Mathematical Classification of the Modes of Tumour Evolution

Veselin Manojlović

Doctor of Philosophy

CITY UNIVERSITY
LONDON

School of Science and Technology

Department of Mathematics

September 2023

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisor, Rob Noble. Needless to say that none of the results presented in this thesis would have been possible without his guidance. More than that, however, I would like to thank him for his patience, as I came to him with a possibly negative amount of knowledge about mathematical biology in the middle of the worst pandemic in a century. I cannot imagine myself seeing through a project of this scale without his help, and could not have asked for a better supervisor.

I would also like to thank my comrades at City maths, who let me drag them away from their work and would listen to me complain about work, politics, life, and everything in between. The daily poco poco at Cafe Fiori is a ritual I will have a hard time getting used to not having.

Outside City, I am grateful to my flatmates, past and present, who have also listened to hours of my rambling and have been a constant source of support and inspiration through their own work. Alex especially has come in clutch with some of the more obscure references in branching processes that let me bring this thesis to a satisfying conclusion.

And of course, I am grateful to my parents for their daily messages and calls, reminding me that I have the support of a loving family behind me. Finally, a special thanks to Ana, the smartest person I've ever met, and the love of my life, whose turn it is now to be the spoiled PhD student complaining how difficult it is to write a thesis.

Declaration of authenticity

I, Veselin Manojlović, declare that this thesis contains genuine work conducted originally by me. The work presented herein has not been submitted and/or accepted for the award of any other degree or diploma in any university. To the best of my knowledge and belief, this thesis contains no materials previously published or written by other person, except where due references has been made.

Signature: Veselin Manojlović

Date: 7th April 2024

Contributions

The data analysed in chapter 5 was sequenced and initially processed by Dr Shibata.

Figure 1.1 was reproduced from (Davis et al. 2017) under a Creative Commons CC BY 4.0 license.

Associated publications

- Parts of chapter 1 were adapted from my contributions to (Lemant et al. 2022).
- Early results from chapter 3 were presented at MMEE 2022 at the University of Reading and at ECMTB 2022 at the University of Heidelberg.
- Early results from chapters 4 and 5 were presented at the 2023 SMBE conference at the University of Ferrara, as well as Mutations Meeting 2024 at the University of Edinburgh.

Abstract

Determining the mode of tumour evolution is a fundamental question in cancer biology. Knowledge of the evolutionary dynamics of tumours could lead to improved diagnosis and treatment through the identification of key patterns in the data. In this thesis, I present a multi-pronged approach to the study of tumour evolution by further developing methods for the analysis of phylogenetic trees, investigating how different measures of tree properties evolve with time, and narrowing down the consideration of evolutionary models to those that are most relevant in colorectal cancer.

A recently introduced tree balance index, J^1, unlike prior definitions, permits meaningful comparison of trees with arbitrary outdegree distributions and node sizes, thus overcoming the shortcomings of conventional methods. I quantify the accuracy of approximations to the expected values of J^1 for two important null models: the Yule process and the uniform model, and prove that, for the Yule process, the approximation converges to the true expectation in the limit of large trees. I further investigate the minima of J^1 for certain important tree families. These results help establish J^1 as a universal, cross-disciplinary index of tree balance that generalizes and supersedes prior approaches.

As balance is only one of several properties that can be used to characterise phylogenetic trees, I also investigate the evolution of other metrics used in the study of phylogenies. By recapitulating the results of a previous study with a slightly altered methodology, and by expanding the analysis to include a new, more comprehensive set of tree indices, I discuss how these methods could be used to examine the evolutionary dynamics of tumours.

Finally, I develop an agent-based model of colorectal cancer evolution which is informed by multi-site DNA methylation data. I use this model to infer properties of mutliple tumours and draw conclusions about the rate of tumour growth and

strength of selection acting on the tumour. I find that the model is able to reproduce the observed data but not detailed enough to infer the strength of selection within tumour glands.

Chapter 1

Introduction

Cancer remains one of the most formidable challenges in the realm of health and medicine, causing a quarter of all deaths in the UK (UK 2015). Despite advances in cancer research, the survival rates for many cancers remain low, with the disease being an increasing burden on healthcare systems (Institute 2020). The disease's heterogeneity, both within and between patients, is a major obstacle to effective treatment. Understanding the underlying evolutionary processes driving this heterogeneity is crucial to developing new treatments and improving patient outcomes. While having a comprehensive mathematical theory of cancer evolution may not be feasible, concrete mathematical models can provide valuable insights into the disease's dynamics. To this end, I consider different approaches to modelling cancer evolution, which includes the use of phylogenetic trees and agent-based models. Further, I employ methylation data to verify the accuracy of the models using Approximate Bayesian Computation (ABC).

Trees as a mathematical object have found use in a variety of fields, of which biology is my main focus. However, I have found interesting links to methods in computer science via information theory. In chapter 2, I expand upon three points. First, I further establish J^1 as a universal index of tree balance through connections with data structures in computer science. Second, I derive upper bounds on the error of the expected value approximations for the Yule process and the uniform model. Finally, I investigate the minimal values of J^1 in important special cases, with special emphasis on the large tree limit.

In chapter 3, I employ the index J^1, along with two other tree shape indices, to test to what degree one can differentiate between different evolutionary regimes

in cancer by only relying tree shape indices. These results are compared to a new, more comprehensive system of tree shape indices (Noble & Verity 2023) which further generalised the concepts of diversity, evenness and richness. These results lay the groundwork for future analysis of cancer tree data.

In chapter 4, I introduce a tailor-made model for simulating a specific type of molecular data, methylation arrays, obtained from multi-site sequencing of colorectal cancer. I show that the model is able to recapitulate the patterns observed in the data and that it can be used to infer the evolutionary history of the tumour. I further explore how the model can be expanded for more general use due to its modular design. I also demonstrate an approximate Bayesian computation workflow for inferring the parameters of the model from data, and discuss the choice of summary statistics and the performance of the ABC algorithm.

In chapter 5, I verify the utility of the model on the example of colorectal cancer methylation data. I discuss the results of the inference and compare the inferred gland divergence trees to ones generated by the model. I further discuss the implications of the results for the understanding of colorectal cancer evolution and the potential for future research.

1.1 Tumour evolution

1.1.1 Introduction

Cancer emergence and progression is an evolutionary process (Nowell 1976, Merlo et al. 2006). This statement is now widely accepted, and the applications of quantitative methods found in evolutionary biology in cancer research are numerous (Rockne et al. 2019, Yin et al. 2019-10, Kourou et al. 2021). The, now well established, area of mathematical oncology is informed by clinicians, computer scientists, mathematicians of all flavours, and biologists alike (Bull & Byrne 2022), which has led to a rapid development of more specific avenues of research spanning from the initiation of the disease (Paterson et al. 2020) to the optimisation of therapy protocols (West et al. 2023). This is a perfect reflection of the complexity of the disease itself, as its rapid evolution, heterogeneity and constraints on how much information one can obtain from a patient take the combined efforts of thousands of scientists. Mathematics plays its own role in this effort, providing a common language through rigour and methods development, and frameworks for the interpretation of data.

1.1.2 Modes of evolution

Over the years, there have been a number of different definitions of what a mode of evolution is. Initially, it was introduced as the term which covers the way or manner in which a species evolves (Eiseley 1945). Depending on the piece of literature, it could also refer to the model used in the study of a population's evolutionary trajectory (Yotoko et al. 2011), or the mechanism which drives the evolution such as genetic drift (Glassman et al. 1996, Wolf & Koonin 2013). This ambiguity of terminology is present in cancer research as well. In this thesis, I will use the term mode of evolution as originally defined by (Eiseley 1945), and used by (Davis et al. 2017, Noble et al. 2022). That is, the way in which a tumour evolves (figure 1.1).

The specifics of tumour evolution are complex as, while deterministic equations may capture the evolutionary dynamics of a cohort of tumours, the individual tumour's evolutionary history is stochastic (Werner et al. 2013). This only adds to the issue of how the surrounding tissue (West et al. 2021) and the tumour's own spatial organisation will affect its progression (Noble et al. 2022, Li et al. 2023). Therefore, existing models of tumour evolution have had to incorporate both general, large-scale processes and sometimes molecular level events to be able to claim progress towards personalised cancer care informed by quantitative models (Yin et al. 2019-10).

1.1.3 Mathematical models of tumour evolution

As mentioned earlier, the applications of mathematics in oncology are diverse. Thus, my focus over the course of this PhD has been on modelling tumour evolution and progression from its early stages up to and excluding treatment. This makes the problem more of an exercise in populatin dynamics than strict oncology, as underlying assumptions of such models tend to focus less on the microenvironment impact and more on how mutations accumulate and spread in the tumour. A good example of one such model is the Big Bang model of tumour growth (Sottoriva et al. 2015). Informed by multi-site sequencing, the authors' hypothesis was that colorectal cancer evolves neutrally after an initial period of rapid expansion and selection. Much like how cosmic microwave background radiation is unevenly distributed across the observable universe, they observed an asymmetrical distribution of mutations across the tumour spheroid. This inspired a spatial branching process model based on gland fission, with each tumour gland approximated to rapid fixation in the event

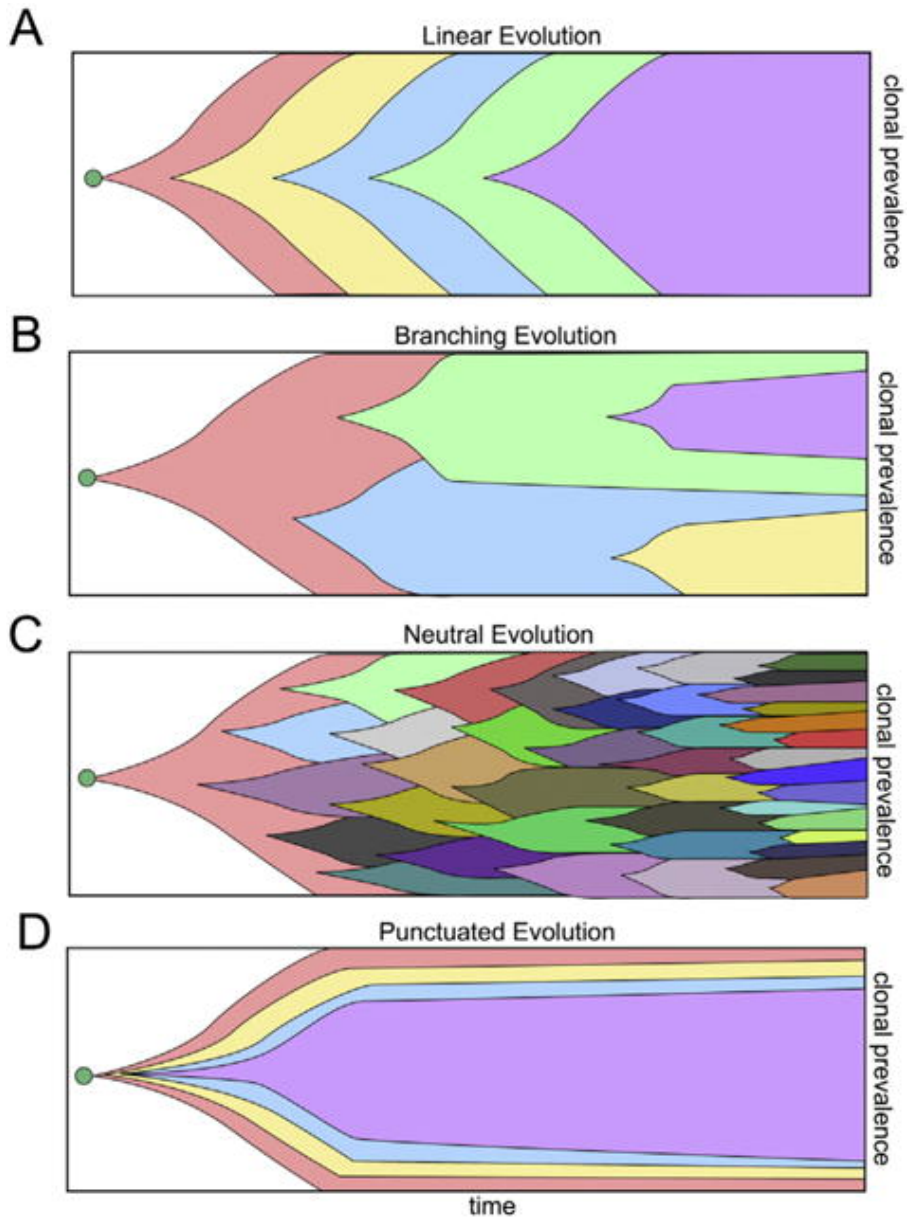

Figure 1.1: Four different modes of tumour evolution represented by their Muller plots. Picture adapted from (Davis et al. 2017) under a CC BY 4.0 license.

of a driver mutation, which showed good agreement with the data. A follow-up paper (Williams et al. 2016) ignited a debate on neutral evolution in exponentially growing tumours within the community (Tarabichi et al. 2018, McDonald et al. 2018, Heide et al. 2018, Bozic et al. 2019). However, theoretical considerations of the two-level model compared to the neutral model did, in fact, show that it is possible to distinguish the two based on mutation frequency spectra (Tung & Durrett 2021). This discussion is far from over, as there are still many questions around the mode of evolution of most cancers.

One would be remiss, however, to only focus on models explicitly designed for

cancer. The abstract nature of mathematical modelling has allowed for the transfer of knowledge between fields, with models developed for other purposes being applicable in cancer. General models which are more easily tested on, for example, bacterial populations (Fusco et al. 2016, Schreck et al. 2023) can be adapted to cancer, as the basic principles of evolution are the same. But digging even deeper, the underlying model of boundary growth dates back to the Eden model of crystal growth (Eden 1961). Among similar examples are uses of the Fisher-Kolmogorov-Petrovsky-Piscounov equation in ecology and its modifications for the study of the spread of mutations in populations with a constant size (Houchmandzadeh & Vallade 2017) as well as growing populations (Wodarz & Komarova 2020). Further, the use of phylogenetic trees and methods in cancer is an emerging field introduced in the following section and expanded upon in chapters 2 and 3.

1.2 Trees and their applications

1.2.1 Introduction

In the most general sense, a tree is a connected graph with no cycles. In this thesis, when a tree is mentioned, I refer to a rooted tree, as formally defined in section 2.2.1. In brief, a tree is rooted if it has a special node called the root, from which all other nodes are reachable. Trees have found use in a variety of fields, including computer science, biology, and linguistics. In computer science, trees are used to represent hierarchical data structures, such as file systems (Nievergelt 1974) or the structure of a program's syntax (Knuth 1968), an approach that computer scientists share with linguists (Chomsky 1957). The concept of search trees, dating back to the mid 20th century, revolutionised the field of computer science with applications in information retrieval in the form of binary search trees and self-balancing trees (Nievergelt & Reingold 1972, Knuth 1997). In evolutionary biology, one of the earliest appearances of trees dates back to the 19th century, when Charles Darwin used them to represent the evolutionary relationships between species. Phylogenetic trees have over time become a key tool in analysing the lineages of species, viral mutations, and cancer evolution. By investigating quantitative summaries of different properties of tree shapes, one can gain insight into the underlying processes driving the evolution of species (Mooers & Heard 1997) or cancer (Scott et al. 2018, Noble et al. 2022). However, most of the inference work so far has been performed using methods which

are not necessarily rooted in sound mathematical theory, but are rather based on heuristics (O'Meara 2012). Specifically, measures of tree balance suffer from a lack of a common framework, with at least 19 different metrics available in literature (Fischer et al. 2021), and few of them being directly comparable. Also, due to the divergent terminology and interest in the use of trees as a tool, there is scarce literature on the transfer of knowledge between the fields of computer science and biology, with certain results being rediscovered nigh on half a century later, as discussed in section 2.2.2.

1.2.2 Quantifying tree balance

Intuitively, a balanced tree should be shaped symmetrically, with equal numbers of nodes on each side of the root. A metric which dictates as much is, for example, the Sackin index (Sackin 1972). However, only considering the shape of the tree ignores the option of having a tree with arbitrarily sized nodes. In this context, a balanced tree is one where the nodes are distributed evenly across the tree. This is the basis of the Colless index, a popular measure of tree balance in biology (Colless 1982). However, the Colless index is limited to bifurcating trees, i.e. trees where each non-terminal (internal) node has exactly two children. This was addressed by a generalisation of the Colless index to multifurcating trees (Mir et al. 2018), allowing for the analysis of tree shapes with more than two children per node. Yet, none of the above indices distinguish between trees where nodes are allowed to have different sizes. This means that conventional balance indices effectively throw away information about the distribution of, for example, access frequencies to nodes in a data storage system or the population size of subclones in a cancer cell population.

In a recent paper (Lemant et al. 2022), Lemant and Noble proposed a new robust, universal index, J^1, for quantifying the balance of rooted trees with arbitrary node degree (number of descendants) and size distributions, going beyond purely the tree shape in measuring its balance. This index is based on Shannon entropy and favours even distributions of node sizes within the tree. This means that not only leaves, but internal nodes are allowed to have any size. This is especially relevant in the case of cancer phylogenetic trees where extant subpopulations will not necessarily be represented by the leaves in the tree. Using the alpha-gamma model (Chen et al. 2009) to generate many such random trees, I showed that J^1 is robust, in the sense that it is insensitive to small changes in node sizes and to the removal of small nodes

(figure 1.2B, C). Noble and I further showed that this index unites and generalises two of the most popular prior approaches to quantifying tree balance in biology, the Colless index and the Sackin index. Applied to evolutionary trees, J^1 outperforms conventional tree balance indices as a summary statistic for comparing model output to empirical data (Noble et al. 2022).

Given any tree shape index, an important task is to obtain its expected and extreme values under standard tree-generating processes, which can then be used as null-model reference points. In (Lemant et al. 2022), Noble and I obtained analytical approximations to the expected values of J^1 under two of the most popular tree-generating processes in biology, the Yule process and the uniform model, and I tested their accuracy numerically for trees with up to 128 leaves (figure 1.2A). In the same study, Noble and I proved that caterpillar trees minimise J^1 among bifurcating trees but not when larger outdegrees are permitted. In chapter 2, I expand on these results.

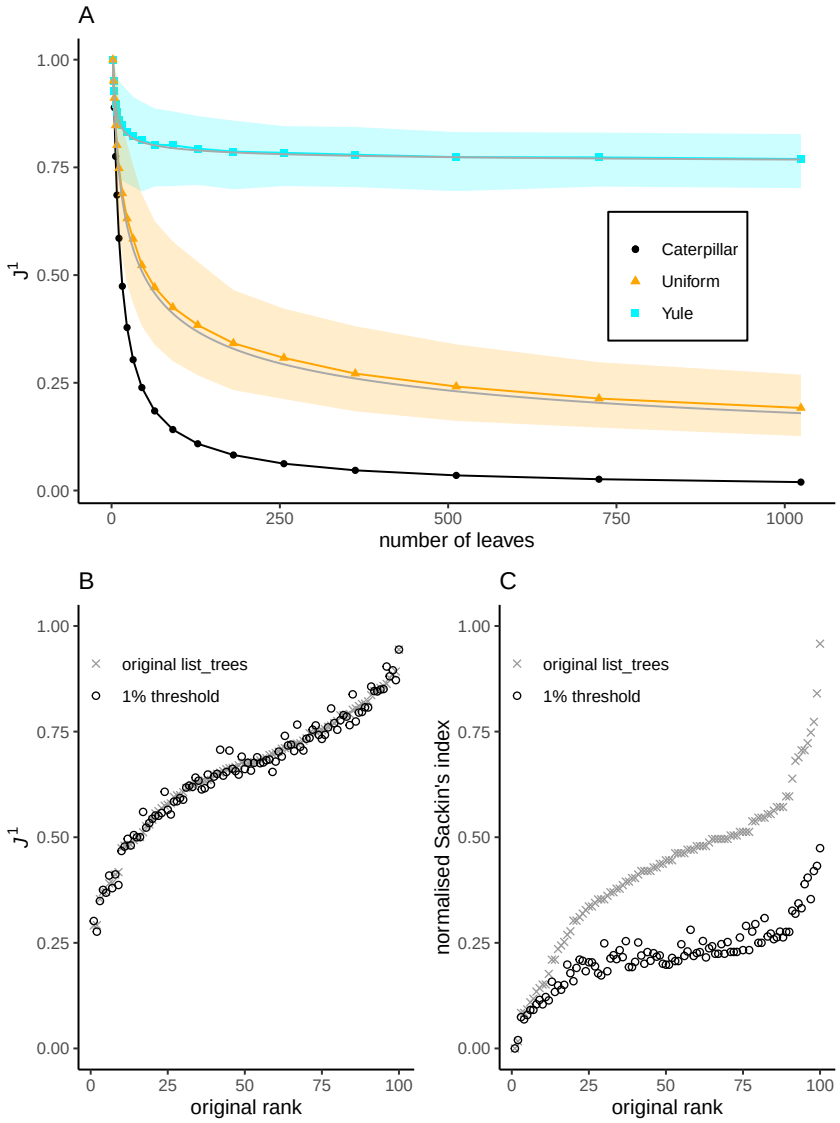

Figure 1.2: **A** Sample mean J^1 values for trees generated under the Yule process and the uniform model. Solid grey curves represent the approximate expected values, and the shaded areas the 5th and 95th percentiles. Each point was averaged over 100 random trees generated under the Yule or uniform models, and caterpillar tree J^1 values were calculated exactly.
B J^1 values for 100 random trees on 16 leaves using the alpha-gamma model, with $\alpha \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Unif}(0, \alpha)$. The values were calculated before and after applying a 1% population threshold, i.e. removing all leaves with sizes smaller than 0.01 times the total population.
C Normalised Sackin index values for the same trees as in **B**.
The Sackin index is much more sensitive to the removal of small nodes as its values change drastically after the threshold is applied.

1.3 Agent-based modelling in oncology

1.3.1 Introduction

Agent-based models (ABMs) are a class of computational models that simulate the actions and interactions of individual agents within a system. These agents can represent anything from cells in a tissue to animals in an ecosystem. ABMs are particularly useful in cancer research, as they can capture the complex interactions happening on the microscale in cancer. Spatial agent-based models (SABMs) are a subclass of ABMs that incorporate spatial information into the simulations. This is particularly useful for modelling solid tumours as it allows for the simulation of things like the spatial heterogeneity of the tumour microenvironment and the effects of spatial constraints on tumour growth. A strength of ABMs is that they can be as simple or as complex as the researcher needs them to be (Colyer et al. 2023). However, therein lies their weakness, as oversimplification of a model can lead to rapid loss of its utility in capturing the behaviour of a complex system such as cancer. On the other hand, a model that is too complex, and attempts to include everything from epigenetic mutations to the effects of the immune system on the tumour, is likely too computationally expensive to be useful for modelling a tumour of reasonable size. This is an organic demonstration of many a researcher's favourite saying *all models are wrong, but some are useful, and some are more useful than others*. In parsing through the literature and developing a new model of my own, I have also been influenced by an alternative wording of this, that is *the best model is its own worst enemy*, by mathematical biologist Philip K. Maini (Maini 2023). My interpretation is that a good model should address the questions it was designed to answer, but also open up new ones which require further investigation, improvements, and research. For example, one can use the `demon-warlock` framework (Bak et al. 2023) to simulate the evolution of a tumour in space and draw conclusions on how spatial organisation will impact intratumour heterogeneity or patient outcomes (Noble et al. 2020, 2022). However, the model does not address the impact of the immune system, spatial heterogeneity in the microenvironment, or the effects of therapy without further modifications. Alternatively, one may want to include diffusion of nutrients and waste products in the model, or the effects of hypoxia on the tumour cells. Tools that would be appropriate for such tasks are, for example, HAL (Bravo et al. 2020) or PhysiCell (Ghaffarizadeh et al. 2018), but

they are not ideal either as simulating a realistically-sized tumour with these models is prohibitively expensive in terms of computational resources. Thus, my preferred approach is to develop a purpose-made model which is informed by the literature and the data, and which has ample room for future expansion and improvement.

1.3.2 The `demon-warlock` framework

In a recent paper (Bak et al. 2023), a new agent-based model for simulating the evolution of a tumour in space was introduced. The model is designed to be versatile and able to simulate a wide range of spatial configurations and evolutionary properties of cancer. Spatially, the model is based on a 2D grid, where each grid cell represents a deme, that is a spatially homogeneous population of cells. The model distinguishes two types of mutations — drivers and passengers. Driver mutations are usually associated with a selective advantage, be it in the form of increased proliferation rate or increased migration rate. Passenger mutations, on the other hand, are usually neutral but can also be deleterious or be used as a marker of cell lineages resistant to therapy. Each cell in the model belongs to a genotype, a unique identifier based on the cell's mutations, and a driver genotype, which differentiates itself from the genotype by not taking into account passenger mutations. Cell migrations in the tumour have multiple modes, including cancer cells invasion of tissue and other demes, and deme fission, the regime where a deme splits into two daughter demes after reaching a certain size. The latter allows for the simulation of tumours with a glandular structure, such as colorectal cancer. Events in the model are scheduled according to the Gillespie algorithm, with the event hierarchy shown in figure 1.3. As the model was written predominantly in plain C, it is highly efficient considering the complexity of the simulations it can run. An accompanying R package, `demonanalysis`, is available for the analysis and visualisation of the model's output, e.g. figure 1.4.

Despite the model's versatility, it is not without its limitations. In its current form, it is not feasible to simulate tumours larger than a few million cells. This leaves out the possibility of simulating realistically-sized glandular tumours which can contain a few million glands containing thousands of cells each at the time of diagnosis. Furthermore, as the main limitation of the model's scalability is tied to the inherent inefficiency of generating many random numbers in each iteration, it is not well-suited to simulating neutral stochastic markers, such as fluctuating

methylation clocks (Gabbutt et al. 2022) which rely on many independent events. This is further discussed in section 4.1.2.

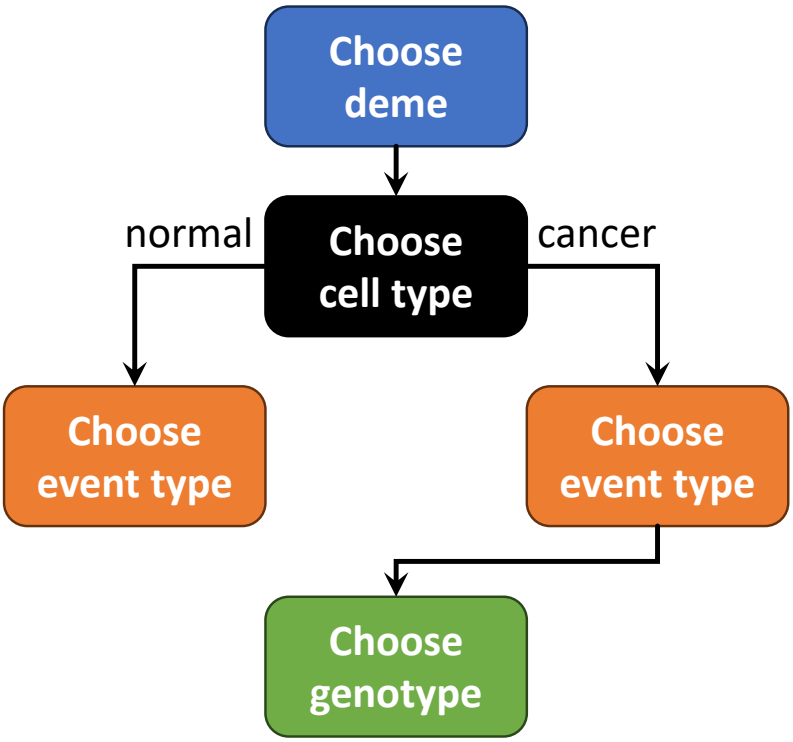

Figure 1.3: Event hierarchy in the `demon-warlock` framework. The algorithm works on a top-down basis, choosing first a deme, then a cell within the deme, and finally an event to perform on the cell. Figure reproduced from (Bak et al. 2023) with the authors' permission.

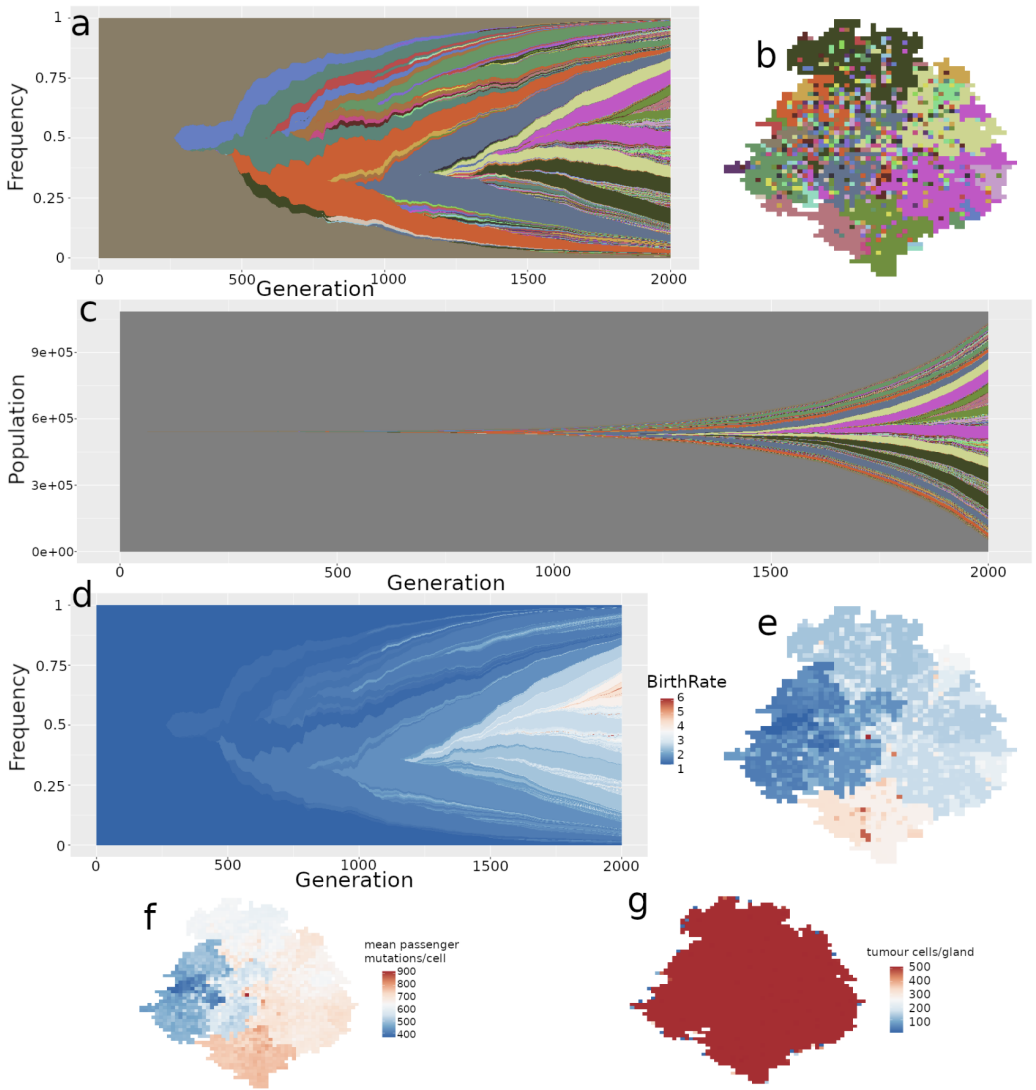

Figure 1.4: Example output from `demon`, visualised using the `demonanalysis` package. **a** Muller plot of clonal dynamics over time. Each colour represents a clone with a distinct combination of driver mutations. **b** Final proportions and spatial plot of clones. **c** Fish plot of clone populations over time using the same colours as in **a**. **d** Muller plot showing evolution of tumour cell division rate. **e** Final spatial distribution of cell division rates. **f** Final spatial distribution of the mean numbers of passender mutations per cell. **g** Final spatial distribution of the number of tumour cells per gland.

1.4 Likelihood-free inference

1.4.1 Introduction

To verify whether a model predicts behaviour of the observed system, a common approach is comparing its output to measurements. The way this is done depends on the complexity of the model and the data. Here we discuss the general framework of likelihood-free inference, and more specifically the use of Approximate Bayesian Computation (ABC).

Models based on differential equations can be compared to data using likelihood-based methods. In the frequentist tradition, the likelihood function is used to estimate the parameters of the model under the assumption that there is a correct, or "true" value of those parameters. An alternative approach is Bayesian statistics, which uses random variables (θ) to represent the uncertainty in the parameters. The distribution of these random variables before observing the data is called the prior distribution ($P(\theta)$). After performing measurements in the system and obtaining data (D) which has an associated likelihood function ($P(D|\theta)$), the prior distribution is update to the posterior distribution ($P(\theta|D)$) using Bayes' theorem (Bayes & Price 1763):

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{1.1}$$

The likelihood function is thus a key component of Bayesian statistics, quantifying the probability of observing the data given the parameters of the model. However, its greatest asset is also its greatest weakness. Depending on the complexity of the model and the data, the likelihood function can be difficult or impossible to calculate analytically, or can be too computationally expensive to calculate numerically. This is especially true for stochastic models, such as agent-based models, where the likelihood function is often intractable.

In the case of intractable likelihoods, a common approach is to use likelihood-free inference methods, designed to approximate the posterior distribution without the need to calculate the likelihood function. These methods rely on the generation of simulated data from the model, and the comparison of these simulations to the observed data. Instead of calculating the likelihood function, these methods often involve a process of simulation and rejection. A common drawback of likelihood-

free inference methods is that they can be computationally expensive, as they often require a large number of simulations to obtain a good approximation of the posterior distribution. However, as the computational power of modern computers increases, these methods are becoming more and more feasible for a wide range of models and data.

1.4.2 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a likelihood-free inference method that has gained popularity in the last three decades (Tavare et al. 1997, Sottoriva & Tavaré 2010, Jagiella et al. 2017). The basic idea behind ABC is to approximate the posterior distribution of the parameters of a model by comparing simulated data to observed data.

In the most general form of ABC, the algorithm proceeds as follows:

1. Sample a set of parameters from the prior distribution, $\hat{\theta}$.
2. Simulate data from the model using the parameters.
3. Compare the simulated data ($\hat{D}$) to the observed data (D) using a distance function, $d(\hat{D}, D)$.
4. If the distance between the simulated and observed data is less than a certain threshold ϵ, accept the parameters. Otherwise, reject them.
5. Repeat steps 1-4 until a sufficient number of accepted parameters have been obtained.

The distance threshold must be strictly positive, and is often chosen to be a small value. Alternatively, in the case of high-dimensional data, the distance function can be replaced by a summary statistic, S, which is a function of the data, i.e. $d'(S(\hat{D}), S(D))$.

ABC does not come without its own set of challenges. As it relies on comparing relevant features of the simulated data to the observed data, the choice of summary statistic or distance metric is crucial, as it determines how the data is reduced before the comparison. Fortunately, there are methods for reducing dimensionality of the data which narrows in on its informative aspects (Blum et al. 2013). The choice of the distance threshold is also important, as it determines the acceptance rate

of the algorithm. Setting the threshold too high or too low can lead to biased or inefficient estimates of the posterior distribution. However, this can be mitigated by using a dynamic threshold, which is adapted during the course of the algorithm (Prangle 2017). Finally, as the algorithm relies on repeated simulations of potentially complex models. This can require a large amount of computational resources, raising questions about the method's scalability and practicality. An obvious way to circumvent this is to use a model which is as lightweight as possible. Even in the case of infinite compute available, one must be mindful of the fact that the more complex the model, the more complex the inference problem, and the more complex the inference problem, the more complex the model. This is a feedback loop which can render both the model and subsequent analysis uninterpretable. Therefore, I believe that the best practice as a mathematician and applied scientist is to abstract the model enough to be able to draw conclusions from it, but not so much that it becomes uninformative.

In chapter 4, I introduce a simplified model of colorectal cancer gland fission and the accompanying ABC workflow. I discuss the choice of summary statistics and the performance of the ABC algorithm, and subsequently demonstrate the model's utility in inferring the evolutionary history of a tumour from methylation data in chapter 5.

1.5 Fluctuating methylation clocks

The concept of the molecular clock is commonplace in molecular evolutionary biology. It is based on the idea that the rate of evolution of a particular gene or set of genes is constant over time, and can be used to estimate the time of divergence between species or the time of a particular event in the evolutionary history of a species. The most famous example of a molecular clock is the mitochondrial DNA clock, which is used to estimate the time of divergence between species (Hasegawa et al. 1985). The key principle behind molecular clocks is that closely related species or individual will have more similar sequences than distantly related ones. This also translates to individual cells in cancer. The issue with using molecular clocks in cancer is that "slowly ticking" molecular clocks, i.e. ones with a low mutation rate, are not informative enough on the timescale of cancer evolution, limiting their utility to cell lineages which diverged too far in the past, with recent events remaining

undetectable. On the other hand, "fast ticking" molecular clocks can reveal recent evolutionary events but also have their own limitations, such as independent mutations in the same site (Kuipers et al. 2017).

Recently, a new type of molecular clock has been proposed, the fluctuating methylation clock, based on the observation that the methylation status of certain CpG sites in the genome is heritable but fluctuates stochastically over time (Gabbutt et al. 2022, 2023). A CpG site is a cytosine nucleotide followed by a guanine nucleotide in the linear sequence of bases along its 5' $\rightarrow$ 3' direction, and is a common site of methylation in the genome. As citosine and guanine are complementary, each CpG site in the 5' $\rightarrow$ 3' direction has a corresponding pair in the 3' $\rightarrow$ 5' direction. This means that each CpG pair can be in one of three states: both methylated (homozygously methylated), both unmethylated (homozygously unmethylated), or one methylated and the other unmethylated (heterozygously methylated). Depending on tissue type, the fluctuating CpG (fCpG) sites can number somewhere between 1000 and 2000, which means each cell has a potentially unique barcode in its fCpG array. As the array is not constant, with methylations and demethylations of fCpG sites happening over the course of cell divisions, the authors of the two papers covering fCpGs so far have been able to reconstruct the evolutionary history of healthy colonic crypts and lymphoid malignancies with high accuracy. This is a promising development, as sequencing methylation arrays is a cheaper method than genome sequencing, but may offer finer temporal resolution. In chapter 5, I investigate whether multi-site methylation array sequencing can be used to reconstruct clonal dynamics of colorectal cancer.

1.6 Aims

1.6.1 Hypotheses

1. The J^1 index can be used, in conjunction with other tree shape indices, to differentiate between evolutionary modes in cancer.

2. SABMs recapitulate molecular data observed in solid tumours (or sth like that)

3. These methods are useful for inferring the evolutionary history of colorectal cancer based on multi-site methylation array sequencing.

1.6.2 Aims

1. Calculate or approximate important properties of the J^1 index, such as its expected value under standard tree-generating processes.
   (a) Contextualise J^1 within the broader field of tree shape indices in biology and computer science.
   (b) Investigate extreme and expected values of J^1 under standard tree-generating processes.
2. Determine the utility of sets of tree shape indices for differentiating between evolutionary modes in cancer.
   (a) Recapitulate the classification of evolutionary modes in cancer using a set of three tree shape indices.
   (b) Extend the discussion to a more interpretable and general system of tree shape indices.
3. Extend the fluctuating methylation clock model to multi-site sequencing of solid tumours on the example of colorectal cancer.
   (a) Develop an agent-based model for simulating the evolutionary dynamics of colorectal adenocarcinoma.
   (b) Develop an ABC workflow for inferring the evolutionary parameters of the model, specifically the gland fission rate, methylation and demethylation rates, driver mutation rate, selective advantage, and the effective number of lineages per tumour gland.
   (c) Apply the model to colorectal cancer data and compare the inferred phylogenies to the trees generated by the model.

Chapter 2

Expected and extreme values of universal tree balance index J^1

2.1 Introduction

Broadly speaking, the balance of a tree is the extent to which its terminal nodes (leaves) are evenly distributed among its branches. Despite the abundance of metrics of tree balance (Fischer et al. 2021), universal indices are hard to come by. This limits practical applications of tree balance indices.

Following the J^1 index paper (Lemant et al. 2022), where a universal index was proposed, shown to be robust to the removal of small nodes and to outperform conventional tree balance indices as a summary statistic for comparing model output to empirical data, I examined several important properties of J^1. Given any new tree shape index, the expected value under standard tree-generating processes and the extreme values need to be known for the index to be useful in practice. In figure 1.2A, I showed the sample mean of J^1 up to 128 leaves under the Yule and uniform models, which appears to be close to the inverse Sackin index expression derived by Noble in (Lemant et al. 2022). Additionally, as a consequence of this relationship, the caterpillar trees minimises J^1 for bifurcating trees. However, I showed in (Lemant et al. 2022) that the caterpillar topology does not minimise J^1 for multifurcating trees by providing a couterexample on 6 leaves.

In this chapter, I will further show the universality of J^1 by identifying fundamental connections to classical results in computer science, related to Huffman coding and self-balancing tree data structures. I will also derive upper bounds on

the error of the expected value approximations for the Yule process and uniform model. For the Yule process, I show that the approximation rapidly converges to the true expected value in the large tree limit. Finally, I will investigate the minimal values of J^1 in important special cases, obtaining a counter-intuitive result in the large tree limit.

In addition to furthering the understanding of the universality of Shannon entropy-based tree balance indices, the results of this chapter are essential for establishing the index J^1 as one of the most useful metrics for tree shape analysis by providing a solid theoretical foundation of its properties and behaviour. This in turn allows for more accurate and robust comparisons of trees generated by models and empirical data, and thus new insights into the underlying evolutionary processes.

2.2 Prerequisites

2.2.1 Preliminary definitions from systematic biology

Definition 2.2.1 (Rooted tree)**.** A **rooted tree** T is a connected acyclic graph with node set $V(T)$ and edge (or branch) set $E(T)$, in which one node is designated the root. Parent-child and ancestor-descendant relationships in a rooted tree are assigned along paths directed away from the root.

Definition 2.2.2 (Node size and tree magnitude, Lemant et al. (2022))**.** We assign to every node a non-negative size. The **magnitude** of a tree T, denoted $S(T)$, is then the sum of its node sizes.

Definition 2.2.3. (Leafy tree, Lemant et al. (2022)) A **leafy tree** is one with only zero-sized internal nodes.

Definition 2.2.4. (Node depth) As we will consider only trees with uniform edge lengths, we define the **depth** of a node as the number of edges in the shortest path from that node to the root.

Definition 2.2.5 (Sackin index, Sackin (1972))**.** The **Sackin index** of rooted tree T is the sum of its leaf depths:

$$I_S(T) = \sum_{l \in L(T)} \nu(l), \tag{2.1}$$

where $L(T)$ is the set of all leaves (terminal nodes) of T, and $\nu(l)$ is the depth of leaf l.

Definition 2.2.6 (Generalised Sackin index, Lemant et al. (2022))**.** The Sackin index can be generalised to account for arbitrary node sizes:

$$I_{S,\text{gen}}(T) = \sum_{i \in V(T)} S_i^*, \tag{2.2}$$

where $V(T)$ is the set of all internal nodes (non-leaves), and S_i^* is the magnitude of the subtree rooted at node i, excluding i. If T is a leafy tree in which all leaves have unit size then $I_{S,\text{gen}}(T) = I_S(T)$.

Definition 2.2.7 (Robust balance index, Lemant et al. (2022))**.** The robust balance index J^1 of tree T is

$$J^1(T) = \frac{1}{I_{S,\text{gen}}(T)} \sum_{i \in \tilde{V}(T)} S_i^* \sum_{j \in C(i)} W_{ij}^1, \tag{2.3}$$

where $\tilde{V}(T)$ the set of all internal nodes whose descendants are not all of zero size, $C(i)$ is the set of children of node i, and W_{ij}^1 is the node balance score, defined as the normalised Shannon entropy of the daughter subtree magnitudes:

$$W_{ij}^1 = \begin{cases} -\frac{S_j}{S_i^*} \log_{d^+(i)} \frac{S_j}{S_i^*}, & \text{for } d^+(i) > 1 \\ 0, & \text{otherwise,} \end{cases} \tag{2.4}$$

where S_i is the magnitude of the subtree rooted at node i, including i, and $d^+(i)$ is the outdegree of i.

Definition 2.2.8 (Binary tree and bifurcating tree)**.** A **binary tree** is a rooted tree in which no node has more than 2 children. A **bifurcating tree** (or full binary tree) is a rooted tree in which each internal node has exactly 2 children.

Definition 2.2.9 (Cherry)**.** A tree consisting of only a root and two leaves is a **cherry**.

Definition 2.2.10 (Caterpillar tree)**.** A **caterpillar tree** is a bifurcating tree in which every internal node except one has exactly one child leaf.

Definition 2.2.11 (Fully symmetric tree)**.** If, for every internal node i, the subtrees

rooted at the children of i all contain the same number of leaves then the tree is **fully symmetric**.

2.2.2 Preliminary definitions from computer science

Definition 2.2.12 (Root balance and tree balance scores, Nievergelt et al. (1972))**.** The **root balance score** of a bifurcating leafy subtree T_i rooted at i and containing at least three nodes is

$$\rho(T_i) = \frac{\min(S_{i_1}, S_{i_2})}{S_i} \in \left[0, \tfrac{1}{2}\right], \tag{2.5}$$

where S_i is the magnitude of T_i, and S_{i_1} and S_{i_1} are the magnitudes of the subtrees rooted at the children of i. The balance score of a bifurcating leafy tree T is then defined as

$$\beta(T) = \min\left(\rho(T_i)_{i \in V(T)}\right). \tag{2.6}$$

For any given leaf count, the balance score is minimal for the caterpillar tree and maximal for the fully symmetric bifurcating tree.

Definition 2.2.13 (Total and weighted path lengths, Nievergelt et al. (1972))**.** In computer science, the Sackin index is better known as the **total path length**. Let T be a rooted tree in which each node i is assigned a weight (or access frequency) w_i. Then the **weighted path length** of T is

$$|T| = \sum_{i \in V(T)} w_i \nu(i). \tag{2.7}$$

2.3 Results

2.3.1 J^1 unites and generalises prior notions of tree balance

In computer science, tree balance is effectively a binary property: a tree is considered balanced if its weighted path length is sufficiently small, given its leaf count (Nievergelt & Reingold 1972). In biology, where comparisons between trees are more relevant, researchers instead use a normalised form of the Sackin index or various other indices to assign balance values on a continuum (Colless 1982, Shao & Sokal 1990, Mir et al. 2013, 2018, Fischer et al. 2021). I will show that J^1 uniquely connects these two historically separate notions of tree balance. Let us note first that the weighted path length is equivalent to the generalised Sackin index:

$$|T| = \sum_{i \in V(T)} \alpha_i \nu(i) = \sum_{i \in V(T)} S_i^* = I_{S,gen}(T). \tag{2.8}$$

Consider then the following proposition.

Proposition 2.3.1 (Lemant et al. (2022)). *Let T be a leafy tree with $d^+(i) = m > 1$ for all internal nodes i. Then*

$$J^1(T) = \frac{H_m(T)S(T)}{I_{S,gen}(T)}, \tag{2.9}$$

where $H_m(T)$ is the Shannon entropy (base m) of the proportional leaf sizes.

Corollary 2.3.1. We can rewrite equation (2.9) for bifurcating trees as

$$J^1(T) = \frac{H_2(T)S(T)}{|T|}. \tag{2.10}$$

Hence, for any given set of leaf sizes, minimising the weighted path length is equivalent to maximising J^1.

Theorem 2.3.1 (Section 5 of Nievergelt et al. (1972)). *Let T be a bifurcating leafy tree with balance score β_T. Then the total path length $|T|$ satisfies the inequality*

$$|T| \leq \frac{S(T)\log_2 S(T) + H_2(T)}{H_2(\beta_T)}. \tag{2.11}$$

If the node sizes sum to unity then this simplifies to

$$|T| \leq \frac{H_2(T)}{H_2(\beta_T)}. \tag{2.12}$$

A special case of this theorem is considered as Theorem 2 in Wong & Nievergelt (1973): If the tree has n leaves, all of size 1 then

$$|T| \leq \frac{n \log n}{H(\beta_T)}. \tag{2.13}$$

The proof of this theorem defines the *average entropy* of a general tree T, corrected for typo in original paper, as

$$\bar{H}(T) = \frac{1}{|T|} \sum_{k \in \tilde{V}(T)} \sum_{j \in C(k)} n_j \log_2 \frac{n_k}{n_j}, \tag{2.14}$$

which is identical to the definition of J^1, equation (2.3), up to the base of the logarithm in the expression for the entropy of internal node k.

Remark 2.3.1. We can trivially expand the definition of the balance score to m-furcating trees, by considering all m descendants of internal nodes in the root balance score. The root balance score of subtree T_j rooted in node j of m-furcating leafy tree T can be defined as

$$\rho_m(T_j) = \frac{\min(S_{j_1}, \ldots, S_{j_m})}{S_j}, \tag{2.15}$$

where $j_1, \ldots, j_m \in C(i)$ are the children of node j. By extension, we define

$$\beta_m(T) = \min(\rho_m(T_i)_{i \in V(T)}), \tag{2.16}$$

the balance score of m-furcating leafy tree T.

Corollary 2.3.2. There is a lower bound on J^1 for an m-furcating leafy tree T on n leaves, with balance score β_T, and it equals

$$J^1_{\text{lower}} = \frac{H_m(T)S(T)}{|T|_{\text{upper}}} = \frac{n \log_m n}{(H_m(\vec{\beta}_T))^{-1} n \log_m n} = H_m(\vec{\beta}_T). \tag{2.17}$$

where $\vec{\beta}_T = (S_{1,min}, \ldots, S_{m,min})$ is the vector of magnitudes of subtrees rooted in the children of the node with the smallest root balance score.

The connections between J^1 and measures of tree balance and entropy in computer science show that these properties are universally important. However, the similarities may well end at this point, as evolutionary biologists and computer scientists use these measures for different purposes and take their research in opposite directions directions, for example inferring evolutionary processes which produced the tree shape (Mooers & Heard 1997) versus shaping the tree to optimise data storage and retrieval (Nagaraj 1997).

2.3.2 J^1 is maximised by Huffman coding

Definition 2.3.1 (Binary search tree)**.** A **binary search tree** (or BST) T_n over n entries (w.l.g. numbers) $x_1, \ldots, x_n$ is a labelled binary tree, each of whose nodes have been labelled with a distinct number chosen from $x_1, \ldots, x_n$ such that for each node N, all nodes in the left subtree of N have a smaller x_i as their label than x_N,

and all nodes in the right subtree of N have a larger number as their label than node N (e.g. figure 2.1).

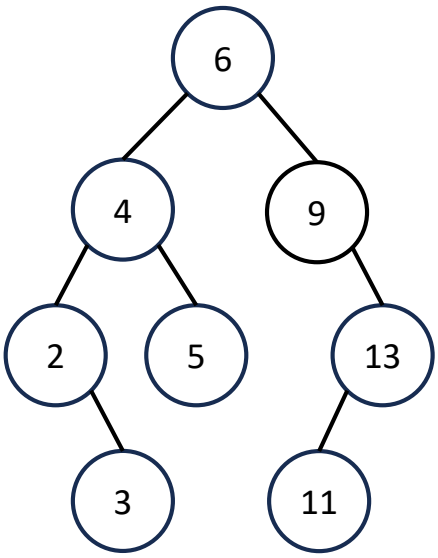

Figure 2.1: A simple example of a binary search tree over the set of labels $S = \{2, 3, 4, 5, 6, 9, 11, 13\}$.

Remark 2.3.2. Each node i in a binary search tree can have an associated non-negative number called access frequency (or weight, size, probability) w_i.

To construct an optimal binary search tree, that is one with a minimal weighted path length, we can use Huffman coding.

Definition 2.3.2 (Huffman coding, Huffman (1952))**.** Let $A = (\alpha_1, \ldots, \alpha_n)$ be a tuple of non-negative numbers. To construct an optimal binary tree on n leaves with sizes given by A we choose the two smallest ones, w.l.g. α_1 and α_2, and join them in a cherry, so that their parent node has size $\alpha_1 + \alpha_2$. We now have $A' = (\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_n)$ as our set of $n - 1$ nodes. By repeating this procedure until we have only one node left, an optimal binary search tree is obtained.

Proposition 2.3.2. *The Huffman method maximises J^1 on bifurcating trees for a given set of node sizes.*

Proof. By corollary 2.3.2, the Huffman method maximises J^1 as it minimises the weighted path length. □

As Huffman coding is an optimisation algorithm, J^1 can be used to measure how close a tree constructed using a given set of node sizes is to the maximally balanced

binary tree on the same set. This means we can quantify how close an alternative algorithm which runs in a faster time complexity, such as arithmetic coding (Pasco 1977), gets to the optimal solution.

2.3.3 Expected value of J^1 under simple evolutionary processes

For applications in evolutionary biology, an important property of balance indices is their expected value under an evolutionary process. This quantity helps us compare the trees generated under a null model to the observed data, and is a valuable part of inferring the underlying evolutionary properties. Two of the simplest, and most widely studied, processes of tree generation are the uniform model (Rosen 1978) and the Yule model (Yule 1925), which generate bifurcating trees and are useful null models in evolutionary biology. The Yule model, also known as the pure birth or coalescent model, is used when considering speciation rates and patterns (Aldous 2001, Steel & McKenzie 2001). The uniform model is used as a null model for comparing neutral evolutionary patterns against ones which include more complex biological mechanisms (Mooers & Heard 1997, McKenzie & Steel 2000). While in section 2.2.2 I discussed the static calculation of a balance index for a given tree, I am also interested in how balanced binary search trees generated under some stochastic process are. The Yule model is also useful for these considerations as it is connected to the BST martingale, a statistical tool used to analyze and predict the behavior of binary search trees, via L_1 convergence (Chauvin & Rouault 2004). From here, one can extend the discussion to AVL (Adelson-Velsky-Landis) and red-black trees (Knuth 1997) in a similar way to more complex evolutionary processes as self-balancing trees will by definition have higher expected values of J^1 than those generated under the Yule process.

The expected value of a few indices, and even some higher moments in certain cases, are known for both Yule and uniform models (Mir et al. 2013, M. Coronado et al. 2020, Goh et al. 2022). Among these is the Sackin index, which is particularly useful for our purposes.

Definition 2.3.3 (Yule model, Yule (1925))**.** Consider a bifurcating tree T on n leaves. To obtain the probability of generating T under the Yule model, start with a single node and replace it with a cherry. Then, at each step, choose one leaf uniformly at random and replace it with a cherry, until the tree has n leaves (figure

2.2A). The sum of probabilities of generating T in all possible ways is the probability of generating T under the Yule model.

Definition 2.3.4 (Uniform model, Rosen (1978))**.** Under the **uniform model** of tree generation, every bifurcating tree on n leaves has an equal probability of being generated, which is equal to $n\binom{2n-2}{n-1}^{-1}$ (figure 2.2B).

Remark 2.3.3. We only consider leafy trees with equal leaf sizes generated by the processes in definitions 2.3.3 and 2.3.4.

Remark 2.3.4. We calculated the exact values of the expectation of J^1 under both the Yule and uniform models semi-manually by creating all possible $(n+1)$-leaf trees given a set of n-leaf trees, eliminating duplicates and assigning appropriate probabilities in the Yule case, and thus computing the exact value of the expectation. The process is inefficient for large trees, and we limited our search to $n \leq 11$, the exact and approximate expected values for which are found in table 2.1.

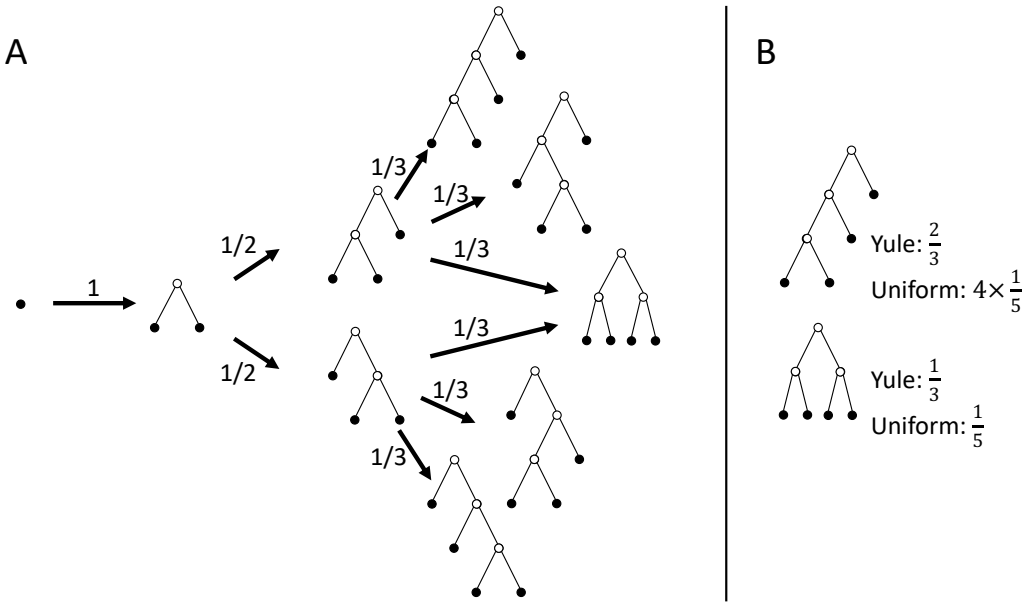

Figure 2.2: Comparison of probabilities for generation of trees on 4 leaves under the Yule and uniform models. **A:** Arrows show generation under the Yule model. Each of the trees shown on 4 leaves has the same probability under the uniform model. **B:** Comparison of probabilities of tree topologies on 4 leaves under the Yule and uniform models.

Under the Yule model, the expected value of the Sackin index for trees on n leaves is

$$\mathbb{E}_Y^n(I_S) = 2n \sum_{i=2}^{n} \frac{1}{i}, \tag{2.18}$$

as shown in Kirkpatrick & Slatkin (1993). Equation (2.9) implies then that the

expected value of J^1 for a tree on n leaves is

$$\mathbb{E}_Y^n(J^1) = \mathbb{E}_Y^n(\frac{n\log_2 n}{I_S}) = n\log_2 n\mathbb{E}_Y^n(1/I_S), \tag{2.19}$$

where $\mathbb{E}_Y^n(1/I_S)$ is the harmonic mean of the Sackin index. The harmonic mean under the Yule process is not a standard result in literature, nor have I been able to obtain a closed-form solution for this problem so far. It is possible, however, to compare the harmonic and arithmetic means of I_S by considering the Jensen gap

$$\mathcal{J}(f, X) = \mathbb{E}[f(X)] - f(\mathbb{E}[X]), \tag{2.20}$$

with $f(x) = 1/x$.

Theorem 2.3.2 (Liao & Berg (2017))**.** *Let X be a one-dimensional random variable with mean μ, and $P(X \in (a,b)) = 1$, where $\infty \leq a < b \leq \infty$. If $f(x)$ is a twice differentiable function on (a,b), and*

$$h(x;\nu) = \frac{f(x) - f(\nu)}{(x-\nu)^2} - \frac{f'(\nu)}{x-\nu}, \tag{2.21}$$

then

$$\inf_{x\in(a,b)}\{h(x;\mu)\}\mathrm{Var}(X) \leq \mathbb{E}[f(X)] - f(\mathbb{E}[X]) \leq \sup_{x\in(a,b)}\{h(x;\mu)\}\mathrm{Var}(X). \tag{2.22}$$

Proposition 2.3.3. *Let $\mathbb{E}_Y(J^1)$ and $\mathbb{E}_U(J^1)$ be expectation values of J^1 under the Yule and uniform models respectively. Then:*

(i) $\mathbb{E}_Y(J^1) \to \frac{n\log_2 n}{\mathbb{E}_Y(I_S)}$,

(ii) $\mathbb{E}_U(J^1) - \frac{n\log_2 n}{\mathbb{E}_U(I_S)}$ *is bounded from both sides,*

as $n \to \infty$.

Proof. (i) Let μ_Y be the expected value of the Sackin index under the Yule process for trees on n leaves, and $f(x) = \frac{1}{x}$. By theorem 2.3.2

$$h(x;\mu_Y) = \frac{f(x) - f(\mu_Y)}{(x-\mu_Y)^2} - \frac{f'(\mu_Y)}{x-\mu_Y} = \frac{1}{x\mu_Y^2}. \tag{2.23}$$

We can then substitute this into the inequality given in the theorem

$$\frac{n\log_2 n}{\frac{(n-1)(n+2)}{2}\mu_Y^2}\mathrm{Var}_Y(I_S) \leq \mathbb{E}[J^1] - \frac{n\log_2 n}{\mathbb{E}[I_S]} \leq \frac{n\log_2 n}{\mu_Y^2 n\log_2 n}\mathrm{Var}_Y(I_S), \tag{2.24}$$

where the supremum and infimum of $h(x,\mu)$ are substituted with extremal values of the Sackin index on bifurcating trees (Fischer 2021). The expectation of the Sackin index under the Yule process is given in equation (2.28), and its variance is calculated as (Cardona et al. 2012)

$$\mathrm{Var}_Y(I_S) = 7n^2 - 4n^2\sum_{i=1}^{n}\frac{1}{i^2} - 2n\sum_{i=1}^{n}\frac{1}{i} - n. \tag{2.25}$$

Substituting these expressions into equation (2.24), we find limits

$$\begin{aligned}\frac{n\log_2 n}{\frac{(n-1)(n+2)}{2}\mu_Y^2}\mathrm{Var}_Y(I_S) &\overset{n\to\infty}{\sim} \frac{\log n\left(7n^2 - 4n^2\sum_{i=2}^{n}\frac{1}{i^2} - 2n\sum_{i=2}^{n}\frac{1}{i} - n\right)}{4n^3\left(\sum_{i=2}^{n}\frac{1}{i}\right)^2}\\ &\sim \frac{\log n}{n} \to 0\end{aligned}$$

for the lower bound on the gap, and

$$\begin{aligned}\frac{n\log_2 n}{\mu_Y^2 n\log_2 n}\mathrm{Var}_Y(I_S) &= \frac{7n^2 - 4n^2\sum_{i=2}^{n}\frac{1}{i^2} - 2n\sum_{i=2}^{n}\frac{1}{i} - n}{4n^2\left(\sum_{i=2}^{n}\frac{1}{i}\right)^2}\\ &\overset{n\to\infty}{\sim} \frac{1}{\left(\sum_{i=2}^{n}\frac{1}{i}\right)^2} \to 0\end{aligned}$$

for the upper bound on the gap. The upper bound reaches a maximum at $n = 13$ and is approximately 0.008, while the lower bound reaches a maximum at $n = 8$ and is approximately 0.005.

(ii) Let μ_U be the expected value of the Sackin's index under the uniform model for trees on n leaves, and $f(x) = \frac{1}{x}$. By theorem 2.3.2

$$h(x;\mu_U) = \frac{f(x) - f(\mu_U)}{(x-\mu_U)^2} - \frac{f'(\mu_U)}{x-\mu_U} = \frac{1}{x\mu_U^2}. \tag{2.26}$$

We can then substitute this into the inequality given in the theorem as in

$$\frac{n\log_2 n}{\frac{(n-1)(n+2)}{2}\mu_U^2}\mathrm{Var}_U(I_S) \leq \mathbb{E}[J^1] - \frac{n\log_2 n}{\mathbb{E}[I_S]} \leq \frac{n\log_2 n}{\mu_U^2 n\log_2 n}\mathrm{Var}_U(I_S), \tag{2.27}$$

analogously to (2.24). The expectation of Sackin's index under the uniform model

is given by (Cardona et al. 2012)

$$\mathbb{E}_U(I_S) = \frac{4^{n-1} n!(n-1)!}{(2n-2)!} - n, \tag{2.28}$$

and its variance is

$$\mathrm{Var}_U(I_S) = n\frac{10n^2 - 3n - 1}{3} - \frac{(n+1)(n+2)}{2}\frac{(2n-2)!!}{(2n-3)!!} - n^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2. \tag{2.29}$$

For the limit $n \to \infty$ we can use Stirling's approximation

$$n! \overset{n\to\infty}{\sim} \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{2.30}$$

$$n!! \overset{n\to\infty}{\sim} \begin{cases} \sqrt{\pi n}\left(\frac{n}{e}\right)^{n/2}, & n \text{ even}, \\ \sqrt{2n}\left(\frac{n}{e}\right)^{n/2}, & n \text{ odd}, \end{cases} \tag{2.31}$$

to obtain asymptotic behaviour of the expected value and variance of I_S under the uniform model. The expectation reduces to

$$\begin{aligned} \mathbb{E}_U(I_S) &\overset{n\to\infty}{\sim} \frac{4^{n-1}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \sqrt{2\pi(n-1)}\left(\frac{n-1}{e}\right)^{n-1}}{\sqrt{2\pi(2n-2)}\left(\frac{2n-2}{e}\right)^{2n-2}} - n \\ &\sim \frac{\sqrt{\pi n} n^n}{e(n-1)^{n-1}} - n \\ &\sim \sqrt{\pi}\exp\left[(n+\frac{1}{2})\log n - (n-1)\log(n-1)\right] - n \\ &\sim \sqrt{\pi} n^{\frac{3}{2}} - n \end{aligned}$$

and the variance

$$\begin{aligned} &\mathrm{Var}_U(I_S) \overset{n\to\infty}{\sim} \frac{10}{3}n^3 - n^2 - \frac{1}{3}n - \frac{n^2+3n+2}{2}\frac{\sqrt{\pi(2n-2)}\left(\frac{2n-2}{e}^{n-1}\right)}{\sqrt{2(2n-3)}\left(\frac{2n-3}{e}^{n-1/2}\right)} \\ &- n^2\left(\frac{\sqrt{\pi(2n-2)}\left(\frac{2n-2}{e}^{n-1}\right)}{\sqrt{2(2n-3)}\left(\frac{2n-3}{e}^{n-1/2}\right)}\right)^2 \\ &\sim \frac{10}{3}n^3 - n^2 - \frac{1}{3}n - \frac{n^2+3n+2}{2}\sqrt{\frac{e\pi}{2}}\exp\left[(n-1)\log\frac{2n-2}{2n-3} + \frac{1}{2}\log(2n-3)\right] \\ &- n^2\exp\left[\log(2n-3)\right] \\ &\sim \frac{4}{3}n^3 + 2n^2 - \frac{1}{3}n - \frac{n^{\frac{5}{2}} + 3n^{\frac{3}{2}} + 2n^{\frac{1}{2}}}{2}\sqrt{e\pi}. \end{aligned}$$

□

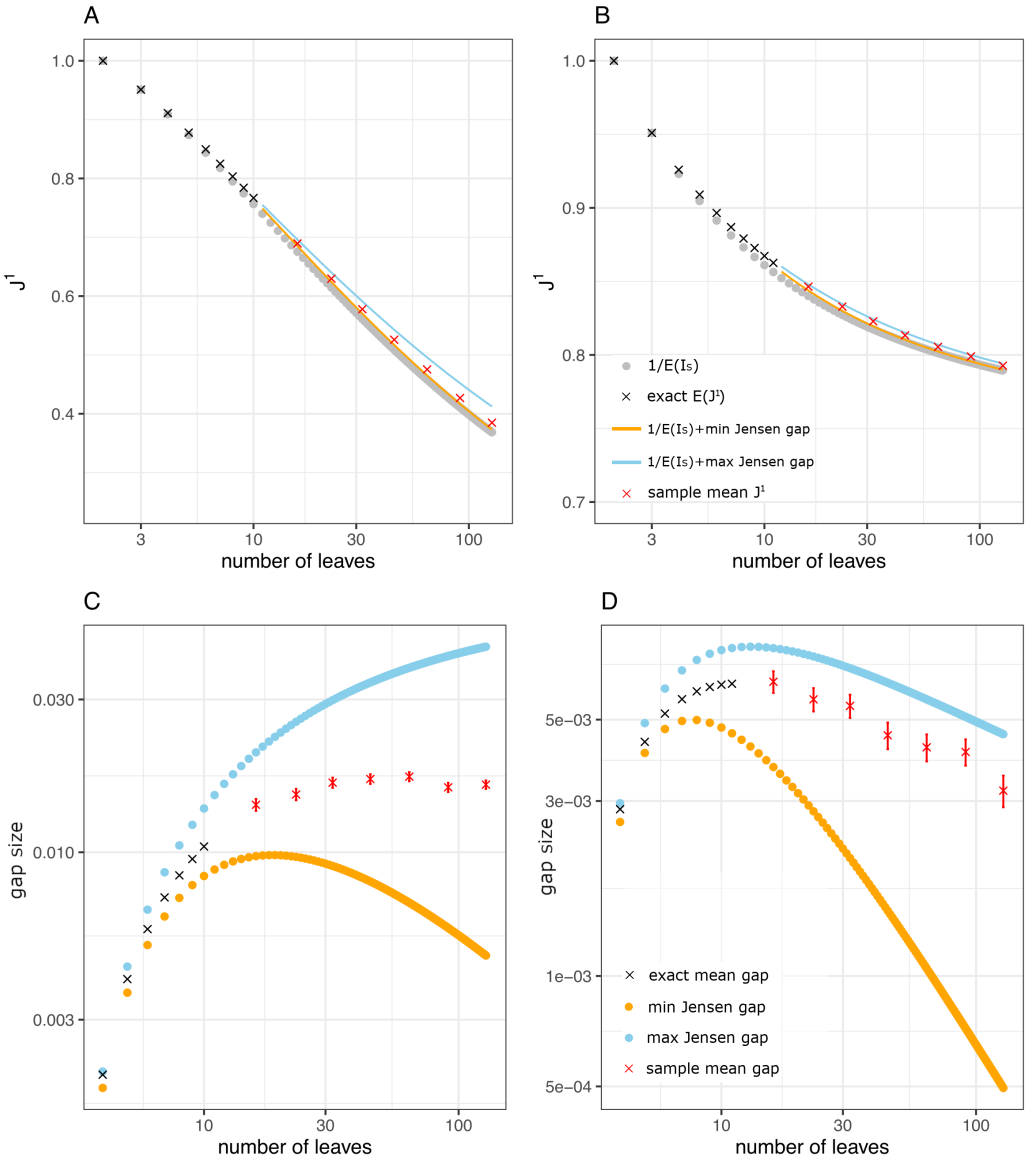

Figure 2.3: Top row: True values of $\mathbb{E}(J^1)$ for up to 10 leaves were calculated manually, and the approximations up to 128 leaves were calculated as $n \log_2 n/\mathbb{E}(I_S)$. **A** — uniform model, **B** — Yule model.
Bottom row: The Jensen gap of $\mathbb{E}(J^1)$ calculated for trees up to 128 leaves under the uniform model (**C**), and the Yule model (**D**). The size of the gap is calculated as the difference between the true and approximate expected value, with the gaps for 2 and 3 leaves equal to zero as there is only one possible bifurcating tree shape for each of those values. Refer to table 2.1 for numerical values of the gap size for the first several values of n. The red crosses in **A** and **B** represent sample mean J^1 values for 100000 trees generated under the uniform model and Yule process, and the difference between the approximate gap size and the sample mean, with standard error represented by error bars, in **C** and **D**.

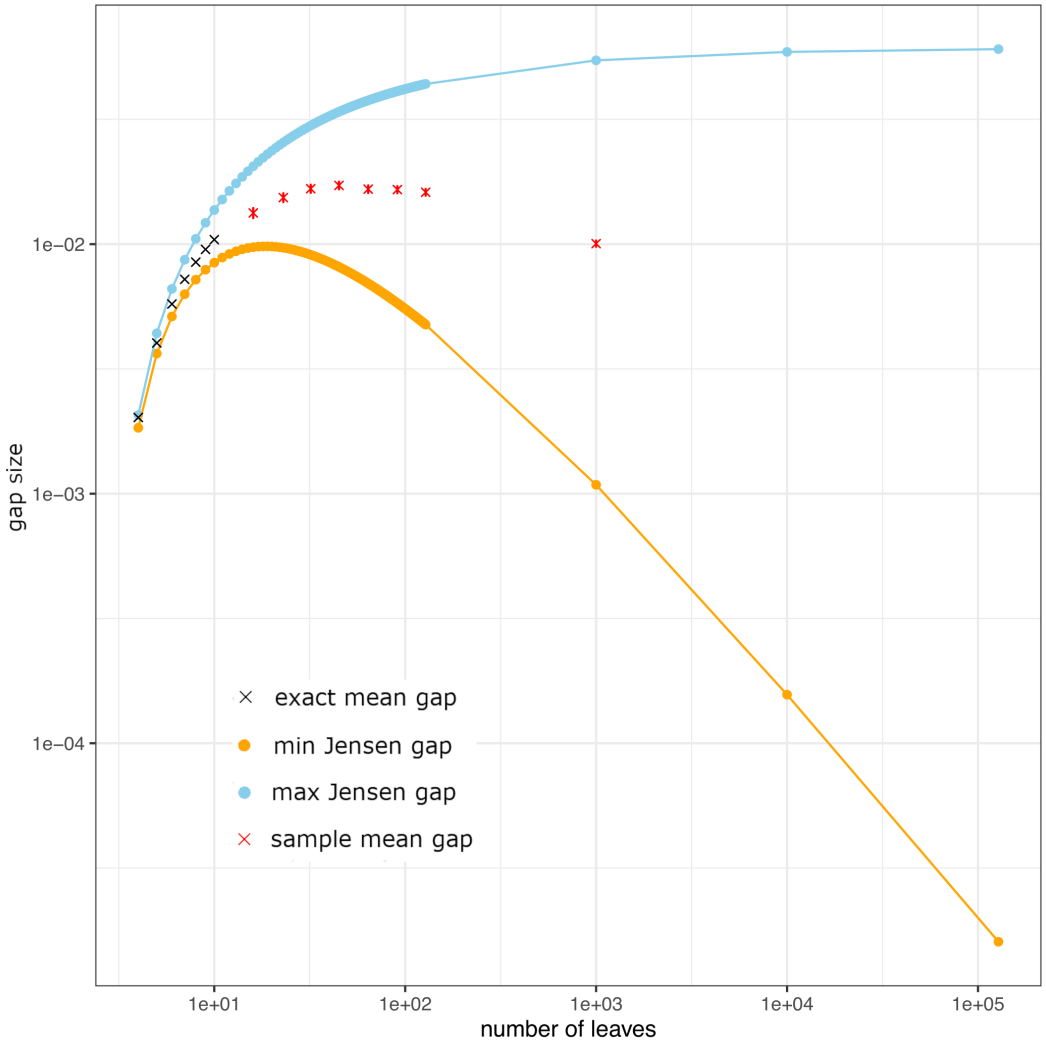

Figure 2.4: Extended figure 2.3C — uniform process. The convergence of the upper bound to $\frac{4}{3}\pi$ is much slower than the convergence of the lower bound to 0, and the maximum it reaches over the plotted range is 0.0604 for $n = 128000$. The red crosses, as in figure 2.3, suggest convergence of the gap size.

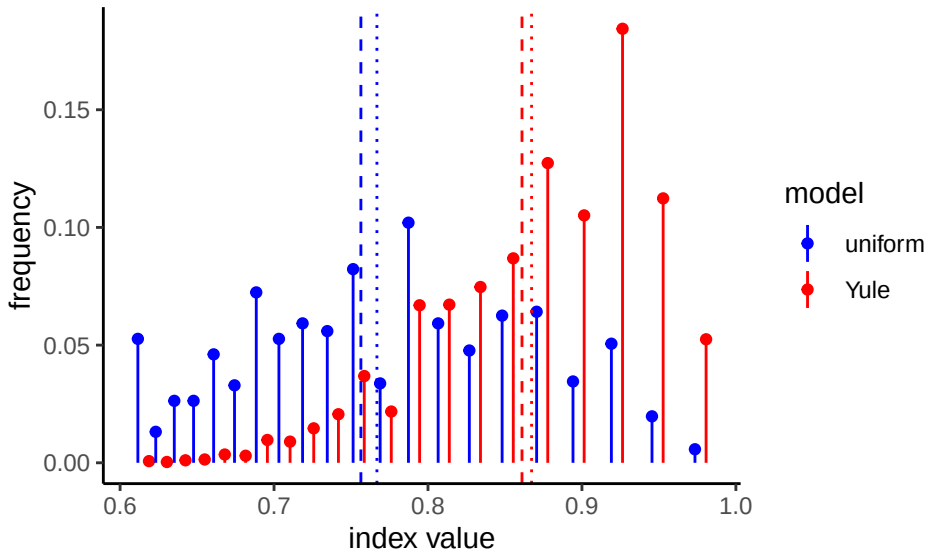

Figure 2.5: Higher variance in the uniform model leads to a non-zero upper bound on the Jensen gap. Shown are frequencies of J^1 values on 10-leaf trees generated under the Yule and uniform models. The dashed lines represent the true expected value of J^1, and the dotted lines the approximate value calculated as $\frac{n \log_2 n}{\mathbb{E}(I_S)}$

.

The lower bound of $\mathbb{E}_U(J^1) - \frac{n \log_2 n}{\mathbb{E}_U(I_S)}$ goes to 0 as $\frac{\log n}{n}$, while the upper bound tends to $\frac{4}{3\pi}$ for $n \to \infty$. This is a consequence of high variance in the uniform model (figure 2.5), as each tree on n leaves is selected with equal probability while the number of trees on n grows exponentially with n, the number of leaves. While I cannot prove analytically that the size of the Jensen gap in this case tends to 0, I can generate random trees using the uniform model and compare the sample mean to the approximation using the expected value of the Sackin index. In figure 2.3, I show behaviour of the Jensen gap and its bounds for J^1 under the Yule and uniform models. The red crosses in figures 2.3C and 2.4 indicate that the gap size does converge for $n \to \infty$. Therefore, I propose the following conjecture:

Conjecture 2.3.1. *For trees generated under the uniform model on $n \to \infty$ leaves, the following holds*

$$\mathbb{E}_U(J^1) \to \frac{n \log_2 n}{\mathbb{E}_U(I_S)}. \tag{2.32}$$

n	$n \log_2 n / \mathbb{E}_Y(J^1)$	$\mathbb{E}_Y(I_S)$	$n \log_2 n / \mathbb{E}_U(J^1)$	$\mathbb{E}_U(I_S)$
2	2	2	2	2
3	5	5	5	5
4	216/25	26/3	360/41	44/5
5	728/57	77/6	3822/289	93/7
6	1162800/67217	87/5	18.25643	386/21
7	199806750/9017743	223/10	23.81979	793/33
8	27.29901	962/35	29.87282	12952/429
9	32.68993	4609/140	36.38201	26333/715
10	38.30246	4861/126	43.31989	106762/2431
11	44.11464	55991/1260	n/a	n/a

Table 2.1: Comparison of exact and approximate expected values of J^1 and I_S under the Yule and uniform models.

2.3.4 Analytic properties of the J^1 index

The index J^1 is normalised in a way which makes comparison of its values on trees of different sizes valid (Lemant et al. 2022). As J^1 was defined to take into account node sizes, it can take any value between 0 and 1 for any given tree topology (figure 2.6). Furthermore, since J^1 is defined to be 0 on linear trees, finding its minimal value on a given number of nodes is trivial. In this section I investigate extremal values on trees where I impose restrictions to both topology and node size distributions, i.e. consider only leafy trees with out-degree of each internal node greater than 1.

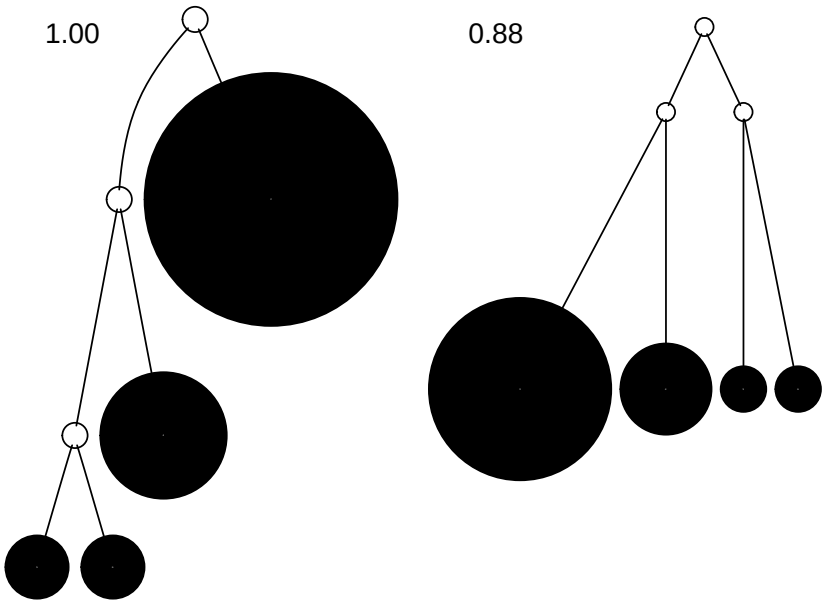

Figure 2.6: By including the node-balance function W^1 in J^1, we allow for the possibility of perfectly balanced caterpillars (left) and less balanced fully symmetric trees (right) based on the node size distribution in the tree. The leaf sizes in these two trees are identical, with a ratio $4:2:1:1$ from largest to smallest.

2.3.5 Properties of J^1 on different tree families

For most balance indices in use in evolutionary biology, the least balanced tree for a given number of leaves n is the binary caterpillar tree. I have previously derived a general expression for leafy trees of this topology (Lemant et al. 2022)

$$J^1(T_C) = \frac{2n \log_2 n}{(n-1)(n+2)}. \tag{2.33}$$

Most balance indices in literature define the caterpillar topology as the least balanced one (Fischer et al. 2021). Intuitively, this makes sense as balance is often associated with symmetry, and the caterpillar is the most asymmetric bifurcating tree. However, in the context of the J^1 index, tree topology is just one of a few factors which contribute to the balance score of a tree, especially since the index does not limit the space of trees to bifurcating ones. Also important to consider are node sizes and, more specifically, how the population is split across different subtrees in the tree of interest. Let us consider a slightly altered caterpillar topology.

Definition 2.3.5. Let T_B be a leafy tree on n leaves. Let every internal node of T_B except for the most distant one from the root have out-degree 2 such that one of its

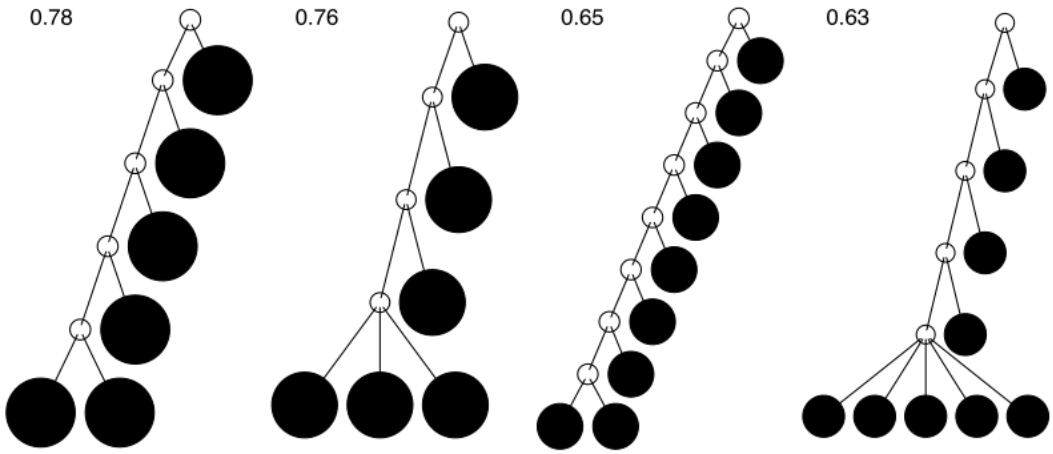

Figure 2.7: If we limit our search to leafy trees with equal leaf sizes, the least balanced tree on a given number of leaves is not necessarily the caterpillar. Pictured are the caterpillar trees on 6 and 9 leaves, as well as minimally balanced brooms for 6 and 9 leaves, with corresponding J^1 values.

descendants is a leaf, and the other an internal node. Further, let the internal node most distant from the root have out-degree k. Then we call tree T_B a **broom tree**. We call the leaves attached to the internal node with the highest out-degree the **broom head**, and the remaining leaves are attached to the **handle**.

A general expression of J^1 for this family of trees is then derived.

Proposition 2.3.4. *The value of J^1 for a broom tree T_B on n leaves, of which k in the broom head is*

$$J^1(T_B) = \frac{2\left(n\log_2 n - k\log_2 k + k\right)}{(n+k)(n-k+1)}. \tag{2.34}$$

Proof.

$$
\begin{aligned}
J^1(T_B) &= \frac{1}{\sum_{l=k}^{n} l} \sum_{i \in \tilde{V}} S_i^* \sum_{j \in C(i)} W_{ij}^1 \\
&= \frac{-2}{(n+k)(n-k+1)} \sum_{i \in \tilde{V}} S_i^* \sum_{j \in C(i)} \frac{S_j}{S_i^*} \log_{d^+(i)} \frac{S_j}{S_i^*} \\
&= \frac{-2}{(n+k)(n-k+1)} \left(\sum_{\substack{i \in \tilde{V} \\ d^+(i)=2}} S_i^* \sum_{j \in C(i)} \frac{S_j}{S_i^*} \log_2 \frac{S_j}{S_i^*} + k \cdot k \cdot \frac{1}{k} \log_k \frac{1}{k} \right) \\
&= \frac{-2}{(n+k)(n-k+1)} \left(\sum_{\substack{i \in \tilde{V} \\ d^+(i)=2}} S_i \left(\frac{S_i - 1}{S_i} \log_2 \frac{S_i - 1}{S_i} + \frac{1}{S_i} \log_2 \frac{1}{S_i} \right) - k \right) \\
&= \frac{2}{(n+k)(n-k+1)} \left(\sum_{i=k+1}^{n} i \left(\frac{i-1}{i} \log_2 \frac{i}{i-1} + \frac{1}{i} \log_2 i \right) + k \right) \\
&= \frac{2}{(n+k)(n-k+1)} \left(\sum_{i=k+1}^{n} \left((i-1) \log_2 \frac{i}{i-1} + \log_2 i \right) + k \right) \\
&= \frac{2}{(n+k)(n-k+1)} \left(log_2 \frac{n^n k!}{k^k n!} + \log_2 \frac{n!}{k!} + k \right) \\
&= \frac{2}{(n+k)(n-k+1)} \left(\log_2 \frac{n^n}{k^k} + k \right) \\
&= \frac{2}{(n+k)(n-k+1)} \left(n \log_2 n - k \log_2 k + k \right)
\end{aligned}
$$

□

The result of proposition 2.3.4 is directly generalisable in the following way.

Proposition 2.3.5. *For a broom tree T_{Bq} on n leaves, of which k in the broom head, such that the sizes of leaves in the head sum to $q \in \mathbb{R}$, and the leaves in the handle all of equal size 1, the value of J^1 is*

$$
\begin{aligned}
J^1(T_{Bq}) = &\frac{1}{(n-k+1)(q+(n-k)/2)} \\
&\times \left(q \log_k q - \left(\sum_{i=1}^{k} l_i \log_k l_i \right) + (q+n-k) \log_2 (q+n-k) - q \log_2 q \right),
\end{aligned}
\tag{2.35}
$$

where $l_1, \ldots, l_k$ are the leaf sizes which add up to q.

In figure 2.7 I show that the caterpillar is not the minimally balanced leafy tree

for a few tree sizes. To take it a step further, consider the following proposition.

Proposition 2.3.6. *For leafy trees on n leaves and no linear parts, the caterpillar minimises J^1 iff $n < 5$.*

Proof. Let $J^1_B(n,k)$ denote the value of J^1 on a broom tree with n leaves, of which k in the broom head. Then

$$J^1_B(n,2) = \frac{2n\log_2 n}{(n+2)(n-1)}, \tag{2.36}$$

$$J^1_B(n,3) = \frac{2}{(n+3)(n-2)}(n\log_2 n - 3\log_2 3 + 3). \tag{2.37}$$

Consider the case when $J^1_B(n,2) < J^1_B(n,3)$. Plugging in equations (2.36) and (2.37), we can rearrange the inequality to find

$$8n\log_2 n - 6(n^2+n-2)\log_2 3 + 6(n^2+n-2) > 0, \tag{2.38}$$

which changes sign at $0, 0.667, 1$ and 4.168. Setting the first derivative of this expression to zero

$$8\log_2 n + \frac{8}{\log 2} - (12n+6)\log_2 3 + 12n + 6 = 0$$

we find solutions around $n = 0.822$ and $n = 2.888$, the latter of which signifies a local maximum. Therefore, as n can only take positive integer values, valid solutions for which the caterpillar is less balanced than the broom with 3 leaves in the broom head according to the index J^1 are 3 and 4, with the $k = 3$ broom being less balanced otherwise. □

This proposition gives us a threshold for the number of leaves at which the caterpillar is no longer the minimally balanced tree for the given number of leaves, which sets J^1 apart from conventional balance indices (figure 2.8A). However, I am yet to prove the following statement.

Conjecture 2.3.2. *For leafy trees on n leaves and no linear parts, the tree that minimises J^1 belongs to the broom family.*

2.3.6 Behaviour as $n \to \infty$

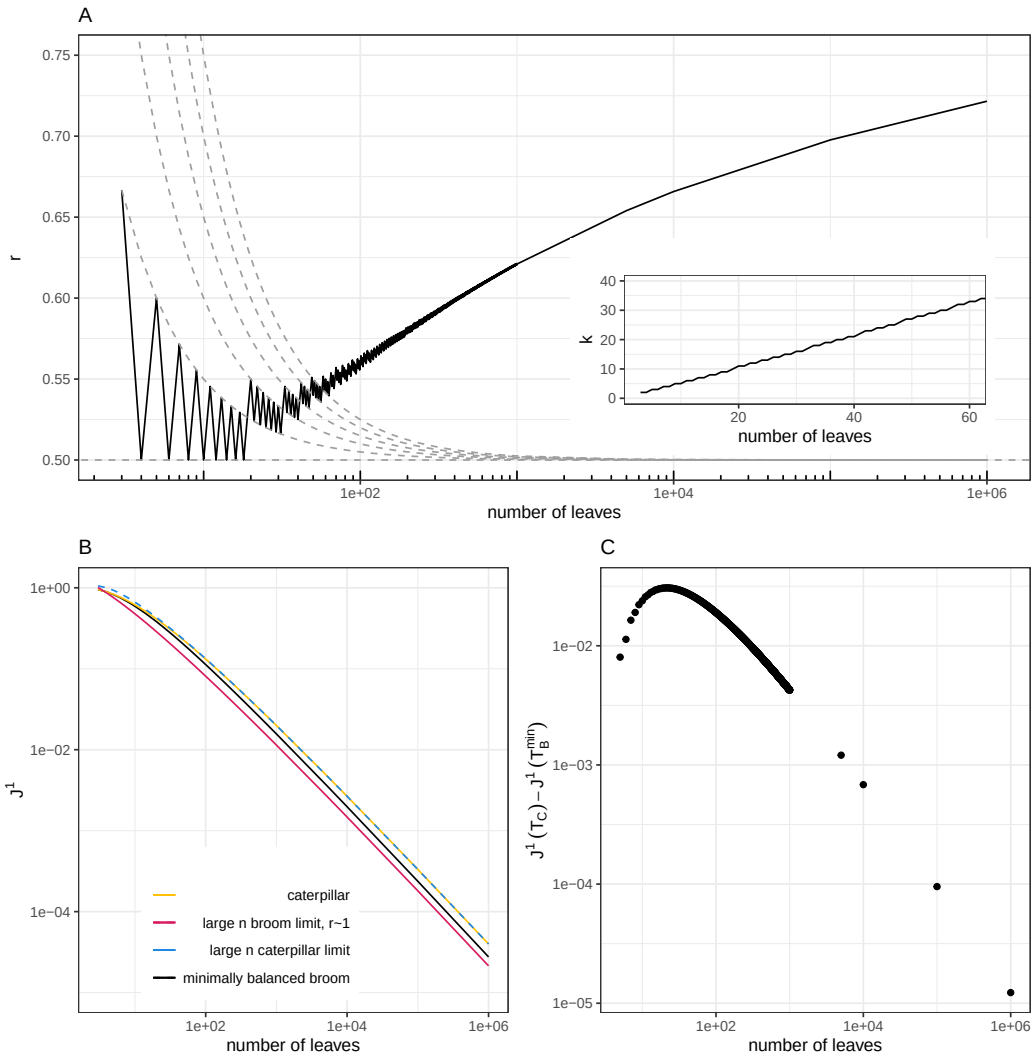

Figure 2.8: The labels used in the figures are as above - n for number of leaves, k for number of leaves in the broom head, $r = n/k$. **A:** Value of r for which the minimum value of J^1 is obtained on leafy trees. Trees on n leaves which satisfy $r = \frac{n+a}{2n}$, for $a = 0, 1, 2, \ldots$ lie on the dashed grey lines. The inset plot shows $k = rn$, the number of leaves attached at the broom head. **B:** Comparison of true and approximate values of J^1 for the caterpillar and minimally balanced broom trees as a function of n. **C:** The difference between values of J^1 of the minimally balanced broom and the caterpillar trees.

I have derived general behaviour of J^1 on broom and caterpillar trees for a given number of leaves n, showing that caterpillar trees are not necessarily minimally balanced for a given number of leaves. If we let $n \to \infty$, the value of J^1 for the caterpillar from equation (2.33) will behave like

$$\lim_{n\to\infty} J^1(T_C) = \frac{2\log_2 n}{n}. \tag{2.39}$$

As J^1 is not limited to trees with equal leaf sizes, there is a threshold we can impose on the broom tree beyond which the caterpillar is less balanced.

Proposition 2.3.7. *Let $T_B(n)$ be a broom tree on n leaves such that the leaves on the handle and head have sizes f and fp respectively, and $T_C(n)$ be a caterpillar tree on n leaves of equal sizes f. Then*

$$J^1(T_B) > J^1(T_C) \quad \textit{iff} \quad p < \frac{1}{2}, \tag{2.40}$$

as $n \to \infty$.

Proof of proposition 2.3.7. Let $n \to \infty$. The the value of J^1 for the caterpillar tree tends to

$$J^1(T_C) = \frac{2\log_2 n}{n},$$

and for the broom tree with equally sized leaves of size p in the broom head

$$J^1(T_{B,p}) = \frac{2}{n(1-r)(1+r(2p-1))}\left[(r(p-1)+1)\log_2 n(r(p-1)+1) - rp\log_2 nrp\right].$$

We can evaluate the difference between these expressions:

$$\begin{aligned} J^1(T_C) - J^1(T_{B,p}) \sim & (1-r)(1+r(2p-1))\log_2 n + rp\log_2 nrp \\ & - (r(p-1)+1)\log_2 n(r(p-1)+1) \\ \sim & ((1-r)(1+r(2p-1)) - (r(p-1)+1) + rp)\log_2 n + o(\log_2 n). \end{aligned}$$

The difference is dominated by the term containing $\log_2 n$ which is always positive. The term in the brackets preceding it can be negative, however:

$$(1-r)(1+r(2p-1)) - r(p-1) + 1 + rp = r(1-r)(2p-1).$$

As $r = k/n$, with k the number of leaves in the broom head, it is always positive. Thus, the expression is negative only when $2p - 1 < 0$ or $p < \frac{1}{2}$ □

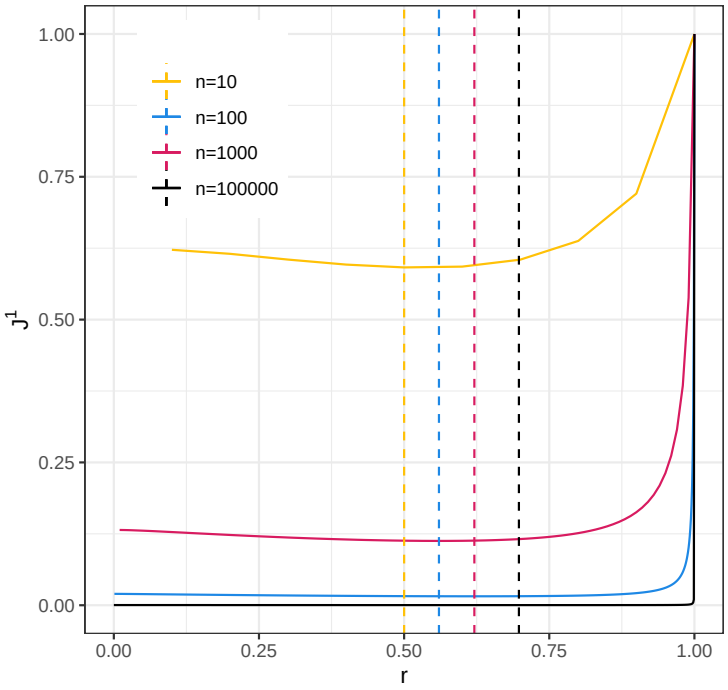

Figure 2.9: Values of J^1 on trees of different sizes calculated using equation (2.34) for different values of $r = k/n$. The dashed lines are at values of r which minimise J^1.

For broom trees, the behaviour is a little more complicated and, perhaps, counter-intuitive (figure 2.9). Consider the following.

Proposition 2.3.8. *Let $\mathcal{T}_B(n)$ be the set of all leafy broom trees with equal leaf sizes on n leaves, $r = \frac{k}{n}$ where k is the number of leaves in the broom head for a given tree, and r_{opt} the value of r which minimises J^1 for a given n. Then $r_{opt} \to 1$ as $n \to \infty$.*

Proof. Let $r = k/n$ and $J^1_B(n, r)$ the value of J^1 for a broom tree on n leaves, of which k in the head. Then

$$J^1 \overset{n\to\infty}{=} \frac{2}{n(1-r)(r+1)}((r+1)\log_2 n(r+1) - r\log_2 nr) \tag{2.41}$$

which is minimised for $r \to 1$. □

The proposition says that most leaves on a minimally balanced broom tree will be concentrated in the head, with comparatively few on the handle, resembling a

star tree more closely than a caterpillar tree. However, one must take into account how imbalanced the nodes above the broom head are, since one of their subtrees contains most of the tree's leaves, whereas the other is a single leaf. For practical purposes, the difference between the J^1 values of the minimally balanced broom and the caterpillar for the number of leaves $n \to \infty$ is small and decreases rapidly as n grows (figure 2.8B, 2.8C).

Finding the true value of k which minimises $J^1(T_B)$ analytically is difficult. The derivative with respect to k of equation (2.34) yields a transcendental equation which is not analytically solvable. I also cannot analytically determine whether broom trees minimise J^1 for a given number of leaves. However, I have exhaustively checked whether the broom minimises J^1 up to 12 leaves — which it does. Beyond that, the number of possible trees grows too rapidly for a similar verification to be computationally feasible without an efficient tree generating algorithm for trees with arbitrary node degree distributions.

2.4 Discussion

The aim of this chapter was to explore deeper analytic properties of the universal balance index J^1 and carve its place in the broader context of tree balance by extending past results and uncovering new connections.

In the chapter I focussed on trees with uniform branch lengths, as J^1 was not defined with defined branch lengths under consideration. A further generalisation of metric describing tree properties is therefore the logical next step (Noble & Verity 2023).

I calculated an approximate expectation of J^1 under the most common null models used in evolutionary biology. Having a good approximation for the expected value of J^1 is a crucial result in the development of this index, as it allows us to employ it in the analysis of evolutionary processes on phylogenetic trees. The next step in this direction would be to obtain a closed-form solution for the expectation of J^1, as well as its variance.

Finally, I only touched upon directly obtainable relationships without considering different real-world use cases of the index and the implications of equation (2.9). This is another avenue of future research as there may exist a relationship between the way indices vary with time and the underlying evolutionary process growing the

associated tree.

Chapter 3

Tracking modes of cancer evolution *in silico* via tree shape indices

3.1 Introduction

A trajectory is a path described by any object (or indeed point) in some space according to some parameter, usually time. Intuitively then, an evolutionary trajectory refers to the changes that a lineage or population undergoes over time — the series of genetic, morphological, and behavioral transformations that occur as organisms evolve and diversify. I am interested in the evolutionary trajectory of cancers but reliably obtaining time-series data is, at the time of writing, not feasible at a larger scale. This stems from multiple issues. Firstly, at time of diagnosis, solid tumours have likely already been growing for long enough to reach a size visible in standard medical imaging (Patrone et al. 2011). This means that even initial data obtained in the clinic represents a relatively late stage in the cancer's evolutionary history most of the time. Secondly, solid tumours are just that — clumps of cells organised in some way in space — meaning that taking a sample from one point in the tumour is not necessarily representative of the rest of the cell population. Finally, a biopsy is an invasive procedure which can cause considerable discomfort to patients, depending on where the tumour is situated. Therefore, having a reasonable estimate of a tumours evolutionary trajectory based on the data that is available at time of sequencing would allow for a more informed treatment strategy.

This begs the question — how can we distinguish between different ways tumours evolve? Is it necessary to wade through sequences of genetic data to do so or is it possible to abstract the key properties of a tumour's evolutionary trajectory into a few numerical summaries?

In this chapter, I will examine the utility of two different sets of tree shape indices for tracking the evolution of tumours on the example of agent-based simulations. By comparing the trajectories of these indices over time and with respect to each other, I aim to determine whether different modes of tumour evolution indeed occupy distinct regions of the index space. As this is a novel approach to condensing the information contained in cancer phylogenetic trees, inspired by (Noble et al. 2022), it raises the question of how much data is needed to reliably reproduce the results in the lab or even in the clinic.

3.2 Preliminaries

3.2.1 Why even bother with indices?

Before introducing the sets of indices used to analyse properties of trees, let us consider a simpler question — is it possible map the set of all possible trees to the set of real numbers? For this purpose I had to decide how to define the set of trees. The number of nodes in a tree is a natural number, $n \in \mathbb{N}$, as is the number of possible tree topologies for a given n. We denote with $T(n)$ the set of enumerated tree topologies (Nakano 2016). Each node then has a corresponding size, giving us an n-tuple of real numbers $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$, and each edge (branch) has a corresponding length or $(l_i, \ldots, l_{(n-1)}) \in \mathbb{R}^{(n-1)}$. This means we would need a family of maps

$$f_n : A(n) \times \mathbb{R}^n \times \mathbb{R}^{n-1} \to R. \tag{3.1}$$

It would be easy to construct a mapping which would allow us to "enumerate" each possible tree with a real number. The problem with this approach, however, is that it would not be very informative. The real numbers are not ordered in any way that would allow us to meaningfully compare trees. The lack of interpretability would render any application of such a mapping useless. This is where tree shape indices come in as a way to summarise key properties of a tree in a way that is both

interpretable and mathematically sound.

3.2.2 A 3-dimensional index space — trees with uniform branch lengths

Shannon diversity

Shannon entropy is a fundamental concept in information theory, that quantifies the uncertainty or randomness of a system (Shannon 1948). By considering a system where diversity represents the variety of elements, such as intra-tumour heterogeneity, we can define the Shannon diversity as the exponential of the Shannon entropy,

$$ {}^1D = \exp\left[{}^1H\right] = \exp\left[-\sum_{i=1}^{N} p_i \log p_i\right], \tag{3.2} $$

where N is the total number of categories (or elements, species, etc.), and p_i the frequency of category i. The Shannon diversity was chosen because of the nice property that it is maximised and equal to the number of categories when all categories are equally represented, and minimised when only one category is present. Previous work on a similar topic (Noble et al. 2022) was based on the Simpson index (Simpson 1949). However, the Shannon index was chosen for this work because it is more sensitive to changes in the frequencies of subclones, better interpretability, and the fact that the J^1 index is based on the Shannon entropy.

Mean number of drivers per cell — distance from the root

Each speciation event in phylogenetics or driver mutation in cancer evolution is associated with a change in the corresponding tree's topology. To capture the average number of these events, I use the mean number of drivers per cell. This is defined as the average of distances from all nodes to the root (with the root distance from itself defined as 1) weighted by the frequencies of the subclones,

$$ n = \sum_{i=1}^{N} p_i \nu(i), \tag{3.3} $$

where $\nu(i)$ is the root distance of node i.

Balance index

As discussed in chapter 2, the balance index J^1 is a weighted average of the evenness of the population distribution within a tree. We use it as the third index in this space.

3.2.3 A general set of indices — any rooted tree

Expanding upon the 3-dimensional space defined above, a new comprehensive set of interpretable robust indices based on Hill numbers was introduced recently (Noble & Verity 2023). The authors expanded and improved upon the existing quantifiers of tree shape properties by deriving methods for trees with arbitrary node size, node degree, and branch length distributions. The methods for calculating all of the indices are included as part of an R package (Verity 2023).

Each generalised index has three components, depending on which part of the tree it is applied — the longitudinal mean, node-wise mean, star mean.

Richness — ${}^0D$

Richness in the context of phylogenetics is simply the number of extant species, i.e. the number of tips in a phylogenetic tree. The generalised richness's three components are:

1. ${}^0D_L$ — the average number of branches across the tree;
2. ${}^0D_N$ — the average effective outdegree, ignoring branch sizes;
3. ${}^0D_S$ — the effective number of non-root nodes.

Diversity — ${}^qD$, $q > 0$

The generalised diversity index represents an extension of the Shannon diversity index. Its three components are:

1. ${}^qD_L$ — the effective number of maximally distant nodes (leaves);
2. ${}^qD_N$ — the average effective outdegree, accounting for branch sizes, i.e. bushiness;
3. ${}^qD_S$ — the effective numbering of branches, accounting for branch sizes.

Evenness — ${}^qJ$, $q > 0$

Finally, the extension of the robust universal balance index J^1, this set of indices generalises tree balance in the following way:

1. ${}^qJ_L$ — evenness of branch sizes across the tree, or tree symmetry for leafy and ultrametric trees;

2. ${}^qJ_N$ — tree balance, or evenness of the node size distribution;

3. ${}^qJ_S$ — evenness of all branch sizes.

3.3 Tree resolution

I rejected the idea of simply mapping trees to real numbers due to the lack of interpretability. Tree shape indices are nominally better, as they summarise properties of a tree, but they may have limitations in the form of a lack of resolution for certain types of trees.

Starting simple, we examine leafy trees with all leaves of equal size in the 3-dimensional index space. The first thing to note is that the Shannon diversity will simply equal the number of leaves in the tree. This already takes away a degree of freedom. The next thing to consider is the value of J^1. If we limit our search, for now, to perfectly balanced trees, we are left with symmetric trees on a fixed number of leaves N. To make the final index equal between two trees, they need to have equal average depths of their leaves. As we are only looking at perfectly symmetric trees, that means that the average depth will be exactly equal to the individual leaf depths. We can then show the following

Proposition 3.3.1. *Let T be a symmetric leafy tree on N leaves with equal leaf sizes. If the canonical factorisation of N is*

$$N = \prod_{i=1}^{k} \alpha_i^{l_i}, \tag{3.4}$$

then there are

$$\frac{\left(\sum_{i=1}^{k} l_i\right)!}{\prod_{i=1}^{k} l_i!} \tag{3.5}$$

distinct trees with the same values of J^1, ${}^1D$, and n, including T.

Proof. First, the values of indices J^1 and D for a symmetric leafy tree with N equally-sized leaves are

$$J^1 = 1,$$
$$D = N,$$
$$n = 1 + \sum_{i=1}^{k} l_i.$$

The result is then a simple combinatorial problem of placing $n-1$ balls (α_i's) into $n-1$ bins, with each α_i repeated l_i times. Therefore, the number of distinct trees is

$$\frac{\left(\sum_{i=1}^{k} l_i\right)!}{\prod_{i=1}^{k} l_i!}. \tag{3.6}$$

□

This case may be interesting mathematically, but is not too relevant for practical purposes as it is highly unlikely that sequencing data would yield a perfectly symmetric leafy tree. As the space of trees is so large, there is little point in performing a grid search, especially when arbitrary node sizes are considered. In testing, there have been no cases of trees with the same values of indices, but different topologies. This is a good sign, as it means that the indices are able to distinguish between different trees. However, the question remains whether there is a set of indices which can differentiate between any two trees.

3.4 Computational methods

3.4.1 Agent-based modelling framework - *warlock/demon*

There is no shortage of agent-based models of tumour evolution (Colyer et al. 2023), and the can range from purpose-built complex frameworks to more stripped-down and abstract ones. Since each model should be "as simple as possible but no simpler", the appropriate framework for our purposes must satisfy certain requirements — flexibility, efficiency, and reproducibility. The first requirement is deceptively specific. As the main inspiration behind this work stems from cancer evolution, I wanted the simulations to have parameters for controlling aspects of the cell population's physical properties which would in turn imply a different way in which it

evolves. This would, for example, include spatial arrangement of cells, mutation rates, migration rates, and selective advantage. Furthermore, while the goal is to simulate large populations of cells, I also need a large number of simulations over which more general deterministic properties can be deduced. Stochastic effects could make vastly different evolutionary modes look more similar than expected in theory. Finally, reproducibility allows us to share parameters of our models for verification by peers, and possible further investigation.

The agent-based modelling framework I decided to use is `warlock` (Bak et al. 2023), a `snakemake` wrapper written for `demon` (Noble 2020). It satisfies the requirements above, with a few associated comments. Firstly, it is a flexible agent-based model of tumour evolution as it does have parameters which control for spatial arrangement, mutation rates and selective advantage, as well as migration. While it is able to simulate spatial structure, `demon` covers at most two spatial dimensions. This is not an issue since we approximate the cell population to undergo stochastic isotropic growth, that is the tumour has equal probability of expanding in all directions in space. This implies approximate spherical symmetry of simulated solid tumours, which allows us to effectively consider the two-dimensional simulation as a cross-section of a tumour spheroid. In terms of efficiency, `demon` was written mainly in C++, and conceptualised so that instead of tracking individual cells, it simulates unique cell genotypes on a two-dimensional grid comprised of demes, or well-mixed patches of cells. The procedure for simulation cell events is based on the Gillespie algorithm (Gillespie 1977), and follows the steps of selecting a deme, then cell type, event type, and finally cell genotype. This approach sacrifices micro-scale interactions between cells to benefit efficiency and the feasibility of mathematical analysis of the model using, for example, diffusion approximations. Finally, all associated code is free and open source (Noble 2020, Manojlović 2023*a*, Verity 2023), which allows reproducibility using identical parameters and random seeds. Parameter values for different batches can be found in appendix A. Each batch of simulations, that is a distinct set of parameters, was run over 50 replicates distinguished by random seeds. Plots including individual and average trajectories are included in appendix A.

3.4.2 Spatial configurations

The simulations were run for four different spatial configurations of tumour evolution. Due to the difference in parameters, the aim was to have each run reach the target final population of 10^6 cells in between 500 and 1000 cell cycle events or generations. The spatial configurations were as follows:

- **Gland fission** — the first tumour cell forms the first deme which, once it reaches a certain size (carrying capacity), splits into two daughter demes. This process continues until the final population is reached.
- **Invasive glandular** — the first tumour cell forms the first deme but, as contrasted with gland fission, the daughter demes are formed by individual cells migrating away from the parent deme, rather than fission. This process continues until the final population is reached.
- **Non-spatial** — no spatial structure is imposed on the tumour, and cells are well-mixed. This is the simplest spatial model.
- **Boundary growth** — the simulation tracks individual cells in space, but cells are only allowed to divide if they are not fully surrounded by other cells.

3.5 Results

Index	Description	Range
$n(T)$	Mean number of drivers per cell — effective tree depth	$[1, N]$
$D(T)$	Shannon diversity — effective number of types	$[1, N]$
$J^1(T)$	Balance index — node distribution in the tree	$[0, 1]$
${}^qD_L(T)$	effective number of leaves	$[1, N]$
${}^qD_N(T)$	average effective node outdegree	$[1, N]$
${}^qD_S(T)$	effective number of branches, accounting for branch sizes	$[1, N]$
${}^qJ_L(T)$	evenness of branch sizes	$[0, 1]$
${}^qJ_N(T)$	evenness of node size distribution	$[0, 1]$
${}^qJ_S(T)$	evenness of all branch sizes	$[0, 1]$

Table 3.1: Summary of indices used in this section. N is the number of nodes in tree T.

3.5.1 Trajectories in 3-dimensional index space

As shown in (Noble et al. 2022), using three indices to track the evolutionary trajectory of a tumour can be quite informative. Depending on the choice of parameters,

the chosen spatial configurations occupy distinct sections of the index space (such as figure 3.1). This is a good sign, in the "reasonable" parts of parameter space. However, the question here would be — what are "reasonable" parameters? The answer to this is not straightforward, as it not only depends on the type of tumour and its clonal structure, but also varies between patients even within the same type of cancer. For example, high mutation rates even with low selective advantage could lead to a tumour expected to follow a roughly branching trajectory to resemble neutral growth (figure 3.2). While I used a different diversity metric from (Noble et al. 2022), the results are consistent with their findindgs. The trajectories in figures 3.1 and 3.2, while averaged over 50 replicates each, still have a considerable amount of noise. This is due to the stochastic nature of the model, with the same set of parameters leading to evolutionary trajectories which can differ at certain points in time (figure 3.3).

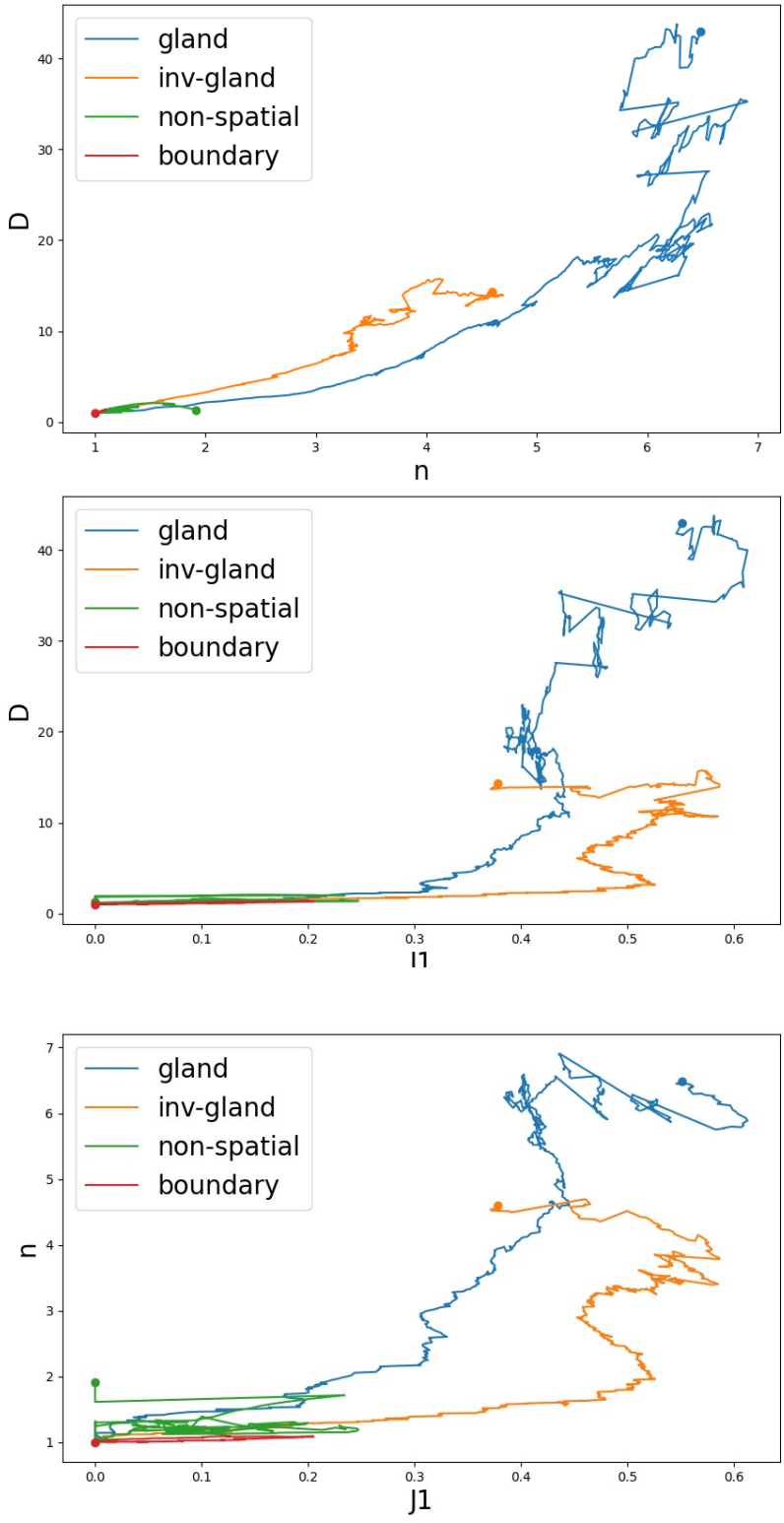

Figure 3.1: The average trajectories in 3-dimensional index space for four different spatial configurations of tumour progression (gland fission, invasive glandular, non-spatial, and boundary growth) are distinct and their final states (circles) lie in separate regions of index space. This example is averaged over 50 replicates for each trajectory. Parameters: mutation rate $\mu = 10^{-5}$, selective advantage $s = 0.1$.

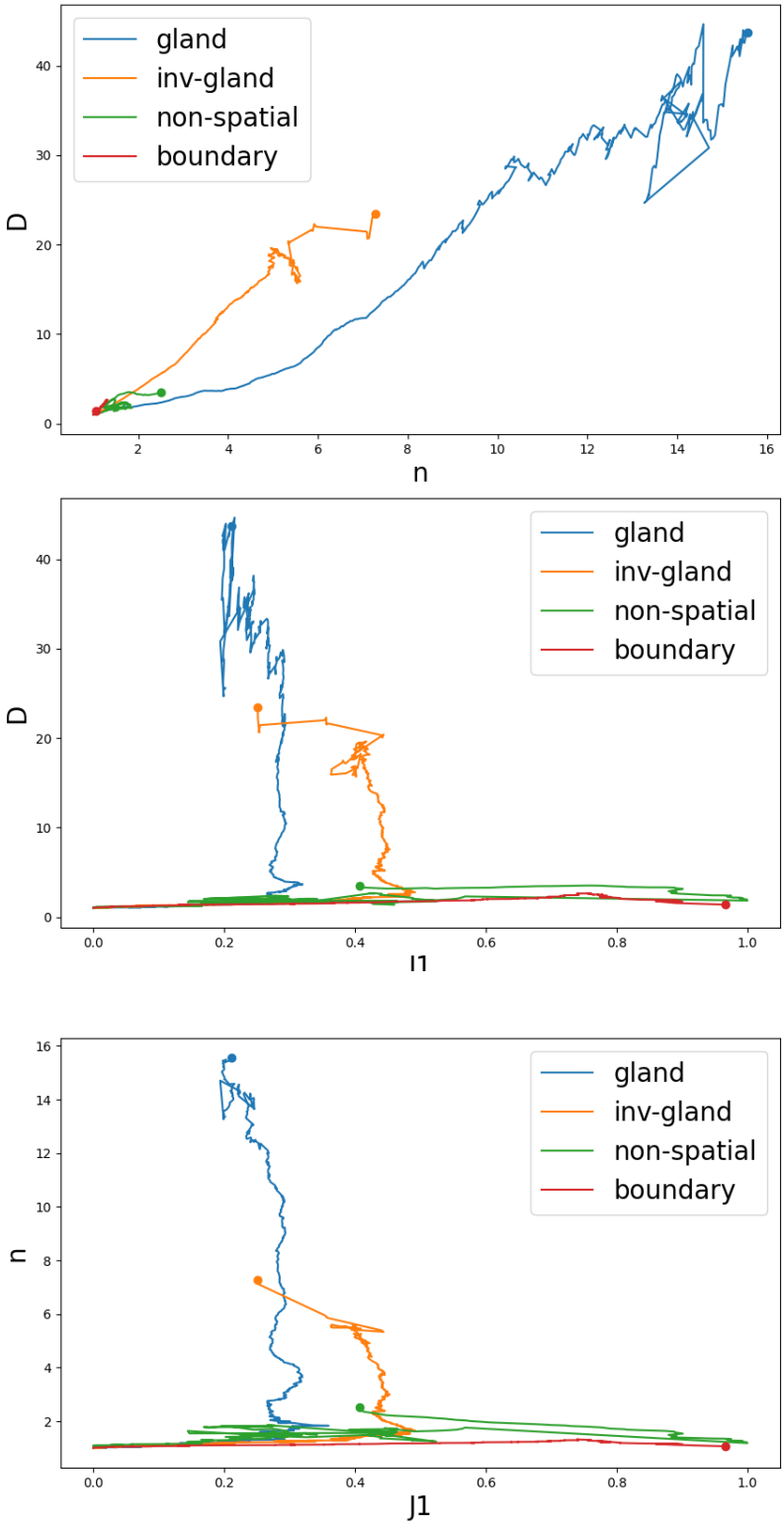

Figure 3.2: The average trajectories for a slightly different set of parameters from figure 3.1. The trajectories and final states are still distinct, with the final states lying in separate regions of index space. Parameter values: mutation rate $\mu = 10^{-4}$, selective advantage $s = 0.05$.

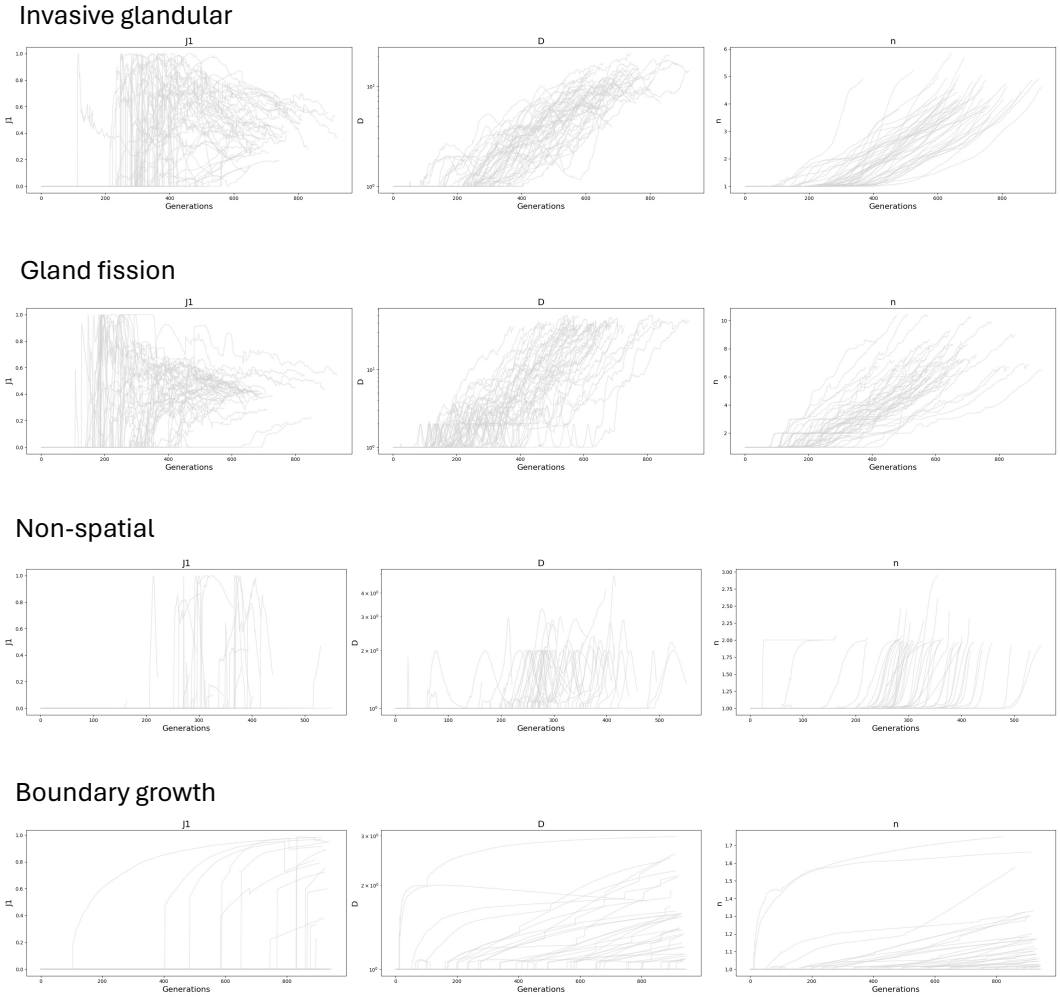

Figure 3.3: Individual replicates' trajectories for the parameters used in figure 3.1. While the shapes of individual trajectories are similar, as expected, there is still a lot of variation in the time at which the tumour reaches the final population size, leading to noisy average trajectories.

3.5.2 Trajectories in the new index space

As the 3-dimensional index space does not distinguish between topologically identical trees with different branch lengths, I decided to re-run the analysis on the same set of simulations using a recently introduced system of indices (Noble & Verity 2023). Using an early version of the R package for calculating the indices (Verity 2023), I was able to construct the average trajectories in this new larger index space. However, bigger is not always better, and there are some redundancies in the new set of indices. For example, in all cases, the longitudinal and node evenness indices were highly correlated across runs (figure 3.4). Furthermore, the richness indices were also correlated with the diversity indices, which is not surprising as the diversity indices are a generalisation of the richness indices. I will narrow down the the set of indices to just the diversity indices and one of the evenness indices in this section, with the full set of trajectories included in appendix A.

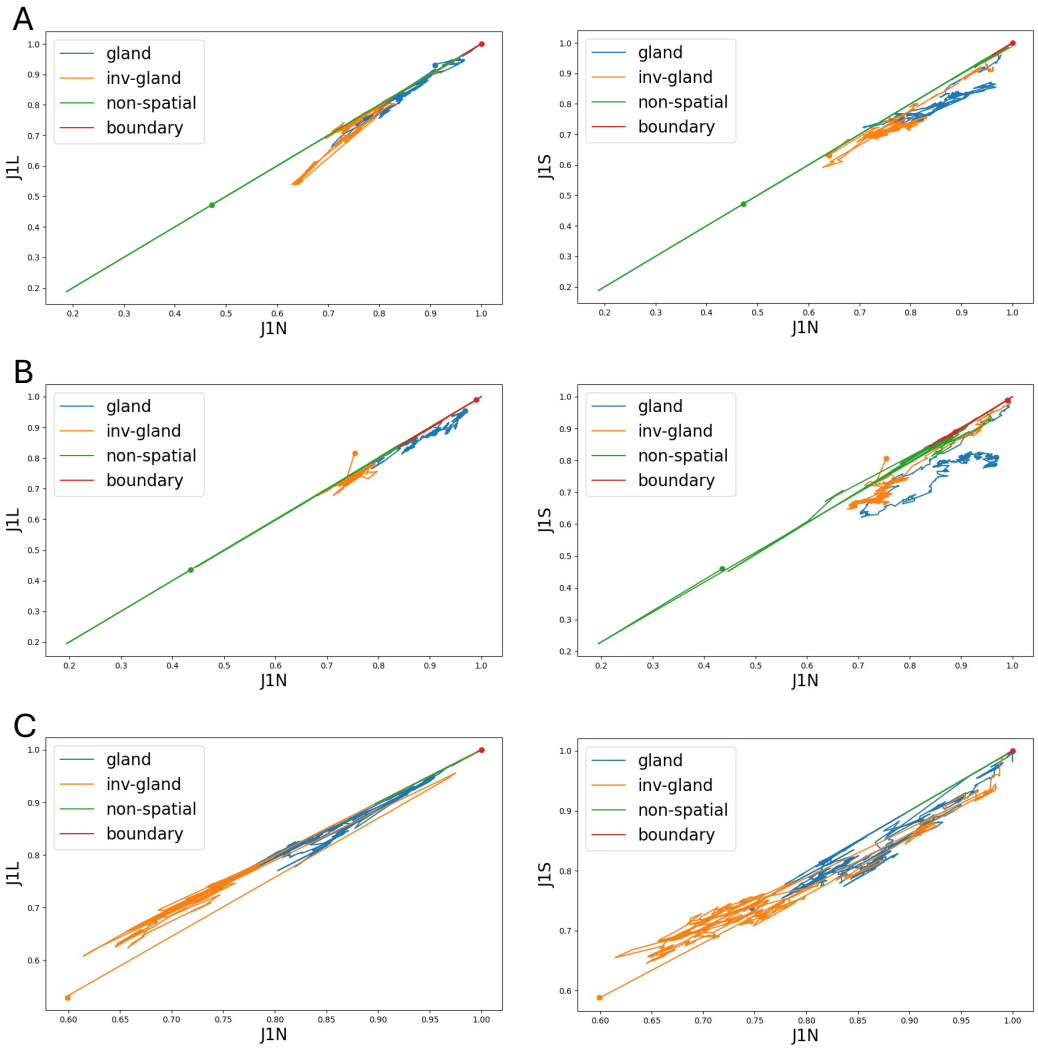

Figure 3.4: Trajectories of different evenness indices encode almost identical information. Shown above are evenness trajectories for different spatial configurations and sets of parameters averaged over 50 replicates each. Mutation rate (μ) and selective advantage (s) values:
A — $\mu = 10^{-5}$, $s = 0.1$; **B** — $\mu = 10^{-4}$, $s = 0.05$; **C** — $\mu = 10^{-6}$, $s = 0.2$.

During the analysis, I found that patterns observed in the new index space are similar to those of the old set of indices. Spatial configurations tended to separate into their own sections of the index space, with overlap present depending on the choice of parameters. This is not surprising, as the new set of indices is a generalised version of the old one. The main difference between the new and old trajectories is the amount of noise present in the new set, even when the outputs are averaged over multiple runs. This is likely due, in part, to the inclusion of branch lengths in the new set of indices, which may vary greatly between runs. The trajectories for the same data sets as above are shown in figures 3.5 and 3.6. An example of trajectories of each individual replicate is shown in figure 3.7, with other spatial configurations' figures included in appendix A. The reason for the noise in the new set of indices is the same as in the old, as the same data was used in the analysis — there is a lot of variation in the time at which the tumour reaches the final population size.

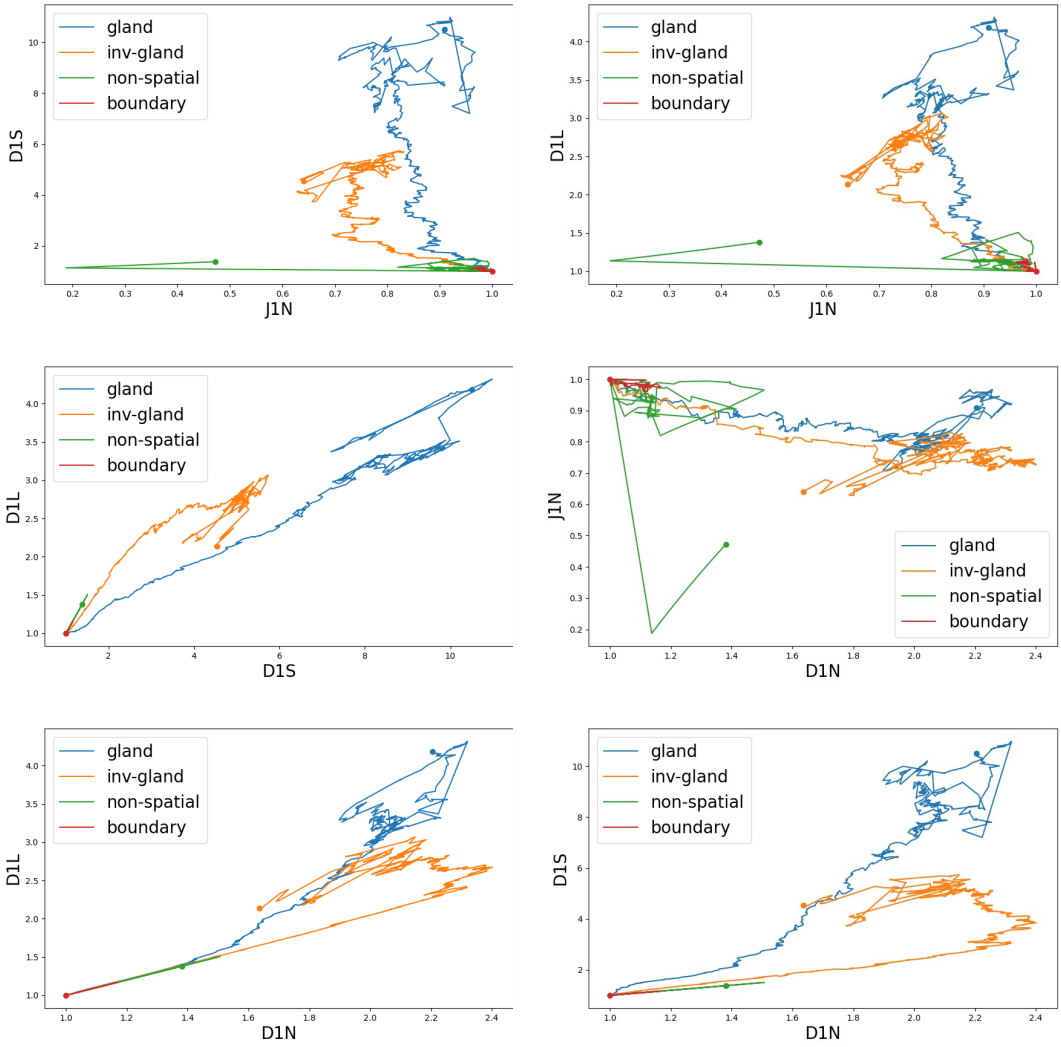

Figure 3.5: Introducing more dimensions to the index space does not change the broad conclusions of the analysis. The average trajectories are distinct between spatial configurations, with the final states lying in separate regions of index space. Parameters: mutation rate $\mu = 10^{-5}$, selective advantage $s = 0.1$.

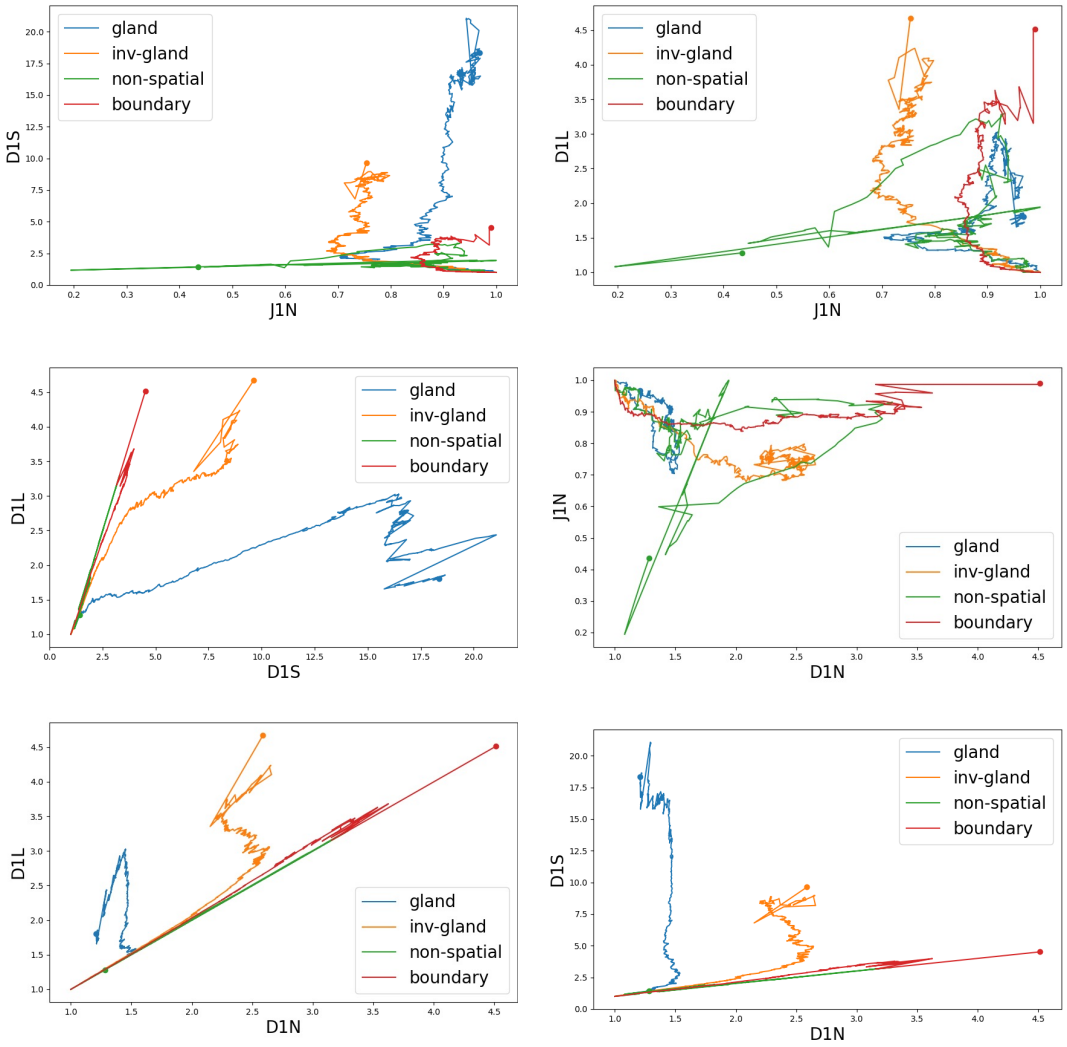

Figure 3.6: Changing the key parameters in the new index space has a similar effect to the old one. The trajectories and final states are still distinct, with the final states lying in separate regions of index space. Parameters: mutation rate $\mu = 10^{-4}$, selective advantage $s = 0.05$.

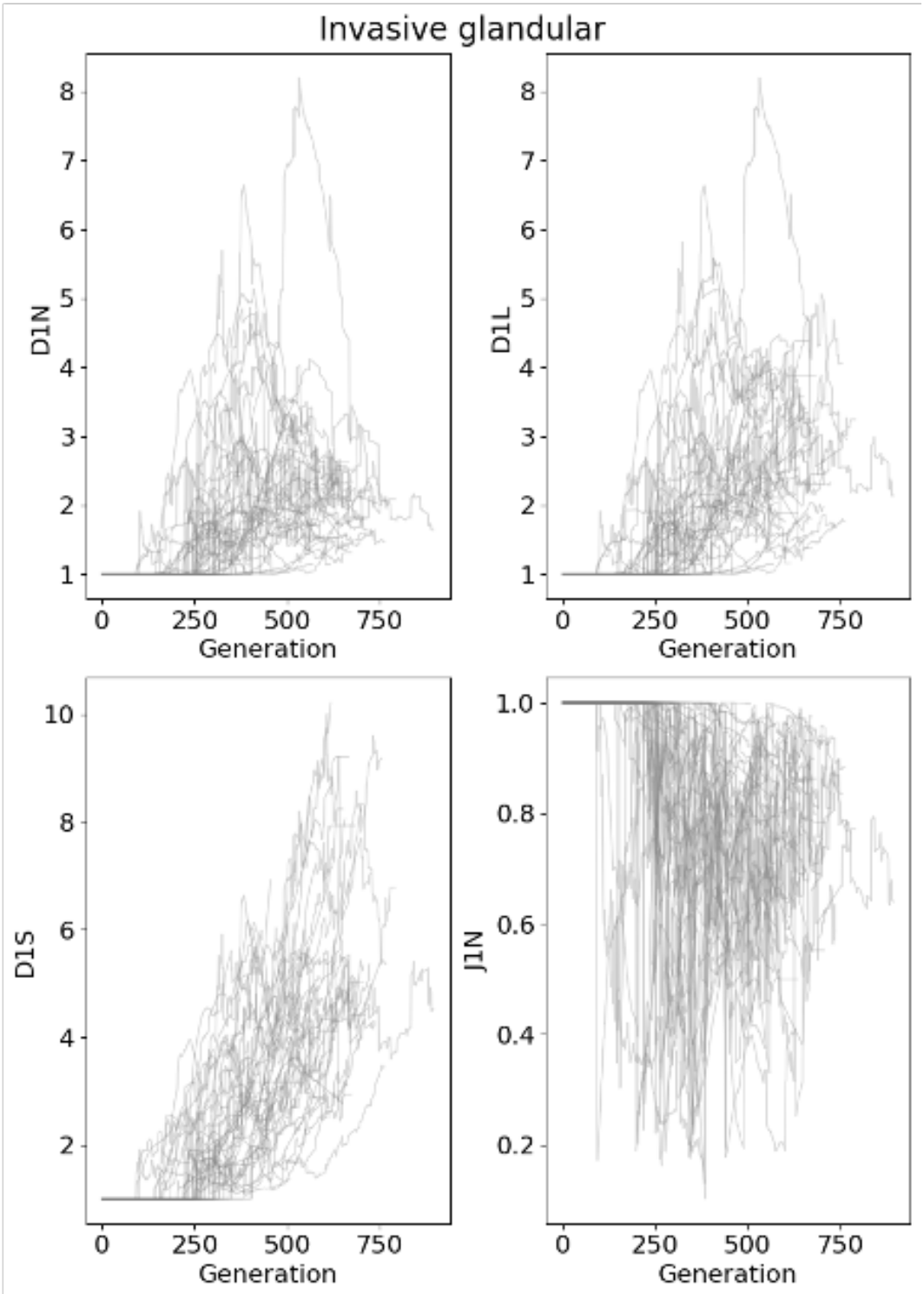

Figure 3.7: Individual replicates' index trajectories of invasive glandular expansion for the parameters used in figure 3.5. As before, there is a lot of variation in the time at which the tumour reaches the final population size, leading to noisy average trajectories.

3.6 Discussion

In this chapter I examined two sets of tree shape indices for tracking the mode of tumour evolution on the example of agent-based simulations. The results suggest that the indices are able to distinguish between different modes of evolution depending on parameter choice, but that further work is needed to establish ranges of parameter values which are useful. While a moderate mutation rate and weak selection seem to be a good starting point, applying the indices to real data should be the next step in order to validate the approach. Furthermore, the new set of indices introduced in (Noble & Verity 2023) is more informative due to its inclusion of branch lengths, but also contains some redundancy as pairs of indices can display high correlation across runs. This redundancy may not necessarily be present in all individual simulations, but averaging even a few dozen runs shows that pairs of indices encode similar information. However, in individual runs, this behaviour is not always the case. For this reason, further work is needed to develop mathematically sound methods for the comparison of individually simulated trajectories. This would allow for data analysis focussed on what could be called "typical" behaviour, rather than the average of a large number of replicates, which would further take advantage of the stochastic nature of the model. This would also address the stochasticity resulting from individual runs, as I propose developing

The new index set is, as the name suggests, new, and there are still some technical wrinkles to iron out. Namely, large trees with many nodes and branches can take a long time to calculate, as the indices from the set are more complex than the old ones. Additionally, there is a precision issue which arises when some of the smallest nodes are included in the tree. This is easily resolved by employing a similar approach as in 1.2, which effectively simulates sampling error. Finally, the R package itself does not directly support the data structure used in the modelling workflow. The conversion, which was purpose-written for this work, may also contribute to the aforementioned technical difficulties. Standardising the data types in a future study would improve the efficiency of the analysis, and possibly reduce the amount of noise between simulated time points.

The main contribution of this chapter is the exploration of different sets of indices, not just over time, but in relation to each other. The first time a similar approach was used was in (Noble et al. 2022), and here I expanded the analysis to

a new set of indices. Additionally, in the context of the new index set, I was able to expand the discussion to include larger trees than previously analysed with the indices (Noble & Verity 2023). This is important as the typical cancer phylogeny may not have thousands of nodes, but there are other applications of large trees which could benefit from this method. Recently, there has been work considering a similar approach for the classification of evolutionary processes in biogeography, which further supports the utility of the indices (Freitas et al. 2024). Personally, I am most interested in how this approach would fare when applied to time series data of real cancer cell populations. Unfortunately, such data is rare and difficult, but not impossible (Salehi et al. 2021), to generate.

Having focussed on numerical summaries of the ABM outputs, I also believe that there may be a way to heuristically derive approximate general properties of a tumour's evolutionary trajectory. This could be done by deriving equations for the simplest evolutionary trajectories, such as progressive selective sweeps which follow a cycloid in the n, D-plane in the 3-dimensional index space. Additionally, as my main focus in this chapter has been the generation, processing, and visualisation of synthetic cancer data, there is a good scope for the application of more novel statistical learning methods to the data, such as evolutionary simulation (Herald et al. 2022), or deep learning

Chapter 4

Agent-based model of fluctuating methylation arrays in growing fragmented cancer cell populations

4.1 Introduction

In chapter 3, I used a general spatial agent-based model to investigate broad evolutionary patterns as related to spatial organisation. While the model was capable of simulating the dynamics of tumour growth, its utility is limited by the computational cost of simulating a large number of cells. This means that using the model's outputs in comparison to or to draw inference from real data is not feasible, as real tumours have billions of cells.

There are a few ways to address this issue. For example, rather than simulating all clones in a tumour, one could take the approach of (Sottoriva et al. 2015) and use demes (tumour glands) as the principal agent of our simulation. This would allow for a realistically-sized tumour to be generated as the number of glands would be around the right order of magnitude. A problem with this approach is that it loses resolution since a gland's population is assumed to be clonal, undergoing rapid fixation in the case of an emerging mutant. If one wanted to study evolutionary dynamics on a finer scale it would be necessary to at least simulate the dynamics of cell lineages, if not individual cells, as performed earlier.

The first effort to model clonal dynamics of cells based on fluctuating methylation clocks (FMCs) was made in (Gabbutt et al. 2022). Working with colon and small intestine samples, the authors found that somatic cell birth and death dynamics are measurable based on the FMCs, and began transferring this knowledge to the context of cancer. In (Gabbutt et al. 2023), the authors employ a stochastic model for an expanding cell population to model the behaviour of fluctuating CpG sites in blood cancers. The model is capable of simulating the dynamics and the corresponding fluctuating methylation arrays of lymphoid malignancies at scale. However, this model is not spatially explicit, which is a feature that has to be distinguished between different glands in a solid tumour. In this chapter, I present a purpose-written agent-based model, `methdemon`, which reduces the computational cost of simulating a tumour's growth and models the fluctuating methylation arrays in colorectal cancer.

4.1.1 Fluctuating methylation arrays

As mathematicians new to biology quickly learn, perfectly clean data containing detailed information about the population structure of a tumour is non-existent. In fact, most data is noisy and at best measures a decent proxy for the properties which can be described by a mathematical model. Therefore, one learns very quickly to adapt their thinking when working with biological data. Specifically, when it comes to cancer, a compromise has to be made between resolution and scale. Where single-cell data can provide a detailed view of the mutations accumulated in the genome, it is not feasible to obtain it for a whole tumour. On the other hand, bulk data gives a high-level view of the tumour's population structure, but a lot of the details get lost in the process.

However, DNA sequencing is not the only way to obtain information about a tumour's population structure. Early work with methylation arrays in colorectal cancer has shown potential for inferring the ancestry and age of a tumour (Hong et al. 2010, Siegmund et al. 2011). In a way the genome shows more mutations in older populations, methylation arrays will also be more diverse as time goes on. Current techniques allow for the sequencing of some $850,000$ CpG sites which, while a small fraction of the genome, is still enough to proide valuable insight into the underlying dynamics of the cell population. Initial studies on methylation as a tracker of evolution made use of the whole array (Siegmund et al. 2008, Sottoriva & Tavaré 2010). However, more recent work has shown that just a small subset of

CpG sites is enough to infer the evolutionary dynamics of a cell population (Gabbutt et al. 2022, 2023). This is the set of fluctuating CpG (fCpG) loci, which is also the topics of chapter 5.

As fCpG loci seem to be a neutral marker of cell population dynamics, I will define the following set of assumptions for modelling their behaviour:

(i) **Each cell has a corresponding fCpG array inherited from its parent cell.**

(ii) **Upon cell division, each methylated fCpG site has an independent and equal probability of being demethylated, and vice-versa.**

(iii) **The rates of methylation and demethylation do not change over time.**

These assumptions are based on the findings of (Gabbutt et al. 2022, 2023).

4.1.2 A comment on using existing models

As discussed in section 1.3, it is preferable to use established frameworks and models for simulating tumour growth and evolution. Therefore, with the assumptions outlined in the previous section, my initial approach was to employ a general agent-based model with small modifications, to simulate the behaviour of fCpG loci in cancer. The first model I considered was `demon`, with which I had already worked in chapter 3, as its light weight and reasonable wall times (total execution time) had shown promise. A naive approach to simulating methylation arrays is to use the model's passenger mutations as a proxy for epigenetic changes. This way, the model could be run as usual with modified passenger mutation rates, and methylation arrays could be assigned to the cells post-hoc. The main issue with this approach is memory management, as the output files tend to be large and thus difficult to handle in the post-processing steps. This meant that applying any sort of inference workflow would take a long time, making the approach impractical. This stands for other SABMs as well, due to the amount of data produced when simulating a whole tumour's growth.

4.2 An ABM of fluctuating methylation arrays in cancer

To tackle the issue of models generating too much unused data, I wrote a new ABM, `methdemon`, capable of simulating a growing cell population and their corresponding fCpG arrays. The model can be run as a well-mixed population expansion, or in a 1D spatial setting, with the latter being relevant in the case of multi-site sequencing of a tumour spheroid. The model is written in C++, with an emphasis on execution time and dynamic memory management.

4.2.1 Model structure

Inspired by the `demon` model, event scheduling happens according to the Gillespie algorithm, with a deme being chosen first, followed by a cell within that deme. Events are then chosen based on the sum of rates in the tumour, between cell birth, cell death, and deme fission. Upon cell division, both the parent and daughter cells have the same probability of acquiring a driver mutation, and each cell's fCpG array is updated according to the rules outlined in the previous section. A list of relevant model parameters is given in table 4.1, with source code and examples available in the model's github repository (Manojlović 2023*b*). Consider a tumour consisting

Parameter	Description	Units
`deme_carrying_capacity`	The maximum number of cells in a gland	cell
`init_migration_rate`	The probability of a gland undergoing fission	$\text{cell}^{-1}(\text{cell division})^{-1}$
`mu_driver_birth`	The probability of a cell acquiring a driver mutation	$\text{cell}^{-1}(\text{cell division})^{-1}$
`s_driver_birth`	The selective advantage of a mutant cell	n/a
`meth_rate`	The probability of an unmethylated fCpG site changing state	$(\text{cell div})^{-1}(\text{fCpG site})^{-1}$
`demeth_rate`	The probability of a methylated fCpG site changing state	$(\text{cell div})^{-1}(\text{fCpG site})^{-1}$

Table 4.1: Parameters used in the `methdemon` model.

of N demes at the end of growth. Each deme, during growth, has a probability p of undergoing fission, and the final mean number of fissions per deme is $\log_2 N$.

Further, the expected number of descendant demes of deme i is given by

$$\frac{N_i}{N} = \frac{(2pt)^i}{i!} e^{-2pt}, \tag{4.1}$$

i.e. the Poisson distribution with mean $2pt$, as derived in (Kharlamov 1969). This condition governs how fissions are handled in the model. As a real tumour can consist of millions of glands, the model narrows the focus to the subset of demes sequenced at the end of growth. Fissions which produce untracked demes, untracked fissions from here, are the main way of accumulating fission events in a deme. Each untracked fission simply discards half of the deme's population, stochastically rounded. Tracked fissions are implemented by assigning a probability to each fission event of being tracked (figure 4.1), equal to

$$\phi = \frac{p}{\mathbb{E}[\text{fissions per deme}]/2}. \tag{4.2}$$

By using this discrete uniform distribution, I ensure that the expected number of fissions before a tracked fission event in a deme is about half of the mean total number of fissions. Were the individual probabilities equal to $1/\mathbb{E}[\text{fissions per deme}]$, the expected number of fission events before a tracked fission would be equal to $\mathbb{E}[\text{fissions per deme}]$. This would, on average, lead to a lot of simulations with the mean number of fissions considerably above the target value. Ignoring untracked fissions, this is equivalent to having a fission rate of $p \times \phi$. Untracked fissions are important, however, not just for the purpose of tracking the mean number of fission events, but also for population dynamics of the demes. Depending on the deme carrying capacity, that is the maximum number of cells in a deme, and mutation and epimutation rates, a fission can impact the population structure of a deme in different ways. Let K be the carrying capacity of a deme, μ the driver mutation rate, γ the epimutation rate, and L the number of fCpG sites per cell (assuming equal methylation and demethylation rates for simplicity). Upon fission, the population of the deme is divided into two, and there need to be $K/2$ cell division events before the deme is back to its carrying capacity. The expected number of mutations is then simply $K/2 \times \mu$, and the expected number of epimutations is $K/2 \times \gamma \times L$. Depending on the rates, these numbers can be quite different. Realistically, the number of driver mutations during repopulation of a deme is likely negligible. While epimutations are

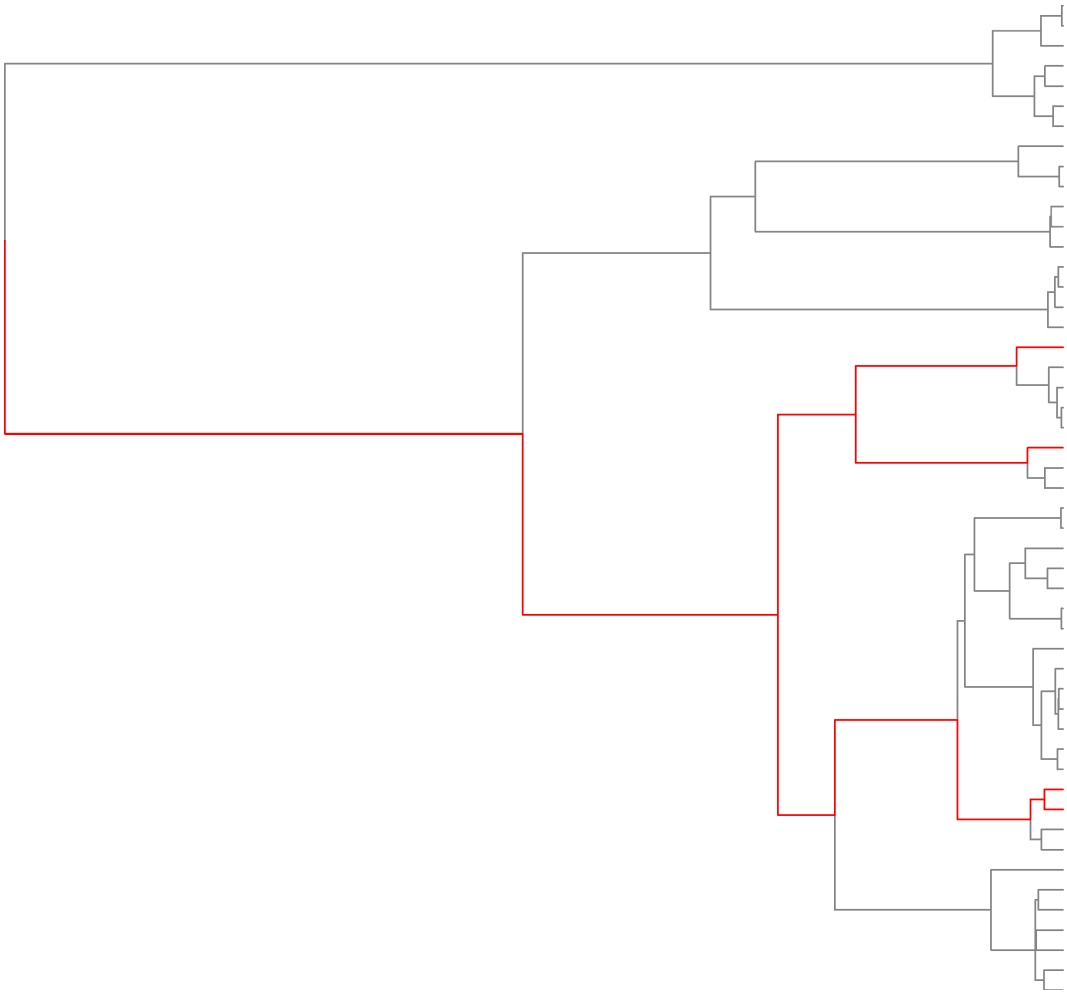

Figure 4.1: A toy example of how fissions are handled in the model. The red branches represent tracked fissions, and the grey branches are hypothetical untracked fissions occurring under a regular branching process.

more probable, fCpG arrays are inherited with a high degree of fidelity, and the fCpG distribution in a deme post-fission will resemble the state just before. When a deme is at carrying capacity, its dynamics are equivalent to a Moran process, meaning the probability of neutral fixation for a mutant population of m in a deme of carrying capacity K is equal to m/K, ignoring fissions. However, as cells grow into empty space post-fission until carrying capcity is reached, fissions increase the probability of neutral fixation at smaller deme sizes as they will offset some of the effects of genetic drift which is more pronounced at smaller population sizes.

4.2.2 Stopping conditions

When simulating the whole tumour, imposing a stopping condition based on the total cell population is one of the most straightforward ways to end the simulation.

However, as this model is focussed on a subset of relevant cells, I decided on a few different options for stopping conditions. If the model is run in the non-spatial setting, the simulation can be terminated after a maximum number of cells is reached. In the spatial setting, more concretely in the case of deme fission, the main indicator of a tumour's growth is the mean number of fissions per deme. This condition is based on the assumption that a tumour's progression by fission is equivalent to a birth-only branching process, similar to the model of (Sottoriva et al. 2015). The model also has the option of simulating steady-state turnover of cells in the simulated demes after the initial growth phase. This is an approximation of a saturation growth regime, as real tumours are not expected to grow indefinitely.

4.2.3 Sensitivity analysis

To check whether the model's behaviour is consistent with the expectations, I wrote a `snakemake` workflow to test many combinations of the parameters from table 4.1. The code is available at the github repository for the workflow, `walter` (Manojlović 2024). A summary of the results is shown in figures 4.2, and 4.3, 4.4, and 4.5. Further details are included in the appendix.

Carrying capacity

By carrying capacity, I mean the maximum number of cells with proliferative potential in a gland. This is effectively equivalent to the maximum number of lineages allowed in the gland. A more likely situation is that some of the simulated cells are closely related. The could be considered as cancer stem, i.e. cells with infinite proliferative potential which maintain the population of a cancer gland. I considered three different carrying capacities: 10, 100, and 1000. A real tumour gland contains about 10^5 cells, but the current understanding of how colorectal cancer evolves suggests that not all of them are able to divide indefinitely, rather following a similar hierarchical structure to real colonic crypts (Cernat et al. 2014). The results follow intuition: higher carrying capacity leads to less likelihood of neutral fixation but also more diversity in the presence of selection.

Fission rate

The fission rate in this model is tracked per cell, with a gland's fission rate being the sum of the fission rates of all cells in the gland. In testing, I considered per cell fission

rates of 10^{-5}, 10^{-4}, and 10^{-3}. As a consequence, the per deme fission rates ranged from 10^{-4} to 1. This led to a complication in the form of unfinished simulations, as with too low of a fission rate, the first deme never splits and the simulation never ends. This is important to note when choosing priors in the ABC workflow. As we only consider tumours which have grown in a reasonable time, the prior distribution of fission rates will be chosen with the carrying capacity taken into account. The results, as expected, show that tumours with a higher fission rate grow faster than those with a lower fission rate. Further, fCpG array diversity depends on the fission rate, epigenetic switching rates and the carrying capacity, as a high fission rate for smaller demes can still produce a diverse array with a high epimutation rate, where slower fissions still lead to a diverse array with a relatively low epimutation rate.

Driver mutation rate and selective advantage

The driver mutation rate is the probability of a cell acquiring a driver mutation upon division. The selective advantage is the proliferative advantage of cells carrying a driver mutation. In testing, I considered driver mutation rates of 10^{-5}, 10^{-4}, and 10^{-3}, and selective advantages of 0, 0.1, and 1. For larger deme sizes, the presence of selection seems to lead to more desynchronised arrays than the neutral case, as neutral fixation is less likely. However, at very strong selection, array diversity again decreases as the mutant quickly fixes in the population. Higher mutation rates lead to more diverse arrays, but an overly high mutation rate leads to the emergence of many mutants, effectively voiding the effect of selection in some cases due to clonal interference.

Epimutation rates

The epimutation rates are the probabilities of a fCpG site changing state upon cell division. These events may not necessarily be constrained to division events in reality. In fact, models developed in (Gabbutt et al. 2022, 2023) consider epimutation events independent from divisions. For the purpose of this model, I have chosen to keep the epimutation rates tied to cell division as it simplifies the assignment of parameters and maintains the temporal units of cell division^{-1}.

In testing, I considered epimutation rates of 10^{-5}, 10^{-4}, and 10^{-3}. The results are consistent with the findings of (Gabbutt et al. 2022), where too slow switching means less diversity, too fast means complete desynchronisation and regression to a

normal distribution around 0.5.

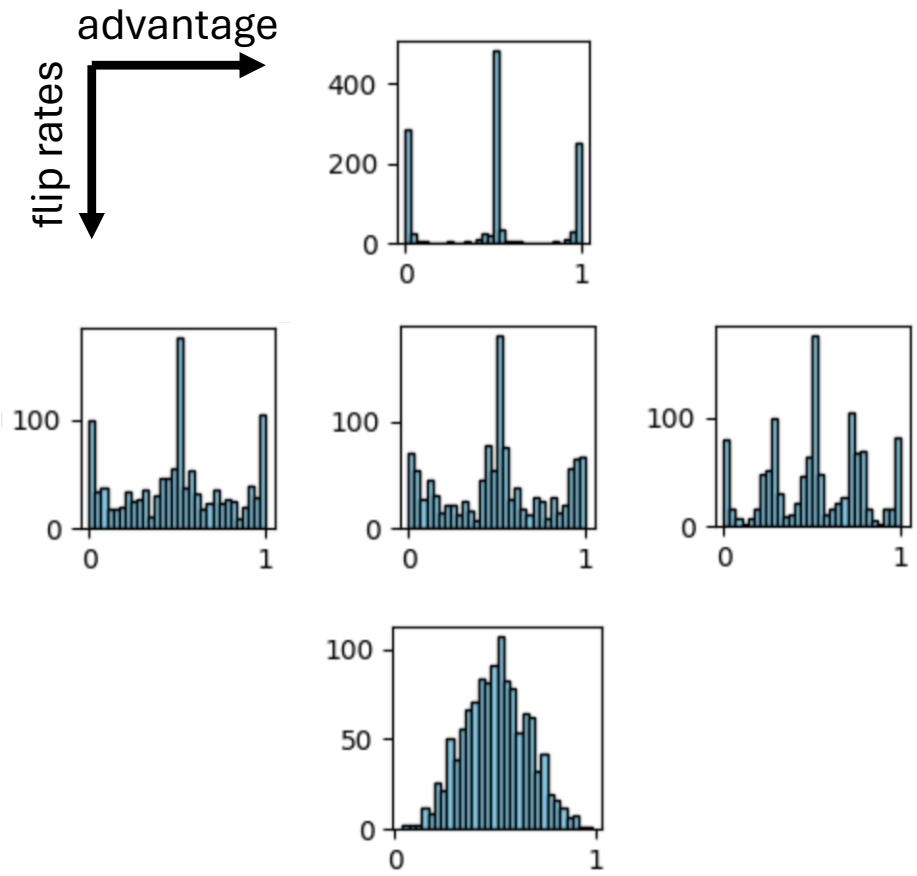

Figure 4.2: Epigenetic mutation rates and strength of selection impact the fCpG distribution within a gland. **x-axis**: selective advantage of driver mutations from neutral to weak ($s = 0.1$) to strong ($s = 0.5$). Strong selection leads to clonal interference and fewer dominant lineages, reflected in the peaks between 0 and 0.5, and 0.5 and 1. Neutral and weak selection have similar signatures in the simulations, with small intermediate peaks emerging occasionally due to the stochastic nature of the model and the probability of neutral fixation. **y-axis**: epimutation rates from slowest (10^{-4}) to medium (10^{-3}) to fastest (10^{-2}). Slower switching shows very little deviation from the progenitor cell's fCpG array, while too fast switching makes the fCpG distribution tend to a Gaussian around 0.5.

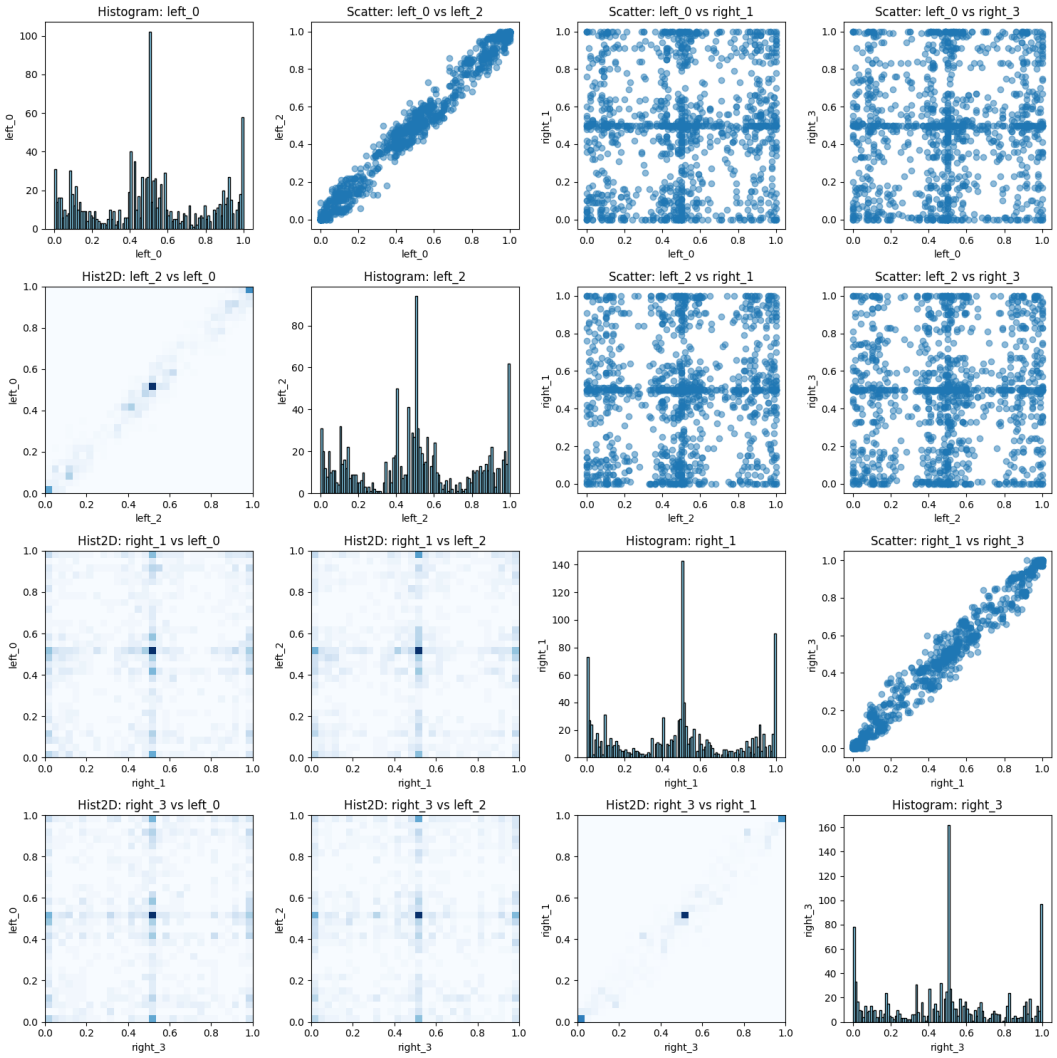

Figure 4.3: Slower fission rates lead to more different fCpG arrays across the sides of the simulated tumour. **diagonal:** Histograms of each gland's fCpG array at the end of the simulation. **above diagonal:** Pairwise scatter plots of the glands' fCpG arrays. **below diagonal:** Pairwise 2D histograms of the scatter plots showing the density of points.

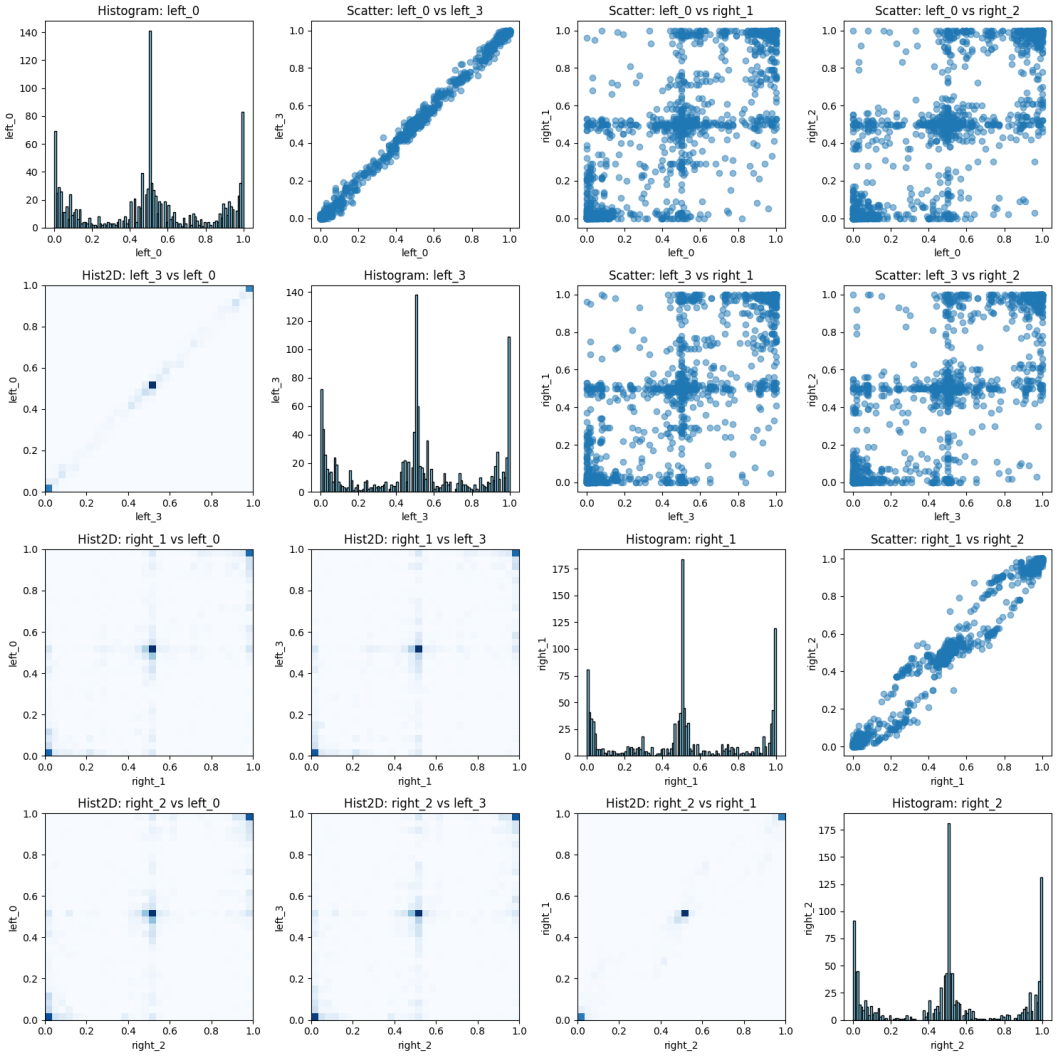

Figure 4.4: Increasing the fission rate leads to more closely related fCpG arrays.

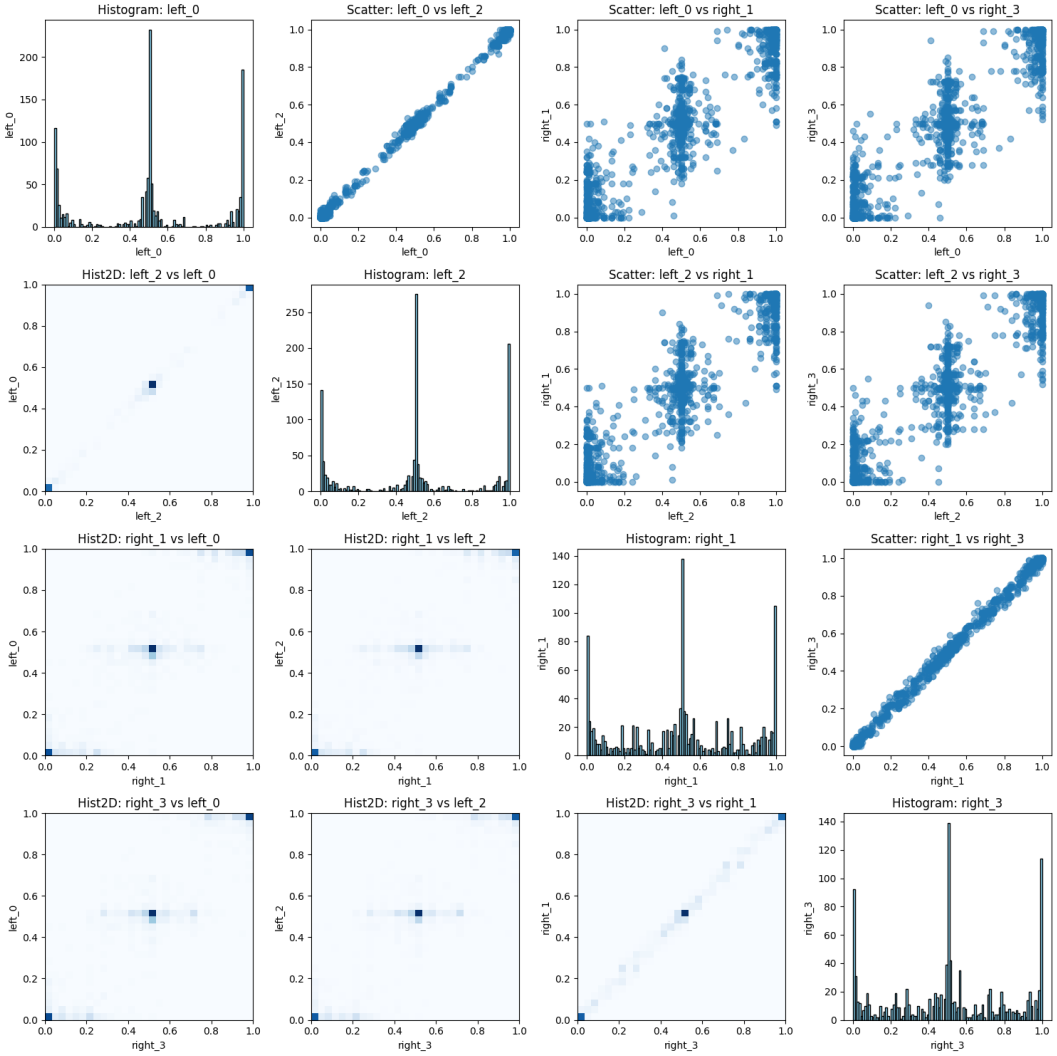

Figure 4.5: Too high a fission rate leads to much less time spent in independent turnover, and thus the most closely related fCpG arrays.

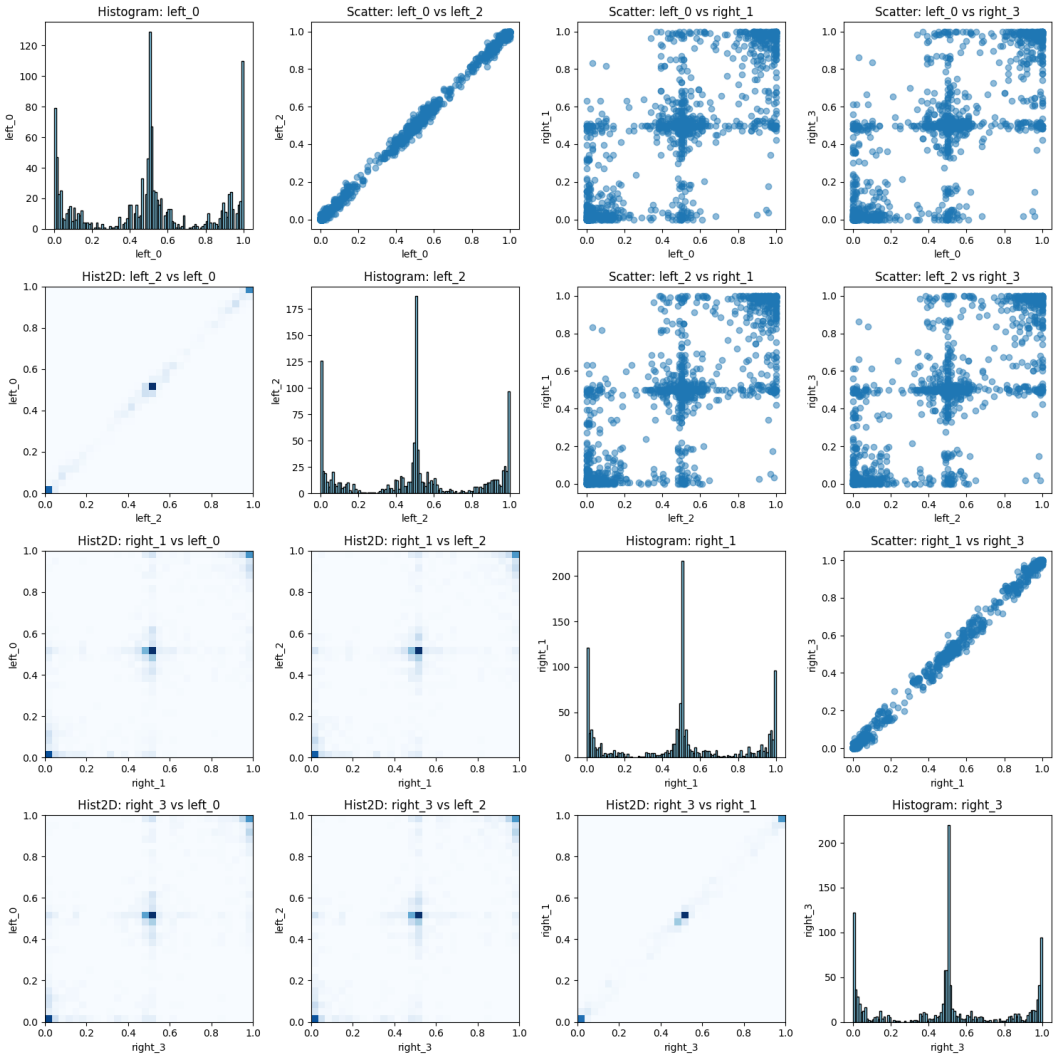

Figure 4.6: High driver mutation rate and strong selection can "compensate" for fast fission rates, leading to slightly more diverse fCpG arrays. While not necessarily a realistic scenario in a real tumour, this example shows how different parts of parameter space can lead to similar results.

4.2.4 Efficiency and memory requirements

Apart from standard C++ libraries, the model makes use of the `boost` library for random number generation and reading in parameters from a config file.

Time complexity

The model's time complexity is dominated by the fCpG array updates, both in individual calls and demes. Each fCpG site has an independent probability of flipping upon division. This means that, if a cell has L fCpG loci, each time a cell division occurs $2L$ random numbers have to be generated. Consider a tumour consisting of N demes, each with a carrying capacity of K, and with a target of F mean fissions per deme. If the fission rate per cell is ϕ, the expected number of cell births before fission is $1/\phi$. The time complexity of the model is then

$$O(F \times N \times K \times L \times \phi^{-1}). \tag{4.3}$$

Memory usage

The bulk of the program's memory usage comes from the tracking of each cell's and deme's fCpG array, with other memory usage being negligible in comparison. A cell's fCpG array is a $1 \times 2L$ vector of integers, where L is the number of fCpG loci per cell. A deme's fCpG array is a $1 \times L$ vector of floats, calculated as the average of the fCpG arrays of all cells in the deme. For a tumour of N demes, each with a carrying capacity of K, the total memory used by the program is approximately $2NKL \times 4\text{bytes} + NL \times 4\text{bytes}$, for a total memory complexity of

$$\theta(NKL). \tag{4.4}$$

Output files

The model writes to at least one csv file, and at most two. The compulsory file contains essential information about the simulated demes and is written at the end of the simulation or every 10 generations. This file consists of the columns:

- `Generation` - time of writing the row
- `Deme` - unique identifier of the deme

- `Parent` - unique identifier of the parent deme
- `Population` - number of cells in the deme at time of writing
- `OriginTime` - time of the deme's birth
- `AverageArray` - the average fCpG array of the deme at time of writing

The other file contains information about all cells in the simulation and is written every 10 generations. This file can be quite large, depending on deme size, and is not necessary for the inference workflow described in the next section.

4.3 ABC workflow for inferring `methdemon` parameters

4.3.1 Overview

For black-box simulations, likelihood-free inference is the most popular method of parameter estimation. Of these, ABC is preferred by most judging by its representation in the literature (Tavare et al. 1997, Sottoriva & Tavaré 2010, Sottoriva et al. 2015, Wang et al. 2024, Bondi et al. 2023). In its most basic form, ABC is a rejection algorithm which draws parameter values from a prior distribution, simulates data using a given model, and compares the simulated data to the observed data. If the distance between the two is less than a given threshold, the parameter values are accepted. This is repeated until a sufficient number of accepted parameter values is obtained. The main issue with this approach is that a completely random search of the parameter space is not efficient, and the number of simulations required to obtain a sufficient number of accepted parameter values blows up as the dimension of the parameter space increases. To address this, the `pyabc` package was developed (Klinger et al. 2018, Schälte et al. 2022). The package uses a sequential Monte Carlo algorithm to sample the parameter space, with the option of using dynamic sampling, thresholds and particle population sizes for improved efficiency. I decided to use this package for the inference workflow because of its robust implementation and intuitive communication with high-performance infrastructure, such as the City, University of London's cluster, Hyperion.

While ABC is a powerful tool for complex models, it is limited by the way data is compared to simulation outputs. This includes using a summary statistic to reduce the dimensionality of the data and summarise the most important features of the

output. This also minimises the computational cost of the comparison step of the workflow. Further, the choice of summary statistic can introduce a bias, as does the non-zero tolerance threshold. The bias can be reduced with a smaller threshold, but this increases the overall complexity of the workflow as more simulations are required to obtain a sufficient number of accepted parameter values. Additionally, it is difficult to say whether any bias observed during the inference is due to the inference method or the model itself.

4.3.2 Distance functions

In the case of the `methdemon` model, the relevant output data is a set of average fCpG arrays, one for each deme, at the end of the simulation. As the fCpG sites fluctuate independently and stochastically, considering the absolute value of individual fCpG loci is not meaningful. Instead, my focus is on the way arrays differ from each other at the end of growth. There are two parts to this approach.

Inter-gland distance matrix

To reduce the dimensionality of a single tumour's output, I have defined the inter-gland distance matrix as a pairwise distance matrix of the average fCpG arrays of the demes.

Definition 4.3.1. Let a simulated tumour consist of N demes, with the array of deme i given by $\mathbf{a}_i$. The inter-gland distance matrix is then defined as

$$D_{ij} = \frac{1}{L}\sum_{k=1}^{L}(\mathbf{a}_i^k - \mathbf{a}_j^k)^2, \tag{4.5}$$

where L is the number of fCpG sites in the array.

The idea behind this distance matrix is to emphasise larger differences between sites, and reduce the impact of small differences. This is done to mitigate the impact of the noise when comparing the arrays. To compare two distance matrices, I use the Frobenius norm of their difference.

Definition 4.3.2. The distance between two inter-gland distance matrices is defined as

$$\delta(D_1, D_2) = \|D_1 - D_2\|_F/\sqrt{2} = \left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^2\right)^{\frac{1}{2}}, \tag{4.6}$$

where a_{ij} is the element of the difference matrix $D_1 - D_2$ at the i-th row and j-th column. The factor of $\sqrt{2}$ is included to account for the fact that the matrix is symmetric, and therefore each pairwise distance is counted twice.

Distance of fCpG distributions

While the inter-gland distance matrix is an overall measure of the differeces between two tumours, having independent methylation and demethylation rates means that the fCpG distribution in a deme can be skewed in different ways. To make sure that the model is capable of capturing the right relationship between methylation and demethylation rates, I also compare individual demes' fCpG distributions using the Wasserstein distance (or the Kantorovich-Rubinstein metric) (Kantorovich 1960):

Definition 4.3.3. The Kantorovich-Rubinstein or Wasserstein distance between two probability distributions P and Q is defined as

$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \int_{\mathbb{R}^2} d(x,y)\gamma(x,y)dxdy, \tag{4.7}$$

where $\Pi(P,Q)$ is the set of all joint distributions with marginals P and Q, and $d(x,y)$ is the distance between x and y.

Intuitively, the Wasserstein distance is the minimum cost of transforming one distribution into another.

4.3.3 Example inference

To demonstrate the utility of the `methdemon` model and the ABC workflow, I have performed an example inference using synthetic data generated from the model. The simulated tumour consists of 8 demes, each with a carrying capacity of 100. The ground truth parameter values and the prior and posterior distribution of the parameters are shown in table 4.2. The inference was performed over 3 generations

Parameter	Ground truth	Prior	Posterior (mean±std dev)
Demethylation rate	0.0018	$U[0, 0.5]$	0.0091 ± 0.0046
Methylation rate	0.0022	$U[0, 0.5]$	0.010 ± 0.005
Fission rate per cell	0.009	$U[0.001, 0.1]$	0.056 ± 0.02
Driver mutation rate	0.0001	$U[0, 0.01]$	0.0044 ± 0.0027
Selective advantage	0.1	$U[0, 0.5]$	0.31 ± 0.12

Table 4.2: Broad priors lead to acceptance of multiple parts of parameter space, resulting in broad posterior distributions.

using 200 particles per generation, with a dynamic tolerance threshold calculated using the `SilkOptimalEpsilon` method in the `pyabc` package. The prior distribution of the parameters was chosen to be uniform and broad. The results of the inference are shown in figure 4.7.

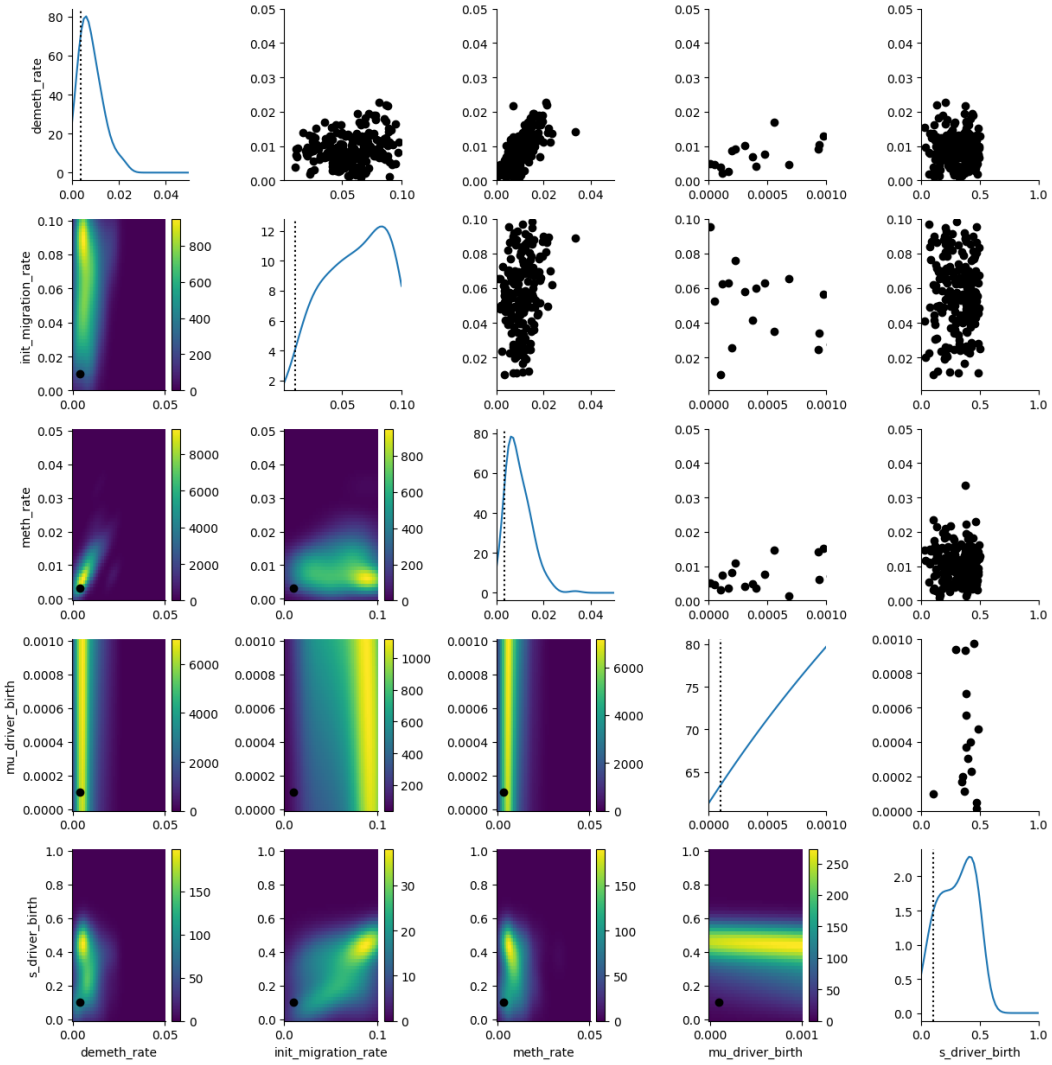

Figure 4.7: Results of the example inference of the `methdemon` model. The ground truth parameter values are shown as dotted vertical lines in the plots on the diagonal. Posterior distributions of the epigenetic mutation rates narrow down close to the ground truth values, but other parameters' posteriors remain broad.

The results show a few interesting things. Firstly, the posterior distribution of the epimutation rates has narrowed down to a small range, close to the ground truth values. This is likely due to the way in which epimutation rates are expressed in the fCpG array - too fast and the distribution becomes Gaussian, too slow and few changes occur. The Wasserstein distance favours simulations with a similar bias to the observed data, meaning that the ratio of the two is likely to stay preserved. Additionally, the magnitude of the epimutation rates is encoded in the inter-gland distance matrices in a similar way to the individual deme fCpG distributions, with the difference between glands being more pronounced close to the sweet-spot of the rates.

The posterior distributions of the other three parameters are less informative, as they have remained broad. This could be the case for a few reasons. Firstly, the data itself could be uninformative, meaning that the model is not capable of capturing the relationship between the input parameters and the output data. Despite prior testing showing that the outputs are sensitive to the input parameters, this is a complex model and it is possible that the relationship between the parameters and the output data is not straightforward. For example, a slowly growing tumour (lower fission rate) with no to weak selection and a moderate mutation rate could produce similar outputs to a fast-growing tumour with strong selection and a high mutation rate. This is not a problem with the model, but a feature of the tumour's growth dynamics. The priors in this test were chosen to be broad, covering both realistic and unrealistic behaviour of cancer. According to mathematical models, weak selection or effectively neutral dynamics fit the growth of many tumours (Williams et al. 2016). Secondly, the distance between observed and simulated data could be uninformative. The choice of summary statistics was a natural one, but it is possible that weaker signatures in the data are not captured by the distance metrics. This is a common problem in ABC, and is usually addressed by refining the rejection step, in addition to a more informative prior. Finally, having multiple possible parameter combinations which produce similar outputs could lead to a broad posterior distribution. This is an issue commonly debated in the literature, as selection is difficult to quantify in real tumour data. None the less, the results of the example inference are useful in that they show the ability of ABC to narrow down at least parts of the parameter space even in the case of broad priors in a complex model.

4.4 Discussion

In this chapter, I presented a novel agent-based model, `methdemon`, which is capable of simulating the growth of a tumour and the corresponding fluctuating methylation arrays. The model is developed with efficiency and dynamic memory management in mind, and is based on well-established models of tumour growth and evolution. The model is capable of simulating the growth of a tumour in a reasonable time, and the outputs are sensitive to the input parameters to varying degrees. However, the simplifying assumptions made should be put to the test in future work, such as epimutations being possible only at cell division. An example of this could be using tau leaping instead of the Gillespie algorithm for event scheduling in the simulations, as its units of time are not necessarily tied to the cell cycle.

I also developed and tested an ABC workflow for inferring the parameters of the model from observed data. The inference workflow is capable of narrowing down the parameter space, and the results of a toy model inference show important features which need to be addressed when expanding the discussion to real data. As is often the case with ABC, the inference workflow can be hindered by the choice of overly broad priors. Therefore, an informative prior distribution should be chosen when considering real data.

Chapter 5

Modelling colorectal cancer methylation data with `methdemon`

5.1 Introduction

The main motivation for the development of `methdemon` is to infer evolutionary properties of colorectal cancer (CRC) from methylation data. CRC is the third most common cancer worldwide, with over 40000 new cases diagnosed in the UK each year on average (*Bowel cancer statistics* 2015). The scale of this issue is not the only motivating factor, as insights into the evolutionary dynamics of colorectal cancer could lead to the better understanding of how other adenocarcinomas progress. CRC is characterised by the accumulation of genetic and epigenetic mutations in colonic cells (Fleming et al. 2012). The most common type of CRC is adenocarcinoma, which arises from the epithelial cells lining the colon, covering the majority of cases. The tumour forms hierarchical cell structures similar to those of normal tissue, organising into crypt-like glands (Ponz de Leon & Di Gregorio 2001). The tumour spreads by the process of gland fission (Preston et al. 2003), which is similar to the branching processes seen in normal crypts (Almet et al. 2018).

In this chapter, I will use the inference workflow introduced in chapter 4 to infer evolutionary properties of colorectal cancer from methylation data. My approach, similar to that of (Gabbutt et al. 2022, 2023), is not the first investigation of whether methylation arrays can uncover evolutionary properties of colorectal cancer, as there have been efforts to infer strength of selection (Siegmund et al. 2011) and the evolutionary history (Hong et al. 2010) of the primary tumour. The model presented

in chapter 4 is informed by the assumptions used in these works, as well as other recent models of colorectal cancer evolution and progression.

5.2 Data collection

The data used in this study were provided by Dr Darryl Shibata from the Keck School of Medicine at the University of Southern California. The data consist of DNA methylation arrays sequenced from multiple glands within colorectal tumours post-surgery. All samples are anonymised. The arrays were obtained from bulk samples of tumour glands, which means that the data are nominally not single-cell resolved. Each tumour sample consists of 8 glands, with each gland's array containing some 850000 CpG sites. The arrays were obtained using the Illumina Infinium MethylationEPIC BeadChip array. The data were pre-processed by Dr Shibata to remove low-quality samples and normalise the arrays. The sample purity was high, with the vast majority of cells in the samples being tumour cells.

5.3 Results

5.3.1 Identification of fCpG loci in colorectal cancer

The first step in the analysis was to identify the fCpG loci in the data. As multiple samples come from the same tumour, a larger cohort of samples is needed to reliably identify fCpG loci using the methods described by (Gabbutt et al. 2023), i.e. isolating a set of CpG loci which are the least informative about the methylation state across the cohort. Dr Gabbutt ran the analysis on colorectal cancer data from the Cancer Genome Atlas (TCGA) and identified 1258 fCpG loci. For comparison, I ran a similar analysis only on the data provided by Dr Shibata and identified a set of some 950 loci. Of these, only 120 were common to both sets. The discrepancy is likely due to the small sample size of the data provided by Dr Shibata. The fCpG loci identified in one of the samples are shown in figure 5.1, with additional figures in appendix B.

5.3.2 Spatial proximity predicts similarity between fCpG arrays

With the assumed hierarchical structure of the tumour in mind, it should make sense intuitively that glands which are spatially close to each other likely diverged

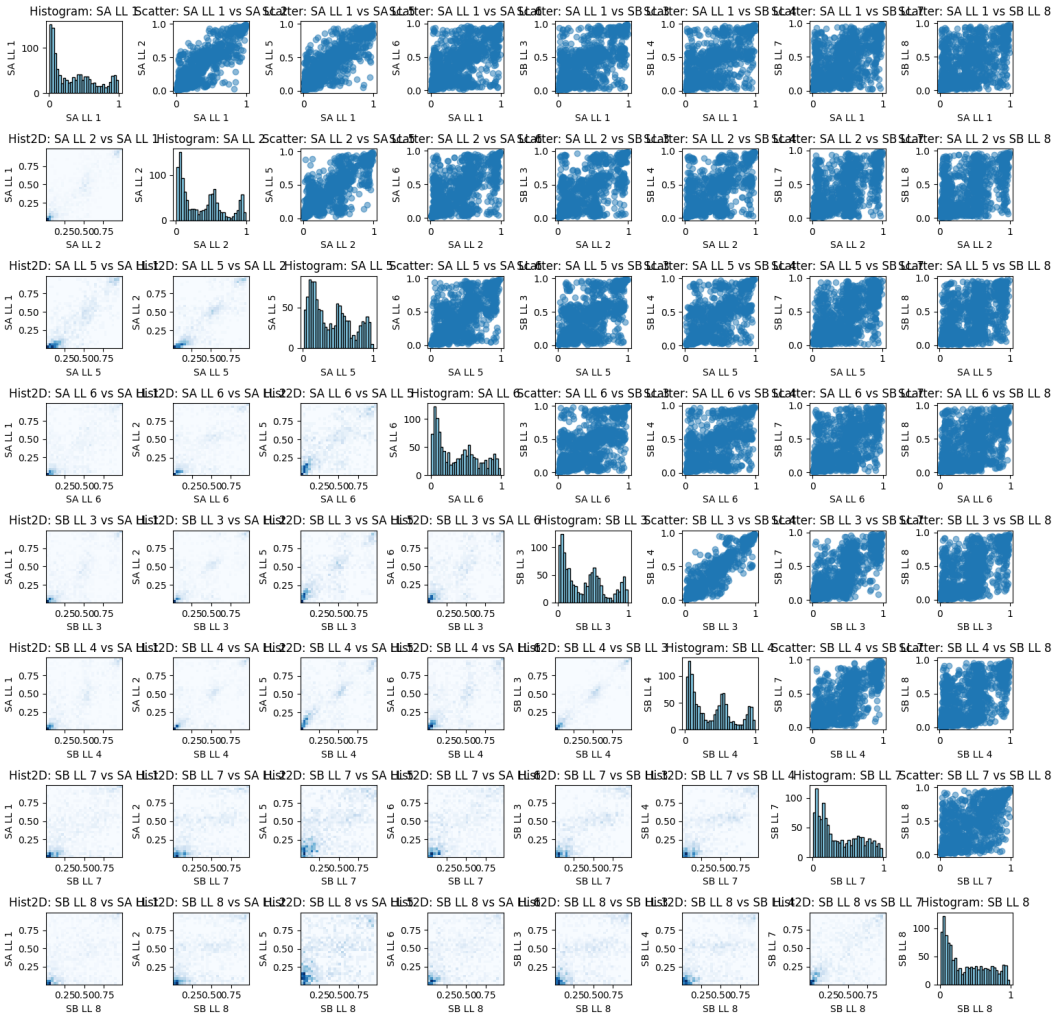

Figure 5.1: Visualisation of the set of fCpGs for tumour samples from patient S. **diagonal** — histograms of fCpG arrays for each gland; **above diagonal** — scatter plots of correlations between glands; **below diagonal** — 2D histograms of the above-diagonal plots, showing the density of points.

more recently than glands which are further apart. As a result, they have spent less time evolving independently and should have more similar fCpG arrays. To test this hypothesis, I calculated the inter-gland distance matrix for each tumour sample. The resulting matrices show a clear correlation between side and distance values. The distance matrix for one of the samples is shown in figure 5.2 for tumour S, and in appendix B for the other samples.

5.3.3 Development of the `methdemon` model

The `methdemon` model was developed for the purpose of simulating the data provided by Dr Shibata. The model's assumptions are based on the general understanding of colorectal cancer evolution and, translated into the language of an agent-based

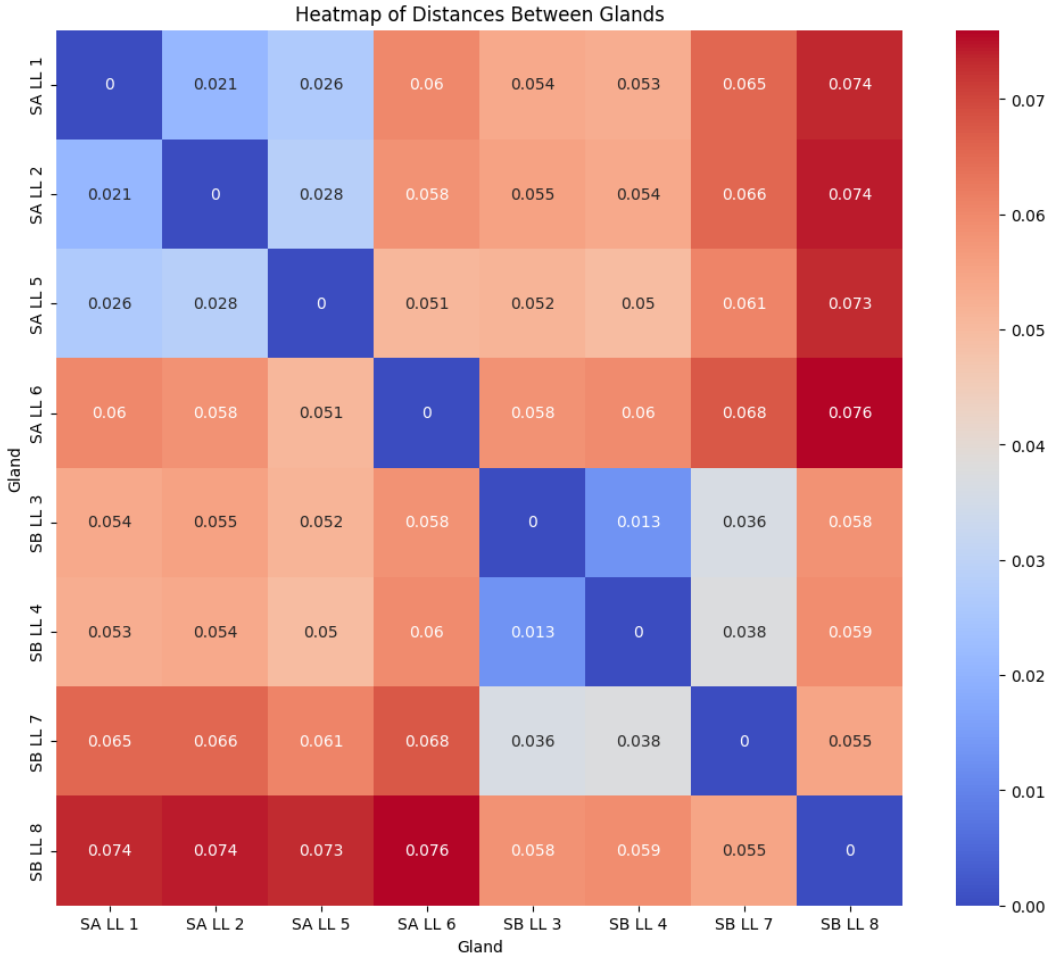

Figure 5.2: The inter-gland distance matrix for tumour S. The distance values are, broadly speaking, higher between distant glands than ones from the same side of the tumour (A or B).

model, are as follows:

(i) **A single cell forms the first gland and initiates tumour growth.** This assumption skips over the process of tumorigenesis, during which a cell accumulates mutations and becomes malignant (Tariq & Ghias 2016). This is a simplification to be sure, but a reasonable one, given that the focus of this work is on the evolutionary dynamics of the tumour rather than its initiation.

(ii) **The rate of driver mutations is Poisson distributed and identical for all cells.** This assumption is consistent with most models of tumour evolution (Metzcar et al. 2019, Niida et al. 2021).

(iii) **The cell population within a gland grows exponentially and is well-mixed.** While not necessarily consistent with the biology of a solid tumour, this assumption allows for more efficiency in the simulation as opposed to

a multi-level spatial model. Further, as the data discussed in chapter 5 is obtained from bulk samples of tumour glands, this assumption is not unreasonable.

(iv) **Once a gland reaches a certain size, which we call the carrying capacity, the population undergoes steady-state turnover according to the Moran process.**

(v) **At carrying capacity, a gland has a certain probability of undergoing fission, which splits the gland's population randomly into two.** As a consequence of assumption (iii), fissions do not take into account a gland's spatial organisation.

(vi) **Gland fission occurs as a neutral spatial branching process.** The previous two assumptions and this one together form the basis of the model's spatial dynamics. While there are other mechanisms of colorectal adenocarcinoma progression, gland fission is the principal way in which the tumour grows (Preston et al. 2003). The assumption of neutrality in the spatial branching process is consistent with the findings of (Sottoriva et al. 2015). Additionally, this assumption is based on the fact that the data used in this study only contains information about whether a gland was sample from side A or B, without any further spatial information other than the approximate size of the full tumour.

5.3.4 Higher deme carrying capacity requires stronger selection to recapitulate the data

To begin the analysis of cancer data using the `methdemon` model, I tested the ranges of parameters based on the assumption that each cancer cell has infinite proliferative potential. This would mean setting the carrying capacity of a gland to about 10000 cells, which is consistent with the size of the glands in the literature (Sottoriva et al. 2015) and our data. Due to the glandular structure of the tumour, this is an effectively neutral model, as selection acts within glands but not between them, leading to progressive diversification of the population, as discussed in chapter 3 and (Noble et al. 2022).

As a first test, I ran the model with weak selection, $s = 0.1$. The resulting outputs are shown in figures 5.3 and 5.4. There are a few notable features about the output

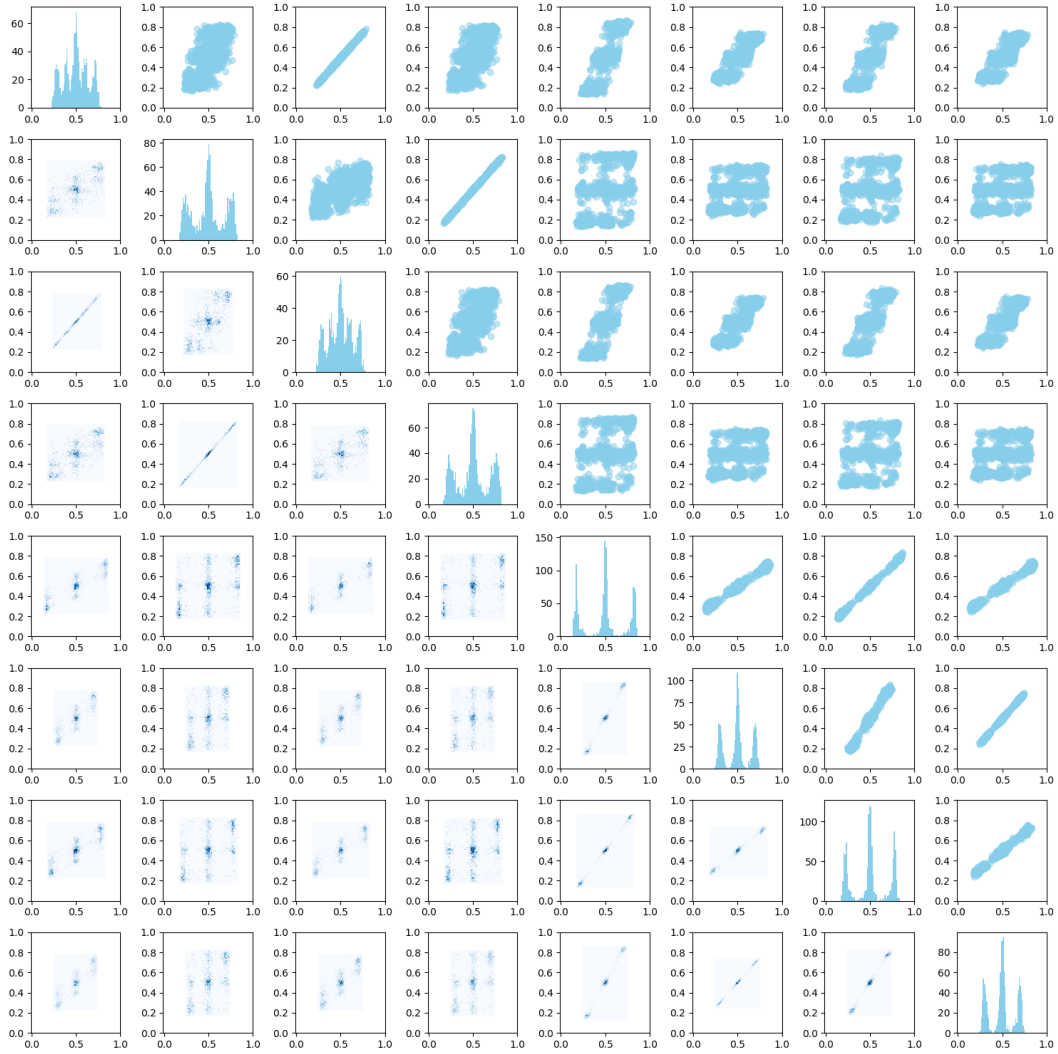

Figure 5.3: Visualisation of the output fCpG arrays from the `methdemon` model with weak selection ($s = 0.1$) and deme carrying capacity 10000. While The individual gland distributions are trimodal and the inter-gland correlation plots show epigenetic switching between sides, the distributions have narrowed down towards the mean (0.5) considerably. This happens in the case when the epimutation rate outpaces the tumour growth rate.

fCpG arrays and the distance matrix. The most obvious is that the peaks associated with homozygous methylation and demethylation states have moved towards the middle. This is to be expected, as we are treating all 10000 cells in a deme as being able to divide ad infinitum, leading to a lot of stochastic noise. This is a consequence of the time spent in turnover, and is adjusted by increasing the fission rate. However, the distance matrix shows that the glands are still very similar to each other, especially when compared to the data. This is likely due to the fact that there is only a small probability of a partial or full sweep of a lineage within a gland. The similarity is a consequence of the large deme carrying capacity, which allows for a lot of stochastic noise to accumulate over time but lowers the probability of fixation

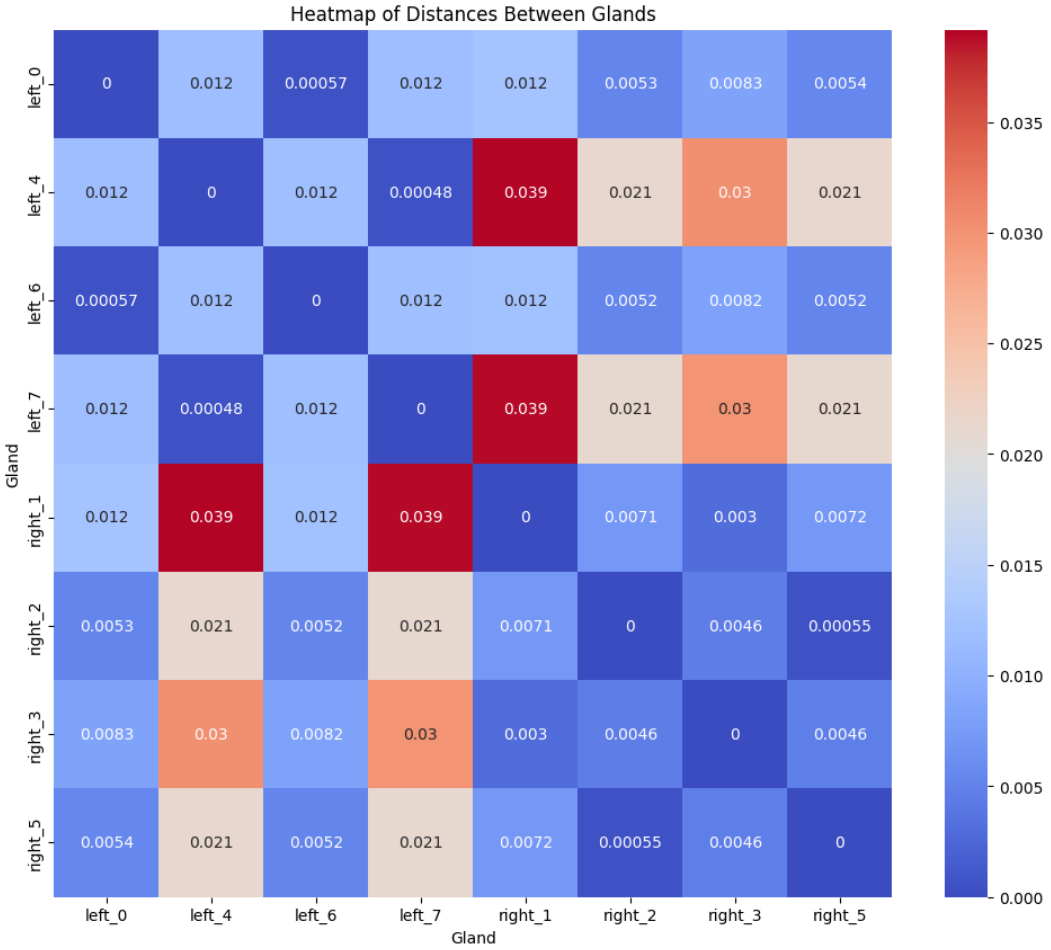

Figure 5.4: Inter-gland distance matrix for the output fCpG arrays from the `methdemon` model with weak selection ($s = 0.1$). While the distance values between glands on opposite sides of the tumour are still on average higher than within one side, the numerical values are around an order of magnitude off those observed in data.

in the weak selection regime. Increasing the selection coefficient to $s = 0.3$ leads to more divergence between the glands, since emerging lineages are more likely to fully or partially sweep the gland's population and establish more distinct fCpG arrays between glands. However, as discussed in chapter 4, strong selection can quickly become problematic in an ABM like this due to the accumulation of advantageous drivers. An example of strong selection at deme size 10000 is in appendix B.

Considering that the number of stem cells in a normal crypt is on the order of 10 (Gehart & Clevers 2019, Gabbutt et al. 2022), with the total number of cells in a crypt being on the order of 1000, I next tested the model with a deme carrying capacity of 100 i.e. about 1% of the total cell population in the gland. The currently available data supports this percentage as a reasonable estimate of the proportion of cancer stem cells (CSCs) in colorectal cancer (O'Brien et al. 2007, Munro et al.

2018). In this case, the model's outputs are as expected, with the fCpG arrays diverging over time even with no or weak selection. Examples are shown in figures 5.5 and 5.6.

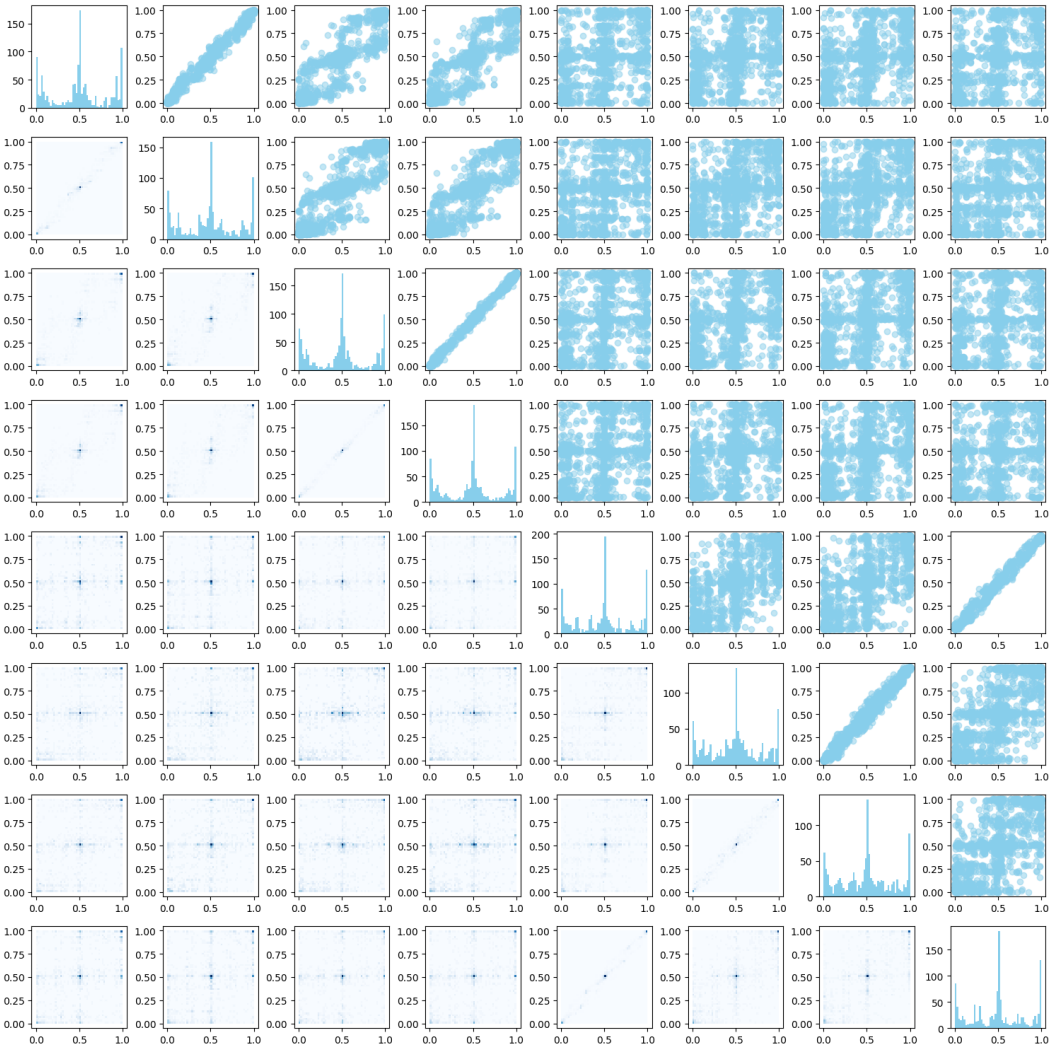

Figure 5.5: Output fCpG arrays from the `methdemon` model with weak selection ($s = 0.1$) and deme carrying capacity 100. The outputs of runs with a lower deme carrying capacity reflect the data better than larger deme carrying capacity.

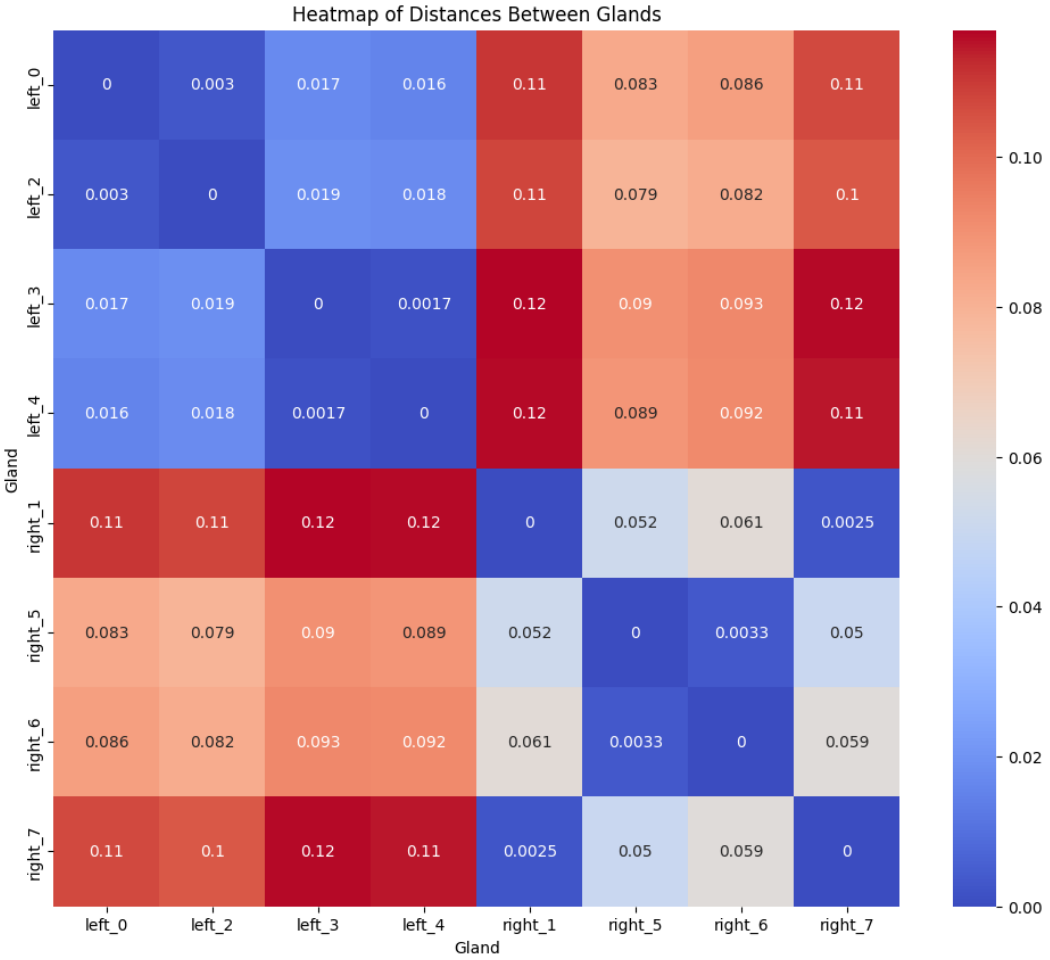

Figure 5.6: Inter-gland distance matrix corresponding to the run from figure 5.5. The values in the distance matrix are comparable to those seen in the molecular data sets.

5.3.5 Parameter inference from colorectal cancer data

Regular model

Having established `methdemon`'s ability to output data which resembles observations, I next attempted to fit the model to the colorectal cancer data. For this, I used the ABC workflow described in chapter 4. Considering the results of the previous section, I set the deme carrying capacity to 100 for all inference runs. I made this choice as there is no evidence to suggest different numbers of stem cells across tumours or even glands within a tumour. Furthermore, having the deme carrying capacity as a free parameter would slow down the inference process significantly. In the first instance, I ran the inference with parameter ranges given in table 5.1. The results of the inference are shown in figure 5.7, with more details in appendix B.

Parameter	Prior
methylation rate	$U(0, 0.1)$
demethylation rate	$U(0, 0.1)$
fission rate	$U(10^{-4}, 10^{-2})$
driver mutation rate	$U(0, 10^{-2})$
selective advantage	$U(0, 0.2)$

Table 5.1: Parameter priors for the first inference run.

The first inference run did produce some results, but the posterior distributions of some parameters remained broad. There are a few possible reasons for this. Firstly, I intentianally used overly broad priors to see how well the ABC workflow would delineate the parameter space. This choice makes the inference process more dificult, as the prior distributions are not informative and the parameter space is large. However, it allows for future runs with more informative priors. The second reason may just be that the model is not able to recover all of the parameters in its current form. Selective advantage and driver mutation rate in particular seem to be difficult to infer for the model. This could be due to the signature of selection being too weak in the model or data (or both). It could also be down to the distance functions in the ABC rejection step not being able to capture the differences between the arrays well enough. Either way, the results prompted further testing with the important change of log-transforming the parameters. This change allows for more efficient exploration of the parameter space, as parameters are sampled on the same scale and steps between generations will cover more of the space.

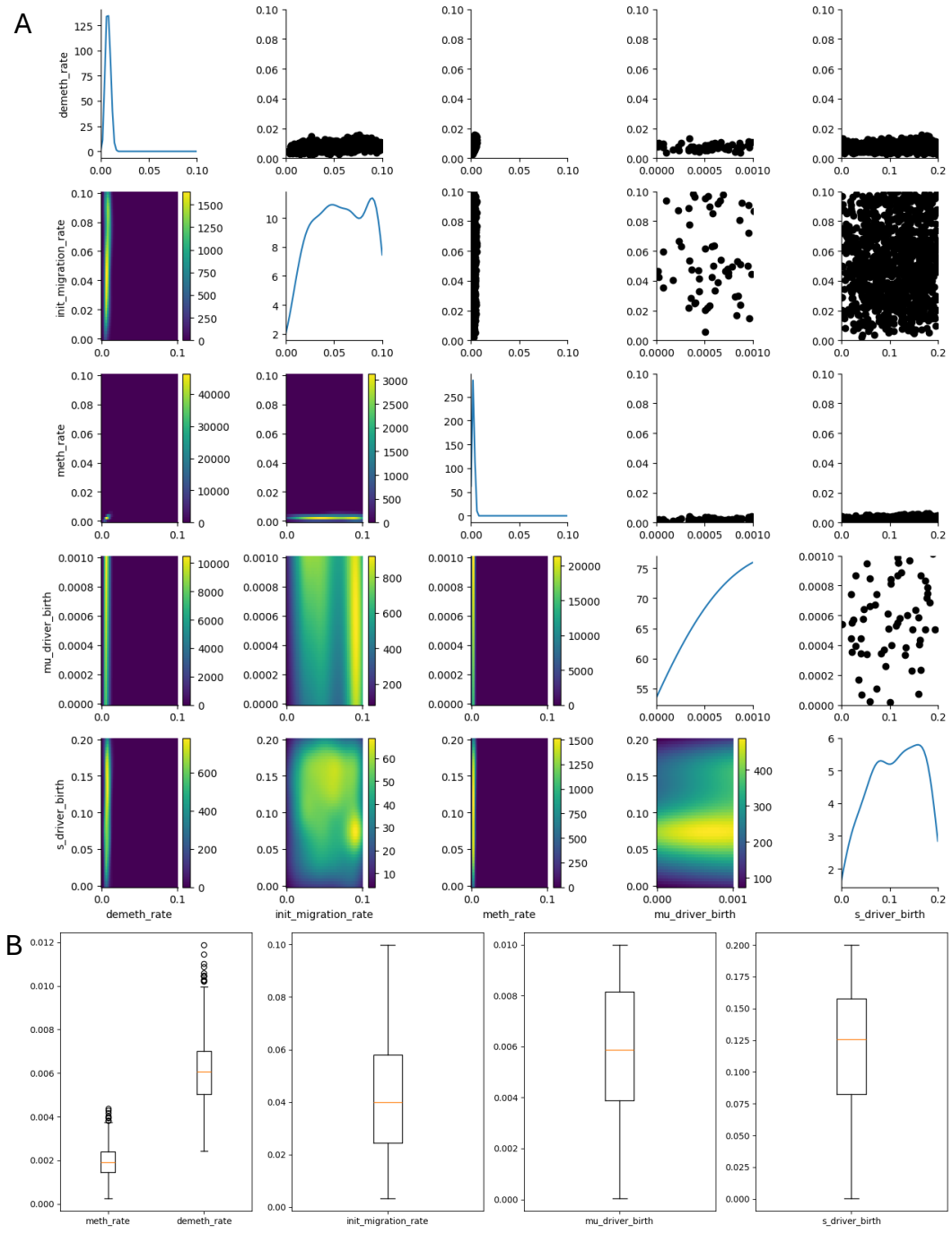

Figure 5.7: Inference outputs from the first run, performed by sampling parameters from uniform priors on the original scale. While the epimutation rates' posteriors narrow down considerably, other parameter distributions remain broad - likely due to too coarse traversal of the parameter space. **A** — posterior distributions of fCpG fluctuation rates have narrowed down rapidly, but other parameters' posteriors remain broad. **B** — box plots of the posteriors show that the model is not able to resolve the effects of selection from the data, and leaves a lot of uncertainty in the fission rates.

Log-transformed model

The second inference run was done with similar parameter ranges as the first. The log-transformed parameter ranges are give in table 5.2 and the results of the inference are shown in figure 5.8, with more details in appendix B. All log transformations were done with base 10.

Parameter	Prior
log methylation rate	$U(-4, -2)$
log demethylation rate	$U(-4, -2)$
log fission rate	$U(-3.3, -1)$
log driver mutation rate	$U(-5, -2)$
selective advantage	$U(0, 0.2)$

Table 5.2: Log-transformed priors for the second inference run.

The results of the second inference run are more promising than the first, with the fission rate posterior distribution being considerably narrower. However, selection and driver mutation rate are still difficult to narrow down. This further supports the idea that the model is not able to detect weak selection at the gland level.

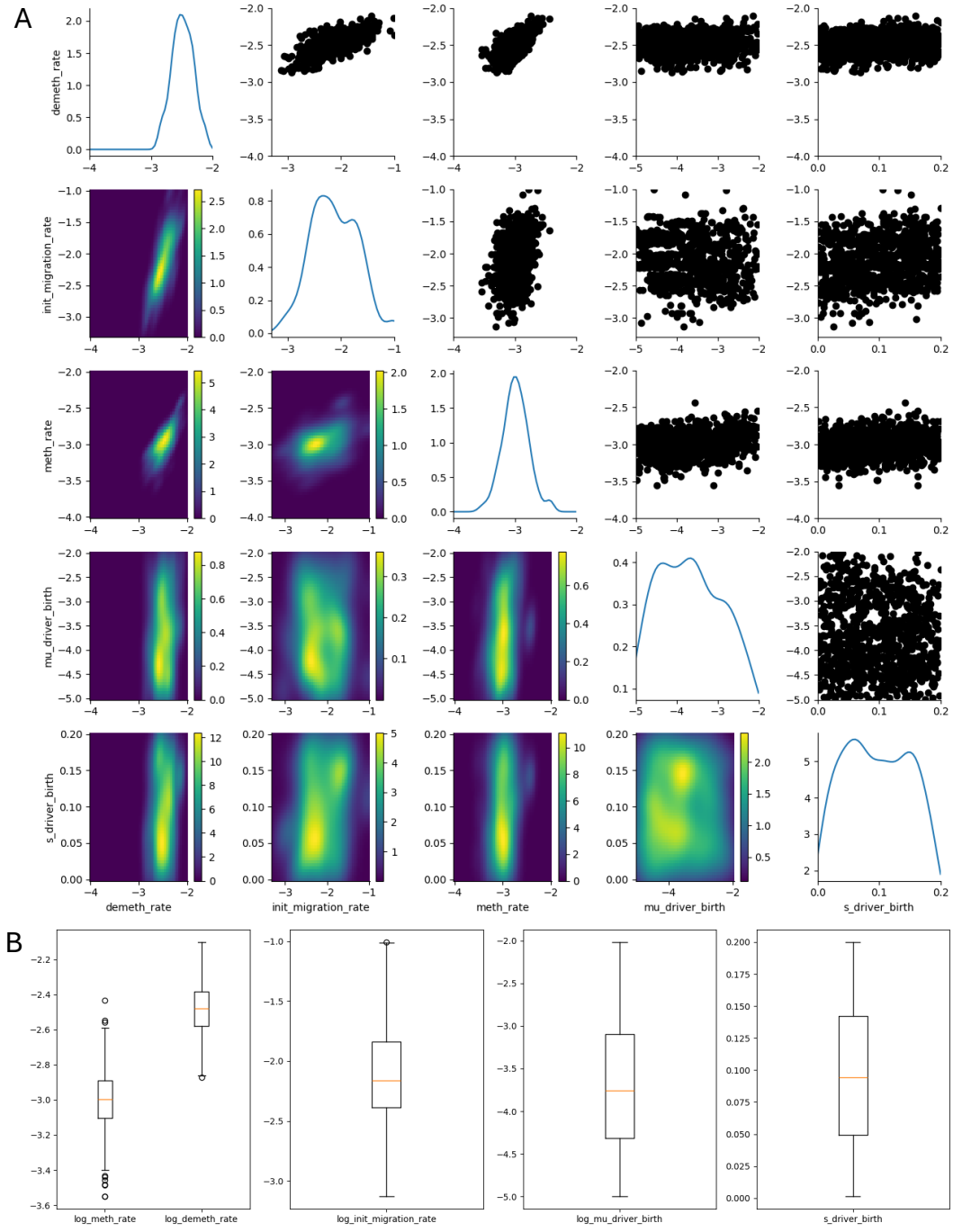

Figure 5.8: Inference outputs from the alternative run, performed by sampling parameters from uniform priors on the log-transformed scale. The posterior distribution of deme fission has now narrowed down in a similar way to the epimutation rates, indicating a more efficient traversal of parameter space. **A** — posterior distributions of fCpG fluctuation rates have narrowed similar to before, but now the fission rate's posterior is also narrower than before. Mutation rate and selective advantage are still not inferred by the model. **B** — box plots of the posteriors.

5.3.6 Fast- and slow-growing tumours

The data provided by Dr Shibata contains samples from different patients, with each tumour being potentially at a different stage of growth. From exploratory analysis, it seems that some tumours are growing faster than others, if we work under the assumption that gland fission is the only mechanism of growth. This hypothesis is supported by the inference results, as samples with higher values in the inter-gland distance matrix tend to grow slower, and ones where gland fCpG arrays are more similar have grown faster. The inferred median fission rates are shown in table 5.3. The inferred epimutation rates are included in appendix B, in table **??**.

Tumour	Inferred median fission rate [cell^{-1}cell div^{-1}]	L_2 half-norm	Tumour size [cm]	Stage
E	0.0025	0.481	6.1	I
I	0.017	0.176	3.6	III
J	0.003444	0.421	5	III
S	0.007	0.295	6	n/a
X	0.0032	0.338	2.5	n/a

Table 5.3: Inferred median fission rates for different tumours and their sizes. The L_2 half-norm of a tumour's inter-gland distance matrix appears to be inversely correlated with the inferred fission rate.

As the inferred fission rates from table 5.3 are set per cell per cell division, the fission rate per gland would be 100 times higher, or roughly in the interval $(0.1, 1)$ — one order of magnitude less, or about the same as the stem cell division rate. In my model, the tumour grows as a pure birth branching process, meaning that the growth is exponential. Let ϕ be the fission rate and $N(t)$ the number of glands in the tumour at time t. Then

$$N(t) = e^{\phi t} N(0). \tag{5.1}$$

As the tumour grows from a single gland, the time τ to reach a certain size N_τ is

$$\tau = \frac{1}{\phi} \log N_\tau. \tag{5.2}$$

The literature does not provide clear estimates for cancer stem cell division rates, but I think a reasonable estimate would be from about 1 per month to 1 per week, or about 0.034 to 0.1 per day. This would mean that the time to reach a size of 10^7 glands would lie in the interval $(230, 7000)$ days, or about 8 months to 20 years. This range is broad, and is not meant as a precise estimate, but rather a sanity

check of the model's outputs. Considering that the orders of magnitude are correct, the model seems to be able to describe the observed data well.

5.4 Discussion

In this chapter, I modelled colorectal cancer methylation data using `methdemon` and the associated ABC workflow. A notable feature of the data is that some samples show a clear bias towards either hyper- or hypomethylation after filtering with the help of TCGA data. This was also the case for fCpG arrays filtered using only the samples provided by Dr Shibata. This suggests that the progenitor cell's methylation state is not necessarily random, but could be affected by internal or external factors. Furthermore, I developed the model under the assumption that methylation and demethylation rates are constant and independent across all cells and fCpG loci. While this may be reasonable on average, there might be mutations which affect these rates, leading to the observed distributions of fCpG loci states. Despite biases in the data, the model was able to recapitulate the patterns observed in the data, and the inferred parameters lay within reasonable ranges. I think it is clear that the methylation and demethylation rates are inferred based on the Wasserstein distance between individual glands in the data and the model output, and the fission rates from the overall similarity between the inter-gland distance matrices. Additionally, the L_2 half-norm of the inter-gland distance matrix is inversely correlated with the inferred fission rate. This is consistent with the idea that faster-growing tumours have had less time for the glands' arrays to diverge.

However, the model struggled to infer the selection coefficient and the driver mutation rate. This could be due to the signature of selection being too weak in the model or data, or the distance functions in the ABC rejection step not being fine-grained enough to detect the differences between evolutionary modes. Having said that, the results do support an effectively neutral evolutionary mode of colorectal cancer growth, with the effects of selection constrained to within glands. Furthermore, it seems clear that multi-region sequencing of methylation arrays from solid tumours can be used to draw inferences about the evolutionary dynamics of the tumour, and warrants further research both in terms of obtaining more data and refining the model to address the evolutionary properties not recovered in this chapter.

An additional point to consider is that, as the `methdemon` model is based on a branching process, it should be straightforward to recover trees from the model's output. In (Gabbutt et al. 2023), the authors used a tree reconstruction workflow to infer phylogenies from blood cancer data with high degrees of accuracy. This would be a useful extension of the current work as, in conjunction with the methods discussed in chapters 2 and 3, it could provide a more complete picture of the mode of evolution of colorectal cancer.

Chapter 6

Discussion

6.1 Summary

Modelling cancer evolution is difficult. Like with any data-driven approach, straying too far into general laws may lead to underfitting of patient-specific data. Focussing too much on the individual, however, makes overfitting the main issue. As novel sequencing techniques, or at least public data sets which use them, have become more available, there is an increased emphasis on multi-pronged approaches (Heide et al. 2022, Househam et al. 2022). In my view, a similar conundrum extends one level of abstraction above this, as focussing on data too much may lead to missing the proverbial forest. Much like some branches of mathematics grew out of necessity to quantify physical phenomena, so has mathematical theory informed experiments decades after development. I believe a similar approach is essential in mathematical oncology as well, with robust mathematical frameworks translating into the laboratory or even clinical trials. A good example would be adaptive therapy, which has the potential to revolutionise patient care and is based on optimising drug scheduling (West et al. 2023).

In chapter 2, I investigated properties of the universal tree balance index J^1. Originally introduced to analyse cancer phylogenies, I show that the roots of the idea for such a metric go surprisingly deep, with pioneering works in computer science defining an effectively identical formula. The expected value of the index under even the simplest tree generation processes was too complex to calculate analytically, so I relied on the relationship derived in the original J^1 paper (Lemant et al. 2022) which connects it with the Sackin index to obtain an accurate approximation. Another

challenge came from deriving tree families which minimise the index. While I proved that there exists a family of trees on which the value of J^1 is lower than on the caterpillar tree, traditionally the tree that minimises balance indices, I could only conjecture that this family indeed minimises the index across all trees on a given number of leaves.

Balance is only one property of many trees posses. This meant that the next logical step was to consider the values of a set of indices during a tree generating process. Inspired by (Noble et al. 2022), I used a modified set of evolutionary indices on the outputs of an agent-based model to test how well one can distinguish different evolutionary trajectories in the space described by the indices. These results I compared to a new set of tree shape indices (Noble & Verity 2023), which generalise J^1 further on trees with defined branch lengths. While these indices have not produced different results from the smaller set, the fact that they still show a clear separation between different evolutionary modes is promising.

In chapter 4, I narrowed down my consideration of the modes of tumour evolution to the specific case of effectively neutral evolution, which manifests itself via progressive differentiation. The model I developed in this chapter restricts the effects of selection to patches of cells, with the patches themselves not interacting or interacting neutrally. I demonstrated that ABC could be employed to recover parameters of the model with varying degrees of accuracy.

Having established a workflow inspired by a concrete data set, in chapter 5 I recovered aspects of evolutionary dynamics of colorectal cancer using its methylation array data. Building on prior work on fluctuating methylation clocks, I modelled multi-site bulk sequences of tumour glands. By considering the differences between spatially close and distant glands, as well as individual gland fCpG arrays, I was able to approximate the epimutation and gland fission rates relative to cancer stem cell division rates. By inferring the rate of tumour growth, it may be possible to more accurately estimate the age of individual tumours in future studies.

6.2 Hybrid modelling and inference

ABM is a powerful tool for studying the evolutionary dynamics of cancer, but the stochastic nature of small-scale events can make it difficult to obtain results which are closely aligned with the data, and `methdemon` is no exception. Prior work in

fCpG modelling was done using a likelihood-based approach (Gabbutt et al. 2022, 2023), and has shown promising results. Due to the complexity of colorectal cancer, an exclusively likelihood-based approach may not be feasible, but a hybrid model which leverages likelihood-based inference locally for detecting potentially small effects of selection, with ABC on the global scale could be a good compromise. The main hurdle in this approach is reconciling fissions with the steady-state turnover process. An approach I am currently exploring is combining tau-leaping for intra-gland dynamics and Gillespie's algorithm for gland fissions.

6.3 Gland phylogenies

Another piece of the colon cancer fCpG puzzle is the reconstruction of gland phylogenetic trees. In (Gabbutt et al. 2023), the authors used a custom BEAST pipeline, a Bayesian phylogenetic inference tool (Bouckaert et al. 2019), to infer clone phylogenies from blood cancer data. The main differences between the two data sets include the fact that the blood cancer data is non-spatial, and therefore possible for meaningful sequencing at multiple points in time. In the case of colorectal cancer, it is not possible to obtain multiple samples from the same gland over time. Further, the spatial nature of the data means that sequencing it in the first place may not be possible before the tumour has been removed. This means that the clock rate of the gland phylogenies is unknown. However, sequencing multiple glands does allow for accurate reconstruction of tree topologies, with the clock rate being a nuisance parameter. Having discussed the potential for tree shape indices being used in evolutionary mode inference, the next logical step would be testing whether they point to signs of global selective pressures in the data. Because of the way `methdemon` is written, phylogenies are easily constructed as a byproduct of the simulation, making it easy to test effectively neutral growth as a null model. Resources permitting, it would be informative to use a larger-scale spatial model to test different ways of gland organisation in space, which could result in different modes of evolution, and thus tree shapes.

6.4 Conclusion

In this thesis, I have taken a multi-faceted approach to modelling cancer evolution. Chapters 2 and 3 go down the path of tree shape indices, which are emerging as a powerful tool for analysing driver phylogenies, as well as trees more generally. Condensing the information contained in a tree into an array of indices allows for a meaningful comparison of trees, which could easily be integrated in popular statistical packages for tree inference. On the other hand, chapters 4 and 5 explore evolution of colorectal cancer using agent-based modelling and approximate Bayesian computation. There is a rich literature on the topic of colorectal cancer ranging from tumourigenesis to metastasis, but this is the first time that the evolutionary dynamics of the disease have been modelled based on fluctuating methylation clocks. As discussed in chapter 5, the model is based on a certain set of assumptions, which may at times be too restrictive. Despite this, the model has been capable of recovering evolutionary properties of the disease, and seems to reliably estimate how fast the tumour has grown.

The research presented in this thesis, based on mathematical and computational modelling of evolutionary dynamics of cancer, contributes to the better understanding of the ways in which cancer can evolve. From differentiating evolutionary trajectories to leveraging the mode of evolution of a tumour in developing a model to describe it, the work presented here offers ample scope for applying and extending these methods in future work.

Appendix A

Trajectories

Parameter	Values
Deme carrying capacity	$1, 512, 8192, \infty$
Driver mutation rate	$10^{-6}, 10^{-5}, 10^{-4}$
Selection coefficient	$0.05, 0.1, 0.2$
Baseline death rate (non-spatial)	0.98
Baseline death rate (spatial)	0

Table A.1: Parameters used for the simulations. The deme carrying capacity is varied across spatial configurations (boundary growth, invasive glandular, gland fission, non-spatial respectively), while other parameter variations are common to all simulations.

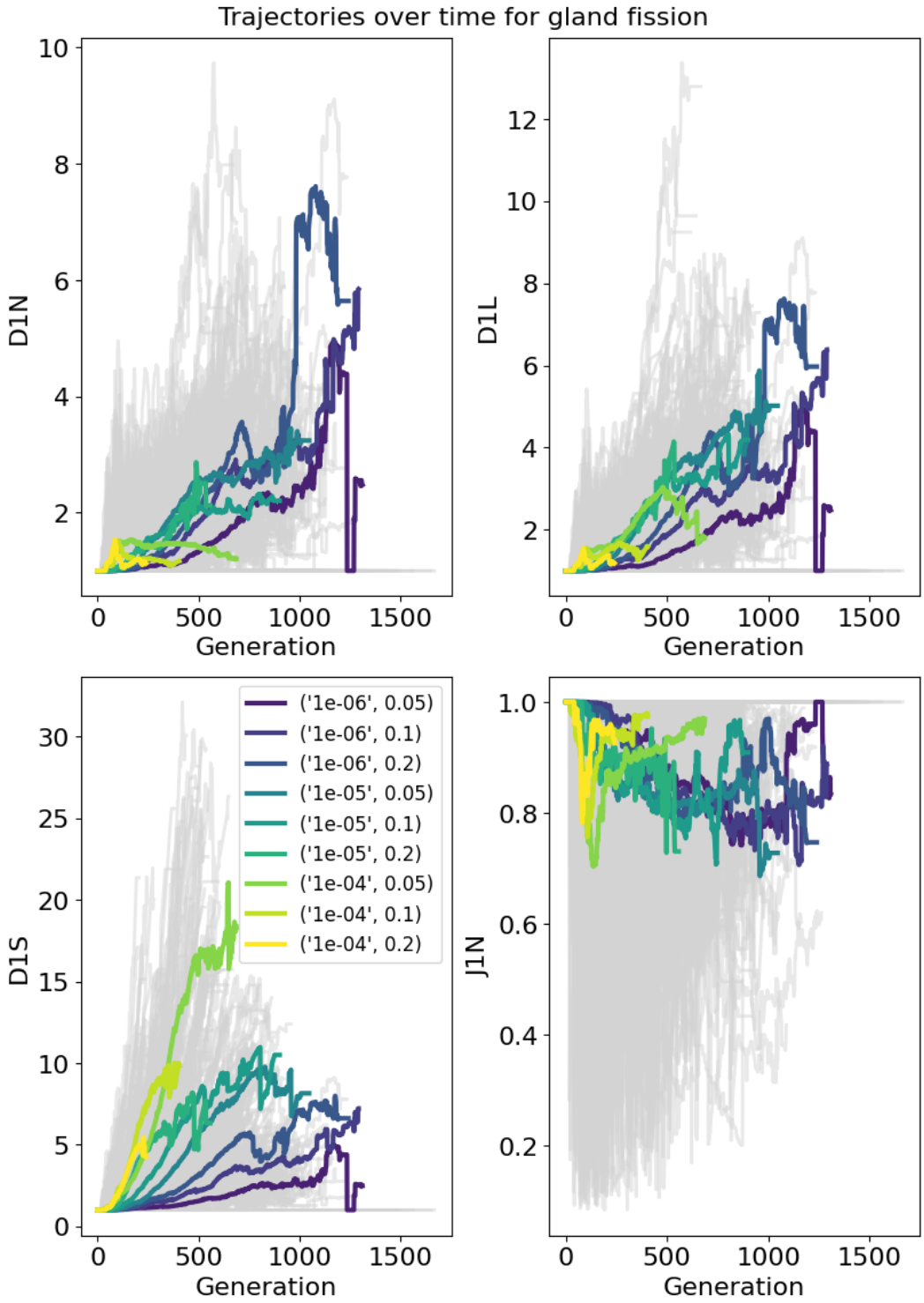

Figure A.1: All trajectories in time for the new set of indices plotted for gland fission with the average trajectories for different sets of indices plotted in colour.

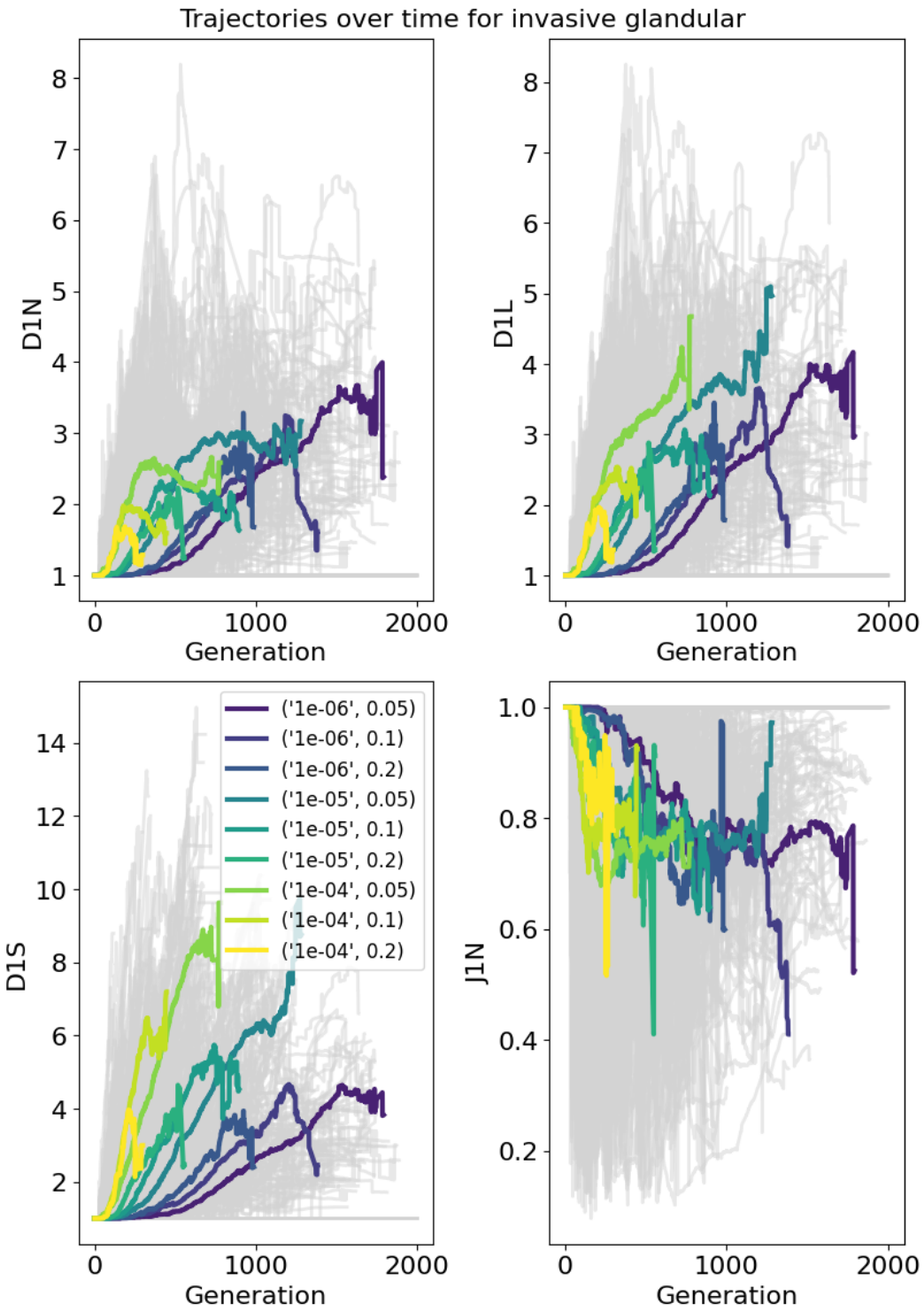

Figure A.2: All trajectories in time for the new set of indices plotted for invasive glandular evolution with the average trajectories for different sets of indices plotted in colour.

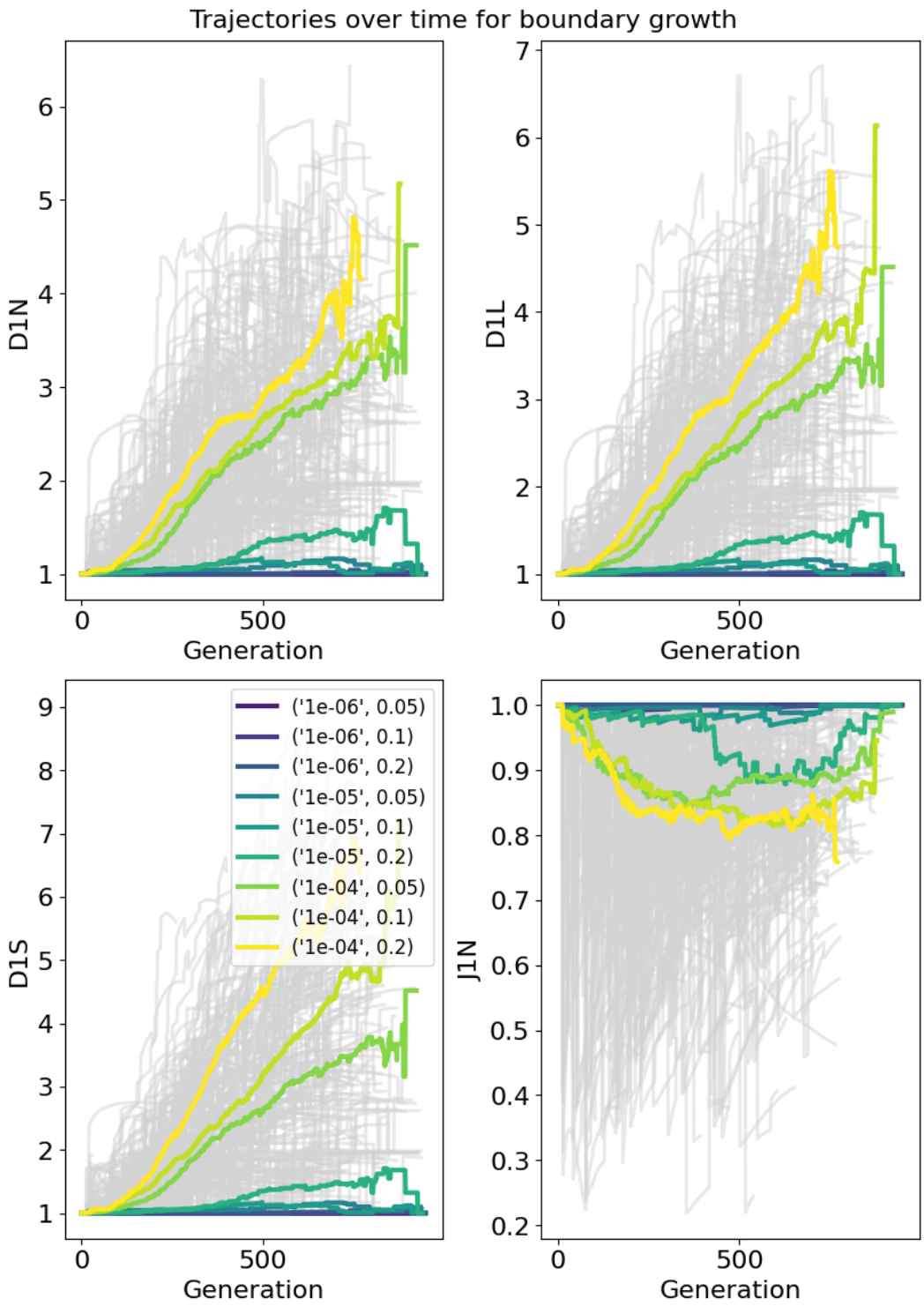

Figure A.3: All trajectories in time for the new set of indices plotted for boundary growth with the average trajectories for different sets of indices plotted in colour.

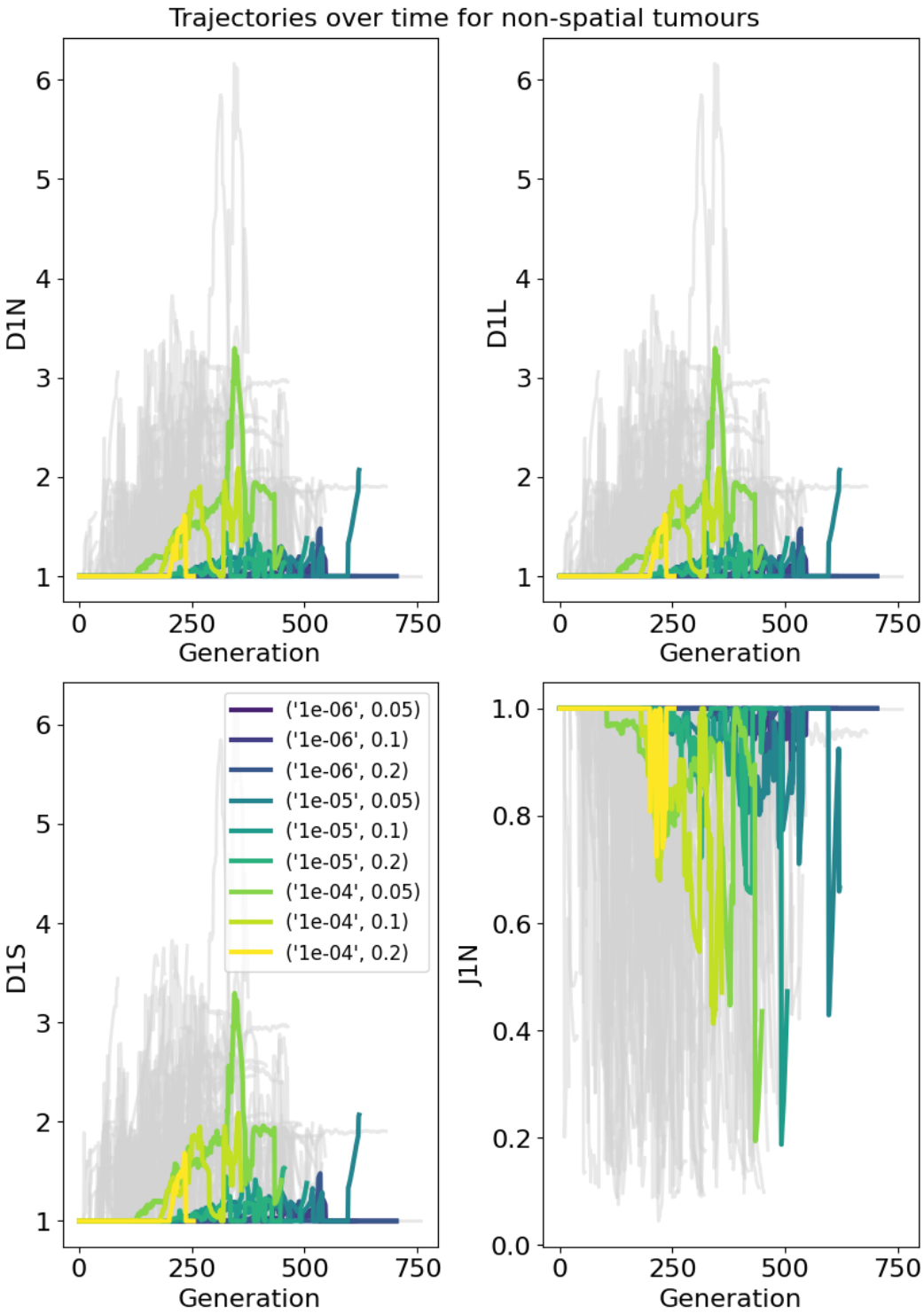

Figure A.4: All trajectories in time for the new set of indices plotted for non-spatial tumours with the average trajectories for different sets of indices plotted in colour.

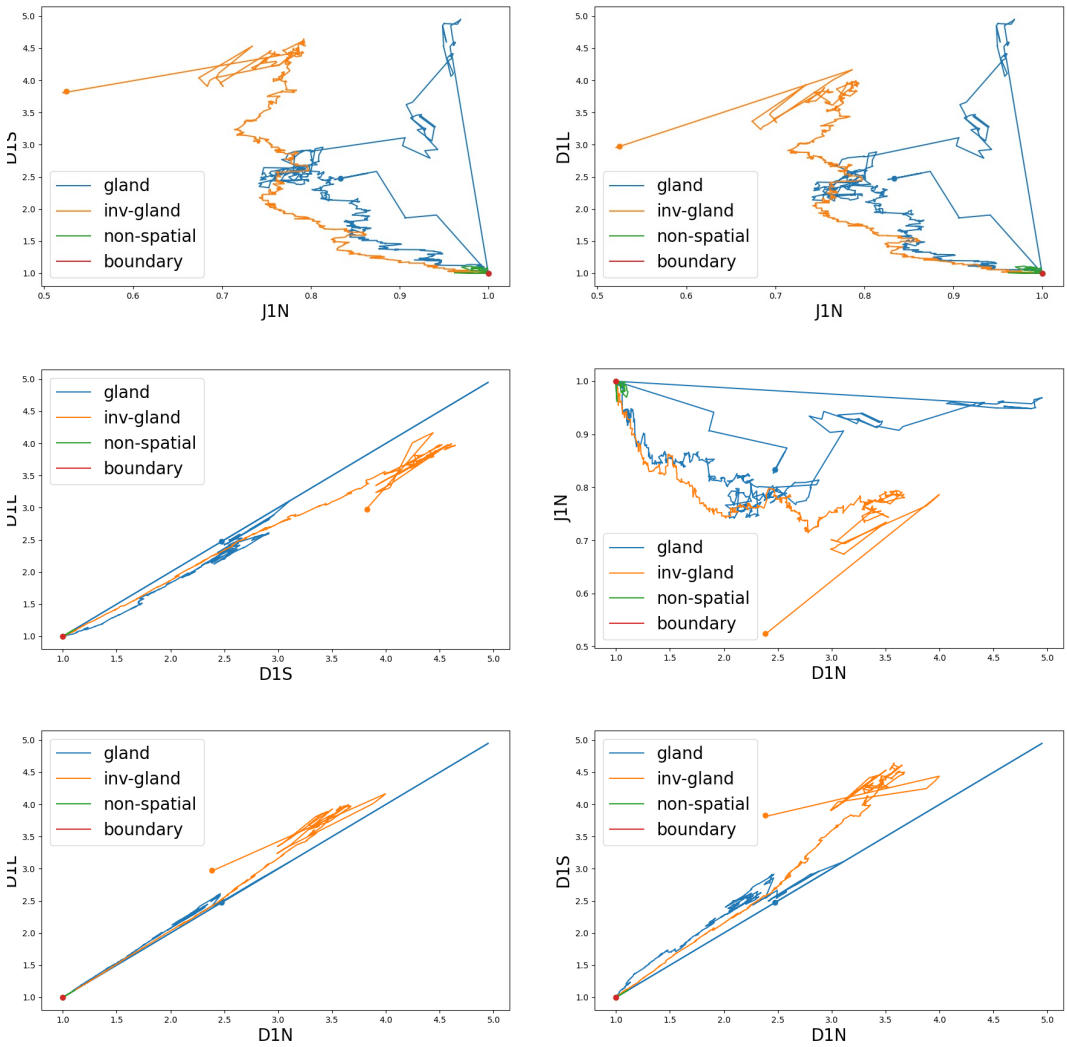

Figure A.5: Trajectories plotted for the four different spatial configurations for the driver mutation rate $\mu = 10^{-6}$, and selective coefficient $s = 0.05$.

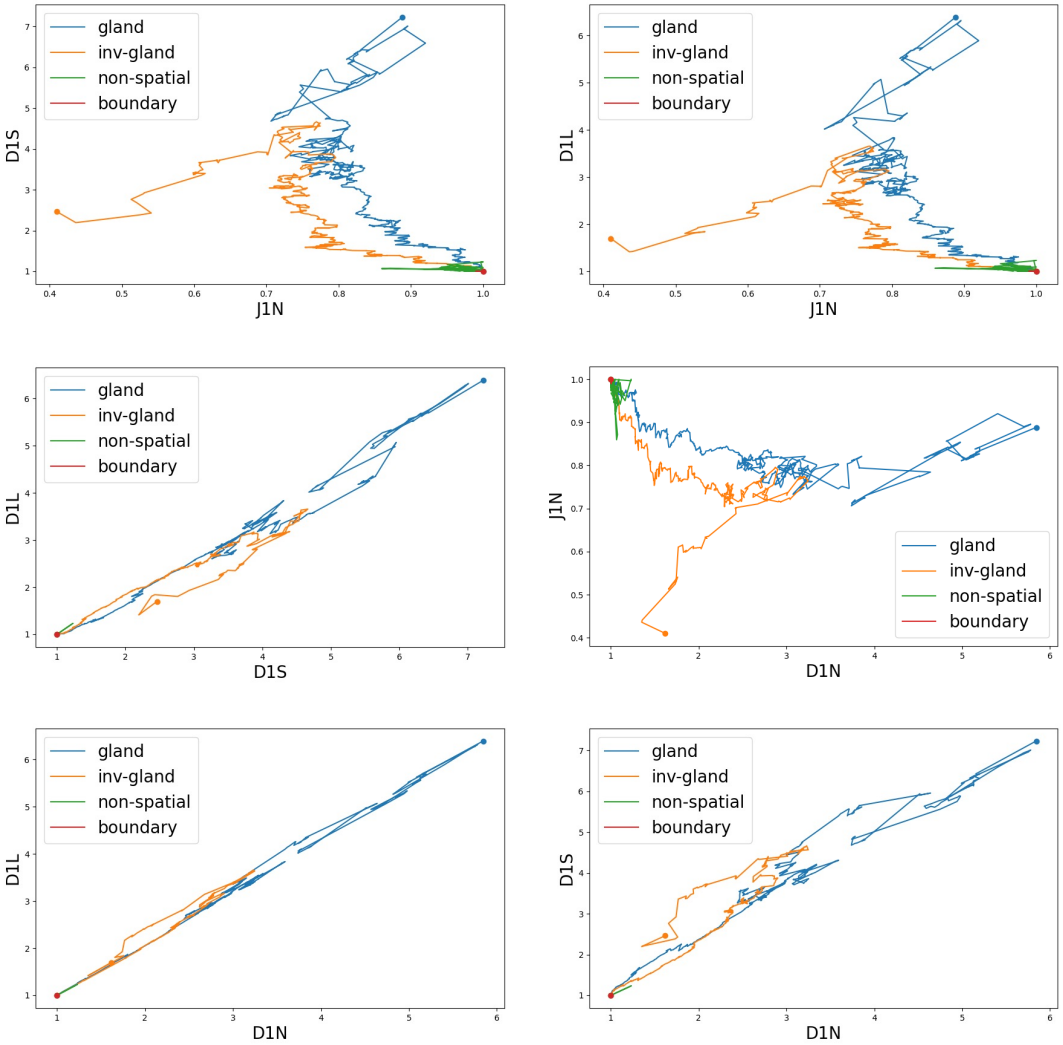

Figure A.6: Trajectories plotted for the four different spatial configurations for the driver mutation rate $\mu = 10^{-6}$, and selective coefficient $s = 0.1$.

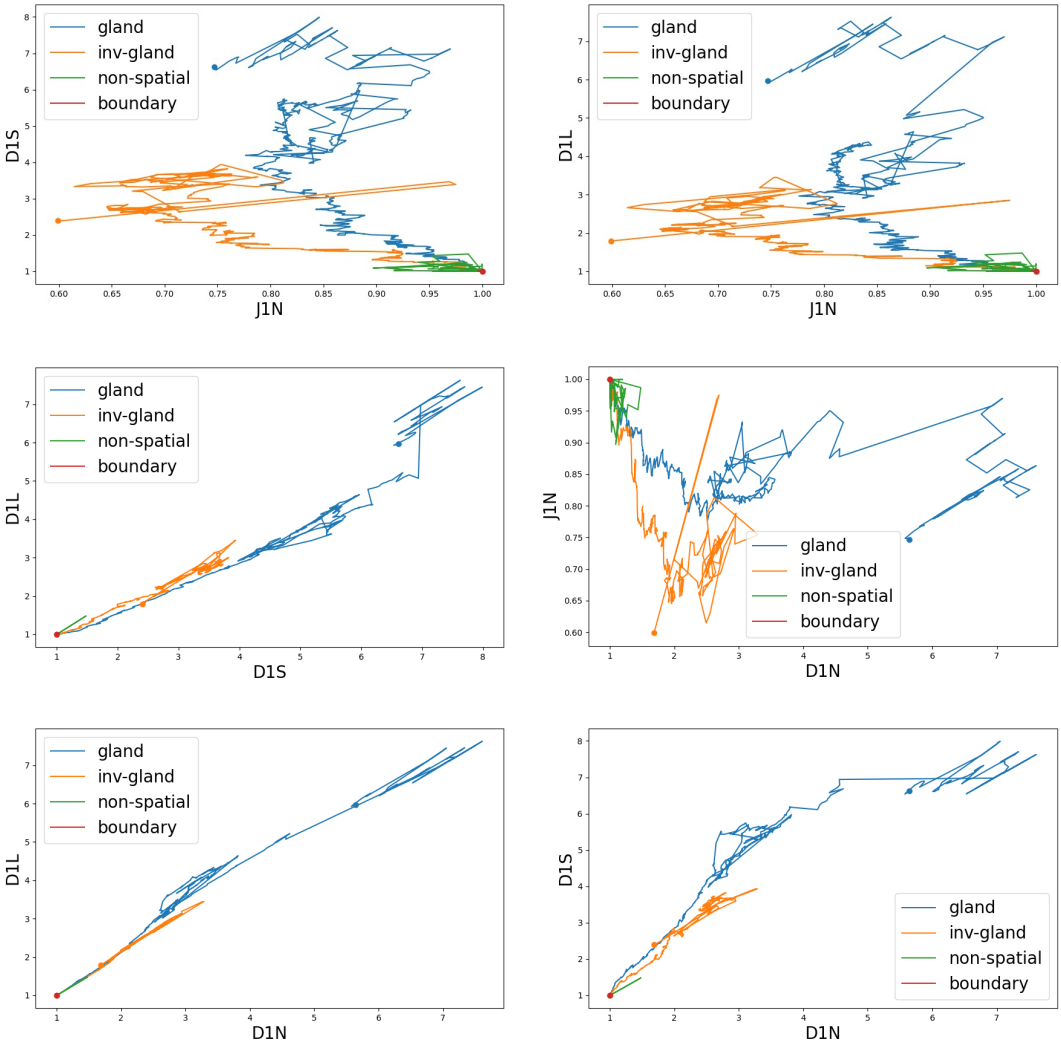

Figure A.7: Trajectories plotted for the four different spatial configurations for the driver mutation rate $\mu = 10^{-6}$, and selective coefficient $s = 0.2$.

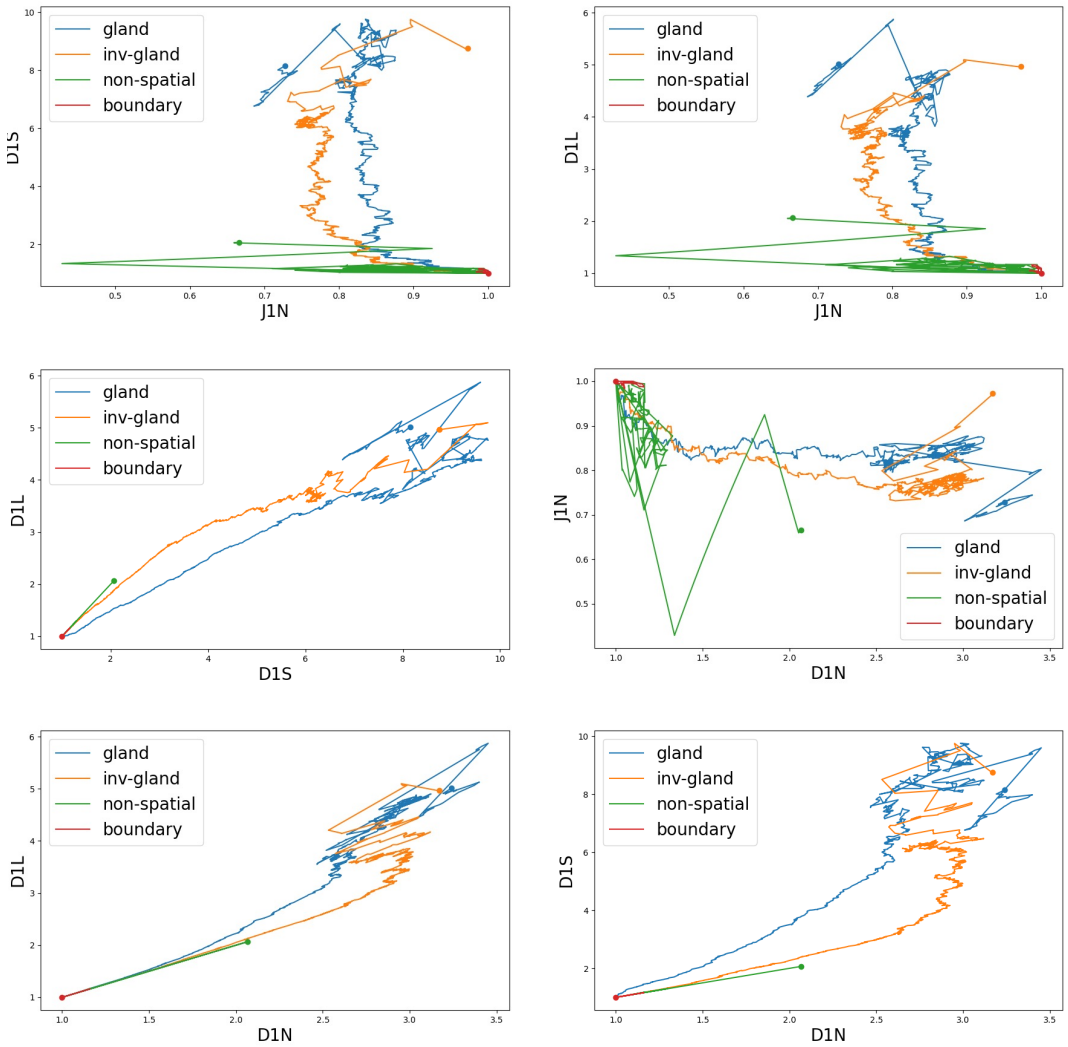

Figure A.8: Trajectories plotted for the four different spatial configurations for the driver mutation rate $\mu = 10^{-5}$, and selective coefficient $s = 0.05$.

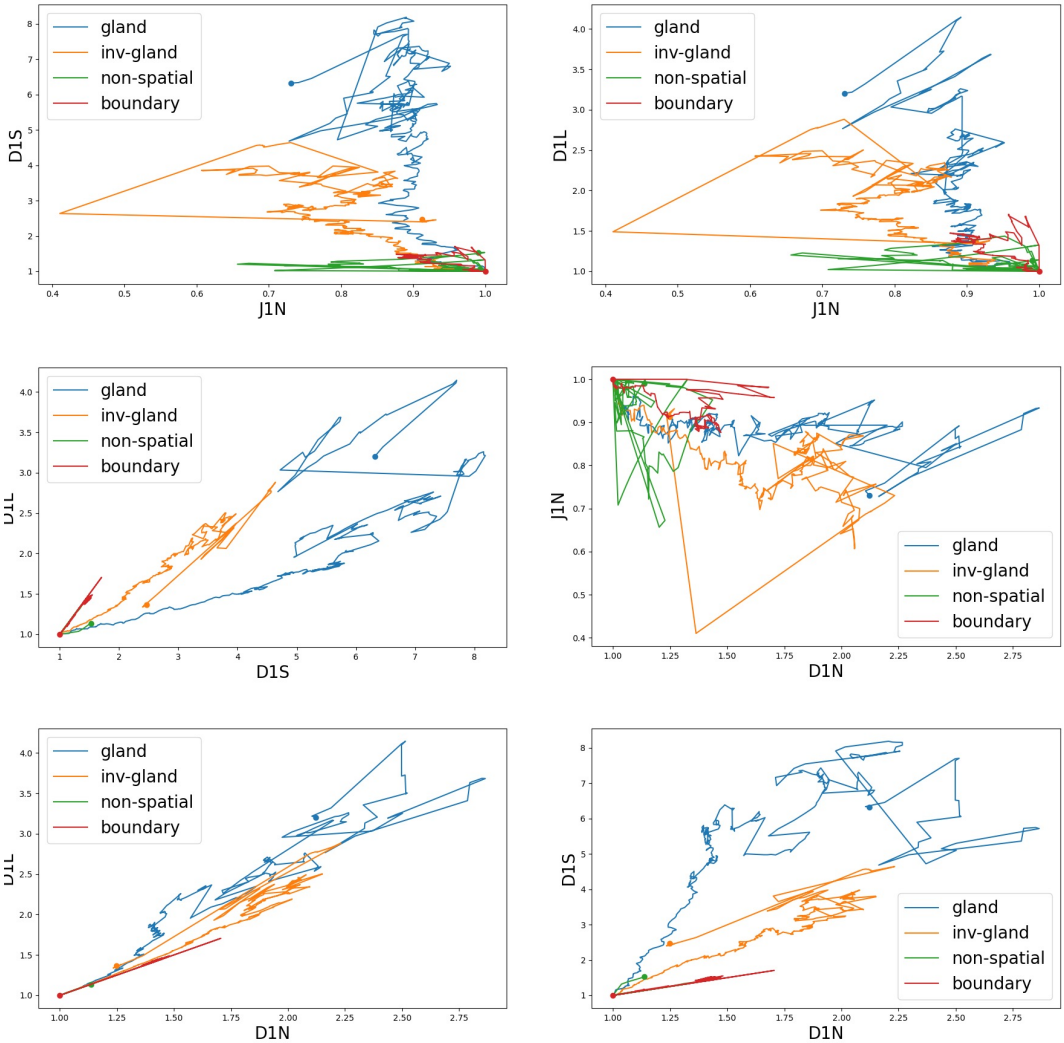

Figure A.9: Trajectories plotted for the four different spatial configurations for the driver mutation rate $\mu = 10^{-5}$, and selective coefficient $s = 0.2$.

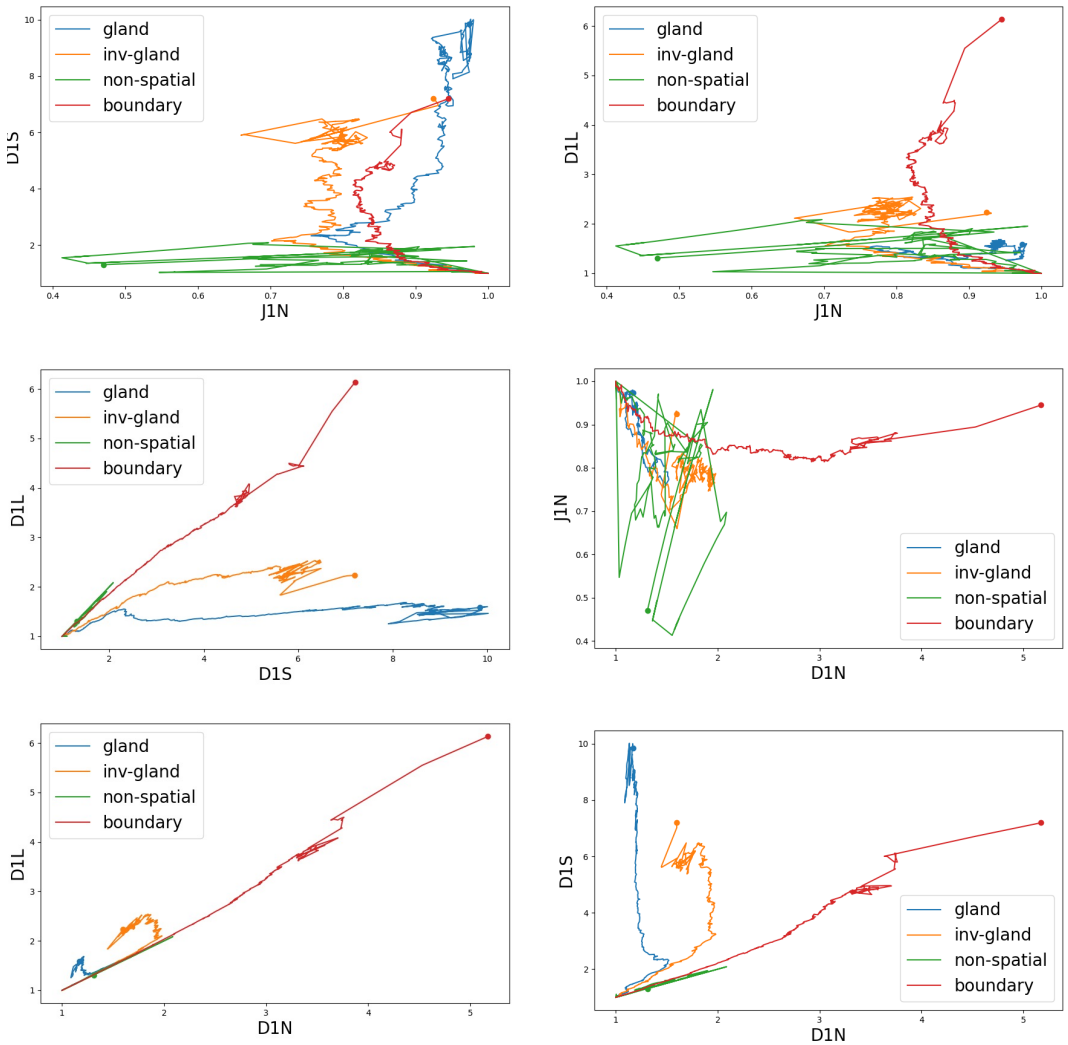

Figure A.10: Trajectories plotted for the four different spatial configurations for the driver mutation rate $\mu = 10^{-4}$, and selective coefficient $s = 0.1$.

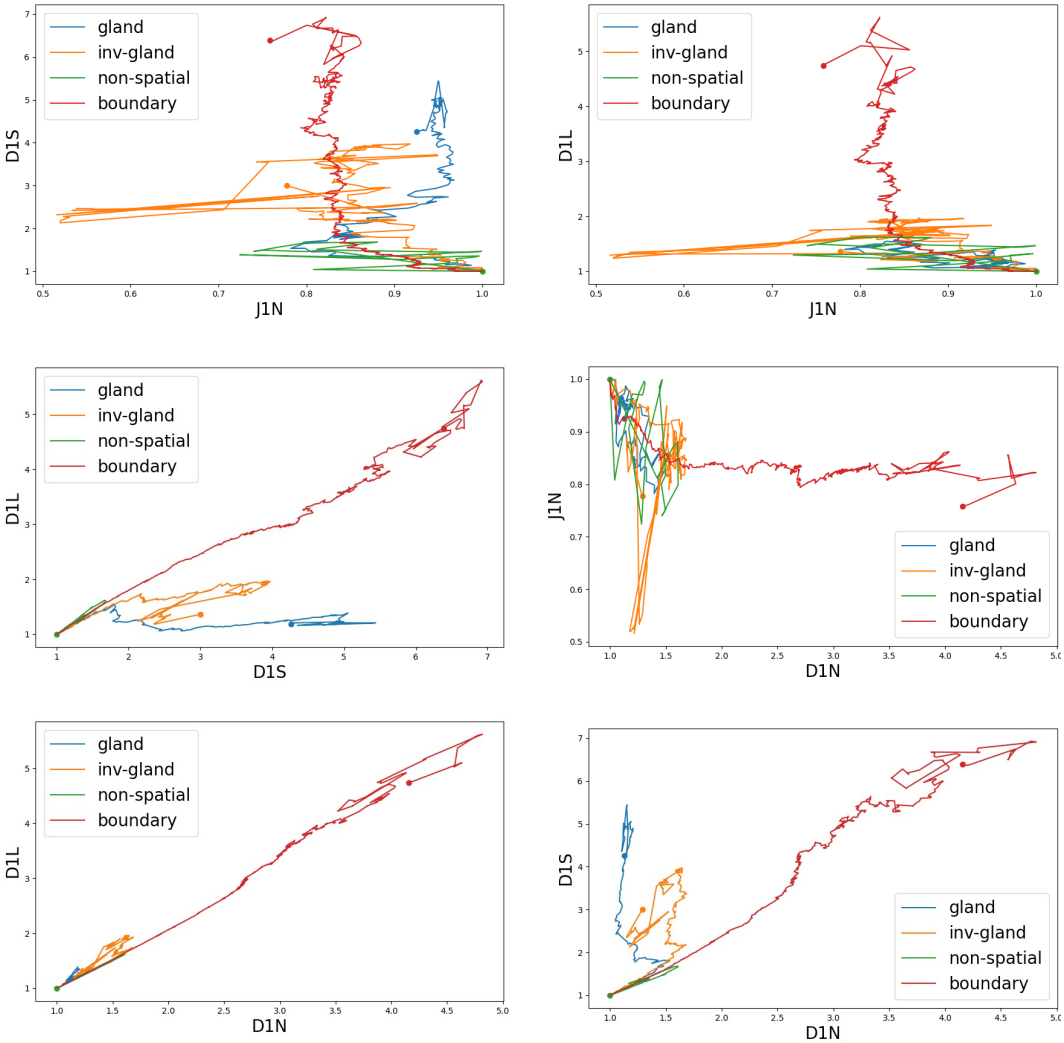

Figure A.11: Trajectories plotted for the four different spatial configurations for the driver mutation rate $\mu = 10^{-4}$, and selective coefficient $s = 0.2$.

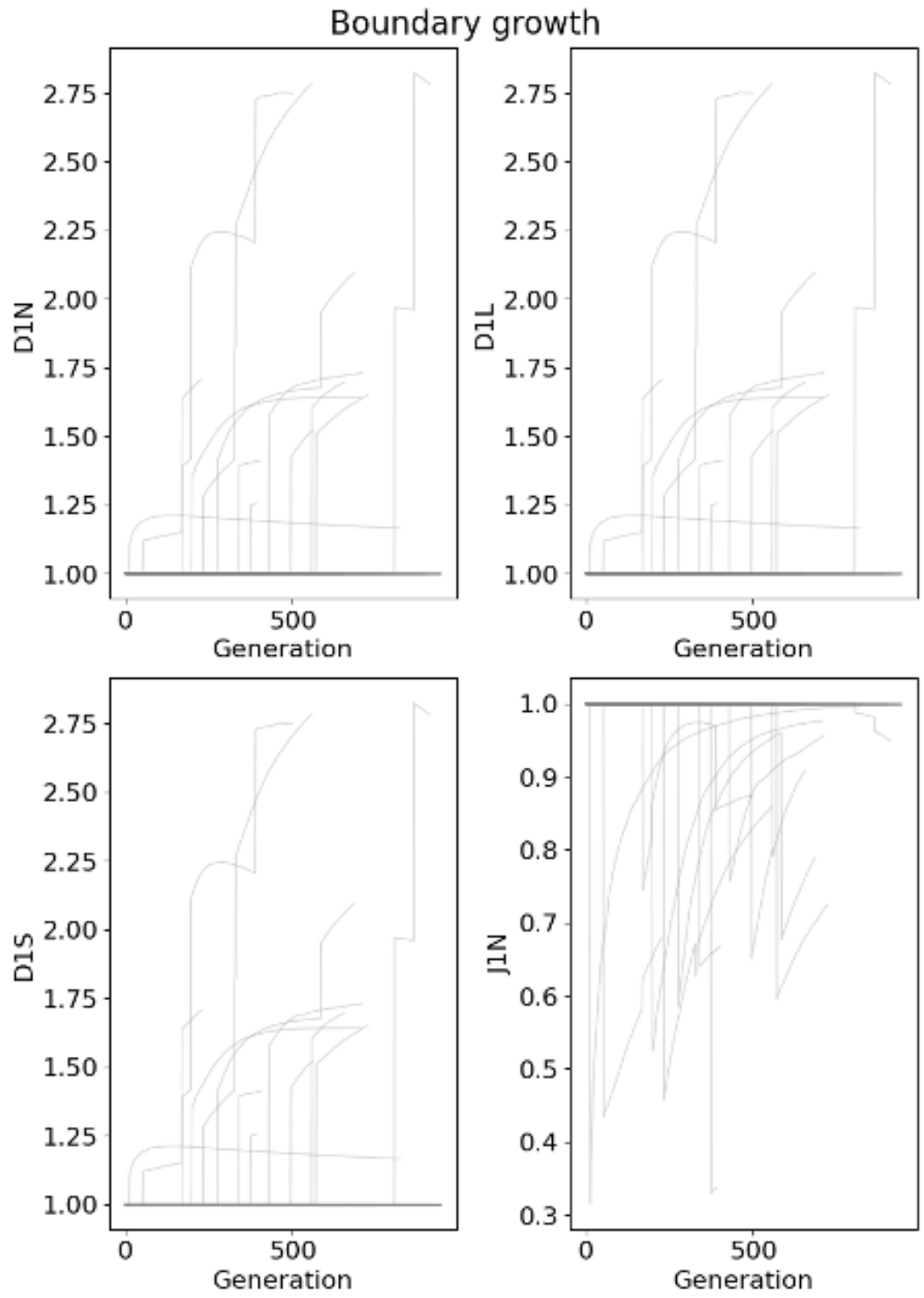

Figure A.12: Individual replicates' index trajectories of boundary growth for the parameters used in figure 3.5.

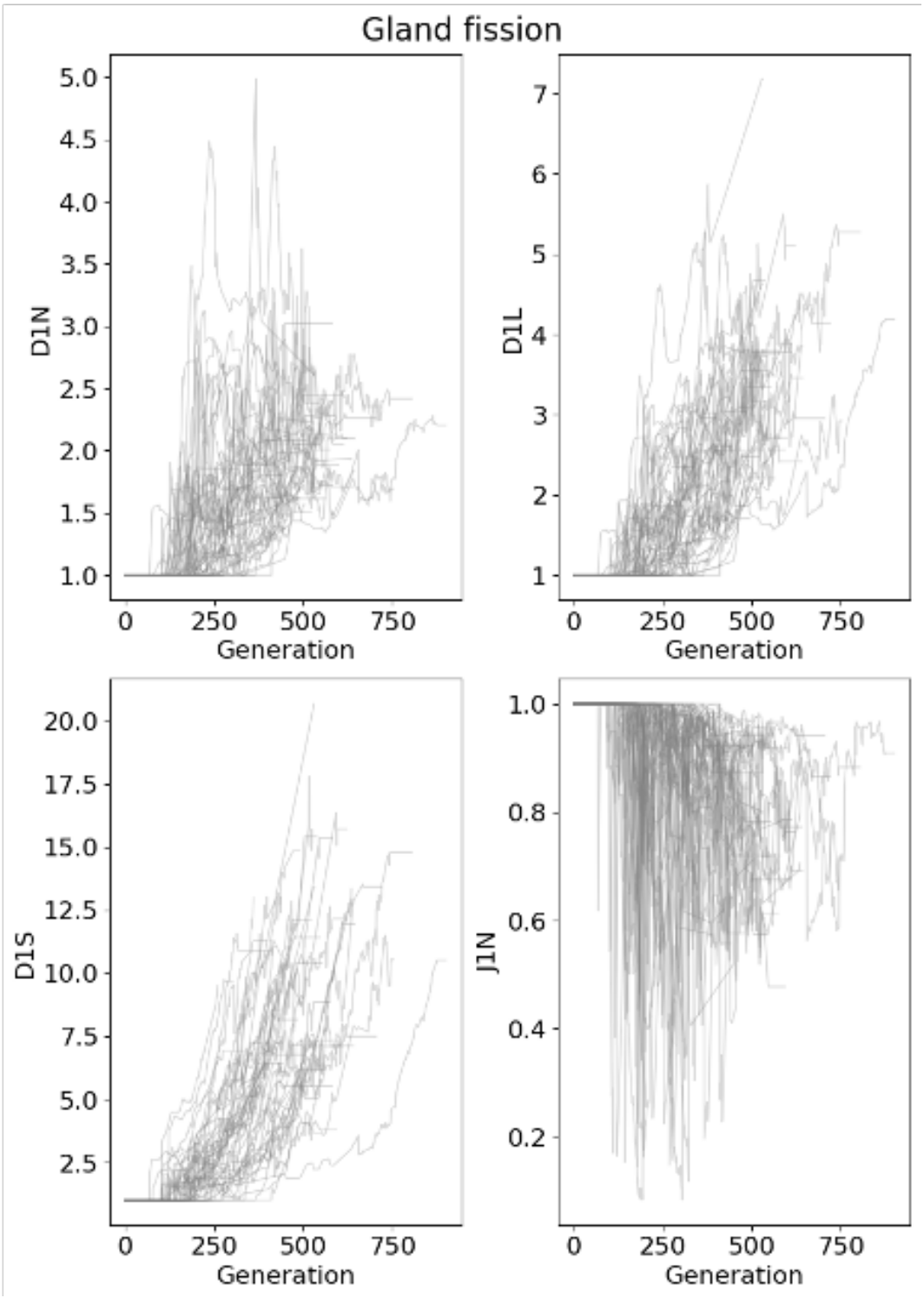

Figure A.13: Individual replicates' index trajectories of gland fission for the parameters used in figure 3.5.

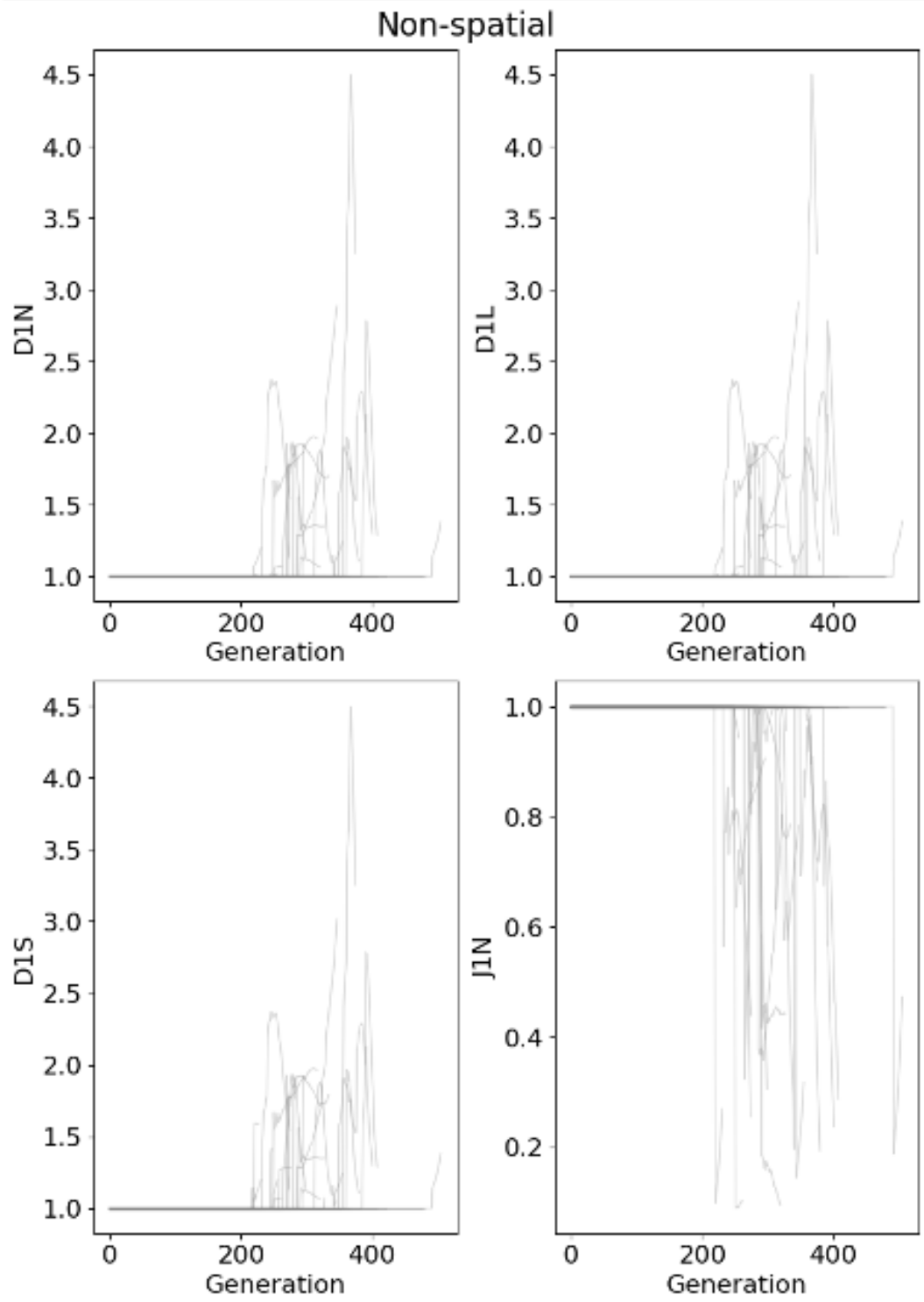

Figure A.14: Individual replicates' index trajectories of non-spatial growth for the parameters used in figure 3.5.

Appendix B

Parameter inference

Tumour	Median methylation rate	Median demethylation rate
E	0.0021	0.002
I	0.005	0.0009
J	0.0012	0.0018
S	0.001	0.0033
X	0.0008	0.0025

Table B.1: Inferred epimutation rates for the tumour samples modelled in this thesis.

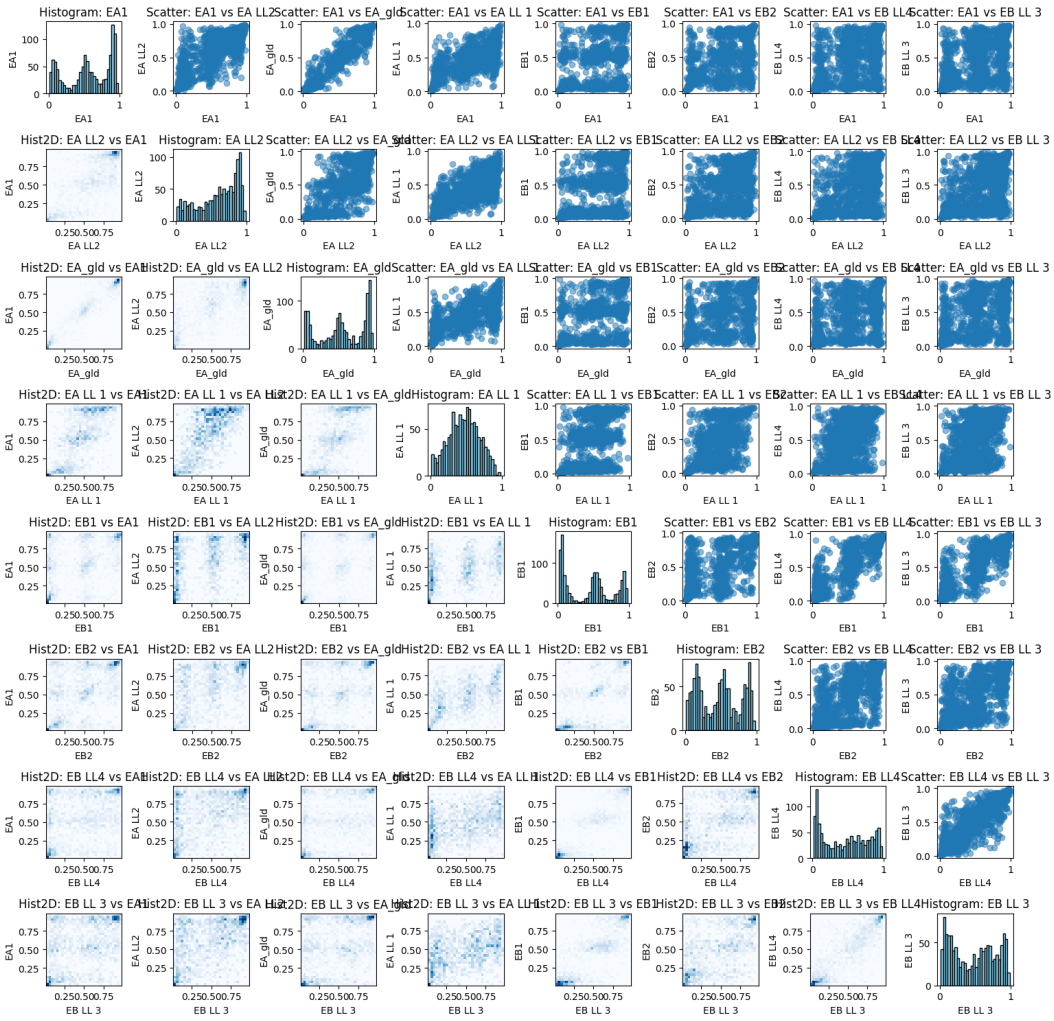

Figure B.1: Visualisation of fCpG arrays for patient E and the inter-gland correlation plots.

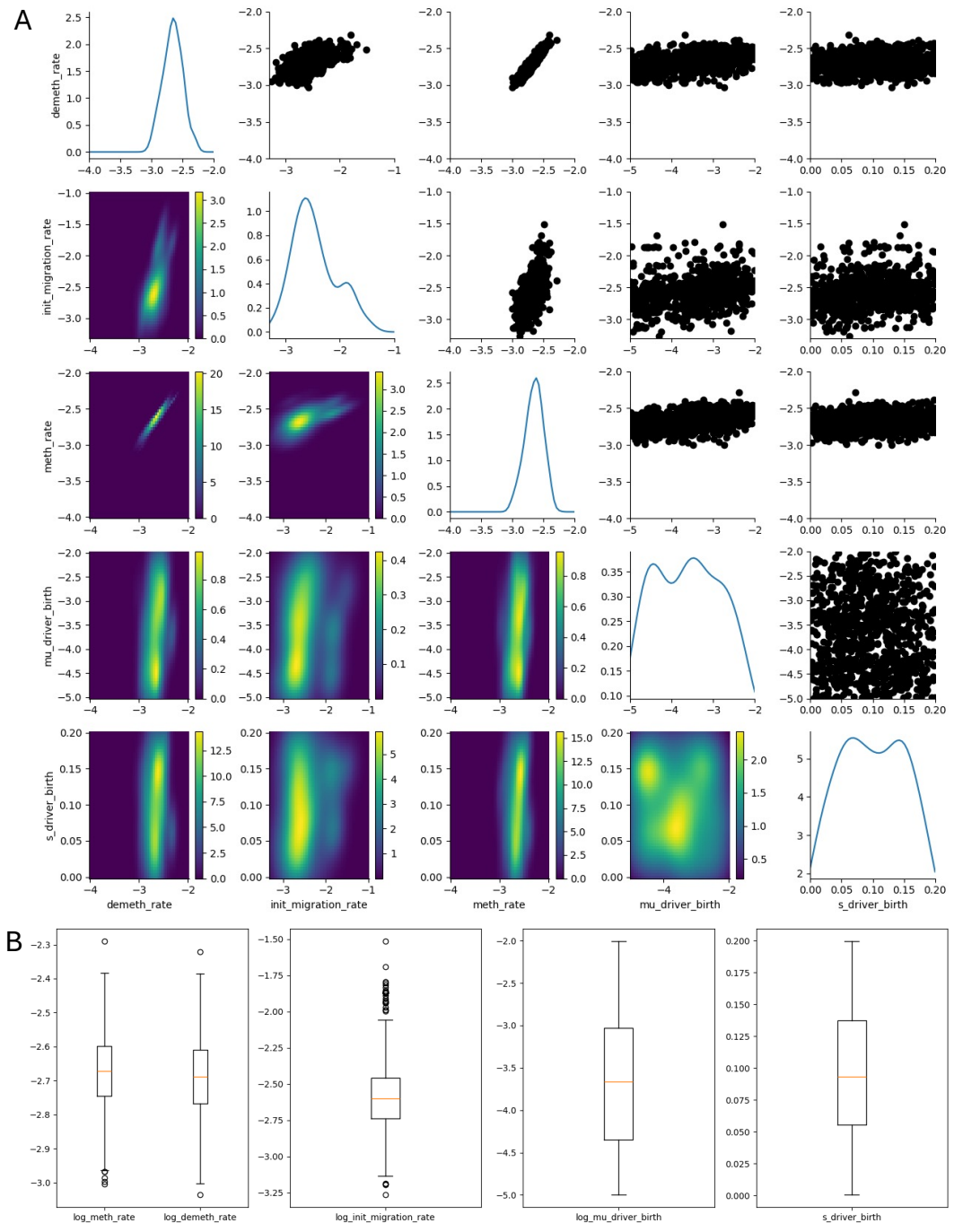

Figure B.2: Parameter inference plots for patient E.

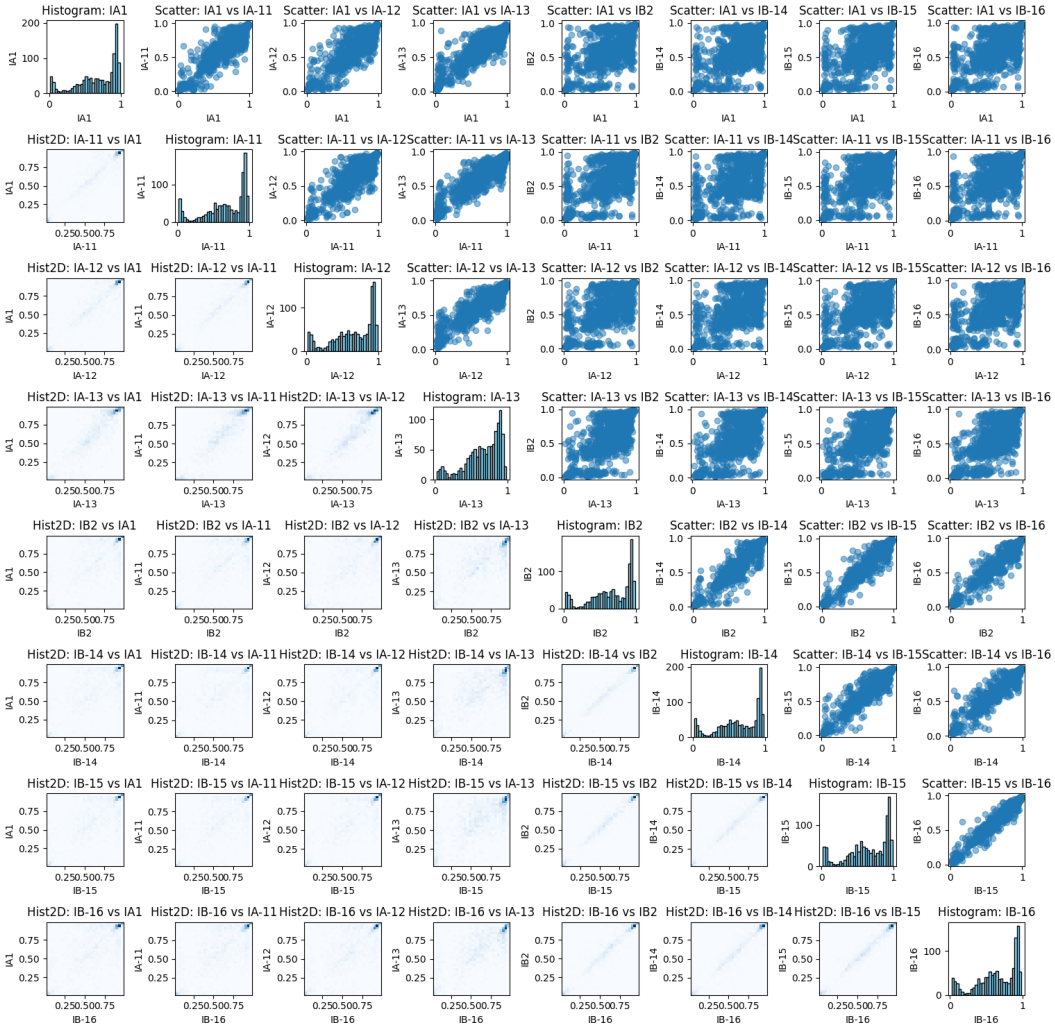

Figure B.3: Visualisation of fCpG arrays for patient I and the inter-gland correlation plots.

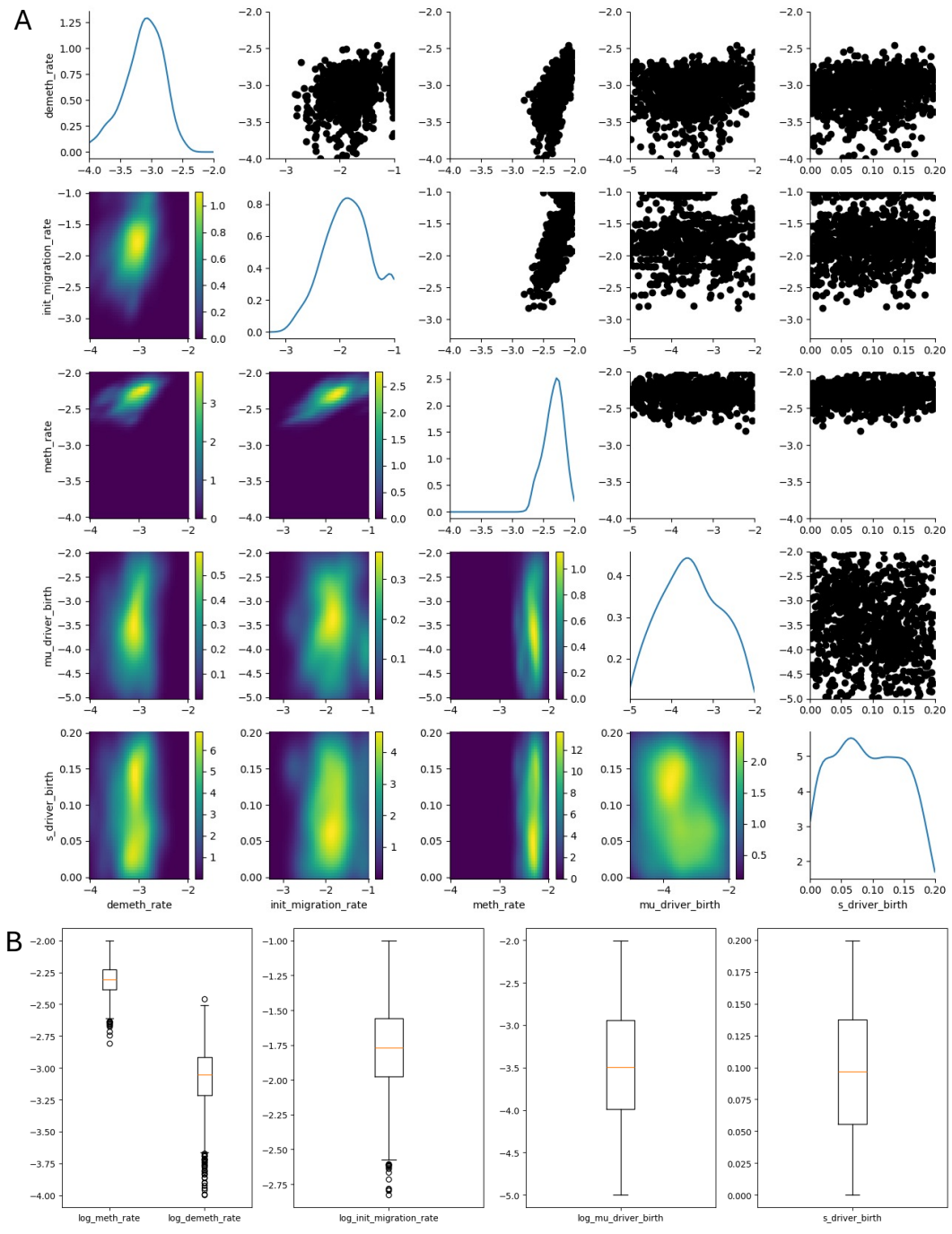

Figure B.4: Parameter inference plots for patient I.

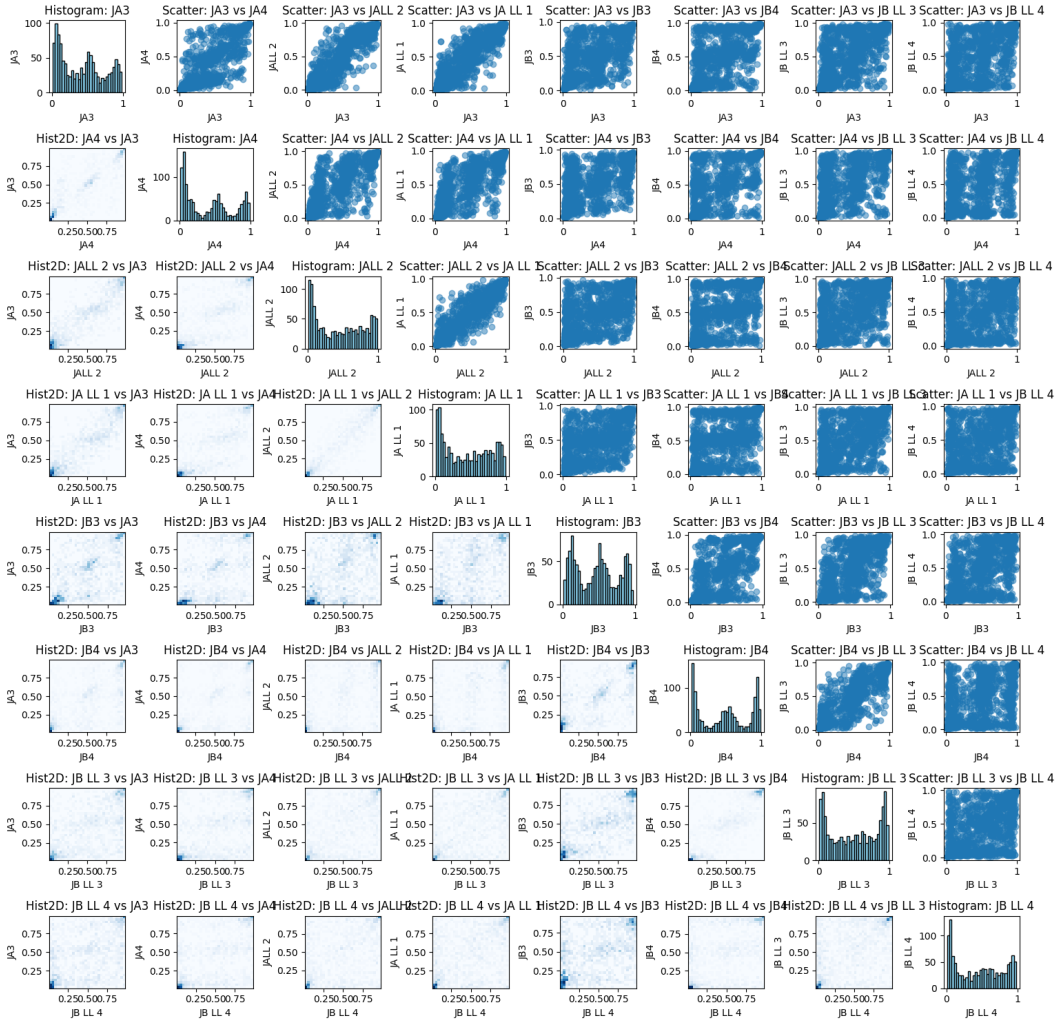

Figure B.5: Visualisation of fCpG arrays for patient J and the inter-gland correlation plots.

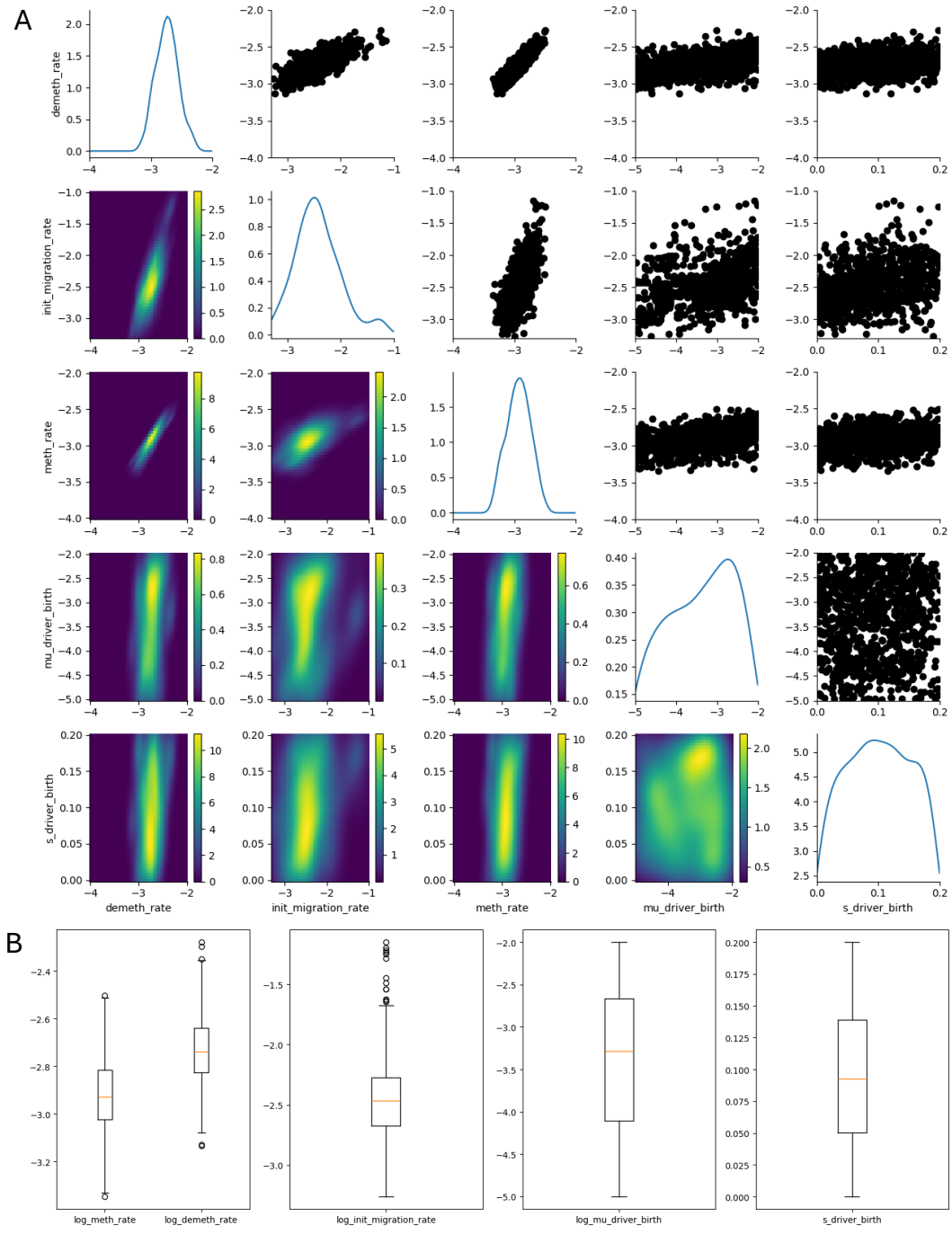

Figure B.6: Inference of the parameters for patient J.

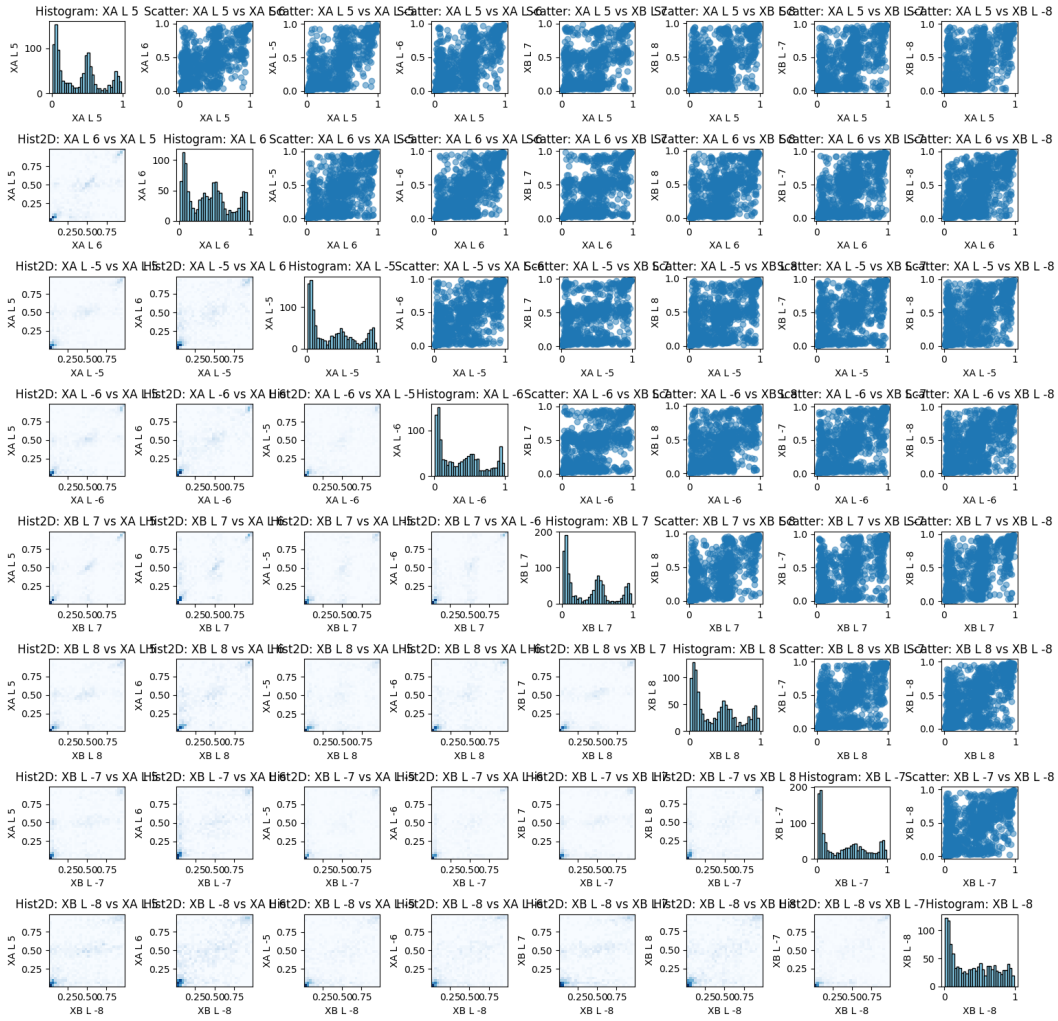

Figure B.7: Visualisation of fCpG arrays for patient X and the inter-gland correlation plots.

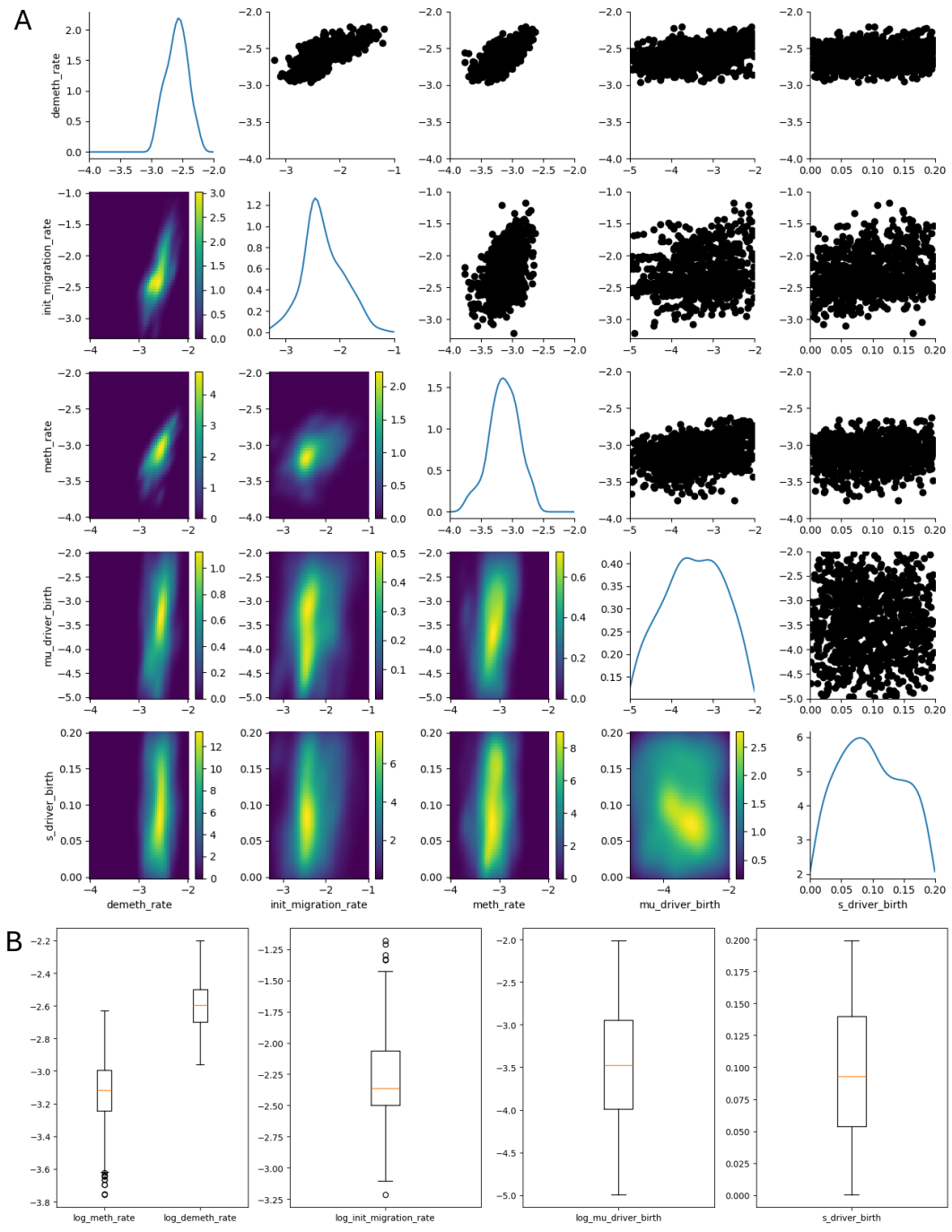

Figure B.8: Inference of the parameters for patient X.

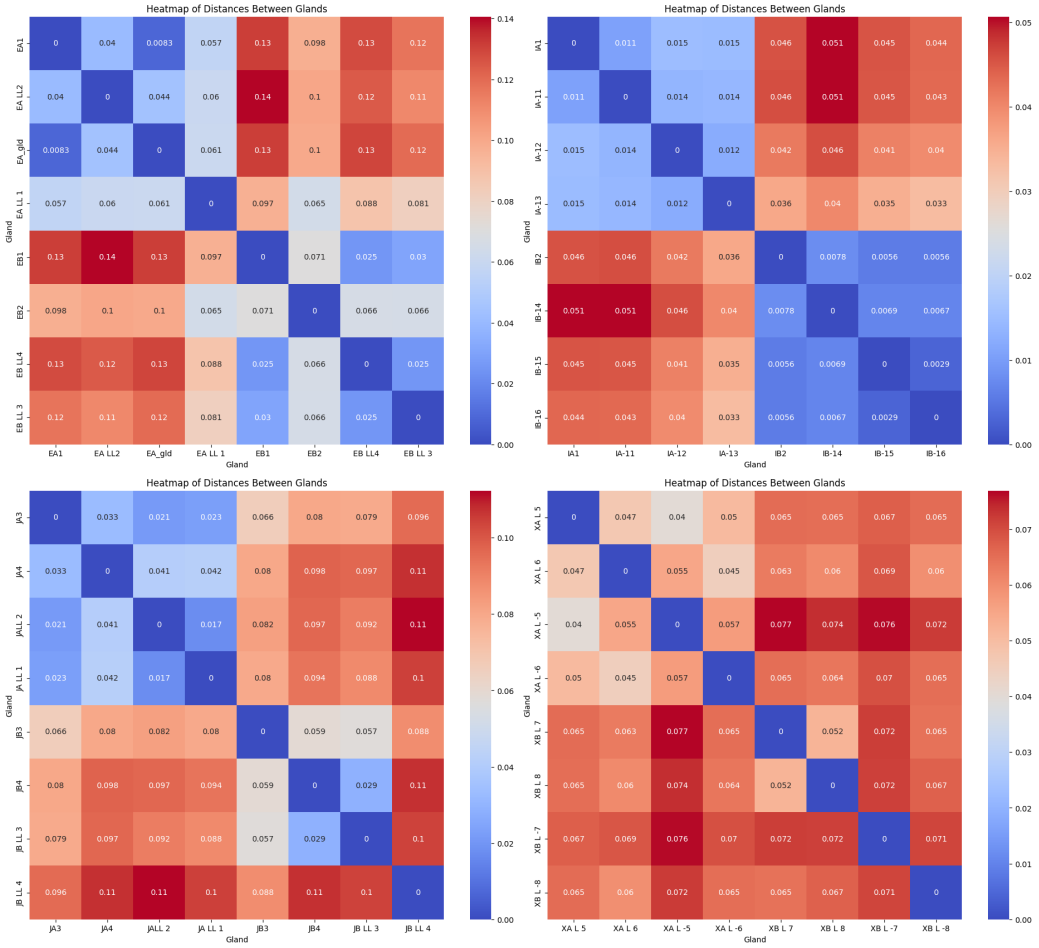

Figure B.9: Inter-gland distance matrices for the tumours modelled in this thesis. Clockwise from top left: E, I, X, J.

Bibliography

Aldous, D. J. (2001), 'Stochastic models and descriptive statistics for phylogenetic trees, from yule to today', **16**(1), 23–34.
URL: *https://projecteuclid.org/journals/statistical-science/volume-16/issue-1/Stochastic-models-and-descriptive-statistics-for-phylogenetic-trees-from-Yule/10.1214/ss/998929474.full*

Almet, A. A., Hughes, B. D., Landman, K. A., Näthke, I. S. & Osborne, J. M. (2018), 'A multicellular model of intestinal crypt buckling and fission', **80**(2), 335–359.
URL: *http://link.springer.com/10.1007/s11538-z*

Bak, M., Colyer, B., Manojlović, V. & Noble, R. (2023), 'Warlock: an automated computational workflow for simulating spatially structured tumour evolution'.
URL: *http://arxiv.org/abs/2301.07808*

Bayes, M. & Price, M. (1763), *An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.*, Royal Society of London.
URL: *http://archive.org/details/philtrans09948070*

Blum, M. G. B., Nunes, M. A., Prangle, D. & Sisson, S. A. (2013), 'A comparative review of dimension reduction methods in approximate bayesian computation', **28**(2), 189–208.
URL: *https://projecteuclid.org/journals/statistical-science/volume-28/issue-2/A-Comparative-Review-of-Dimension-Reduction-Methods-in-Approximate-Bayesian/10.1214/12-STS406.full*

Bondi, L., Bonetti, M., Grigorova, D. & Russo, A. (2023), 'Approximate bayesian computation for the natural history of breast cancer, with application to data from a milan cohort study', **42**(18), 3093–3113.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Plessis, L. d., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T. & Drummond, A. J. (2019), 'BEAST 2.5: An advanced software platform for bayesian evolutionary analysis', **15**(4), e1006650.
URL: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006650*

Bowel cancer statistics (2015).
URL: *https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer*

Bozic, I., Paterson, C. & Waclaw, B. (2019), 'On measuring selection in cancer from subclonal mutation frequencies', **15**(9), e1007368.
URL: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007368*

Bravo, R. R., Baratchart, E., West, J., Schenck, R. O., Miller, A. K., Gallaher, J., Gatenbee, C. D., Basanta, D., Robertson-Tessi, M. & Anderson, A. R. A. (2020), 'Hybrid automata library: A flexible platform for hybrid modeling with real-time visualization', **16**(3), e1007635.
URL: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007635*

Bull, J. A. & Byrne, H. M. (2022), 'The hallmarks of mathematical oncology', pp. 1–18.

Cardona, G., Mir, A. & Rossello, F. (2012), 'Exact formulas for the variance of several balance indices under the yule model'.
URL: *http://arxiv.org/abs/1202.6573*

Cernat, L., Blaj, C., Jackstadt, R., Brandl, L., Engel, J., Hermeking, H., Jung, A., Kirchner, T. & Horst, D. (2014), 'Colorectal cancers mimic structural organization of normal colonic crypts', **9**(8), e104284.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4128715/*

Chauvin, B. & Rouault, A. (2004), 'Connecting yule process, bisection and binary search tree via martingales'.
URL: *https://arxiv.org/abs/math/0410318v1*

Chen, B., Ford, D. & Winkel, M. (2009), 'A new family of Markov branching trees: the alpha-gamma model', *Electronic Journal of Probability* **14**(none), 400 – 430.

Chomsky, N. (1957), *Syntactic structures*, Syntactic structures, Mouton.

Colless, D. H. (1982), 'Review of phylogenetics: The theory and practice of phylogenetic systematics.', **31**(1), 100–104.
URL: *https://www.jstor.org/stable/2413420*

Colyer, B., Bak, M., Basanta, D. & Noble, R. (2023), 'A seven-step guide to spatial, agent-based modelling of tumour evolution'.
URL: *http://arxiv.org/abs/2311.03569*

Davis, A., Gao, R. & Navin, N. (2017), 'Tumor evolution: Linear, branching, neutral or punctuated?', **1867**(2), 151–161.
URL: *https://www.sciencedirect.com/science/article/pii/S0304419X17300197*

Eden, M. (1961), 'A two-dimensional growth process', **4.4**, 223–240.
URL: *https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fourth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/A-Two-dimensional-Growth-Process/bsmsp/1200512888*

Eiseley, L. C. (1945), 'Tempo and mode in evolution. by george gaylord simpson. columbia university press, new york city, 1944. 217 pp. of text, bibliography, and index. 36 figures, 19 tables. price $3.50', **3**(2), 208–209.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.1330030215*

Fischer, M. (2021), 'Extremal values of the sackin tree balance index', **25**(2), 515–541.
URL: *https://link.springer.com/10.1007/s00026-2*

Fischer, M., Herbst, L., Kersting, S., Kühn, L. & Wicke, K. (2021), 'Tree balance indices: a comprehensive survey'.
URL: *http://arxiv.org/abs/2109.12281*

Fleming, M., Ravula, S., Tatishchev, S. F. & Wang, H. L. (2012), 'Colorectal carcinoma: Pathologic aspects', **3**(3), 153–173.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3418538/*

Freitas, O., Campos, P. R. A. & Araujo, S. B. L. (2024), 'Patch biogeography under intermittent barriers: macroevolutionary consequences of microevolutionary processes', *Journal of Evolutionary Biology* p. voae035.
URL: *https://doi.org/10.1093/jeb/voae035*

Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. (2016), 'Excess of mutational jackpot events in expanding populations revealed by spatial luria-delbrück experiments', **7**, 12760.

Gabbutt, C., Duran-Ferrer, M., Grant, H., Mallo, D., Nadeu, F., Househam, J., Villamor, N., Krali, O., Nordlund, J., Zenz, T., Campo, E., Lopez-Guillermo, A., Fitzgibbon, J., Barnes, C. P., Shibata, D., Martin-Subero, J. I. & Graham, T. A. (2023), 'Evolutionary dynamics of 1,976 lymphoid malignancies predict clinical outcome'.
URL: *https://www.medrxiv.org/content/10.1101/2023.11.10.23298336v1*

Gabbutt, C., Schenck, R. O., Weisenberger, D. J., Kimberley, C., Berner, A., Househam, J., Lakatos, E., Robertson-Tessi, M., Martin, I., Patel, R., Clark, S. K., Latchford, A., Barnes, C. P., Leedham, S. J., Anderson, A. R. A., Graham, T. A. & Shibata, D. (2022), 'Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues', **40**(5), 720–730.
URL: *https://www.nature.com/articles/s41587-w*

Gehart, H. & Clevers, H. (2019), 'Tales from the crypt: new insights into intestinal stem cells', **16**(1), 19–34.
URL: *https://www.nature.com/articles/s41575-y*

Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M. & Macklin, P. (2018), 'PhysiCell: An open source physics-based cell simulator for 3-d multicellular systems', **14**(2), e1005991.
URL: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005991*

Gillespie, D. T. (1977), 'Exact stochastic simulation of coupled chemical reactions', **81**(25), 2340–2361.
URL: *https://doi.org/10.1021/j100540a008*

Glassman, M. L., De Groot, N. & Hochberg, A. (1996), 'Cancer, evolution and birth:

Reliving our ancestral past', **46**(1), 13–16.
URL: *https://www.sciencedirect.com/science/article/pii/S0306987796902273*

Goh, G., Fuchs, M. & Zhang, L. (2022), 'Two results about the sackin and colless indices for phylogenetic trees and their shapes', **85**(6), 69.
URL: *https://link.springer.com/10.1007/s00285-2*

Hasegawa, M., Kishino, H. & Yano, T. (1985), 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA', **22**(2), 160–174.

Heide, T., Househam, J., Cresswell, G. D., Spiteri, I., Lynn, C., Mossner, M., Kimberley, C., Fernandez-Mateos, J., Chen, B., Zapata, L., James, C., Barozzi, I., Chkhaidze, K., Nichol, D., Gunasri, V., Berner, A., Schmidt, M., Lakatos, E., Baker, A.-M., Costa, H., Mitchinson, M., Piazza, R., Jansen, M., Caravagna, G., Ramazzotti, D., Shibata, D., Bridgewater, J., Rodriguez-Justo, M., Magnani, L., Graham, T. A. & Sottoriva, A. (2022), 'The co-evolution of the genome and epigenome in colorectal cancer', **611**(7937), 733–743.
URL: *https://www.nature.com/articles/s41586-1*

Heide, T., Zapata, L., Williams, M. J., Werner, B., Caravagna, G., Barnes, C. P., Graham, T. A. & Sottoriva, A. (2018), 'Reply to 'neutral tumor evolution?'', **50**(12), 1633–1637.
URL: *https://www.nature.com/articles/s41588-z*

Herald, M., Nicuşan, A., Wheldon, T. K., Seville, J. & Windows-Yule, C. (2022), 'Autonomous digitizer calibration of a monte carlo detector model through evolutionary simulation', **12**(1), 19535.
URL: *https://www.nature.com/articles/s41598-x*

Hong, Y., Marjoram, P., Shibata, D. & Siegmund, K. (2010), 'Using DNA methylation patterns to infer tumor ancestry', **5**, e12002.

Houchmandzadeh, B. & Vallade, M. (2017), 'Fisher waves: An individual-based stochastic model', **96**(1), 012414.
URL: *http://link.aps.org/doi/10.1103/PhysRevE.96.012414*

Househam, J., Heide, T., Cresswell, G. D., Spiteri, I., Kimberley, C., Zapata, L., Lynn, C., James, C., Mossner, M., Fernandez-Mateos, J., Vinceti, A., Baker, A.-M., Gabbutt, C., Berner, A., Schmidt, M., Chen, B., Lakatos, E., Gunasri,

V., Nichol, D., Costa, H., Mitchinson, M., Ramazzotti, D., Werner, B., Iorio, F., Jansen, M., Caravagna, G., Barnes, C. P., Shibata, D., Bridgewater, J., Rodriguez-Justo, M., Magnani, L., Sottoriva, A. & Graham, T. A. (2022), 'Phenotypic plasticity and genetic control in colorectal cancer evolution', **611**(7937), 744–753.
URL: *https://www.nature.com/articles/s41586-x*

Huffman, D. A. (1952), 'A method for the construction of minimum-redundancy codes', **40**(9), 1098–1101.
URL: *https://ieeexplore.ieee.org/document/4051119*

Institute, N. C. (2020), 'Financial burden of cancer care | cancer trends progress report'.
URL: *https://progressreport.cancer.gov/after/economic%5Fburden*

Jagiella, N., Rickert, D., Theis, F. J. & Hasenauer, J. (2017), 'Parallelization and high-performance computing enables automated statistical inference of multi-scale models', **4**(2), 194–206.e9.

Kantorovich, L. V. (1960), 'Mathematical methods of organizing and planning production', **6**(4), 366–422.
URL: *https://pubsonline.informs.org/doi/10.1287/mnsc.6.4.366*

Kharlamov, B. P. (1969), 'On the generation numbers of particles in a branching process with overlapping generations', **14**(1), 44–50.
URL: *http://epubs.siam.org/doi/10.1137/1114005*

Kirkpatrick, M. & Slatkin, M. (1993), 'Searching for evolutionary patterns in the shape of a phylogenetic tree', **47**(4), 1171–1181.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1993.tb02144.x*

Klinger, E., Rickert, D. & Hasenauer, J. (2018), 'pyABC: distributed, likelihood-free inference', **34**(20), 3591–3593.
URL: *https://doi.org/10.1093/bioinformatics/bty361*

Knuth, D. E. (1968), 'Semantics of context-free languages', **2**(2), 127–145.
URL: *https://doi.org/10.1007/BF01692511*

Knuth, D. E. (1997), *The art of computer programming, volume 1 (3rd ed.): fundamental algorithms*, Addison Wesley Longman Publishing Co., Inc.

Kourou, K., Exarchos, K. P., Papaloukas, C., Sakaloglou, P., Exarchos, T. & Fotiadis, D. I. (2021), 'Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis', **19**, 5546–5555.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S2001037021004281*

Kuipers, J., Jahn, K., Raphael, B. J. & Beerenwinkel, N. (2017), 'Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors', **27**(11), 1885–1894.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5668945/*

Lemant, J., Le Sueur, C., Manojlović, V. & Noble, R. (2022), 'Robust, universal tree balance indices', **71**(5), 1210–1224.
URL: *https://academic.oup.com/sysbio/article/71/5/1210/6567363*

Li, H., Yang, Z., Tu, F., Deng, L., Han, Y., Fu, X., Wang, L., Gu, D., Werner, B. & Huang, W. (2023), 'Mutation divergence over space in tumour expansion', **20**(208), 20230542.
URL: *https://royalsocietypublishing.org/doi/10.1098/rsif.2023.0542*

Liao, J. G. & Berg, A. (2017), 'Sharpening jensen's inequality'.
URL: *http://arxiv.org/abs/1707.08644*

M. Coronado, T., Mir, A., Rosselló, F. & Rotger, L. (2020), 'On sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index', **21**(1), 154.
URL: *https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-1*

Maini, P. K. (2023), Plenary talk, Mathematical Biology Conference, UCL.

Manojlović, V. (2023*a*), 'demon_trajectories'.
URL: *https://github.com/vesmanojlovic/demon_trajectories*

Manojlović, V. (2023*b*), 'vesmanojlovic/methdemon'.
URL: *https://github.com/vesmanojlovic/methdemon*

Manojlović, V. (2024), 'vesmanojlovic/walter'.
URL: *https://github.com/vesmanojlovic/walter*

McDonald, T. O., Chakrabarti, S. & Michor, F. (2018), 'Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution', **50**(12), 1620–1623.
URL: *https://www.nature.com/articles/s41588-6*

McKenzie, A. & Steel, M. (2000), 'Distributions of cherries for two models of trees', **164**(1), 81–92.

Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. (2006), 'Cancer as an evolutionary and ecological process', **6**(12), 924–935.

Metzcar, J., Wang, Y., Heiland, R. & Macklin, P. (2019), 'A review of cell-based computational modeling in cancer biology', (3), 1–13.
URL: *https://ascopubs.org/doi/10.1200/CCI.18.00069*

Mir, A., Rosselló, F. & Rotger, L. (2013), 'A new balance index for phylogenetic trees', **241**(1), 125–136.
URL: *https://www.sciencedirect.com/science/article/pii/S0025556412002076*

Mir, A., Rotger, L. & Rosselló, F. (2018), 'Sound colless-like balance indices for multifurcating trees', **13**(9), 1–27.
URL: *https://doi.org/10.1371/journal.pone.0203401*

Mooers, A. O. & Heard, S. B. (1997), 'Inferring evolutionary process from phylogenetic tree shape', **72**(1), 31–54.
URL: *https://www.journals.uchicago.edu/doi/10.1086/419657*

Munro, M. J., Wickremesekera, S. K., Peng, L., Tan, S. T. & Itinteang, T. (2018), 'Cancer stem cells in colorectal cancer: a review', **71**(2), 110–116.
URL: *http://jcp.bmj.com/lookup/doi/10.1136/jclinpath*

Nagaraj, S. V. (1997), 'Optimal binary search trees', **188**(1), 1–44.
URL: *https://www.sciencedirect.com/science/article/pii/S0304397596003209*

Nakano, S.-i. (2016), Tree enumeration, *in* M.-Y. Kao, ed., 'Encyclopedia of Algorithms', Springer, pp. 2252–2254.
URL: *https://doi.org/10.1007/978%5F726*

Nievergelt, J. (1974), 'Binary search trees and file organization', **6**(3), 195–207.
URL: *https://dl.acm.org/doi/10.1145/356631.356634*

Nievergelt, J., Pradels, J., Wong, C. K. & Yue, P. C. (1972), 'Bounds on the weighted path length of binary trees', **1**(6), 220–225.
URL: *https://www.sciencedirect.com/science/article/pii/0020019072900154*

Nievergelt, J. & Reingold, E. M. (1972), Binary search trees of bounded balance, *in* 'Proceedings of the fourth annual ACM symposium on Theory of computing - STOC '72', ACM Press, pp. 137–142.
URL: *http://portal.acm.org/citation.cfm?doid=800152.804906*

Niida, A., Mimori, K., Shibata, T. & Miyano, S. (2021), 'Modeling colorectal cancer evolution'.
URL: *http://www.nature.com/articles/s10038-0*

Noble, R. (2020), 'demon'.
URL: *https://github.com/robjohnnoble/demon%5Fmodel*

Noble, R., Burley, J. T., Le Sueur, C. & Hochberg, M. E. (2020), 'When, why and how tumour clonal diversity predicts survival', **13**(7), 1558–1568.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/eva.13057*

Noble, R., Burri, D., Le Sueur, C., Lemant, J., Viossat, Y., Kather, J. N. & Beerenwinkel, N. (2022), 'Spatial structure governs the mode of tumour evolution', **6**(2), 207–217.
URL: *https://www.nature.com/articles/s41559-9*

Noble, R. & Verity, K. (2023), 'A new universal system of tree shape indices'.
URL: *https://www.biorxiv.org/content/10.1101/2023.07.17.549219v2*

Nowell, P. C. (1976), 'The clonal evolution of tumor cell populations', **194**(4260), 23–28.

O'Meara, B. C. (2012), 'Evolutionary inferences from phylogenies: A review of methods', **43**(1), 267–285.
URL: *http://www.annualreviews.org/doi/10.1146/annurev-ecolsys*

O'Brien, C. A., Pollett, A., Gallinger, S. & Dick, J. E. (2007), 'A human colon cancer cell capable of initiating tumour growth in immunodeficient mice', **445**(7123), 106–110.
URL: *http://www.nature.com/articles/nature05372*

Pasco, R. (1977), 'Source coding algorithms for fast data compression (ph.d. thesis abstr.)', **23**(4), 548–548.
URL: *https://ieeexplore.ieee.org/document/1055739*

Paterson, C., Clevers, H. & Bozic, I. (2020), 'Mathematical model of colorectal cancer initiation', **117**(34), 20681–20688.
URL: *http://www.pnas.org/lookup/doi/10.1073/pnas.2003771117*

Patrone, M. V., Hubbs, J. L., Bailey, J. E. & Marks, L. B. (2011), 'How long have i had my cancer, doctor? estimating tumor age via collins' law', **25**(1), 38–43, 46.

Ponz de Leon, M. & Di Gregorio, C. (2001), 'Pathology of colorectal cancer', **33**(4), 372–388.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S1590865801800955*

Prangle, D. (2017), 'Adapting the ABC distance function', **12**(1), 289–309.
URL: *https://projecteuclid.org/journals/bayesian-analysis/volume-12/issue-1/Adapting-the-ABC-Distance-Function/10.1214/16-BA1002.full*

Preston, S. L., Wong, W.-M., Chan, A. O.-O., Poulsom, R., Jeffery, R., Goodlad, R. A., Mandir, N., Elia, G., Novelli, M., Bodmer, W. F., Tomlinson, I. P. & Wright, N. A. (2003), 'Bottom-up histogenesis of colorectal adenomas: Origin in the monocryptal adenoma and initial expansion by crypt fission', p. 8.

Rockne, R. C., Hawkins-Daarud, A., Swanson, K. R., Sluka, J. P., Glazier, J. A., Macklin, P., Hormuth, D. A., Jarrett, A. M., Lima, E. A. B. F., Tinsley Oden, J., Biros, G., Yankeelov, T. E., Curtius, K., Al Bakir, I., Wodarz, D., Komarova, N., Aparicio, L., Bordyuh, M., Rabadan, R., Finley, S. D., Enderling, H., Caudell, J., Moros, E. G., Anderson, A. R. A., Gatenby, R. A., Kaznatcheev, A., Jeavons, P., Krishnan, N., Pelesko, J., Wadhwa, R. R., Yoon, N., Nichol, D., Marusyk, A., Hinczewski, M. & Scott, J. G. (2019), 'The 2019 mathematical oncology roadmap', **16**(4), 041005.
URL: *https://iopscience.iop.org/article/10.1088/1478-3975/ab1a09*

Rosen, D. E. (1978), 'Vicariant patterns and historical explanation in biogeography', **27**(2), 159–188.
URL: *https://www.jstor.org/stable/2412970*

Sackin, M. J. (1972), '"good" and "bad" phenograms', **21**(2), 225–226.
URL: *https://doi.org/10.1093/sysbio/21.2.225*

Salehi, S., Kabeer, F., Ceglia, N., Andronescu, M., Williams, M. J., Campbell, K. R., Masud, T., Wang, B., Biele, J., Brimhall, J., Gee, D., Lee, H., Ting, J., Zhang, A. W., Tran, H., O'Flanagan, C., Dorri, F., Rusk, N., de Algara, T. R., Lee, S. R., Cheng, B. Y. C., Eirew, P., Kono, T., Pham, J., Grewal, D., Lai, D., Moore, R., Mungall, A. J., Marra, M. A., Consortium, I., McPherson, A., Bouchard-Côté, A., Aparicio, S. & Shah, S. P. (2021), 'Clonal fitness inferred from time-series modelling of single-cell cancer genomes', *Nature* **595**, 585–590.

Schreck, C. F., Fusco, D., Karita, Y., Martis, S., Kayser, J., Duvernoy, M.-C. & Hallatschek, O. (2023), 'Impact of crowding on the diversity of expanding populations', **120**(11), e2208361120.
URL: *https://www.pnas.org/doi/full/10.1073/pnas.2208361120*

Schälte, Y., Klinger, E., Alamoudi, E. & Hasenauer, J. (2022), 'pyABC: Efficient and robust easy-to-use approximate bayesian computation', **7**(74), 4304.
URL: *https://joss.theoj.org/papers/10.21105/joss.04304*

Scott, J. G., Maini, P. K., Anderson, A. R. A. & Fletcher, A. G. (2018), 'Inferring tumour proliferative organisation from phylogenetic tree measures in a computational model'.
URL: *http://biorxiv.org/lookup/doi/10.1101/334946*

Shannon, C. E. (1948), 'A mathematical theory of communication', **27**(3), 379–423.
URL: *https://ieeexplore.ieee.org/document/6773024*

Shao, K.-T. & Sokal, R. R. (1990), 'Tree balance', **39**(3), 266–276.
URL: *https://www.jstor.org/stable/2992186*

Siegmund, K., Marjoram, P. & Shibata, D. (2008), 'Modeling DNA methylation in a population of cancer cells', **7**, Article 18.

Siegmund, K., Marjoram, P., Tavaré, S. & Shibata, D. (2011), 'High DNA methylation pattern intratumoral diversity implies weak selection in many human colorectal cancers', **6**, e21657.

Simpson, E. H. (1949), 'Measurement of diversity', **163**(4148), 688–688.
URL: *https://www.nature.com/articles/163688a0*

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D. & Curtis, C. (2015), 'A big bang model of human colorectal tumor growth', **47**(3), 209–216.
URL: *https://www.nature.com/articles/ng.3214*

Sottoriva, A. & Tavaré, S. (2010), Integrating approximate bayesian computation with complex agent-based models for cancer research, *in* Y. Lechevallier & G. Saporta, eds, 'Proceedings of COMPSTAT'2010', Physica-Verlag HD, pp. 57–66.
URL: *https://link.springer.com/10.1007/978%5F5*

Steel, M. & McKenzie, A. (2001), 'Properties of phylogenetic trees generated by yule-type speciation models q', p. 22.

Tarabichi, M., Martincorena, I., Gerstung, M., Leroi, A. M., Markowetz, F., Spellman, P. T., Morris, Q. D., Lingjærde, O. C., Wedge, D. C. & Van Loo, P. (2018), 'Neutral tumor evolution?', **50**(12), 1630–1633.
URL: *https://www.nature.com/articles/s41588-x*

Tariq, K. & Ghias, K. (2016), 'Colorectal cancer carcinogenesis: a review of mechanisms', **13**(1), 120–135.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4850121/*

Tavare, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997), 'Inferring coalescence times from DNA sequence data', **145**(2), 505–518.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1207814/*

Tung, H.-R. & Durrett, R. (2021), 'Signatures of neutral evolution in exponentially growing tumors: A theoretical perspective', **17**(2), e1008701.
URL: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008701*

UK, C. R. (2015), 'Cancer mortality statistics'.
URL: *https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality*

Verity, K. (2023), 'kimverity/RUIindices'.
URL: *https://github.com/kimverity/RUIindices*

Wang, X., Jenner, A. L., Salomone, R., Warne, D. J. & Drovandi, C. (2024), 'Calibration of agent based models for monophasic and biphasic tumour growth using

approximate bayesian computation', **88**(3), 28.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10869399/*

Werner, B., Dingli, D. & Traulsen, A. (2013), 'A deterministic model for the occurrence and dynamics of multiple mutations in hierarchically organized tissues', **10**(85), 20130349.
URL: *https://royalsocietypublishing.org/doi/10.1098/rsif.2013.0349*

West, J., Adler, F., Gallaher, J., Strobl, M., Brady-Nicholls, R., Brown, J., Roberson-Tessi, M., Kim, E., Noble, R., Viossat, Y., Basanta, D. & Anderson, A. R. (2023), 'A survey of open questions in adaptive therapy: Bridging mathematics and clinical translation', **12**, e84263.
URL: *https://doi.org/10.7554/eLife.84263*

West, J., Schenck, R. O., Gatenbee, C., Robertson-Tessi, M. & Anderson, A. R. A. (2021), 'Normal tissue architecture determines the evolutionary course of cancer', **12**(1), 2060.
URL: *https://www.nature.com/articles/s41467-1*

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. (2016), 'Identification of neutral tumor evolution across cancer types', **48**(3), 238–244.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4934603/*

Wodarz, D. & Komarova, N. L. (2020), 'Mutant evolution in spatially structured and fragmented expanding populations', **216**(1), 191–203.
URL: *https://academic.oup.com/genetics/article/216/1/191/6065597*

Wolf, Y. I. & Koonin, E. V. (2013), 'Genome reduction as the dominant mode of evolution', **35**(9), 829–837.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201300037*

Wong, C. K. & Nievergelt, J. (1973), 'Upper bounds for the total path length of binary trees', **20**(1), 1–6.
URL: *https://dl.acm.org/doi/10.1145/321738.321739*

Yin, A., Moes, D. J. A., van Hasselt, J. G., Swen, J. J. & Guchelaar, H. (2019-10), 'A review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors', **8**(10), 720–737.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6813171/*

Yotoko, K. S. C., Dornelas, M. C., Togni, P. D., Fonseca, T. C., Salzano, F. M., Bonatto, S. L. & Freitas, L. B. (2011), 'Does variation in genome sizes reflect adaptive or neutral processes? new clues from passiflora', **6**(3), e18212.
URL: *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018212*

Yule, G. U. (1925), 'II.—a mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f. r. s', **213**(402), 21–87.
URL: *https://royalsocietypublishing.org/doi/10.1098/rstb.1925.0002*