. B., & Brown, J. H. (2005). The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proceedings of the National Academy of Sciences, 102*(1), 140-145.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., . . . Dudas, G. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science, 345*(6202), 1369-1372.

Gompertz, B. (1825). XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c. *Philosophical transactions of the Royal Society of London*(115), 513-583.

Griffiths, R. C., & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of computational biology, 3*(4), 479-502.

Griffiths, R. C., & Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 344*(1310), 403-410.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell, 100*(1), 57-70. doi:10.1016/s0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell, 144*(5), 646-674.

Hasegawa, M., Kishino, H., & Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution, 22*, 160-174.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Heesterman, B. L., Bokhorst, J.-M., de Pont, L. M., Verbist, B. M., Bayley, J.-P., van der Mey, A. G., . . . Jansen, J. C. (2019). Mathematical models for tumor growth and the reduction of overtreatment. *Journal of Neurological Surgery Part B: Skull Base, 80*(01), 072-078.

Hohna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., . . . Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Syst Biol, 65*(4), 726-736. doi:10.1093/sysbio/syw021

Hong, W. S., Shpak, M., & Townsend, J. P. (2015). Inferring the Origin of Metastases from Cancer Phylogenies. *Cancer Res, 75*(19), 4021-4025. doi:10.1158/0008-5472.CAN-15-1889

Jäckel, P. (2002). *Monte Carlo methods in finance* (Vol. 5): John Wiley & Sons.

Joy, D. A., Feng, X., Mu, J., Furuya, T., Chotivanich, K., Krettli, A. U., . . . Suh, E. (2003). Early origin and recent expansion of Plasmodium falciparum. *Science, 300*(5617), 318-321.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism, 3*, 21-132.

Junttila, M. R., & de Sauvage, F. J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature, 501*(7467), 346-354. doi:10.1038/nature12626

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences, 78*(1), 454-458.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications, 13*(3), 235-248.

Kishino, H., & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the

evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of molecular evolution, 29*, 170-179.

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), 1-30.

Kuhner, M. K., Yamato, J., & Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics, 149*(1), 429-434.

Kuhner, M. K., Yamato, J., & Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics, 156*(3), 1393-1401.

Leung, M. L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., . . . Navin, N. E. (2017). Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res, 27*(8), 1287-1299. doi:10.1101/gr.209973.116

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.

Mendelsohn, M. L. (1963). Cell proliferation and tumor growth. *Cell proliferation*, 190-210.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics, 21*(6), 1087-1092. doi:10.1063/1.1699114

Moya, A., Holmes, E. C., & González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. *Nature Reviews Microbiology, 2*(4), 279-288.

Norton, L. (1988). A Gompertzian model of human breast cancer growth. *Cancer research, 48*(24_Part_1), 7067-7071.

Norton, L., & Massagué, J. (2006). Is cancer a disease of self-seeding? *Nature medicine, 12*(8), 875-878.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science, 194*(4260), 23-28. doi:10.1126/science.959840

Ogilvie, H. A., Bouckaert, R. R., & Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular biology and evolution, 34*(8), 2101-2114.

Patt, H. M., & Blackford, M. E. (1954). Quantitative studies of the growth response of the Krebs ascites tumor. *Cancer Res, 14*(5), 391-396.

Penny, D. (2004). Inferring Phylogenies.—Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts. *Systematic Biology, 53*(4), 669-670. doi:10.1080/10635150490468530

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C., & Harvey, P. H. (2001). The epidemic behavior of the hepatitis C virus. *Science, 292*(5525), 2323-2325.

Pybus, O. G., Rambaut, A., & Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics, 155*(3), 1429-1437.

Riester, M., Stadler, P. F., & Klemm, K. (2009). FRANz: reconstruction of wild multi-generation pedigrees. *Bioinformatics, 25*(16), 2134-2139.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences, 53*(1-2), 131-147.

Rodrigo, A. G., & Felsenstein, J. (1999). Coalescent approaches to HIV population genetics. *The evolution of HIV*, 233-272.

Rodriguez-Brenes, I. A., Komarova, N. L., & Wodarz, D. (2013). Tumor growth dynamics: insights into evolutionary processes. *Trends Ecol Evol, 28*(10), 597-604. doi:10.1016/j.tree.2013.05.020

Roman, J., & Palumbi, S. R. (2003). Whales before whaling in the North Atlantic. *Science, 301*(5632), 508-510.

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics, 19*(12), 1572-1574. doi:10.1093/bioinformatics/btg180

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., . . . Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology, 61*(3), 539-542.

Sarapata, E. A., & de Pillis, L. G. (2014). A comparison and catalog of intrinsic tumor growth models. *Bull Math Biol, 76*(8), 2010-2024. doi:10.1007/s11538-014-9986-y

Schwartz, R., & Schaffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet, 18*(4), 213-229. doi:10.1038/nrg.2016.170

Shahriyari, L., & Komarova, N. L. (2013). Symmetric vs. asymmetric stem cell divisions: an adaptation against cancer? *PLoS One, 8*(10), e76195. doi:10.1371/journal.pone.0076195

Spence, J. P., Steinrücken, M., Terhorst, J., & Song, Y. S. (2018). Inference of population history using coalescent HMMs: review and outlook. *Current opinion in genetics & development, 53*, 70-76.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., . . . Schriml, L. (1998). Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes. *The American Journal of Human Genetics, 62*(6), 1507-1515.

Storz, J. F., Beaumont, M. A., & Alberts, S. C. (2002). Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical Bayesian model. *Molecular biology and evolution, 19*(11), 1981-1990.

Sullivan, P. W., & Salmon, S. E. (1972). Kinetics of tumor growth and regression in IgG multiple myeloma. *J Clin Invest, 51*(7), 1697-1708. doi:10.1172/JCI106971

Summers, W. C. (1966). Dynamics of tumor growth: a mathematical model. *Growth, 30*(3), 333-338.

Swofford, D. L. (1996). Phylogenetic inference. *Molecular systematics, 2nd ed.*, 407-514.

Tarabichi, M., Salcedo, A., Deshwar, A. G., Ni Leathlobhair, M., Wintersinger, J., Wedge, D. C., . . . Boutros, P. C. (2021). A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat Methods, 18*(2), 144-155. doi:10.1038/s41592-020-01013-2

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequence. *Lecture of Mathematics for Life Science, 17*, 57.

Thomé, M. T. C., & Carstens, B. C. (2016). Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proceedings of the National Academy of Sciences, 113*(29), 8010-8017.

Vaidya, V. G., & Alexandro, F. J., Jr. (1982). Evaluation of some mathematical models for

tumor growth. *Int J Biomed Comput, 13*(1), 19-36. doi:10.1016/0020-7101(82)90048-4

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondence mathematique et physique, 10*, 113-129.

Williams, A. P., Cook, B. I., & Smerdon, J. E. (2022). Rapid intensification of the emerging southwestern North American megadrought in 2020–2021. *Nature Climate Change, 12*(3), 232-234. doi:10.1038/s41558-022-01290-z

Wodarz, D. (2013). Computational modeling approaches to studying the dynamics of oncolytic viruses. *Math Biosci Eng, 10*(3), 939-957. doi:10.3934/mbe.2013.10.939

Wodarz, D., & Komarova, N. (2009). Towards predictive computational models of oncolytic virus therapy: basis for experimental validation and model selection. *PLoS One, 4*(1), e4271. doi:10.1371/journal.pone.0004271

Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16*(2), 97.

Wu, M., & Swartz, M. A. (2014). Modeling tumor microenvironments in vitro. *J Biomech Eng, 136*(2), 021011. doi:10.1115/1.4026447

Yang, Z., & Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol, 14*(7), 717-724. doi:10.1093/oxfordjournals.molbev.a025811

Yin, X., Goudriaan, J., Lantinga, E. A., Vos, J., & Spiertz, H. J. (2003). A flexible sigmoid function of determinate growth. *Ann Bot, 91*(3), 361-371. doi:10.1093/aob/mcg029

Yule, G. U. (1925). II.—A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character, 213*(402-410), 21-87.

Zuckerkandl, E. (1962). Molecular disease, evolution, and genic heterogeneity. *Horizons in biochemistry*, 189-225.

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (pp. 97-166): Elsevier.